# The effectiveness of query-based hierarchic clustering of documents for information retrieval

**Anastasios Tombros**

**Department of Computing Science**

**Faculty of Computing Science, Mathematics and Statistics**

**University of Glasgow**

**Thesis submitted for the degree of Doctor of Philosophy**

**© Anastasios Tombros, 2002**

**Glasgow, April 2002**

# Abstract

Hierarchic document clustering has been applied to Information Retrieval (IR) for over three decades. Its introduction to IR was based on the grounds of its potential to improve the effectiveness of IR systems. Central to the issue of improved effectiveness is the *Cluster Hypothesis*. The hypothesis states that relevant documents tend to be highly similar to each other, and therefore tend to appear in the same clusters. However, previous research has been inconclusive as to whether document clustering does bring improvements.

The main motivation for this work has been to investigate methods for the improvement of the effectiveness of document clustering, by challenging some assumptions that implicitly characterise its application. Such assumptions relate to the static manner in which document clustering is typically performed, and include the static application of document clustering prior to querying, and the static calculation of interdocument associations.

The type of clustering that is investigated in this thesis is query-based, that is, it incorporates information from the query into the process of generating clusters of documents. Two approaches for incorporating query information into the clustering process are examined: clustering documents which are returned from an IR system in response to a user query (post-retrieval clustering), and clustering documents by using query-sensitive similarity measures.

For the first approach, post-retrieval clustering, an analytical investigation into a number of issues that relate to its retrieval effectiveness is presented in this thesis. This is in contrast to most of the research which has employed post-retrieval clustering in the past, where it is mainly viewed as a convenient and efficient means of presenting documents to users. In this thesis, post-retrieval clustering is employed based on its potential to introduce effectiveness improvements compared both to static clustering and best-match IR systems.

The motivation for the second approach, the use of query-sensitive measures, stems from the role of interdocument similarities for the validity of the cluster hypothesis. In this thesis, an axiomatic view of the hypothesis is proposed, by suggesting that documents relevant to the same query (co-relevant documents) display an inherent similarity to each other which is dictated by the query itself. Because of this inherent similarity, the cluster hypothesis should be valid for any document collection. Past research has attributed failure to validate the hypothesis for a document collection to characteristics of the collection. Contrary to this, the view proposed in this thesis suggests that

failure of a document set to adhere to the hypothesis is attributed to the assumptions made about interdocument similarity.

This thesis argues that the query determines the context and the purpose for which the similarity between documents is judged, and it should therefore be incorporated in the similarity calculations. By taking the query into account when calculating interdocument similarities, co-relevant documents can be "forced" to be more similar to each other. This view challenges the typically static nature of interdocument relationships in IR. Specific formulas for the calculation of query-sensitive similarity are proposed in this thesis.

Four hierarchic clustering methods and six document collections are used in the experiments. Three main issues are investigated: the effectiveness of hierarchic post-retrieval clustering which uses static similarity measures, the effectiveness of query-sensitive measures at increasing the similarity of pairs of co-relevant documents, and the effectiveness of hierarchic clustering which uses query-sensitive similarity measures.

The results demonstrate the effectiveness improvements that are introduced by the use of both approaches of query-based clustering, compared both to the effectiveness of static clustering and to the effectiveness of best-match IR systems. Query-sensitive similarity measures, in particular, introduce significant improvements over the use of static similarity measures for document clustering, and they also significantly improve the structure of the document space in terms of the similarity of pairs of co-relevant documents.

The results provide evidence for the effectiveness of hierarchic query-based clustering of documents, and also challenge findings of previous research which had dismissed the potential of hierarchic document clustering as an effective method for information retrieval.

# Acknowledgements

I would like to thank the following people:

My supervisor Keith van Rijsbergen, for his encouragement and confidence in my work. For advice, moral and practical support when I needed it. Keith's support made it possible for me to complete this work.

Mark Dunlop and Matthew Chalmers, who were involved in my supervision, for helpful discussions and research ideas. Mark Sanderson for "convincing" me to do a Ph.D.

Ian Ruthven for reading this thesis and giving lots of useful feedback. Iain Campbell and Robert Villa, who also read parts of this thesis and drafts of articles containing material from this thesis. I would also like to thank them for their support during the last stages of my Ph.D. Mirna Adriani, for offering useful *Perl* help.

Alan Smeaton and Joemon Jose, for their useful comments and ideas about aspects of this work.

The Glasgow IR group, past and present members included, for providing a stimulating research environment, fun and relaxation. I consider myself lucky to have been part of this group.

All those who offered an opportunity to take my mind off research, for a pleasant reason, during the last four years, and especially during the last nine months. Those include Vangelis, Harold, Iain and Denise, Kostas S., Tony, Mark and *almost* everyone in the "Bball this Saturday" unofficial club. Special thanks go to my sister Lena, and to Kostis.

My parents Ioannis and Anna, for their love and support. I owe everything to them.

# Table of Contents

**ABSTRACT** ..... **i**

**ACKNOWLEDGEMENTS** ..... **iii**

**LIST OF FIGURES** ..... **viii**

**LIST OF TABLES** ..... **ix**

**1. INTRODUCTION** ..... **1**

1.1 INTRODUCTION ..... 1
1.2 MOTIVATION ..... 2
1.3 THESIS STATEMENT ..... 3
1.4 THESIS OUTLINE ..... 4

**2. BASIC CONCEPTS OF INFORMATION RETRIEVAL** ..... **6**

2.1 INTRODUCTION ..... 6
2.2 REPRESENTING DOCUMENTS AND QUERIES ..... 8
*2.2.1 Query operations* ..... *10*
2.3 MATCHING BETWEEN DOCUMENTS AND QUERIES ..... 11
2.4 EVALUATION OF IR SYSTEMS ..... 13
2.5 DOCUMENT CLUSTERING ..... 15

**3. DOCUMENT CLUSTERING FOR IR: BACKGROUND** ..... **17**

3.1 INTRODUCTION ..... 17
*3.1.1 Hierarchic clustering: an outline* ..... *20*
3.2 DOCUMENT REPRESENTATION ..... 21
*3.2.1 Exhaustivity of document representations* ..... *21*
*3.2.2 The effect of term-weighting* ..... *22*
3.3 MEASURING INTERDOCUMENT RELATIONSHIPS ..... 24
*3.3.1 Formal definitions* ..... *24*
*3.3.2 Choice of a particular measure* ..... *26*
3.4 HIERARCHIC CLUSTERING METHODS ..... 28
*3.4.1 Single link* ..... *31*
*3.4.2 Complete link* ..... *31*
*3.4.3 Group average link* ..... *32*
*3.4.4 Ward's method* ..... *33*
*3.4.5 An example* ..... *33*
*3.4.6 Other methods* ..... *34*
*3.4.7 Some remarks* ..... *35*
3.5 CLUSTER REPRESENTATION ..... 36

*3.5.1 Cluster representatives* ... *36*
*3.5.2 Representations of cluster contents* ... *37*
3.6 CLUSTER VALIDITY ... 39
*3.6.1 Clustering tendency and cluster-based effectiveness* ... *43*
3.7 RECENT TRENDS ... 45
*3.7.1 Hypertext and web clustering* ... *45*
3.8 REFLECTIONS ON DOCUMENT CLUSTERING RESEARCH ... 48
3.9 SUMMARY ... 49

**4. ON THE EFFECTIVENESS OF CLUSTER-BASED INFORMATION RETRIEVAL ... 51**

4.1 INTRODUCTION ... 51
4.2 TESTING FOR THE VALIDITY OF THE CLUSTER HYPOTHESIS ... 52
*4.2.1 Separation of frequency distributions* ... *52*
*4.2.2 The nearest neighbour test* ... *53*
*4.2.3 The density test* ... *54*
*4.2.4 Some notes on the tests* ... *54*
4.3 CLUSTER-BASED RETRIEVAL ... 56
*4.3.1 Top-down search* ... *57*
*4.3.2 Bottom-up search* ... *58*
*4.3.3 Comparative performance of the two searches* ... *59*
*4.3.4 Optimal cluster search* ... *61*
*4.3.5 On optimal effectiveness measurements* ... *62*
4.4 THE EFFECTIVENESS OF HIERARCHIC CLUSTERING IN IR ... 65
*4.4.1 Comparisons of hierarchic methods* ... *65*
*4.4.2 Cluster-based vs. non cluster-based retrieval* ... *69*
4.5 WHAT THIS THESIS ADDRESSES ... 71
4.6 SUMMARY ... 72

**5. QUERY-BASED DOCUMENT CLUSTERING ... 74**

5.1 INTRODUCTION ... 74
5.2 POST-RETRIEVAL CLUSTERING ... 75
*5.2.1 Re-examining the cluster hypothesis for Scatter/Gather* ... *78*
*5.2.2 Some notes on post-retrieval clustering* ... *80*
5.3 REVIEWING THE ROLE OF THE CLUSTER HYPOTHESIS ... 81
*5.3.1 The static nature of interdocument relationships* ... *82*
*5.3.2 Related work* ... *84*
5.4 RESEARCH OBJECTIVES ... 87
*5.4.1 The effectiveness of post-retrieval hierarchic clustering in IR* ... *87*
*5.4.2 Query-sensitive similarity measures for the calculation of interdocument relationships* ... *88*
5.5 THE EXPERIMENTAL ENVIRONMENT ... 89
*5.5.1 Test collections* ... *90*
*5.5.2 IR system* ... *92*

*5.5.3 Clustering methods* ... *93*
*5.5.4 Measuring retrieval effectiveness* ... *95*
5.6 SUMMARY ... 97

**6. THE EFFECTIVENESS OF HIERARCHIC POST-RETRIEVAL CLUSTERING ... 98**

6.1 INTRODUCTION ... 98
6.2 CLUSTERING TENDENCY ... 99
6.3 CLUSTER-BASED EFFECTIVENESS ... 103
*6.3.1 Effectiveness for different numbers of top-ranked documents* ... *106*
*6.3.2 Cluster-based vs. inverted file search effectiveness* ... *108*
*6.3.3 Random cluster-based effectiveness* ... *115*
*6.3.4 Optimal cluster characteristics* ... *117*
*6.3.5 Bottom-level optimal clusters* ... *119*
6.4 COMPARATIVE EFFECTIVENESS OF THE FOUR CLUSTERING METHODS ... 121
*6.4.1 Effectiveness under post-retrieval clustering* ... *121*
*6.4.2 Effectiveness under static clustering* ... *124*
6.5 SUMMARY ... 125

**7. QUERY-SENSITIVE SIMILARITY MEASURES ... 127**

7.1 INTRODUCTION ... 127
7.2 QUERY-SENSITIVE SIMILARITY MEASURES ... 128
*7.2.1 Defining query-sensitive similarity measures* ... *128*
*7.2.2 An example* ... *133*
*7.2.3 Limitations* ... *135*
*7.2.4 Related work* ... *136*
7.3 EXPERIMENTAL ENVIRONMENT ... 138
7.4 EXPERIMENTAL RESULTS ... 140
*7.4.1 Global vs. local query-term weighting* ... *140*
*7.4.2 Selecting parameters for M3* ... *142*
*7.4.3 Comparative effectiveness of the query-sensitive measures and the cosine coefficient* ... *146*
*7.4.4 Comparative effectiveness of M1, M2 and M3* ... *153*
*7.4.5 Effect of query length on the query-sensitive measures* ... *155*
7.5 SUMMARY ... 160

**8. HIERARCHIC DOCUMENT CLUSTERING USING QUERY-SENSITIVE SIMILARITY MEASURES ... 163**

8.1 INTRODUCTION ... 163
8.2 THE EFFECTIVENESS OF HIERARCHIC CLUSTERING USING QSSM ... 165
*8.2.1 Comparatively to clustering using static similarity measures* ... *166*
*8.2.2 Comparatively to IFS effectiveness* ... *170*
*8.2.3 Comparatively to random cluster-based effectiveness* ... *174*
*8.2.4 Discussion* ... *176*

8.3 HIERARCHY CHARACTERISTICS ..... 177
*8.3.1 Size of clusters ..... 177*
*8.3.2 Optimal cluster characteristics ..... 181*
*8.3.3 Discussion ..... 186*
8.4 COMPARISON OF THE QUERY-SENSITIVE MEASURES ..... 187
*8.4.1 Comparative effectiveness of M1, M2 and M3 ..... 187*
*8.4.2 The effect of query length ..... 188*
*8.4.3 Effectiveness for different numbers of top-ranked documents ..... 190*
*8.4.4 Discussion ..... 192*
8.5 COMPARATIVE EFFECTIVENESS OF THE FOUR CLUSTERING METHODS ..... 194
*8.5.1 The group average method ..... 194*
*8.5.2 Complete link and Ward's methods ..... 195*
*8.5.3 The single link method ..... 196*
*8.5.4 Discussion ..... 198*
8.6 SUMMARY ..... 199

**9. CONTRIBUTIONS AND FUTURE WORK ..... 201**

9.1 CONTRIBUTIONS AND CONCLUSIONS ..... 201
9.2 FUTURE WORK ..... 205

**BIBLIOGRAPHY AND REFERENCES ..... 209**

**APPENDIX A ..... A-1**

**APPENDIX B ..... B-1**

**APPENDIX C ..... C-1**

**APPENDIX D ..... D-1**

# List of Figures

FIGURE 2.1: A TYPICAL IR SYSTEM ..... 6
FIGURE 2.2: DOCUMENT AND QUERY REPRESENTATIONS IN THE VECTOR MODEL ..... 11
FIGURE 2.3: CALCULATION OF PRECISION AND RECALL ..... 13
FIGURE 2.4: A RECALL-PRECISION GRAPH ..... 14
FIGURE 3.1: A SIMILARITY MATRIX ..... 25
FIGURE 3.2: A SIMILARITY DENDROGRAM ..... 30
FIGURE 3.3: TRANSFORMATION OF THE SIMILARITY MATRIX BY THE SINGLE LINK METHOD ..... 34
FIGURE 3.4: THE SIMILARITY DENDROGRAM FOR THE EXAMPLE ..... 34
FIGURE 3.5: AN IDEAL SCENARIO: TOTAL SEPARATION OF RELEVANT AND NON-RELEVANT DOCUMENTS . 44
FIGURE 4.1: SEPARATION OF FREQUENCY DISTRIBUTIONS ..... 52
FIGURE 4.2: A SAMPLE BROAD TOP-DOWN SEARCH ..... 57
FIGURE 4.3: A SAMPLE BOTTOM-UP SEARCH ..... 59
FIGURE 4.4: EXAMPLE OF CALCULATION OF OPTIMAL EFFECTIVENESS MEASURES ..... 64
FIGURE 5.1: THE EXPERIMENTAL SYSTEM ..... 90
FIGURE 5.2: A SAMPLE TREC TOPIC ..... 91
FIGURE 6.1: RANDOM VS. ACTUAL VALUES FOR THE NN TEST USING THE WSJ COLLECTION ..... 103
FIGURE 6.2: AVERAGE SIZE OF OPTIMAL CLUSTERS FOR THE LISA COLLECTION ..... 119
FIGURE 6.3: COMPARATIVE EFFECTIVENESS OF THE FOUR METHODS USING THE MEDLINE COLLECTION . 122
FIGURE 6.4: COMPARATIVE EFFECTIVENESS OF THE FOUR METHODS USING THE WSJ COLLECTION..... 122
FIGURE 7.1: THE VARIABLE SIMILARITY $SIM(D_1,D_2,Q)$ ..... 129
FIGURE 7.2: THE SIMILARITY MATRIX FOR THE EXAMPLE ..... 134
FIGURE 7.3: THE EFFECTIVENESS OF M3 AS A FUNCTION OF $\Theta_1$ AND $\Theta_2$ FOR THE WSJ COLLECTION ..... 144
FIGURE 7.4: RANDOM VS. ACTUAL VALUES FOR THE 5NN TEST USING THE WSJ COLLECTION ..... 150
FIGURE 8.1: PRECISION-ORIENTED EFFECTIVENESS USING THE SINGLE LINK METHOD AND THE WSJ COLLECTION ..... 173
FIGURE 8.2: RECALL-ORIENTED EFFECTIVENESS USING THE SINGLE LINK METHOD AND THE WSJ COLLECTION ..... 174
FIGURE 8.3: AVERAGE SIZE OF OPTIMAL CLUSTERS USING THE GROUP AVERAGE METHOD AND THE AP COLLECTION FOR RECALL-ORIENTED SEARCHES ..... 183
FIGURE 8.4: AN EXAMPLE DOCUMENT HIERARCHY ..... 184
FIGURE 8.5: EFFECTIVENESS ACROSS NUMBERS OF TOP-RANKED DOCUMENTS USING THE AP COLLECTION, M3 AND PRECISION-ORIENTED SEARCHES ..... 191
FIGURE 8.6: COMPARATIVE EFFECTIVENESS OF THE FOUR METHODS USING M1 AND AP..... 195
FIGURE 8.7: COMPARATIVE EFFECTIVENESS OF THE FOUR METHODS USING M2 AND MEDLINE..... 196

# List of Tables

TABLE 2.1: IR TEST COLLECTIONS ..... 15
TABLE 3.1: VALUES FOR THE PARAMETERS OF THE LANCE&WILLIAMS COMBINATORIAL EQUATION ..... 35
TABLE 4.1: STUDIES COMPARING HIERARCHIC AGGLOMERATIVE METHODS IN IR ..... 67
TABLE 5.1: COLLECTION STATISTICS ..... 91
TABLE 5.2: INITIAL RETRIEVAL EVALUATION ..... 93
TABLE 5.3: AVERAGE CLUSTER SIZES FOR THE FOUR METHODS USING THE AP AND WSJ COLLECTIONS ..... 94
TABLE 6.1: RESULTS OF THE NN TEST ..... 100
TABLE 6.2: AVERAGE NUMBER OF RELEVANT DOCUMENTS FOR DIFFERENT NUMBERS OF TOP-RANKED DOCUMENTS ..... 101
TABLE 6.3: RESULTS FOR THE NN TEST GENERATED BY RANDOM SIMILARITY VALUES ..... 102
TABLE 6.4: RESULTS USING THE GROUP AVERAGE METHOD ..... 104
TABLE 6.5: RESULTS OBTAINED USING THE RELEVANT AND RETRIEVED DOCUMENTS TO CALCULATE RECALL FOR THE WSJ COLLECTION ..... 105
TABLE 6.6: SIGNIFICANCE LEVELS FOR COMPARISONS ACROSS VALUES OF *N*. RESULTS ARE FOR THE GROUP AVERAGE METHOD AND B=1 ..... 107
TABLE 6.7: COMPARATIVE EFFECTIVENESS OF CLUSTER-BASED AND INVERTED FILE SEARCHES USING THE GROUP AVERAGE METHOD FOR B=0.5 AND 2 ..... 109
TABLE 6.8: AVERAGE OFFSETS FOR THE MK4 MEASURE, FOR THE AP AND MED COLLECTIONS ..... 111
TABLE 6.9: COMPARATIVE MK1 AND MK4 EFFECTIVENESS WHEN USING THE TWO DIFFERENT DEFINITIONS OF RECALL ..... 113
TABLE 6.10: RANDOM VS. ACTUAL EFFECTIVENESS VALUES FOR THE CISI COLLECTION USING THE SINGLE LINK METHOD ..... 116
TABLE 6.11: AVERAGE SIZE AND AVERAGE NUMBER OF RELEVANT DOCUMENTS FOR OPTIMAL CLUSTERS USING THE GROUP AVERAGE METHOD (MK1), AND FOR OPTIMAL IFS (MK4), USING THE LISA AND WSJ COLLECTIONS FOR B=0.5 AND 2 ..... 118
TABLE 6.12: BOTTOM-LEVEL CLUSTER SIZE STATISTICS FOR LISA HIERARCHIES ..... 119
TABLE 6.13: PERCENTAGE OF OPTIMAL BOTTOM-LEVEL CLUSTERS ..... 120
TABLE 7.1: QUERY TERM STATISTICS FOR THE SIX TEST COLLECTIONS ..... 132
TABLE 7.2: GLOBAL VS LOCAL QUERY-TERM WEIGHTING FOR WSJ ..... 141
TABLE 7.3: RESULTS OF THE 5NN TEST FOR THE LISA COLLECTION BY VARYING THE $\Theta_1$: $\Theta_2$ RATIO IN FAVOUR OF $\Theta_1$ ..... 143
TABLE 7.4: RESULTS OF THE 5NN TEST FOR THE LISA COLLECTION BY VARYING THE $\Theta_1$: $\Theta_2$ RATIO IN FAVOUR OF $\Theta_2$ ..... 144
TABLE 7.5: AP AND WSJ RESULTS ..... 147
TABLE 7.6: CACM AND LISA RESULTS ..... 148
TABLE 7.7: CISI AND MEDLINE RESULTS ..... 148
TABLE 7.8: RESULTS OF THE 1NN TEST WHEN USING CISI AND WSJ ..... 152
TABLE 7.9: THE EFFECT OF QUERY LENGTH FOR AP: RESULTS OF THE 5NN TEST ..... 156
TABLE 7.10: THE EFFECT OF QUERY LENGTH FOR WSJ: RESULTS OF THE 5NN TEST ..... 157

TABLE 8.1: OPTIMAL CLUSTER-BASED EFFECTIVENESS USING THE GROUP AVERAGE METHOD ..... 167
TABLE 8.2: COMPARATIVE EFFECTIVENESS OF CLUSTER-BASED AND INVERTED-FILE SEARCHES USING THE GROUP AVERAGE METHOD ..... 171
TABLE 8.3: RANDOM VS. ACTUAL EFFECTIVENESS FOR THE CISI COLLECTION USING THE SINGLE LINK METHOD FOR $B$=2 ..... 175
TABLE 8.4: AVERAGE SIZE OF CLUSTERS GENERATED USING THE COSINE COEFFICIENT AND MEASURE M2 FOR THE WSJ COLLECTION ..... 178
TABLE 8.5: AVERAGE SIZE OF CLUSTERS GENERATED USING THE COSINE COEFFICIENT AND MEASURES M1 AND M3 FOR THE WSJ COLLECTION (WARD AND COMPLETE LINK METHODS) ..... 178
TABLE 8.6: COMPOSITION OF BOTTOM LEVEL CLUSTERS FOR THE COMPLETE LINK METHOD USING THE WSJ COLLECTION ..... 180
TABLE 8.7: AVERAGE SIZE OF CLUSTERS GENERATED USING THE COSINE COEFFICIENT AND MEASURE M2 FOR THE LISA COLLECTION ..... 180
TABLE 8.8: PERCENTAGE OF RELEVANT AND RETRIEVED DOCUMENTS CONTAINED IN AN OPTIMAL CLUSTER USING THE GROUP AVERAGE METHOD AND THE AP COLLECTION ..... 182
TABLE 8.9: AVERAGE SIZE AND AVERAGE NUMBER OF RELEVANT DOCUMENTS IN OPTIMAL SETS, USING THE WSJ COLLECTION FOR PRECISION-ORIENTED SEARCHES ..... 183
TABLE 8.10: HIERARCHY LEVELS AT WHICH OPTIMAL CLUSTERS ARE FORMED. USING THE WSJ COLLECTION AND PRECISION-ORIENTED SEARCHES ..... 185
TABLE 8.11: AVERAGE CLUSTER SIZE FOR WARD'S METHOD, USING EXPANDED QUERIES AND THE WSJ COLLECTION ..... 189
TABLE 8.12: DIFFERENCE IN EFFECTIVENESS BETWEEN SHORT AND ORIGINAL QUERIES USING THE WSJ COLLECTION AND MEASURE M2 ..... 190

# Chapter 1

# Introduction

## 1.1 Introduction

This thesis investigates the effectiveness of query-based hierarchic clustering of documents for the purpose of *Information Retrieval* (IR). Incorporating information from the query into the process of generating clusters of documents is not common practice in IR research. Moreover, the effectiveness of document clustering that incorporates information from the query has not been thoroughly investigated in the past. This work addresses these two issues. The main argument of this thesis will be that the retrieval effectiveness of document clustering can be significantly improved by taking into account the searcher's subject of inquiry.

Document clustering has been applied to IR for over thirty years. Research in the field has undergone a number of significant changes, from focusing on efficiency issues in the early years (Rocchio, 1966; Salton, 1971), to postulating the potential of clustering to increase the effectiveness of the IR process (Jardine & Van Rijsbergen, 1971; Croft, 1978). The literature published in the field covers a number of diverse areas, such as for example the development of efficient algorithms for document clustering (Silverstein & Pedersen, 1997; Larsen & Aone, 1999), the visualisation of clustered document spaces (Allen *et al*, 2001; Leuski, 2001), the application of document clustering to browsing large document collections (Cutting *et al*., 1992; Hearst & Pedersen, 1996), etc. This thesis focuses solely on the retrieval effectiveness of document clustering, and examines a number of issues relating to the effectiveness of hierarchic clustering methods in IR.

In the rest of this chapter I first outline the motivation that led to the undertaking of this research, I then state its aims and achievements, and I outline the structure of the remainder of this thesis.

## 1.2 Motivation

Jardine and Van Rijsbergen (1971) first provided some experimental evidence to suggest that the retrieval effectiveness of an IR system can benefit from the use of document clustering. The effectiveness of an IR system was expected to increase through the use of clustering, since the file organisation, and any strategy to search it, take into account the relationships that hold between documents in a collection (Croft, 1978). Relevant documents that might have otherwise been ranked low in a best-match search, will be (through inter-document associations) grouped together with other relevant documents, thus improving the effectiveness of an IR system.

The *Cluster Hypothesis* is fundamental to the issue of improved effectiveness; it states that relevant documents tend to be more similar to each other than to non-relevant documents, and therefore tend to appear in the same clusters (Jardine & Van Rijsbergen, 1971). If the cluster hypothesis holds for a particular document collection, then relevant documents will be well separated (i.e. grouped separately) from non-relevant ones. The actual effectiveness of hierarchic clustering can be gauged by *cluster-based searches* that retrieve the cluster that best matches the query (Croft, 1978; Voorhees, 1985a).

Clustering has typically been applied statically, over the whole document collection prior to querying (*static clustering*). Document hierarchies are thus static, and do not reflect the user's interest, which is reflected through the query posed to the IR system. Research that has investigated the effectiveness of static clustering has suggested some limitations, mainly in the form of the poor comparative effectiveness of cluster-based and best-match searches (El-Hamdouchi & Willett, 1989).

Clustering has also been applied to the search results of an IR system (*post-retrieval clustering*) (Willett, 1985; Allen *et al*., 1993; Hearst & Pedersen, 1996). This type of clustering takes into account the query, since it only clusters those documents that have a high likelihood of being relevant to the query. In contrast to static clustering, the behaviour and effectiveness of post-retrieval clustering has not been extensively investigated.

The main motivation for this work has been to investigate possibilities for the improvement of the effectiveness of document clustering by challenging its typically static nature. By doing so, I aim to challenge previous findings which have demonstrated a number of limitations regarding the effectiveness of cluster-based retrieval. This work also aims to demonstrate that, by challenging some of the static assumptions that characterise document clustering, it is possible to enhance cluster-based effectiveness. It is through incorporating aspects of the query into the clustering process that this work attempts to move away from the static clustering paradigm.

### 1.2.1 Query-based clustering

In this thesis I investigate *query-based clustering*, which incorporates aspects of the user's query into the clustering process. Two approaches for implementing query-based clustering are investigated. The first one is post-retrieval clustering, for which I present an analytical investigation into a number of issues that relate to its effectiveness. This is in contrast to most of the research which has employed post-retrieval clustering, where it is viewed as a convenient means of presenting documents to users (Allen *et al*., 1993; Eguchi *et al*., 2001; Leuski, 2001), or as a method to improve the efficiency of the clustering process (since less documents are clustered) (Kirriemuir & Willett, 1995).

The investigation of post-retrieval clustering demonstrated a number of positive results regarding the effectiveness of cluster-based retrieval. However, it also demonstrated a number of shortcomings. This led into investigating alternative ways to incorporate the query into the clustering process, and more specifically, into proposing the use of query-sensitive measures for the calculation of interdocument relationships.

The motivation for the use of query-sensitive measures stemmed from the cluster hypothesis, and the role of interdocument similarities for the validity of the hypothesis. In this thesis I propose an alternative, axiomatic view of the hypothesis, by suggesting that documents relevant to the same query (*co-relevant* documents) display an inherent similarity to each other which is dictated by the query itself. Because of this inherent similarity, the cluster hypothesis should be valid for any document collection. Past research has attributed failure to validate the hypothesis for a document collection to characteristics of the collection. Contrary to this, the view proposed in this thesis suggests that failure of a document set to adhere to the hypothesis is attributed to the assumptions made about interdocument similarity.

Motivated by studies in other fields that have demonstrated the dynamic nature of similarity (Goodman, 1972; Tversky, 1977), I challenge its typically static use in IR. I argue that the query determines the context and the purpose for which the similarity between documents is judged, and it should therefore be incorporated in the similarity calculations. By taking the query into account when calculating the similarity between pairs of documents, co-relevant documents can be "forced" to be more similar to each other. If query-sensitive measures are effective in this respect, then a clustering of documents generated by using such measures can be expected to be effective.

## 1.3 Thesis statement

The statement of this thesis is that by incorporating information from the query into the clustering process, the effectiveness of the clustering process can be enhanced. I investigate the effectiveness

of the two query-based clustering methods that I previously described: post-retrieval clustering, and clustering using query-sensitive similarity measures.

I first study the effectiveness of post-retrieval clustering, and I compare it to that of static clustering and to that of inverted file, best-match searches. The results from this study demonstrate that the effectiveness of post-retrieval clustering is significantly higher than that of static clustering, and that it also has the potential to significantly exceed inverted file search (IFS) effectiveness. However, a number of shortcomings are also demonstrated, mainly in the form of poor comparative effectiveness to IFS, and close-to-random effectiveness in a number of experimental conditions.

Post-retrieval clustering can be seen as a first level of incorporating information from the query into the clustering of documents. The use of query-sensitive similarity measures introduces a level of query influence that can complement post-retrieval clustering. The effectiveness of this second form of query-based clustering is also investigated, and is compared to that of post-retrieval clustering using a static similarity measure, and to that of inverted file searches.

The results demonstrate that viewing similarity as a dynamic concept that depends on the query is an effective approach which introduces a number of benefits for IR. This work demonstrates three main benefits. The first benefit is that query-sensitive measures manage to increase the similarity of pairs of co-relevant documents when compared to static similarity measures. The second benefit is that query-sensitive measures generate document hierarchies whose effectiveness is significantly higher than that of hierarchies generated using static measures. The third benefit, is that the effectiveness of cluster-based IR when using query-sensitive measures has the potential to significantly outperform the effectiveness of IFS more consistently and more significantly than when using static similarity measures.

## 1.4 Thesis Outline

This thesis is organised into 9 Chapters (including the present chapter). An outline of the contents of the remaining chapters follows.

**Chapter 2:** in this chapter I discuss some of the main concepts of information retrieval, focusing on issues that are relevant to the experimental work reported in this thesis. The purpose is to establish a basic terminology and coverage of issues that are used in the following chapters.

**Chapter 3:** this chapter provides a detailed review of past work on document clustering. The review is organised around the main steps of the clustering process. The purpose of this chapter is to provide the necessary background on the application of hierarchic document clustering to IR,

and to discuss aspects of the clustering process that are important to the work that I describe in later chapters.

**Chapter 4:** in this chapter I focus on the issue of cluster-based effectiveness, which is central to this work. I review past work dealing with issues of cluster-based retrieval effectiveness, present the methodology that is used to measure retrieval effectiveness, and I present the view taken in this thesis regarding previous findings in this area.

In Chapters 3 and 4 I discuss aspects of research on document clustering in detail. The main reason for doing so, is to give a thorough coverage of issues that relate to document clustering research, and also to provide the necessary background to illustrate a number of decisions that I implement later, when I report the experimental part of this work. One such decision is the use of optimal cluster evaluation. In both chapters, through argument and review of previous work, I demonstrate the appropriateness of this evaluation approach. Consequently, these two chapters are central to the flow of the argument in this thesis.

**Chapter 5:** in this chapter I present the two specific methods of query-based clustering that are pursued in this thesis. I examine the implications of post-retrieval clustering for cluster-based effectiveness. A different view of the cluster hypothesis is proposed, and the use of query-sensitive similarity measures is postulated. The motivation and the intuitions behind the proposed approach are also outlined, and the research objectives of this thesis are stated. In this chapter I also present details of the experimental environment used.

**Chapter 6:** in this chapter I investigate the effectiveness of the first form of query-based clustering, post-retrieval clustering.

**Chapter 7:** I define specific formulas for the calculation of query-sensitive similarity between documents, and I also present experimental evidence for the application of query-sensitive measures to IR.

**Chapter 8:** I investigate the effectiveness and the characteristics of the second form of query-based clustering which uses query-sensitive similarity measures for the calculation of interdocument relationships.

**Chapter 9:** in Chapter 9 I report the main contributions that this work made, and I also point to some issues for future work that follow from this thesis.

# Chapter 2

# Basic Concepts of Information Retrieval

## 2.1 Introduction

*Information Retrieval* (*IR*) is a discipline involved with the organisation, structuring, analysis, storage, searching and dissemination of information. A compact definition of the basic function of an *information retrieval system* (*IRS*) has been given by Lancaster, (1968):

> "An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his enquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request."

Within the few lines of the above definition, the three major parts of an IRS have already been identified: a user with a request (*query*) for information, a collection of documents against which this request is matched, and finally the response of the IR system in relation to the user's request. The task of an IRS is, through its response, to help a user locate those documents that have the potential to satisfy his *information need*.

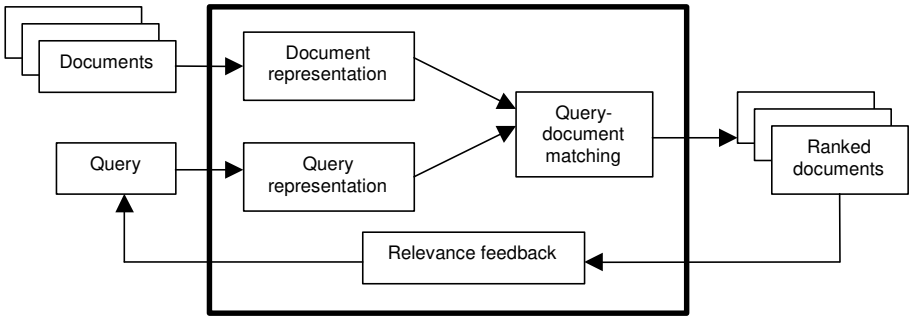


**Figure 2.1.** A typical IR system

Figure 2.1 shows a diagram of a typical IR system. A set of documents (*document collection*) is processed by the IRS in such a way that an internal representation of these documents is derived.

This internal representation can then be further processed by the IRS. A user who wishes to search this document collection expresses an information need in the form of a query that is posed to the system. The IRS represents this query in an internal form that is suitable for further processing. The IRS matches the user query against each document in the collection, and produces a list of documents (that is usually ranked in some way) that is presented to the user. The user can interact with this ranked list by indicating documents that relate to his information need, or he can modify his initial query in the light of the documents returned in the ranked list. Each of these steps is covered in more detail in the remainder of this chapter.

A number of IR models have been developed, which describe the way in which documents and queries are represented, as well as the way in which the document-query matching process is implemented. The most publicised IR models are the Boolean, the vector space (or vector processing) (Salton, 1971; Salton *et al*., 1975), the probabilistic (Robertson, 1977), and the logical (Van Rijsbergen, 1986). The vector space model has been the basis on which most of the experimental work on document clustering has been based (see Chapter 3), and therefore is the model that will be adopted for the experimental work reported in this thesis.

IR is a field that has existed since computers were first able to count words (Belew, 2000). The first IR systems were developed in order to facilitate the automated searching of library material by users. However, due to the advent of powerful computing facilities and the explosive growth of the information available in an electronic form, IR systems gradually expanded their scope over the last few decades. It is now not only research literature that is within the scope of IR systems, but also a wide spectrum of heterogeneous types of information, including multimedia data (e.g. images, audio, video, etc.).

The *Internet*, and more specifically the *World Wide Web*, has become the medium in which increasing numbers of people search for information. IR systems that have been developed on the Internet (*search engines*), aim to make the plethora of available data searchable and easily accessible by users. This has resulted in the development of research efforts towards exploiting features that characterise web pages (e.g. hyperlinked structures, Web page popularity, HTML structure, etc.) (Bharat & Henzinger, 1998; Kleinberg, 1999; Belew, 2000). The research reported in this thesis has been applied to textual information that is stored in the form of *documents*, the set of which constitutes a *collection*, or a *corpus*. The document collection is not assumed to possess a hyperlinked structure.

It is not the aim of this chapter to provide a thorough review of information retrieval research. Readings that aim to do so include books by Van Rijsbergen (1979), Salton and McGill (1983), Frakes and Baeza-Yates (1992), and Belew (2000). Sparck Jones and Willett (1997), have also edited a selection of papers in the field of IR that span over five decades, and cover almost every major aspect of research in the field.

In this chapter, those concepts that are pertinent to the research reported later in this thesis are elaborated, so as to provide the necessary background. In section 2.2 I discuss how documents and queries are represented by IR systems, in section 2.3 I examine the mechanism for matching queries against documents, and in section 2.4 I present issues pertaining to the evaluation of the effectiveness of IR systems. Finally, in section 2.5 I briefly present document clustering and its application to IR.

## 2.2 Representing documents and queries

Documents are traditionally processed by an IRS not in their original form, but in an internal representation which is the outcome of what is called the *indexing* process. This internal representation aims to model the original information content as accurately as possible. The essence of the indexing process is to assign to each document a set of *indexing features*. At their simplest form such features maybe a list of words, known as *terms*, extracted from the text of the original documents. A more complicated approach might involve the extraction of phrasal units, or the use of linguistic, semantic and knowledge-based methods (Lewis & Sparck-Jones, 1993) to build a higher level representation. For the purposes of this overview, as well as for the purposes of this thesis, it will be assumed that document representations are lists of words extracted from the original texts.

Before such words become indexing features however, they usually go through some specific form of lexical processing. For example, they will typically have their case normalised. Moreover, certain high frequency function words (*stop-words*) will not be considered as indexing features (Van Rijsbergen, 1979). Typical examples of stop-words are articles (e.g. 'the') and prepositions (e.g. 'in', 'at'). The benefit of this method is that without losing any significant information it is possible to achieve a reduction of the text volume of up to 50 percent[1].

Another typical lexical processing of the feature set is to remove the suffixes from the remaining words of the input text. This can be achieved through the application of a *stemming algorithm*[2] that will reduce words to a common root form (stem). For example, the words 'manufacture' and 'manufacturing' will be mapped to the same entity, 'manufactur' in the vocabulary of index terms. A stemming algorithm that is widely used by IR researchers was developed by Porter (1980).

Each word that is selected as an indexing feature for a particular document can be thought of as *discriminating* (to a measurable degree at least) between that document and all the other

[1] See (Van Rijsbergen, 1979), pp. 17.

[2] A comprehensive overview of stemming algorithms can be found in (Frakes & Baeza-Yates, 1992), pp 131-151.

documents in the collection. In order to quantify the discriminating power of terms, methods that assign numerical weights to terms have been used. Luhn (1958), associated the discriminating power of an index term with its frequency of occurrence within a document (*term frequency*, *tf*). Luhn postulated that the most discriminating terms are those that occur with medium frequency. High frequency words are discarded as carrying little information, and low frequency words are rejected as being unlikely candidates to appear in a query.

Later research (Sparck Jones, 1972) proposed that a more accurate quantification of term importance can be achieved if one also makes use of information about term usage within the entire document collection. The *inverse document frequency* (*idf*) captures this belief: for a document collection comprising $N$ documents, if term $i$ occurs in $n_i$ documents, then the term's *idf* weight is given by $\log(\frac{N}{n_i})$.

A frequently used term weighting function is a combination of the *tf* and *idf* weights, typically referred to as a *tf-idf* weight (Salton, 1971)[3]:

$$w_{ij} = \frac{\log(freq_{ij} + 1)}{\log(length_j)} \cdot \log(\frac{N}{n_i}),$$

$w_{ij}$ = *tf-idf* weight of term $i$ in document $j$

$freq_{ij}$ = frequency of term $i$ in document $j$

$length_j$ = length (in words) of document $j$

$N$ = number of documents in the collection

$n_i$ = number of documents that term $i$ is assigned to

It should be noted that term weighting methods have been extensively researched. As a consequence, a large number of weighting schemes have been proposed in the IR literature. Salton and Buckley, (1988), provide a comprehensive overview of various weighting schemes and their comparative effect on retrieval effectiveness. Any of the classical IR textbooks, such as for example (Van Rijsbergen, 1979; Salton & McGill, 1983), also provide further details on term weighting methods.

A data structure that is typically used in IR systems to store information about term usage is an *inverted file* structure (Van Rijsbergen, 1979; Frakes & Baeza-Yates, 1992). In this structure, for each index term, the list of documents in which this term occurs is stored. For retrieval purposes this means that given a search keyword, it is possible to immediately locate all the documents in the database that contain this keyword (Van Rijsbergen, 1979). The inverted file structure can be

[3] Note that the *tf* component has been normalised by the length of the document (see Salton & Buckley, 1988).

seen in contrast to the document vector, which contains for a given document all the terms that occur in it.

### 2.2.1 Query operations

The goal of any IRS is to help a user locate those documents that have the potential to satisfy his information need. An information need is expressed by a user in a form that is recognised by a computer system, e.g. by means of keyboard input. Such a formulation of an information need is usually called a *query*.

Some IR systems allow for the formulation of *Boolean* queries, through the use of Boolean operators. An example of such a query would be: (Information AND Retrieval) NOT Evaluation. Boolean systems have been criticised as not allowing non-experienced users to formulate effective queries (Sparck Jones & Willett, 1997, p. 258). They have mostly been displaced by systems in which query formulation can essentially be made in the form of natural language text, with no need to use any specific operators (e.g. "*I want to find out about civil aviation in Greece*"). Such systems are based on *best-match*, or *similarity searching*, where a measure of query-document similarity is calculated for each document and for each query. The matching between documents and queries is discussed in section 2.3. Boolean queries are not further examined in the context of this thesis.

Once a query has been posed to an IRS, a similar processing to that for documents may take place, i.e. lexical processing and term-weighting. A query may also be expanded before retrieval is performed. A way to achieve this is by using a thesaurus to select terms that are semantically related to the ones present in the query. Such terms can then be added to the original query (Voorhees, 1994). A query can also be expanded after retrieval has taken place, by adding extra terms that appear in documents relevant to the query, but which were not included in the original query (Magennis & Van Rijsbergen, 1997). The expansion process can either be automatic, whereby the IRS selects the added terms, or interactive, where the expansion process is controlled by the user.

One form of automatic query expansion is *relevance feedback* (Figure 2.1). This is a technique that allows the user to better specify his information need by augmenting his initial query with information that appears in documents marked as relevant. In the relevance feedback process, the user marks documents returned by the IRS that are relevant to his query. Based on these assessments, the IRS selects terms to add to the query that would potentially retrieve more documents like the ones marked relevant, and less like the rest of the collection. A number of studies have examined the effect of relevance feedback methods on the effectiveness of IR systems, often reporting favourable results (Harman, 1992). However, an often reported problem

with both query expansion and relevance feedback is the unwillingness on behalf of the users to engage in the process (Ruthven *et al*, 2001).

## 2.3 Matching between documents and queries

In Boolean matching, an IRS simply finds the subset of the document collection that satisfies the logical requirements of the query expression, and presents it unranked to the user. Systems based on best match searching, on the other hand, rely on some method of comparing the query to each of the documents in the collection. The outcome of this comparison is a relevance score, that quantifies the likelihood of a specific document to be relevant to the query. The user is subsequently presented with a ranked list of documents, sorted in decreasing order of their relevance scores. It should be noted that in this way any number of documents may be presented to the user, by simply selecting that number of documents from the top of the retrieved list.

In the vector processing model (Salton & McGill, 1983), both documents and queries are represented as vectors in a multidimensional space. The dimensions of the space correspond to the indexing vocabulary of the document collection. Bollmann and Raghavan, (1993) questioned the validity of modelling documents and queries as objects in the same space, by presenting a number of examples that demonstrated potential problems of this modelling. However, it remains to be seen whether the authors' artificial examples may actually occur in realistic retrieval settings.

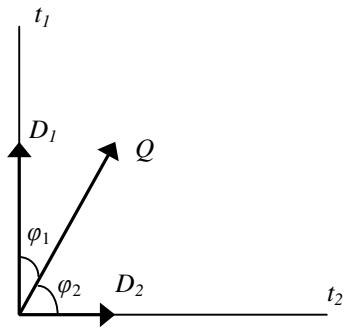


**Figure 2.2.** Document and query representations in the vector model

An example of document and query representation in this model is presented in Figure 2.2. In this simplified example, the vector space is presented in two dimensions corresponding to the two index terms $t_1$ and $t_2$. Documents $D_1$ and $D_2$ are represented in this space using each document's term weights as coordinates. Weights are assumed to correspond to term frequency weights: $D_1=\{t_1, t_1\}$ and $D_2=\{t_2\}$. The query $Q$ is represented in the same way in this space: $Q=\{t_1, t_1, t_2\}$. The angles between the vectors of $D_1$-$Q$ and $D_2$-$Q$ are also presented in this figure ($\varphi_1$ and $\varphi_2$ respectively).

In this space, one can define measures that quantify the similarity (or the distance) between points (i.e. documents and queries). The answer to the question of which documents best match the query, can then be given by those documents that are closest to the query according to a specific similarity (or distance) measure. There is ample literature on specific formulas for similarity and distance measures. A number of such formulas are presented in Appendix A, however, they are modified for the purposes of interdocument relationship calculation. For the use of such measures in IR, the books by Van Rijsbergen (1979), and Salton and McGill (1983), as well as the papers published by Norreault *et al*. (1981), Jones and Furnas (1987), Ellis *et al*. (1993), and Rorvig (1999) provide detailed information.

Perhaps the simplest way of comparing a document to a query is by counting the number of terms they have in common. Assuming that both the query and the document representations are expressed as vectors of length $n$ (where $n$ is the number of terms in the database), then:

$$Sim(D,Q) = \sum_{i=1}^{n} D_i Q_i \qquad (2.1)$$

This measure is called the *coordination level* matching function. Other measures, in contrast to the coordination level function, allow for the similarity to be normalised by the length of document and the query (e.g. Dice coefficient, Appendix A).

A measure that has been widely used in IR systems is the *cosine coefficient*. The reason for its popularity is likely to be based on this measure's consistency with a geometric interpretation of the vector model (Belew, 2000). It should however be noted that Raghavan and Wong (1986) suggested that the theoretical work on which Salton based the derivation of the vector model did not rely on the representation of documents in a geometrical space in which index terms correspond to its dimensions.

$$Sim(D,Q) = \frac{\sum_{i=1}^{n} D_i Q_i}{\sqrt{\sum_{i=1}^{n} (D_i)^2 . \sum_{i=1}^{n} (Q_i)^2}} \qquad (2.2)$$

The cosine measure (Equation 2.2) is a function of the angle between the document and the query vectors, and its value ranges between 0 (angle of $90^o$, unrelated vectors) and 1 (angle of $0^o$, identical vectors). In the example of Figure 2.2 the similarity between documents $D_1$ and $D_2$ is equal to zero, since the angle between the two vectors is $90^o$ (they contain no terms in common).

## 2.4 Evaluation of IR systems

Much effort has gone into the study of evaluation in IR, resulting in a number of methodologies for measuring the usefulness of IR systems. The inherent complexity of the task of evaluation makes it a challenging area. A reason for its complexity is that IRS evaluation combines issues from a number of diverse areas such as cognition, statistics, experimental design, system design, human computer interaction, etc. In this section I will present an outline of evaluation issues with emphasis on those aspects that are central to the research reported in this thesis.

There are a number of aspects of the IR process that can be evaluated. Such aspects may be the speed of an IRS, the level of user interaction it allows for, the style of presentation of information to users, etc. However, the aspect that is mostly used in IR research, and the one that is also central to this thesis, is the evaluation of the quantity of relevant documents an IRS system retrieves in response to a user query. This is one aspect of the *effectiveness* of an IRS (Cleverdon *et al*., 1966), for the computation of which numerous measures have been devised (Van Rijsbergen (1974a, 1979) provides further details on this issue).

| *Documents* | *relevant* | *not relevant* | |
|---|---|---|---|
| *retrieved* | $A \cap B$ | $\neg A \cap B$ | $B$ |
| *not retrieved* | $A \cap \neg B$ | $\neg A \cap \neg B$ | $\neg B$ |
| | $A$ | $\neg A$ | |

**Figure 2.3.** Calculation of precision and recall

The most commonly used measures of effectiveness are *precision* and *recall*. Precision is defined as the proportion of retrieved documents that are relevant, and recall as the proportion of relevant documents that have been retrieved. Referring to Figure 2.3, precision and recall can formally be defined as:

$$\text{Precision} = \frac{|A \cap B|}{|B|}, \quad \text{Recall} = \frac{|A \cap B|}{|A|},$$

where $|A \cap B|$ is the number of relevant and retrieved documents, $|B|$ is the number of retrieved documents, and $|A|$ is the number of relevant documents.

It is apparent from these definitions that the total number of relevant documents in a collection must be known in order for recall to be calculated. However, because of the amount of effort and time required on behalf of users, this is not possible in most operative cases. To facilitate the evaluation of IR systems, a number of test collections have been built for which this value (total number of relevant documents) has been determined (Sparck Jones & Van Rijsbergen, 1976). These are document collections that are accompanied by a set of queries, and a set of *relevance assessments* for each query, i.e. lists of documents in the collection that are judged by domain

experts to be relevant to each query. Test collections have given IR researchers the opportunity to efficiently evaluate their experimental approaches, and furthermore, to compare the effectiveness of their system to that of others.

In Figure 2.4 a typical recall-precision (R-P) graph is presented. Such graphs are typical of the way retrieval effectiveness results are conveyed and publicised. In Chapter 7 of Van Rijsbergen's book, (1979), a number of issues pertaining to the derivation of such graphs are presented in detail. The most typical method of deriving a R-P graph is to calculate precision values for certain recall points, i.e. after certain numbers of relevant documents have been retrieved. The recall points normally used are 0, 0.1, 0.2, 0.3, ..., 1. Recall and precision values are then typically averaged over the set of queries of a test collection, and as such are also presented in Figure 2.4. Although Figure 2.4 only presents the precision and recall values of a single IRS (or of a single IR strategy), it is more typical to plot the effectiveness of multiple IR systems in one graph, so as to be able to compare their effectiveness.

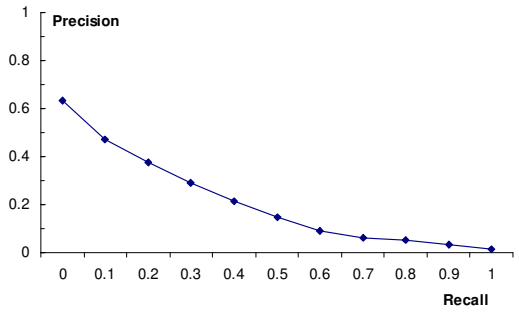


**Figure 2.4.** A recall-precision graph

In Table 2.1 basic statistics for a number of IR test collections are presented. Initially test collections comprised a few thousand documents, and therefore it was possible to consider exhaustive relevance assessments (i.e. for each query, to scan all the retrieved documents to find the relevant ones). However, with the advent of increasingly larger collections, especially since the start of the Text Retrieval Conferences (TREC) (Harman, 1993), exhaustive judgements have become an impossibility. The test collections used for TREC consist of hundreds of thousand documents that occupy many gigabytes of disk space. For a comparison with smaller collections, statistics for the collection used in the first TREC conference (TREC-1) are presented in Table 2.1.

A technique called *pooling* has been extensively used in these cases (Harman, 1993). The essence of this technique is to submit a test collection query to a number of IR systems, to combine the top-ranked documents of each of the systems, and to judge the relevance of this combined set. This technique can be effective when the IR systems that are used retrieve relevant documents that are representative of all relevant documents available (Harman, 1993).

The type of evaluation described above depends on relevance judgements provided by some expert judges, based on *topical* (or a*lgorithmic*) relevance (Saracevic, 1975). Topicality associates relevance with the presence of query terms in documents. Recent research (Schamber *et al*., 1990; Barry, 1994) has demonstrated that relevance is a multidimensional concept, and that topicality is only one such dimension. Driven by this observation, researchers such as Borlund (Borlund & Ingwersen, 1997), and Reid, (2000), have worked on the definition of evaluation methodologies that judge the utility of IR systems on other dimensions of the concept of relevance. A special track of TREC, the interactive track (Hersh & Over, 2001), has also been formed, in an attempt to provide a framework for the evaluation of interactive IR systems.

| *Collection* | *Number of documents* | *Number of queries* |
|---|---|---|
| Cranfield | 1400 | 225 |
| CACM | 3204 | 52 |
| CISI | 1460 | 35 |
| Evans | 2542 | 39 |
| Harding | 2472 | 65 |
| INSPEC | 12,684 | 84 |
| Keen | 800 | 63 |
| LISA | 6004 | 35 |
| MEDLARS | 1033 | 30 |
| NPL | 11,429 | 93 |
| TIME | 423 | 83 |
| TREC-1 | 742,611 | 100 |
| UKCIS | 27,361 | 182 |

**Table 2.1.** IR test collections

It should finally be noted that the extraction of scientific conclusions based on the outcome of IR experiments is an issue that involves a large number of complications, such as the choice of appropriate measures of performance, the presentation of experimental comparisons, the statistical testing of experimental results, etc. In order to address these issues, a number of published articles provide researchers with a methodology on which to base the extraction of scientific inference from IR experiments (Robertson, 1981; Keen, 1992).

## 2.5 Document clustering

According to best-match IR systems, if a document does not contain any of the query terms then its similarity to the query will be zero, and this document will not be retrieved in response to the query. Document clustering offers an alternative file organisation to that of best-match retrieval, and it has the potential to address this issue (and therefore to increase the effectiveness of an IR

system). Documents are organised in clusters based on their similarity, as defined in terms of their content overlap.

The first suggestions that clustering could improve the effectiveness of an IR system were made by Jardine and Van Rijsbergen (1971). The effectiveness of an IRS was expected to increase since the file organisation, and any strategy to search it, takes into account relationships that hold between the documents in a collection (Croft, 1980). For example, a relevant document may be ranked low in a best-match search because it may lack some of the query terms. In a clustered collection, this relevant document may be clustered together with other relevant items that do have the required query terms, and could therefore be retrieved through a clustered search (Croft, 1978).

The *Cluster Hypothesis* is fundamental to the issue of improved effectiveness; it states that relevant documents tend to be more similar to each other than to non-relevant documents, and therefore tend to appear in the same clusters (Jardine & Van Rijsbergen, 1971). If the cluster hypothesis holds for a particular document collection, then relevant documents will be well separated (i.e. grouped separately) from non-relevant ones.

In the following chapter, I discuss the application of document clustering to IR in detail.

# Chapter 3

# Document Clustering for IR: Background

## 3.1 Introduction

C*luster analysis* is a multivariate statistical technique that allows the identification of groups, or clusters, of similar objects in a space that is typically assumed to be multi-dimensional. Groups of objects are formed in such a way that objects in the same cluster are similar to one another and dissimilar to objects in other clusters (Gordon, 1987). Clustering is a task that has been practised by humans for thousands of years (Willett, 1988; Kural, 1999), and it has been fully automated in the last few decades due to the advancements in computing technology (Willett, 1988).

There is often a confusion regarding the usage of the terms cluster analysis and *classification*. Watanabe (1969, p. 381) and Willett (1988), among others, have distinguished between the two. A clustering task involves grouping objects, based on a defined set of properties, into classes according to the strength of interobject relationships (i.e. the classes have to be discovered). For a classification task, on the other hand, a sample set of objects are first placed in some classes (typically manually, as part of a training stage), and then new samples are expected to be placed in the existing classes imitating the classification demonstrated at the training stage. It should however be noted that in the early literature, classification implied the task of assigning objects to clusters, and *diagnosis* implied the task of assigning a new incoming object to one of the existing clusters (Jardine & Sibson, 1971). The terminology has changed over the years, especially in the field of IR. Therefore, the term classification will not be used as an alternative for *clustering* in this thesis.

Cluster analysis techniques have long been applied to scientific fields such as life sciences (biology, zoology), medical sciences (psychiatry, pathology), social sciences (archaeology, sociology, criminology), earth sciences (geography, geology) and engineering sciences (pattern recognition, cybernetics) (Anderberg, 1973). Specific applications of cluster analysis range from clustering DNA structures for gene expression analysis (Hartuv *et al*., 1999) to clustering single-

malt whiskies based on characteristics such as peatttiness, sweetness, etc. (Wishart, 1998). Consequently, the literature on cluster analysis is both voluminous and diverse. A large number of books and review articles, that cover almost any field of endeavour of cluster analysis, have been published. Such readings include (Macnaughton-Smith, 1965; Cole, 1969; Cormack, 1971; Jardine & Sibson, 1971; Anderberg, 1973; Sneath & Sokal, 1973; Hartigan, 1975; Van Ryzin, 1977; Gordon, 1987; Jain & Dubes, 1988; Everitt, 1993; Jain *et al*., 1999; Theodoridis & Koutroumbas, 1999).

As far as its application to information retrieval is concerned, cluster analysis has been used both for *term* (or *keyword*) *clustering*, and for *document clustering*. Term clustering (Doyle, 1964; Sparck Jones, 1971; Lewis, 1992; Wulfekuhler & Punch, 1997) is performed on the basis of the documents in which terms co-occur, and it allows each term in a document, or query, to be replaced by the representation (i.e. collection of indexing terms) describing the cluster to which this term belongs. Application areas for term clustering include *query expansion* (Sparck Jones, 1971; Minker *et al*., 1972; Van Rijsbergen *et al*., 1981), automatic thesaurus construction (Crouch & Yang, 1992), and thesaurus linking (Amba *et al*., 1996). Peat and Willett (1991) have raised questions regarding the effectiveness of the use of keyword co-occurrence data (of the type that term clustering operates upon). Keyword clustering falls beyond the aims of this thesis, and subsequent discussion on cluster analysis in the context of IR will be restricted to document clustering.

Document clustering can be performed on the basis of terms shared between documents, or on the basis of citations shared between documents. The latter form, which is typically referred to as *co-citation analysis* (Small, 1999; Popescul *et al*., 2000), is used in order to provide insights into the nature of the literature of a specific scientific field. In this chapter I will focus on the former type of clustering.

Document clustering typically operates based on the notion of *interdocument similarity.* The set of terms shared between a pair of documents is typically used as an indication of the pair's similarity. According to Van Rijsbergen (1979), one of the first researchers to suggest the use of automatic clustering for IR was Good (1958). Document clustering has traditionally been applied statically, to an entire document collection, before querying (*static clustering*). An alternative application of clustering is to only cluster documents that have been retrieved by an IR system in response to a query (*post-retrieval clustering*) (Preece, 1973). Under post-retrieval clustering the resulting groups of documents are likely to be different for different queries. Two broad types of document clustering have been mainly used in IR, *partitioning* and *hierarchic*.

Partitioning methods cluster a set of $N$ documents into a single organisation of $k$ mutually exclusive clusters, where $k$ is either specified *a priori*, or is determined as part of the clustering method. The computational requirements of partitioning methods are low, typically in the order of

*O*(*N*) to *O*(*NlogN*) for the clustering of *N* documents (Willett, 1988). This had as a consequence that in the early period of research in document clustering, partitioning methods were favoured, as they offered the potential to increase the efficiency of an IR system (Rocchio, 1966; Salton, 1971).

The central idea in most of the partitioning methods is to choose some initial partition of the documents and then to alter cluster memberships so as to obtain a better partition[4] (Anderberg, 1973). The number of possible partitions of *N* documents in *k* clusters (especially for large values of *N*) makes a complete enumeration hard (Willett, 1988) and an optimal solution impossible[5], and thus heuristic methods are employed in order to find an approximate solution. As a consequence, partitioning methods suffer on a theoretical basis as they generally require a great number of arbitrarily determined experimental parameters (e.g. cluster membership, number of clusters, cluster size), and may depend on the order in which the documents are processed (Salton & Wong, 1978; Willett, 1988).

Early experimentation showed that the effectiveness of searches based on document partitions is significantly inferior to that based on searches of the unclustered file[6] (Salton, 1971). Most recent applications of partitioning methods to IR (Cutting *et al*., 1992; Hearst & Pedersen, 1996; Silverstein & Pedersen, 1997; Zamir & Etzioni, 1998) have also focused on efficiency aspects for on-line browsing tasks, rather than on the effectiveness of the methods. Some partitioning type algorithms that are much publicised in the IR literature are the *C3M* algorithm (Can & Ozkarahan, 1990), the *Buckshot* and *Fractionation* algorithms employed by the Scatter/Gather system (Cutting *et al*., 1992; Silverstein & Pedersen, 1997) and the *suffix tree clustering* algorithm (*STC*) proposed by Zamir and Etzioni (1998).

The type of clustering employed in this thesis is hierarchic, perhaps the most commonly used type of clustering in IR (Willett, 1988). This is a choice based on the more sound theoretical basis of hierarchic clustering. Jardine and Sibson (1971), Salton and Wong (1978) and Van Rijsbergen (1979) have identified three strengths of hierarchic methods. Firstly, such methods are theoretically attractive since they do not depend on the order in which documents are processed. Secondly, they are well-formed, in the sense that a single classification will be derived from a given set of documents. And finally, hierarchic methods are stable, since small changes in the original document vectors will result in small changes in the resulting hierarchies. Although the

[4] A typical way of deciding cluster membership is by minimising a cost function. For example, the popular *k*-means method (Kaufman & Rousseeuw, 1990) is based on minimising the sum-of-squares function.

[5] Garey and Johnson, (1979, p 281), have defined the problem as NP-hard.

[6] Scheibler and Schneider (1985), and Milligan and Cooper (1987), report on some studies that compare partitioning and hierarchic methods, though not specifically for IR. Larsen and Aone (1999) and Steinbach et al. (2000) do so for the specific area of data mining.

reported research can also be applied to other types of clustering, this thesis will concentrate on hierarchic clustering alone.

### 3.1.1 Hierarchic clustering: an outline

Hierarchic clustering was introduced to IR by Jardine and Van Rijsbergen (1971), based on its potential to increase the effectiveness of IR systems (section 2.5). The cluster hypothesis is fundamental to the issue of improved effectiveness. A more thorough discussion of the cluster hypothesis and its implications for clustering effectiveness is presented in Chapter 5. I also discuss the cluster hypothesis is this chapter, in section 3.6.1.

The basic steps that characterise the clustering process (not only the hierarchic clustering process) are the following (Rasmussen, 1992; Theodoridis & Koutroumbas, 1999) :

- *Document representation*: The attributes that will represent each document have to be selected and appropriately weighted.
- *Association measure*: Such a measure defines how similar or dissimilar two documents are. The choice of a particular type of measure may affect the clustering output.
- *Clustering method*: A specific method that will try to effectively structure the document space needs to be selected and applied.
- *Cluster representation*: Clusters need to be represented, both for retrieval purposes, and for purposes of succinctly presenting their contents to users.
- *Validation of the results*: Once a clustering structure has been obtained, its correctness needs to be verified. This is usually done using appropriate tests.

The purpose of this chapter is to give an overview of issues that relate to each of these steps, and to emphasise those aspects of the clustering process that are particularly related to the work reported in this thesis. In section 3.2 the selection and weighting of document attributes for the purpose of document clustering is discussed. Section 3.3 deals with the measurement of interdocument relationships, followed by section 3.4 that presents the details of some hierarchic clustering methods. Sections 3.5 and 3.6 discuss cluster representation and cluster validation issues respectively, and section 3.7 presents some recent trends in document clustering. Section 3.8 provides some reflections on clustering research over the past thirty years, and outlines the specific aspects of clustering research on which the work of this thesis focuses. Section 3.9 concludes the chapter by providing a summary of the main issues discussed.

## 3.2 Document representation

The first step in the clustering process is to decide on the type and number of variables that describe each document. Documents are typically represented by a vector in a *n*-dimensional space, where *n* corresponds to the number of terms forming the indexing vocabulary of the database[7]. Single terms have been used as indexing features for clustering in the majority of the research reported in the literature. Recently, Hatzivassiloglou et al. (2000) and Maarek et al. (2000), have augmented document representations with phrasal units by employing different levels of linguistic analysis. Results obtained from these studies have demonstrated some small benefits from the use of linguistically motivated features. In the context of this thesis however, document representations for clustering will be restricted to single-term indexing units.

Assuming such a representation, two questions naturally arise, and will be discussed in this section: which terms does one choose to represent a document, and how does one weight the relative importance of these terms for the purposes of clustering.

### 3.2.1 Exhaustivity of document representations

Traditionally, before clustering is applied, documents in a collection undergo a form of lexical processing similar to that described in section 2.2. Such processing usually involves case normalisation, removal of terms that appear in a stop-list, and application of a stemming algorithm. Typically each document $X$ is then represented as a vector $X = \{x_1, x_2, \ldots, x_n\}$, where $n$ is the number of terms that constitute the indexing vocabulary of the document collection. All terms that belong to the indexing vocabulary of the entire document collection, and that occur within a document, are typically used in that document's indexing representation. This representation is the most *exhaustive* representation of a document (Van Rijsbergen, 1979).

One of the studies that investigated the effect of indexing exhaustivity on clustering effectiveness was carried out by Shaw (1990, 1991, 1993). Shaw clustered the Cystic Fibrosis database, comprising 1239 documents and 100 queries, by using the single link method. He measured the variation of optimal cluster effectiveness (see section 4.3.4) as a function of the exhaustivity of document representations. Different levels of indexing exhaustivity were determined by setting a term weight threshold (TW). A term is retained if its weight exceeds the threshold TW (term weights were *idf* weights, normalised to vary in the range 0-999). Thus, for each representation, as TW is increased the representation becomes less exhaustive and more specific.

[7] Barry (1994) noted a number of factors which affect relevance, and which could be included in document representations (e.g. cost, obtainability, recency). However, such attributes are not commonly measured, and will not be considered in this thesis.

The main conclusion from Shaw's work was that clustering effectiveness, for the single link method, varies significantly as a function of indexing exhaustivity. Retrieval effectiveness, in general, increased as the exhaustivity of the representations decreased from the most exhaustive to an optimal representation. Such optimal representations tended to appear at relatively low levels of indexing exhaustivity, where only a fraction of the indexing terms are present in the representations. The most exhaustive representations, which are commonly used in clustering, displayed low levels of effectiveness.

Burgin, (1995), extended Shaw's work by including four other clustering methods (complete link, group average link, Ward and weighted average methods) in addition to single link, and three new document collections (Medlars, Cranfield, and Time, see Table 2.1). Burgin followed the same experimental procedure as Shaw, aiming to test whether Shaw's results would be applicable to other experimental environments. The results of Burgin's experiments confirmed Shaw's findings, in that the effectiveness of the single link method varies as a function of indexing exhaustivity. However, the results failed to confirm similar patterns for the other four clustering methods. The conclusion from these studies is that only the effectiveness of the single link method varies significantly as a function of indexing exhaustivity.

### 3.2.2 The effect of term-weighting

Once the terms that represent each document have been selected, one must somehow weight their relative importance. It is well established in IR research that weighting terms according to their occurrence within documents and within the entire document collection increases effectiveness (Salton & Buckley, 1988). Binary representations are often associated with poor retrieval effectiveness. However, it has not been established equally clearly whether the advantages of term weighting apply to cluster-based systems. Sneath and Sokal, (1973), advised for the use of binary vectors for feature representation, mainly because of the simplicity of this approach. They did not believe that additional weighting information could significantly affect the quality of the resulting clustering.

Willett (1983) investigated this issue for the field of IR, by considering five term weighting methods (including binary representations), and three document collections (Keen, Cranfield, and Evans, Table 2.1) that were clustered by the single link method. Willett measured the retrieval effectiveness of searches carried out on the resulting hierarchies. Based on the results of this study, Willett concluded that there does not seem to be a consistent and significant improvement in effectiveness introduced by the use of weighted term vectors over the use of binary vectors. It should, however, be mentioned that Willett's study included only a single clustering method, and three test collections of a small size (which is typical of research of that time). His results should

therefore be viewed with caution, and not be extended beyond the specific experimental environment in which they were generated.

Can and Ozkarahan (1990) also offer some insight into the effect of binary vs. weighted document representations. Their study involved the C3M clustering method which is of a partitioning type. In the first part of their study the authors used binary and weighted versions of document vectors. The weighted version of vectors corresponded to term frequency data. The authors estimated the number of clusters expected to be present in each of the two collections that they used in their experiments (INSPEC and TODS214[8]). They noted that the clustered structure generated using binary vectors had a smaller average deviation from the expected values. Based on this observation, the authors conjectured that there is little to gain from the use of term frequency information in document representations.

The second part of their experiments involved measuring the effectiveness of cluster-based searches. A number of experimental parameters were varied (e.g. the term weighting functions for documents, queries, and cluster representatives, the length of cluster representatives), and the effect that the variation had on the effectiveness of the searches was examined. The term weighting functions that were used were ones that had demonstrated the highest effectiveness in non-cluster based retrieval experiments carried out by Salton and Buckley (1988). Binary representations were not included. The results for both databases showed a considerable variation in retrieval effectiveness that depended on the weighting functions. The limited scope of the experiments (two databases), and the confounding effect of a number of experimental parameters (e.g. cluster representatives, etc.) may have affected the reported results.

There does not seem to be substantial experimentally-founded evidence which recommends for, or against, the use of a specific term-weighting function for document clustering. Sneath and Sokal suggested the use of binary weights, on the grounds of simplicity, about thirty years ago. The reasons that may have led them to attribute importance to simplicity for the efficient implementation of clustering are most likely obsolete today - abundance of computing power and time-efficient algorithms are the norm. In the absence of other evidence, there seems to be no reason to reject term-weighting approaches that have proven themselves in non-cluster based retrieval experiments, such as those reported by (Salton & Buckley, 1988). Such measures include variants of the popular *tf-idf* function. This conclusion can only remain valid until future research provides evidence that suggests otherwise.

[8] The TODS214 database contains the papers published by the ACM in the journal Transactions on Database Systems during March 1976 to September 1984. The database consists of 214 documents and 58 queries.

## 3.3 Measuring interdocument relationships

Having established an appropriate representation for the document set to be clustered, one needs to measure the degree of resemblance of all possible pairs of documents that belong to this set. To this end, a large number of measures that quantify the resemblance between objects have been devised. Sneath and Sokal (1973) categorise such measures in four main classes: *association*, *dissimilarity*, *probabilistic*, and *correlation coefficients*. The use of probabilistic and correlation coefficients in document clustering has been limited, and thus the majority of the literature refers to the former two categories, namely association (or similarity) and dissimilarity coefficients.

A large number of these measures have been used in cluster analysis. Cormack (1971), Anderberg (1973), Sneath and Sokal (1973), Hubálek (1982), Kaufman and Rousseeuw (1990), among others, provide detailed discussions on the use of such measures in cluster analysis. For the particular context of document clustering, the most commonly used coefficients can be found in (Van Rijsbergen, 1979; Salton & McGill, 1983; Willett, 1988; Ellis *et al*., 1993). Although not specifically in the context of document clustering, Jones and Furnas (1987) give a geometric interpretation of a number of measures that are commonly used in IR. In Appendix A some of the most popular measures are presented. Before discussing the effect that the choice of a particular measure may have on the effectiveness of document clustering, a formal definition of such measures is appropriate.

### 3.3.1 Formal definitions

For all subsequent discussion in this section it will be assumed that documents are represented as vectors in a *n*-dimensional space, where *n* is the size of the indexing vocabulary. Therefore, a document $x_i$ is assumed to comprise *n* indexing terms that are represented either by binary (absence / presence) or by real-valued weights: $x_i = (x_{i1}, x_{i2}, ..., x_{in})$. If *X* is the set of documents to be clustered, then a *distance* coefficient is a function $d : X \text{ x } X \rightarrow R$, where *R* is the set of non-negative real numbers. Such a function *d*, in general, satisfies the following axioms:

Reflexivity: $d(x,x) = 0$

Symmetry: $d(x,y) = d(y,x)$

Triangular inequality: $d(x,y) \leq d(x,z) + d(z,y)$,

where *x*, *y*, and *z* are all documents belonging to set *X*. A distance is a particular type of dissimilarity function. A distance function that satisfies these three axioms is a *metric* function. An *ultrametric* function $\delta$ is one that satisfies the first two axioms and: $\delta(x,y) \leq \max[\delta(x,z), \delta(z,y)]$ (Diday & Simon, 1976).

A similarity measure $s$ can also be defined as a function $s : X \times X \rightarrow R$. A similarity function $s$ is metric if it satisfies the reflexivity and symmetry axioms, and in addition:

$$s(x, y)s(y, z) \leq [s(x, y) + s(y, z)]\, s(x, z)$$ (Theodoridis & Koutroumbas, 1999)

Typically, similarity and distance functions are normalised so that their values fall within the range of 1 and 0. Intuitively, for similarity functions the greater the similarity value the more similar the two documents are. For distance functions the opposite holds, i.e. the smaller the value of the function the more similar the two documents are. For the rest of this chapter I will only consider similarity measures.

It must be noted that Tversky (1977) questioned the validity of the metric assumptions. He presented experimental evidence to suggest that similarities between objects can be asymmetric, i.e. $s(x,y) \neq s(y,x)$. Tversky argues that humans tend to select the most salient stimulus as a referent and the less salient stimulus as an object when judging inter-stimuli similarities. Thus, we are more likely to say that "the portrait resembles the person" rather than "the person resembles the portrait". To the best of my knowledge, the validity of Tversky's assertions has not been investigated in the context of IR.

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | | | | |
| $x_2$ | 0.6 | 1 | | | |
| $x_3$ | 0.4 | 0.8 | 1 | | |
| $x_4$ | 0.1 | 0.5 | 0.7 | 1 | |
| $x_5$ | 0.1 | 0.2 | 0.2 | 0.3 | 1 |

**Figure 3.1.** A similarity matrix

If a document collection comprises $N$ documents, then a $N$x$N$ matrix $S(X)$ is needed to store all interdocument association values. This matrix is a triangular matrix, whose element $s_{ij}$ is the measure of association between documents $x_i$ and $x_j$. The diagonal elements of the matrix are equal to the maximum value that the used similarity measure can yield. Also, since similarity functions are symmetric ($s_{ij} = s_{ji}$), the calculation of all pairs of associations requires $\frac{N \times (N-1)}{2}$ operations, which makes similarity calculation $O(N^2)$ dependent and consequently computationally expensive for large datasets. Figure 3.1 shows an example of a 5x5 similarity matrix. The elements above the diagonal have been purposefully left blank since the matrix is symmetric.

In an attempt to make similarity calculations efficient for large document collections, Croft (1977) proposed a method for the calculation of coefficients that was based on an inverted file structure (section 2.2). The inverted file was used to determine documents that had no terms in common

with a given document, and thus to avoid calculating the coefficients between such pairs of documents (since the similarity value would be equal to zero).

Willett (1981), introduced an improvement to Croft's algorithm by proposing a method that is able to identify all non-zero-valued coefficients for a given document simultaneously. The time-efficiency of Croft's algorithm is affected by increases in the indexing exhaustivity (i.e. the mean number of terms per document), when a large number of non-zero-valued coefficients may have to be calculated several times. Willett demonstrated that the performance of his algorithm does not deteriorate as the mean number of terms per document increases, as each document description is processed only once for the calculation of all the similarities that involve it.

### 3.3.2 Choice of a particular measure

Given the large number of measures available, the question naturally arises of the choice of the most appropriate one(s) for the purpose of document clustering. Sneath and Sokal (1973), suggested the use of the simplest type of measure possible, out of consideration for ease of interpretation. In his book, Van Rijsbergen (1979) advised against the use of any measure that is not normalised by the length of the document vectors under comparison. A further remark made by Van Rijsbergen, and also by Sneath and Sokal (1973), is that the various association and distance measures are monotone with respect to each other. Consequently, a clustering method that depends only on the rank ordering of the resemblance values would give similar results for all such measures.

The need to normalise resemblance measures by the length of the document vectors was experimentally verified by Willett (1983)[9] in one of the few studies that attempt to establish the relationship between clustering effectiveness and choice of similarity measure. In this study Willett used the single link method, three document collections (Keen, Cranfield, and Evans), four similarity measures (inner product, Tanimoto coefficient, cosine coefficient, and the overlap coefficient) and five term weighting schemes. Experimental results confirmed the poor effectiveness of non-normalised measures, and also showed little variation in the effectiveness of hierarchies obtained with normalised measures. The cosine coefficient generated clusterings that demonstrated a slightly better retrieval effectiveness than other measures.

Some further evidence for the inappropriateness of non-normalised resemblance measures was offered (though not purposefully and explicitly, but rather accidentally) by Griffiths *et al*. (1984). Griffiths and his colleagues used the Hamming distance and the Dice coefficient to measure interdocument relationships. It should be noted that the former measure is not normalised by

[9] This study is part of the research reported in section 3.2.2 on the effect of term weighting on clustering effectiveness.

document length, while the latter is. The quality of the structure of the resulting hierarchies, as well as the effectiveness of cluster-based searches, were measured in a series of experiments. In both cases, the results obtained with the Hamming distance proved to be significantly inferior to those obtained with the Dice coefficient for all experimental conditions. Despite the nature of the results, the authors did not significantly acknowledge the normalisation effect.

Kirriemuir and Willett, (1995), applied hierarchic clustering to the output of a database search using four clustering methods and five similarity and distance measures, among which was the cosine and Jaccard coefficients, and the normalised Euclidean distance. Measures of quality[10] of the resulting hierarchies revealed that the cosine and Jaccard coefficients led to the most effective clusterings. One of the findings of this study was that merging of unrelated documents usually occurred for short documents. This prompted the authors to raise questions about the effectiveness of the normalising factors of the measures they used. However, they do not investigate this issue further in their study.

Some further research that compared different measures for the calculation of interdocument relationships was conducted by Rorvig (1999). Rorvig was mainly interested in investigating how a set of similarity measures would perform as part of a visual information retrieval interface, i.e. how successfully the measures would convey the known structure of the document space. Five TREC topics were selected, and the document sets used for each topic comprised between 421 and 586 TREC documents. The measure of quality of the similarity measures was defined as the visual separation of relevant and non-relevant documents for each topic. For visualisation purposes, multidimensional scaling was employed, using three different scaling assumptions (ordinal, interval, and maximum likelihood)[11]. The study included five similarity measures (Dice, Jaccard, cosine, overlap and asymmetric), and its results suggested that the cosine and overlap coefficients were more successful in recovering structure.

The cosine coefficient and the Euclidean distance are two measures that have been commonly used for the measurement of interdocument proximity in a document vector space. Furnas and Jones (1987) analysed the properties of a large number of similarity measures, including the cosine coefficient. They noted that the comparison of documents based on their angle in a vector space approximates a comparison that is based on their topical content, as this is expressed through within-document term relationships.

Dubin (1996) has noted that angular measures are more sensitive to relative attribute weights, as opposed to distance measures that are more sensitive to absolute weights. The use of Euclidean

[10] The measures were tailored to the specific application in which Kirriemuir and Willett were interested (identification of duplicates in a news database), and will not be discussed here.

[11] The details of the MDS scaling assumptions will not be elicited here, as they fall outside the scope of this chapter.

distance for clustering has been criticised by Willett (1988), who notes that according to this measure two documents can be regarded as highly similar even if they do not have any terms in common. Zhang and Rasmussen (2001) have recently developed a new similarity measure that combines Euclidean distance and angular similarity (i.e. the cosine coefficient). This measure is proposed for the matching of queries against documents, but it could as well be applied to the calculation of interdocument similarities. Its effectiveness in either task remains to be investigated.

Ellis *et al*. (1993) in their comprehensive article on the measurement of interdocument similarity in textual databases, examined the theoretical properties of a large number of commonly used measures. They concluded that the historical attachment to the association coefficients provided by the Dice and cosine formulae seems to be in no need of revision. Given the evidence that the research reviewed in this section has offered, this would seem a valid conclusion to make.

The issue of similarity is central to this thesis, since one of the main aims of this work is to challenge the static use of similarity in IR, and to provide experimental evidence for the applicability of query-sensitive similarity measures that take the query into account when calculating interdocument relationships. In Chapter 5 I extensively discuss this issue (section 5.3), by elaborating on the static nature of interdocument similarities, and by viewing the issue of similarity in relation to the cluster hypothesis.

## 3.4 Hierarchic clustering methods

Hierarchic clustering methods result in tree-like classifications in which small clusters of objects (i.e. documents) that are found to be strongly similar to each other are nested within larger clusters that contain less similar objects.

Let us assume that $X$ is the document set to be clustered, $X = \{x_1, x_2, \ldots, x_N\}$. Each document $x_i$ is a $n$-dimensional vector, where each dimension typically corresponds to an indexing term. In section 3.2 I discussed issues pertaining to the selection of indexing terms, and the effect of term weighting on clustering effectiveness.

A clustering of $X$ in $m$ sets can be defined as $R = \{C_1, C_2, ..., C_m\}$, so that the following conditions are satisfied:

- Each cluster $C_i$ contains at least one document: $C_i \neq \emptyset$, $i=1, \ldots, m$
- The union of all clusters is the set $X$: $\bigcup_{i=1}^{m} C_i = X$
- No two clusters have documents in common: $C_i \bigcap C_j = \emptyset, i \neq j, i, j = 1,..., m$

A clustering $R_1$ that contains $k$ clusters is said to be *nested* in the clustering $R_2$, which contains $r < k$ clusters, if each cluster in $R_1$ is a subset of a cluster in $R_2$, and at least one cluster of $R_1$ is a proper subset of $R_2$ (Theodoridis & Koutroumbas, 1999). For example, the clustering $R_1 = \{\{x_1, x_3\}, \{x_4\}. \{x_2, x_5\}\}$ is nested in $R_2 = \{\{x_1, x_3, x_4\}, \{x_2, x_5\}\}$. On the other hand, $R_1$ is not nested within $R_3 = \{\{x_1, x_4\}, \{x_3\}, \{x_2, x_5\}\}$ (examples taken from Theodoridis and Koutroumbas (1999, p. 403)).

Hierarchic methods are divided into two broad categories, *agglomerative* and *divisive*. An agglomerative strategy proceeds through a series of ($N$-1) merges, for a collection of $N$ documents, and results in clusterings building from the bottom to the top of the structure. In a divisive strategy, on the other hand, a single initial clustering is subdivided into progressively smaller groups of documents (Van Rijsbergen, 1979). Divisive methods (Tanaka *et al*., 1999) normally result in *monothetic* classifications, where documents in a given cluster must contain certain terms in order to gain membership (Sneath & Sokal, 1973; Van Rijsbergen, 1979; Gordon, 1987). In *polythetic* clusterings, on the other hand, no specific terms are required for membership in a cluster, and such structures are usually the result of agglomerative methods. For information retrieval, polythetic clusterings are preferred (Van Rijsbergen, 1979; Willett, 1988), and consequently hierarchic agglomerative clustering methods (HACM) have prevailed in the field (Willett, 1988).

Agglomerative methods can be distinguished on the basis of whether they are founded on concepts of matrix theory, or on concepts of graph theory (Anderberg, 1973; Theodoridis & Koutroumbas, 1999). In this section I will concentrate on methods that are based on matrix theory, as they are the ones most commonly used in IR (Willett, 1988). The input to an HACM of this type is the similarity matrix S(X) that contains the values for all interdocument associations (see section 3.3.1).

Hierarchic agglomerative methods usually follow the following generic procedure (Murtagh, 1983):

1. Determine all interdocument similarities
2. Form a cluster from the two closest objects or clusters
3. Redefine the similarities between the new cluster and all other objects or clusters, leaving all other similarities unchanged
4. Repeat steps 2 and 3 until all objects are in one cluster

The various agglomerative methods available differ on the way that they implement step 3 of the above procedure. At each step $t$ of the clustering process, the size of the similarity matrix $S(X)$ (which initially is $N$x$N$) becomes ($N$-$t$) x ($N$-$t$). The matrix $S_t(X)$ of step $t$ of the process is derived

from the matrix $S_{t-1}(X)$ by deleting the two rows and columns that correspond to the newly merged documents (or clusters), and by adding a new row and column that contain the new similarities between the newly formed cluster and all unaffected (from step $t$ of the process) documents or clusters.

The output of a hierarchic clustering method can be presented in the form of a *dendrogram* (Jardine & Sibson, 1971) (Figure 3.2). A dendrogram is usually represented as a tree with numeric levels associated to its branches. The numeric values are the similarity levels at which clusters are formed. At any similarity level, one can draw a line perpendicular to the similarity axis. In this way, each branch of the tree that is cut by the line represents a cluster consisting of elements in the subtree rooted at that branch. At the lowest level of similarity, all documents are in a single cluster.

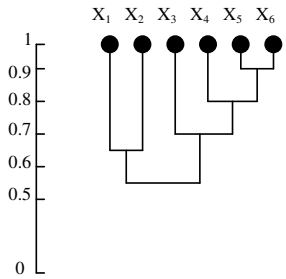


**Figure 3.2.** A similarity dendrogram

Although the efficiency of the various clustering methods is not of primary importance in this thesis, for reasons of completeness I will also refer to the efficiency of commonly used algorithms that implement the various methods. Readings that offer significant amount of detail on aspects of efficiency include (Hartigan, 1975; Croft, 1977; Murtagh, 1984a; Voorhees, 1985a, 1986; Willett, 1988). It should also be noted that efficiency is really a property of the algorithm that implements the clustering method (Jardine & Sibson, 1971). Van Rijsbergen (1979) noted that it is sometimes useful to distinguish the cluster method from its algorithm, but also acknowledged that in the context of IR this distinction is less important since many cluster methods are defined by their algorithms.

Most hierarchic agglomerative algorithms operate on the stored matrix approach (Hartigan, 1975), where the similarity matrix is kept in memory. In section 3.3.1 I discussed issues relating to the efficient calculation of interdocument relationships. Therefore, a typical algorithm that clusters $N$ documents by using the stored matrix approach has storage requirements of $O(N^2)$ (for the storage of the similarity matrix), and time requirements of $O(N^3)$ since the matrix is searched $N$-1 times.

In the following paragraphs I will present four hierarchic clustering methods that have been extensively used in IR research in the past, and that will also be used in the research reported in this thesis. These are the single link, complete link, group average, and Ward's methods. Jardine

and Sibson (1971), Anderberg (1973), Hartigan (1975), and Späth (1980) offer a more in-depth analysis of the various methods. Willett (1988) also presents a number of algorithms that are used in IR to implement the various hierarchic methods.

### 3.4.1 Single link

In the single link method the similarity between two clusters is the maximum of the similarities between all pairs of documents such that one document is in one cluster and the other document is in the other cluster (Voorhees, 1985a). For example, if at some stage clusters *i* and *j* have merged, then the similarity between the new cluster (labelled *p*) and some other cluster *r* is determined as follows: $S_{pr} = max(S_{ir}, S_{jr})$. In graph theoretical terms, the clusters at some similarity level are the connected components of the graph.

The method is known as single linkage because clusters are joined at each stage by the single strongest link between them (Anderberg, 1973). For any cluster produced by the single link method, every member is more similar to some other member of the same cluster than to any other object not in the cluster, and consequently each document must be in the same cluster with its most similar document (or it *nearest neighbour*). However, the minimum similarity between documents in the same cluster can be zero.

The single link method does not succeed in delineating poorly separated clusters, where intermediates are present between clusters (Cormack, 1971). Another characteristic of this method is its tendency to form elongated clusters with little internal cohesion, an effect that is called *chaining* (Jardine & Sibson, 1968). Jardine and Sibson view this chaining effect not as a defect of the method, but rather as a description of what the method does in graph-theoretic terms. The clusters produced by the single link method are described by Hartigan (1975) as "famously strung out in long sausage shapes, in which objects far apart are linked together by a chain of close objects". On the other hand, if clusters are long "sausage"-type with high densities of objects within each cluster, then the single link method will be better than other hierarchic methods in discovering such shapes.

Van Rijsbergen (1971) proposed an implementation of the single link method that has $O(N^2)$ storage and space requirements. The SLINK algorithm (Sibson, 1973) is also commonly found in the literature; it has time and space requirements of $O(N^2)$, and $O(N)$ respectively and has been shown to be an optimally efficient implementation for the single link method.

### 3.4.2 Complete link

The definition of the complete link method is the opposite of the single link: the similarity between two clusters is the minimum of the similarities between all pairs of documents, such that

one document of the pair is in one cluster and the other document in the other cluster. For example, if at some stage of the method clusters *i* and *j* have merged, then the similarity between the new cluster (labelled *p*) and some other cluster *r* is determined as follows: $S_{pr} = min(S_{ir}, S_{jr})$. In graph theoretical terms, complete linkage clustering corresponds to the identification of the maximally complete subgraphs at some similarity threshold.

Because of the way they form, complete link clusters tend to be small and tightly bound, the exact opposite of single link clusters (Voorhees, 1985a). The minimum similarity between documents in the same cluster can never be as low as zero (as in the single link method); instead, it is the similarity level at which the cluster forms. Also contrary to the single link method, the nearest neighbour of a document may be in a different cluster, however mutual nearest neighbours will always be in the same cluster (Voorhees, 1985a).

The main criticism of complete linkage is that it is a *space-diluting* method (Lance & Williams, 1967). The essence of this criticism lies at the heart of the complete link method. Since a document can not join a cluster until it obtains a given similarity level with all members of a cluster, the probability of a cluster obtaining a new member becomes smaller as the size of the cluster increases. In terms of a multidimensional space, the method *dilutes* the space because the larger a particular cluster becomes, the larger the distance between the cluster and some non-member also becomes.

One of the most commonly used algorithms that implement the complete link method is the CLINK algorithm (Defays, 1977) that was devised through a modification of Sibson's (1973) SLINK method, and that has the same time and space complexities ($O(N^2)$ and $O(N)$ respectively). However, document hierarchies produced with this algorithm have displayed poor retrieval effectiveness (El-Hamdouchi, 1987; El-Hamdouchi & Willett, 1989) since it does not generate an exact complete linkage hierarchy (Defays, 1977).

### 3.4.3 Group average link

The similarity between two clusters in the group average link method is the mean of the similarities between all pairs of documents, such that one document of the pair is in one cluster and the other document in the other cluster.

The group average link produces clusters that are neither as loose as the single link clusters, nor as tight as the complete link clusters. In this method clusters are formed on the basis of average similarities, and therefore nothing can be inferred about the minimum or maximum similarities between documents in a cluster (Voorhees, 1985a). This method frequently gives results that are little different from those obtained with the complete link method (Anderberg, 1973).

Sneath and Sokal (1973), based on a number of comparative studies carried out by other researchers, have asserted that average linkage is the most preferable of the hierarchic methods. However, Williams and his co-workers (1971a) have criticised it as being more likely than other methods to form 'non-conformist' groups (i.e. groups whose members share only the property that they are unlike everything else, including each other) as the sizes of clusters increase.

For the group average method, the only $O(N^2)$ time, and $O(N)$ space algorithm known is the one Voorhees (1985a, 1986) used in her Ph.D. thesis. Voorhees noted that for the inner product similarity function, the similarity between a centroid[12] of a cluster and a document is equal to the mean similarity between the document and all the documents in the cluster. Thus, the centroids of a cluster can be used to compute the similarities between the clusters, requiring $O(N)$ space.

### 3.4.4 Ward's method

According to this method proposed by Ward (1963), the merges between clusters at any stage of the method are chosen so as to minimise an objective function that reflects the investigator's interest in the particular problem. Ward illustrated this method with an error sum of squares objective function, and Wishart (1969) showed how Ward's method can be implemented through updating a matrix of squared Euclidean distances between cluster centroids.

Implicitly, Ward's method defines a cluster as a group of documents such that the error sum of squares of Euclidean distances between documents of each cluster is minimal. This method has the tendency to produce clusters of approximately the same size (Milligan *et al.*, 1983). Cormack (1971) criticised Ward's method as being biased towards spherical clusters that may not accurately represent the true shape of groups of data present in the original set.

### 3.4.5 An example

I will demonstrate the single link method by means of an example. The input similarity matrix will be assumed to be the same as in Figure 3.1, and is presented again for ease of reference in Figure 3.3a.

The pair of documents with the highest similarity in Figure 3.3a is $\{x_2, x_3\}$, with a similarity of 0.8. This pair is the first to merge, forming cluster $x_{2,3}$. The similarity matrix in Figure 3.3b results from the one in 3.3a through the deletion of the rows and columns that correspond to $x_2$ and $x_3$, and through the insertion of a row and column that correspond to the newly formed cluster $x_{2,3}$. Moreover, the similarity values in matrix 3.3b have been updated, so that the new similarity

[12] Voorhees defined the centroid of a cluster as the mean of all the vectors in the cluster.

between each of $x_1$, $x_4$, and $x_5$ and the newly formed cluster $x_{2,3}$ is the maximum of the similarities between the respective document and documents $x_2$ and $x_3$.

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | | | | |
| $x_2$ | 0.6 | 1 | | | |
| $x_3$ | 0.4 | 0.8 | 1 | | |
| $x_4$ | 0.1 | 0.5 | 0.7 | 1 | |
| $x_5$ | 0.1 | 0.2 | 0.2 | 0.3 | 1 |

(3.3a)

| | $x_1$ | $x_{2,3}$ | $x_4$ | $x_5$ |
|---|---|---|---|---|
| $x_1$ | 1 | | | |
| $x_{2,3}$ | 0.6 | 1 | | |
| $x_4$ | 0.1 | 0.7 | 1 | |
| $x_5$ | 0.1 | 0.2 | 0.3 | 1 |

(3.3b)

| | $x_1$ | $x_{2,3,4}$ | $x_5$ |
|---|---|---|---|
| $x_1$ | 1 | | |
| $x_{2,3,4}$ | 0.6 | 1 | |
| $x_5$ | 0.1 | 0.3 | 1 |

(3.3c)

**Figure 3.3.** Transformation of the similarity matrix by the application of the single link method

The next stage of the method joins document $x_4$ with the cluster $x_{2,3}$, since this pair displays the largest similarity (0.7) in matrix 3.3b. In a similar way as before, matrix 3.3c can be derived from 3.3b, and the method continues merging the pair of documents, or clusters, with the highest similarity value until only one cluster remains. The final form of the similarity matrix, where only two objects remain in it ($x_1$ and $x_{2,3,4,5}$), is not shown in Figure 3.3 as it is considered trivial.

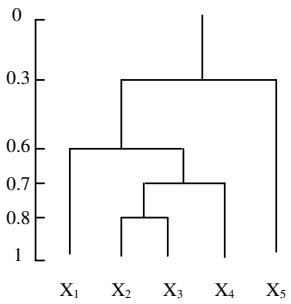


**Figure 3.4.** The similarity dendrogram for the example

In Figure 3.4 the similarity dendrogram that results from the application of the single link method to the input similarity matrix is presented. Only the similarity levels at which clusters are formed in the example are shown in Figure 3.4. At similarity level 0 there is only one cluster present, the entire document set. The other three clustering methods can be applied through a different updating strategy of the similarity matrix.

### 3.4.6 Other methods

Other hierarchic agglomerative methods (e.g. median, centroid, weighted average) have seldom been used for IR applications, and this is the reason for not presenting them in more depth here. Any of the readings that comprehensively cover hierarchic clustering methods (e.g. Anderberg, 1973; Hartigan, 1975; Späth, 1980) offer further details.

### 3.4.7 Some remarks

Lance and Williams (1967) have shown that there exists a general combinatorial equation that can be used to describe how the different agglomerative hierarchic methods update the similarity matrix after a fusion of any two objects. The equation is:

$$s_{hk} = \alpha_i s_{hi} + \alpha_j s_{hj} + \beta s_{ij} + \gamma \mid s_{hi} - s_{hj} \mid$$

In this equation, $s_{ij}$ refers to the similarity between the objects $i$ and $j$ that have merged to form the new cluster $k$. The new similarity between cluster $k$ and any object $h$ is given by $s_{hk}$, and $\alpha_i$, $\alpha_j$, $\beta$, and $\gamma$ are parameters whose values are specified by the hierarchic agglomerative procedure. In Table 3.1, the values of the parameters for each of the four clustering methods previously analysed are presented. For the formulas in this table, $n_r$ corresponds to the number of documents contained in cluster $r$, where $r = i, j, h$.

It should be noted that Wishart (1969) suggested that Ward's method is compatible with Lance and Williams' formula, and designed an efficient algorithm for its implementation. He then suggested that all other agglomerative methods could be implemented through the same algorithm by making use of the update formula given by Lance and Williams. A number of readings in cluster analysis have since adopted this approach for the implementation of hierarchic clustering methods (e.g. Späth, 1980).

| | $\alpha_i$ | $\alpha_j$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|
| Single Link | 1/2 | 1/2 | 0 | -1/2 |
| Complete Link | 1/2 | 1/2 | 0 | 1/2 |
| Group Average | $\frac{n_i}{n_i + n_j}$ | $\frac{n_j}{n_i + n_j}$ | 0 | 0 |
| Ward's Method | $\frac{n_h + n_i}{n_h + n_i + n_j}$ | $\frac{n_h + n_j}{n_h + n_i + n_j}$ | $\frac{-n_j}{n_h + n_i + n_j}$ | 0 |

**Table 3.1.** Values for the parameters of the Lance & Williams combinatorial equation

Of the four methods that were described in the previous paragraphs, single link was first applied to IR by Jardine and Van Rijsbergen (1971). Further research in the 1970s and the early 1980s also focused on the single link method (Van Rijsbergen, 1974b; Van Rijsbergen & Croft, 1975; Croft, 1977, 1978, 1980; Willett, 1983, 1985). It was not until the mid-1980s that the application of other hierarchic agglomerative methods was suggested, almost simultaneously, by Griffiths *et al*. (1984, 1986), Voorhees (1985a), and El-Hamdouchi (1987). Most of the research reported in the 1980s evaluated the comparative effectiveness of the various hierarchic methods (e.g. Griffiths *et al*., 1984; Voorhees, 1985a; Griffiths *et al*., 1986; El-Hamdouchi, 1987). Such studies are presented in detail in Chapter 4.

# 3.5 Cluster representation

The issue of cluster representation is a central one in document clustering. I will divide cluster representation into two forms: *internal* and *external*. Internal refers to the formation of *cluster representatives*, or *centroids*, that attempt to summarise the contents of a cluster for the purposes of cluster-based retrieval. Incoming queries are matched against representatives, and the cluster whose representative is most similar to the query is retrieved (in Chapter 4 I provide more details on cluster-based retrieval). External representation refers to textual or graphical presentations of cluster contents in a manner such that they will support judgements by users on the utility (or relevance) of the clusters. These two types of cluster representation are presented in the following paragraphs.

## 3.5.1 Cluster representatives

Clusters of documents are traditionally represented by some kind of profile, typically called a cluster representative, or a cluster centroid. Representatives are typically used in cluster-based searches, when incoming queries are matched against them. Clusters whose representatives are most similar to the query are subsequently retrieved. Two requirements that a representative should meet are that it should sufficiently describe the contents of the cluster, and that it should sufficiently discriminate between the cluster it describes and all the other clusters of the database. This is because the representative essentially acts as a stand-in for the documents of the actual cluster in the retrieval process. It should also be noted that representatives were initially introduced for efficiency purposes, i.e. to reduce the number of comparisons between the query and objects in the database (Rocchio, 1966).

One can find in the literature a large number of methods for deriving cluster representatives (Croft, 1978; Van Rijsbergen, 1979). Sometimes, the cluster representative can be defined as a document of the cluster itself. For example, clusters can be represented at some level of similarity by a graph. A simple way of finding the representative in such a case is by finding the document that is linked to the maximum number of other documents in the graph (*maximally linked document*) (Jardine & Van Rijsbergen, 1971).

Typically however, the representative is not a document of the cluster itself. For example, Salton (1971) defines the representative as the "centre of gravity" of the documents in the cluster by averaging the descriptions of the members of the clusters. Jardine and Van Rijsbergen (1971) represent clusters by a binary string in which a 1 in the $i$-th position can either indicate the presence of the $i$-th term in more than 1 documents in the cluster, or the presence of the $i$-th term in more than $log_2 C$ documents, where $C$ is the total number of documents in the cluster.

Murray (1972) proposed a method for cluster representation that has subsequently been used by other researchers (Van Rijsbergen & Croft, 1975; Voorhees, 1985a). In short, those terms with the highest frequency within the cluster are included in the centroid, and are assigned weights based on their rank order values of their frequency. The characteristic feature of this method is that it adopts a deletion strategy, so that terms that do not occur frequently enough within the cluster are removed from the list of representative terms. Murray demonstrated that such deletions can be implemented without any loss in retrieval effectiveness.

For his Ph.D. thesis, Croft (1978) proposed a cluster representation model based on Gower's (1974) *maximal predictor* theory. A predictor for a cluster is a binary vector that predicts the characteristics of any document that belongs to the cluster. A maximal predictor is one whose correct predictions are as numerous as possible. In Gower's original maximal predictor theory it is assumed that both types of prediction error (i.e. a 1 in the predictor where there should be a 0, and vice versa) are equally important. In the modified model, Croft suggested that predicting a 0 where there is a 1 in the original document is a more serious error than the reverse, since only few of the vocabulary terms are assigned to a particular document. A relative weight to each type of error can then be attributed by means of a parameter that is user-specified.

The impression that one obtains by reading through the related literature, is that the effect of different representatives on cluster-based retrieval effectiveness has not been extensively investigated (Croft's work for example, investigated different centroid types using a single small database, namely the Cranfield-1400 collection). Most researchers typically adopt a single representation strategy, and calculate retrieval effectiveness based on that selection. Voorhees (1985a) examined the effect of centroid length on retrieval effectiveness. The representative she used in her experiments was the one proposed by Murray (1972). The effectiveness of the searches obtained by varying the centroid's length displayed considerable variability, something that prompted Voorhees to note that (p. 75): "The variability of the effectiveness of searches with varying maximum centroid lengths underscores the necessity of further research into a theory of centroid creation and weighting". The challenge that was raised by Voorhees seventeen years ago still stands unaddressed.

### 3.5.2 Representations of cluster contents

Effective ways by which the contents of a cluster can be summarised are also needed. Such representations are particularly needed during a browsing session, where a number of clusters are presented to a user. The user then has to select the cluster(s) of interest based on the cluster representation that the system displays. One such example is the Scatter/Gather system (Hearst & Pedersen, 1996).

Typical cluster representations include the display of a number of terms that are most heavily weighted within the cluster, or representative document titles from the cluster, as for example in (Allen *et al*., 1993; Hearst & Pedersen, 1996; Neto & Santos, 2000; Roussinov & Chen, 2001). Anick and Vaithyanathan (1997) used phrasal units to represent cluster contents, but do not report any form of evaluation to test the effectiveness of their approach. The same lack of experimental evidence is also noted for Maarek et al. (2000), who also focus on cluster representations that use phrasal descriptions. It is typically the case that in the "Future Work" sections of such research articles the need for more effective cluster representations is emphasised, or the need to conduct user experiments to validate specific approaches is put forward.

An exception to this is the work carried out by Kural (1999, 2001). In this study users were presented with clusters containing the 50 top-ranked documents retrieved in response to a query submitted to an IR system. Cluster representations consisted of the ten most discriminating terms of the clusters and three document titles. The main goal of this research was to investigate whether users would be able to select the 'best' cluster based on the cluster representation provided. The experimental results led Kural to suggest that "clusters can not be relied upon to consistently produce meaningful document groups that can easily be recognised by the users".

However, Kural's work does not escape criticism. She restricted the study to include only the C3M clustering method (Can & Ozkarahan, 1990), one document collection, a fixed number of top-ranked documents, and more importantly, a single cluster representation method. Accordingly, one can argue that the reason for which users were not able to recognise useful clusters was that the chosen clustering method was not effective, or that the chosen cluster representation method was not informative enough, or that the number of top-ranked documents was not large enough. Kural's experimental methodology did not factor out any of these parameters, and so the conclusions one can draw from it are limited.

Wu et al. (2001) report a smaller (in that it involved less users) scale study for the purposes of the TREC interactive track. The effectiveness of cluster representations (10 highest weighted terms of the cluster, 5 most frequent word pairs, and titles of the 3 documents most similar to the query) as relevance clues was one of the many issues the authors were investigating. Users in this study did not express their own information needs. The authors noted that in the majority of the cases users managed to locate the cluster with the most relevant documents. However, a large number of users expressed their dissatisfaction with the way clusters were represented. Again, different methods of representation, or different clustering methods were not investigated.

The area of presenting relevance clues to users by means of different document representations (e.g. abstracts, document titles, indexing terms, list of citations, automatically generated sentence extracts) has been researched by a number of workers such as Rath et al. (1961), Saracevic (1969), Janes (1991), and Barry (1998). Mizzaro (1997) in his review article on relevance

summarises the findings of these and other studies over the course of the last forty years. The one overall conclusion from these studies is that document representations significantly affect users' judgements of the relevance of documents. While the comparative effectiveness of the various representations in these studies does not seem to be uniform, a general trend seems to suggest that abstracts are the most effective indicators of relevance, followed by automatically generated extracts, titles, lists of citations, and indexing terms (Barry, 1998).

It should be noted that these studies were carried out in environments where users had to judge the relevance of a single document in response to a query. The various representations can be seen as a level of abstraction that reduces the amount of information contained in a single document, so that a user can quickly and accurately judge the utility of the original text. In the case of cluster representations the level of abstraction is even higher. Clusters themselves are an abstraction layer on top of single documents. Representations of clusters must succinctly describe the contents of clusters in such a way that users can correctly base their relevance judgements upon such representations; they are a representation of representations (Kural *et al*., 2001). Doyle, as early as 1964, noted that a potential problem with document grouping methods is that "there is no obvious clear-cut way to represent the groups of documents for perusal by literature searchers". The effect of the choice of cluster representation has not been fully acknowledged in the literature so far, and with the exception of Kural's research, the lack of user studies is notable.

In my opinion, more research towards this end is warranted, and such efforts are more likely to originate from the automatic summarisation community. Multiple document summarisation (Mani & Bloedorn, 1999; Radev *et al*., 2000) is an approach that may prove effective as a method of cluster representation, especially query-biased summarisation. The effectiveness of query-biased summaries as indicators of relevance of single documents has been demonstrated in previous research both for textual (Tombros & Sanderson, 1998), and for spoken documents (Tombros & Crestani, 2000).

Based on the current state of research in the area of cluster representation, one should not draw conclusions regarding the ability of users to identify relevant clusters. Until future research systematically investigates the comparative effectiveness of different types of cluster representation, IR researchers should acknowledge the effect that the choice of a particular representation has on users' perception of relevance of document clusters.

## 3.6 Cluster validity

The process of cluster analysis on a set of data is *structure seeking*, i.e. it is attempting to discover structure in the data set. However, the application of cluster analysis is said to be a *structure imposing* process: a clustering method may impose structure on the data even if such structure is

absent from the data itself (Theodoridis & Koutroumbas, 1999). Therefore, measures that can quantitatively evaluate the *clustering tendency* of the original data, as well as the results of clustering methods, are needed. Such measures have been widely used in cluster analysis. Almost every book on cluster analysis dedicates a section on methods for measuring cluster validity (e.g. Jardine & Sibson, 1971; Sneath & Sokal, 1973; Theodoridis & Koutroumbas, 1999). Dubes and Jain (1979) compiled a comprehensive review of such methodologies that is still a point of reference for cluster validity studies.

Three approaches have generally been followed for examining clustering validity: testing for clustering tendency of the original data set, measuring the degree of distortion imposed by the clustering method on the similarity matrix, and measuring the effectiveness of a clustering method at recovering known structure that is present in the original data set.

One way to examine whether a data set exhibits any degree of clustering tendency is by means of the *Random Graph Hypothesis* (RGH) (Ling & Killough, 1976; Dubes & Jain, 1979). Given a dataset *X* comprising *N* objects and the *N*x*N* symmetric similarity matrix *S(X)* for this set, one can create a *N*x*N* symmetric ordinal similarity matrix *R(X)* that contains the numbers 1,2, …, *N*(*N*-1)/2 in the lower triangle without ties; the most similar pair of items has rank 1. The RGH is that all such $[N(N-1)/2]!$ ordinal matrices are equally likely. Based on this hypothesis a number of characteristics of random graphs can be studied. One such characteristic is the number of edges, *V*, needed to connect a random graph. Knowing the distribution of this number of edges permits one to judge how many edges must be observed before deciding that the data are random (Dubes & Jain, 1979). Dubes and Jain explain in detail this process, and Ling and Killough (1976) have calculated tables for the probability of observing specific values of the minimum number of *V* under the RGH given a graph that contains *N* nodes.

Cormack (1971) and Gordon (1987) provide an extensive list of measures for the distortion imposed on the similarity matrix by a clustering method. Such measures typically proceed by comparing the values of interdocument similarities in the input matrix to the corresponding similarity levels of the similarity dendrogram (Figure 3.2). One commonly used distortion measure is the *cophenetic correlation coefficient* (CPCC) (Sokal & Rohlf, 1962). Given a dendrogram whose level values are on the same scale as those in the *N*x*N* original similarity matrix *S(X)*, a *N*x*N* cophenetic matrix *CP(X)* can be defined. Each element $c_{ij}$ of *CP(X)* corresponds to the first level in the dendrogram that objects i and j merge to join the same cluster. The CPCC is then defined as the product-moment correlation between the elements of *S(X)* and *CP(X)*:

$$CPCC = \frac{(1/L)\sum s_{ij}c_{ij} - (\bar{s})(\bar{c})}{\sqrt{(1/L)\sum s_{ij}^2 - \bar{s}}\sqrt{(1/L)\sum c_{ij}^2 - \bar{c}}},$$

where $\bar{s} = (1/L)\sum s_{ij}, \bar{c} = (1/L)\sum c_{ij}, L = N(N-1)/2$. In general, the larger the values of the CPCC are, the better the match between the similarity and the cophenetic matrices. The specific CPCC value that could be deemed as sufficient to suggest that the output dendrogram has not significantly distorted the input similarity matrix has been found to be at least 0.8. However, as Dubes and Jain report (1979), even a value of 0.9 would not guarantee that the output dendrogram is a sufficiently good summary of the original inter-object relationships. It has also been suggested that the CPCC will always yield a high value for the group average method, because of the clustering criterion employed by this method (Farris, 1969).

Another measure of distortion is given by a family of measures developed by Jardine and Sibson (1968). The coefficient is given by:

$$\Delta_\mu = \frac{\left(\sum_{ij} | s_{ij} - c_{ij} |^{1/\mu}\right)^\mu}{\left(\sum_{ij} c_{ij}^{1/\mu}\right)^\mu},$$

where $s_{ij}$ and $c_{ij}$ are the same as before (similarity between objects $i$, $j$ and their cophenetic value respectively), and $\mu$ is an arbitrary parameter, $0 \leq \mu \leq 1$. By varying the value of the parameter $\mu$ it is possible to emphasise between smaller or greater differences between the similarities and the cophenetic values.

Determining the ability of clustering methods to recover cluster configurations that are known to exist in the original data, has also drawn attention from researchers as a method of evaluating cluster validity. To this end, Monte Carlo simulation techniques have been used for generating data sets with known structure (Cunningham & Ogilivie, 1972; Kuiper & Fisher, 1975; Blashfield, 1976; Scheibler & Schneider, 1985; Milligan & Cooper, 1987). These artificial datasets are then analysed by the clustering methods under investigation, and the level of agreement between the actual structure of the dataset and the one discovered by the clustering methods is compared. An advantage of such approaches is that there is no doubt as to the 'ground truth', i.e. the true cluster structure. The main problem that is associated with this approach on the other hand, is the limited degree of generalisation that can be applied to data distributions and structures that are not included in the study (Milligan & Cooper, 1987).

If one attempts to locate similar validity studies in the IR literature, then one will notice that the issue of cluster validity has been somewhat overlooked, or at least, has not been pursued with the same vigour as it has in other fields. This is not to imply that IR researchers are not concerned with the quality of clustering output. Instead, it is merely the case that evaluation of cluster validity in IR has customarily been performed in a different manner. Before seeing what this

manner is, I will report on some studies that have evaluated cluster validity in the terms of the methods previously presented. Willett also presents an overview of such research up to the date of 1988.

W.M. Shaw and his colleagues (1997) tested the RGH in a cluster-based environment that comprised 13 test collections. They constructed random graphs by the single link clustering criterion, where the number of points of the graph corresponds to the number of documents of each test collection. The number of edges was varied from one to a large number $q$. In the resulting random graphs, effectiveness values (a function of precision and recall) were calculated for each component containing two or more documents for each of the queries in the test collection. In this way average effectiveness values were obtained for each test collection, and these values represent expected values of random clustering performance.

Operational retrieval values were compiled by Shaw and his co-workers from a list of nine papers published from 1971 to 1994 by other researchers. By comparing random performance values to operational values, Shaw et al. concluded that operational cluster-based effectiveness is not significantly different from that attained by random structures. This observation led the authors to raise questions regarding the validity of the application of document clustering to IR. They suggested that the structure imposed on a set of documents by topical relatedness may not reliably associate documents relevant to the same query. Quoting the closing sentence from this article, Shaw et al. postulate that: "If cluster-based retrieval is to play a role in IR, it is likely to be demonstrated by adaptive clustering techniques and not by fixed clustering outcomes". This statement by Shaw et al. relates to the issues that this thesis challenges; I will return to this statement at the end of Chapter 4.

Burgin (1995) compared the performance of random clustering to that of operational clustering by following a different experimental procedure. He used four test collections and five hierarchic clustering methods in his experiments (see section 3.2.1). For each test collection Burgin generated 30 random similarity matrices that he subsequently clustered with each of the five clustering methods. Optimal cluster-based retrieval results were obtained and averaged over the 30 iterations to derive random performance values. These were compared to optimal effectiveness values that were obtained by the application of each clustering method to each test collection. The results of Burgin's experiments revealed that the single link method produced hierarchies whose effectiveness was not significantly different to that of random clustering for a number of experimental conditions. The same did not hold for the other clustering methods.

R.J. Shaw and Willett (1993) presented some evidence suggesting that a clustering based on documents and their nearest neighbours (i.e. most similar documents), as the one proposed by (Griffiths *et al.*, 1986; El-Hamdouchi & Willett, 1989; Croft *et al.*, 1989; Wilbur & Coffee, 1994), does not exhibit random behaviour. Their experiments were conducted on four databases. For

each relevant document in each database, they computed a list containing its N most similar documents, and counted the number of relevant documents in this neighbourhood. They repeated the process by randomly assigning nearest neighbours to relevant documents, and by means of statistical analysis determined that the random behaviour was significantly worse than the one achieved with the actual similarities.

The use of distortion measures for the evaluation of cluster validity is not common practice in IR. Griffiths et al. (1984) conducted one of the few studies in which the cophenetic coefficient was used to compare the distortion imposed on similarity matrices by the four hierarchic methods[13] reviewed in section 3.4. The results of this study are further analysed in Chapter 4. However, it is worth mentioning that the average link method, as suggested by (Dubes & Jain, 1979), gave the best results, followed by the single link method. Complete link seemed to impose the largest degree of distortion on the similarity matrix.

Willett (1988) justified the limited application of distortion measures to IR by the observation that the method that imposes the least distortion is not necessarily the most effective one, as shown by Griffiths et al. (1984). Willett also suggested that the distortion of the similarity matrix is not necessarily to be avoided in IR applications: a clustering method should try to discover groupings that are more intense than the ones present in the similarity matrix. Williams and Clifford (1971) have also noted that "…the system is automatically distorted as classification proceeds, and the original similarities are not, and are not intended to be, preserved".

### 3.6.1 Clustering tendency and cluster-based effectiveness

With the exception of these studies, and the ones reported by Willett (1988), the majority of IR researchers have assessed the clustering tendency of a document set by means of the cluster hypothesis[14] (Jardine & Van Rijsbergen, 1971), and the quality of the resulting clusterings by means of cluster-based retrieval. Since both the cluster hypothesis and the evaluation of effectiveness through cluster-based retrieval form focal points of the research reported in this thesis, they are separately discussed in Chapter 4. In this section I will attempt to clarify the reasons that have driven the IR community to adopt these two approaches for evaluating cluster validity.

Van Rijsbergen and Croft in their early work on clustering postulated the potential effectiveness gains that can be achieved by the application of hierarchic clustering to IR (see section 2.5).

[13] No distortion results are reported for Ward's method in the article.

[14] Dubin (1996) proposed an alternative method for measuring clustering tendency based on *skewness* and *elongation*. These measures are tailored for diagnosing document spaces for visualisation, and fall beyond the aim of this thesis. Mather (2000) has also proposed a measure of cluster quality based on linear algebra.

Through interdocument associations relevant documents should 'pull' closer to other relevant documents, and away from non-relevant ones. This assumption has been formally expressed by Jardine and Van Rijsbergen (1971) as the cluster hypothesis.

In brief, the hypothesis postulates that closely associated documents will tend to be relevant to the same queries. Therefore, if for a specific database all relevant documents, on average across all queries, are more similar to each other than to non-relevant ones, then it can be said that this database adheres to the cluster hypothesis. The application of a clustering method to such a document set can be expected to be highly effective, as such a collection is assumed to possess a high degree of clustering tendency. Clustering methods that operate on the similarity matrix will be likely to merge relevant documents first before all non-relevant ones, if the separation between the two is clear.

The ultimate goal of a clustering system is to completely separate relevant from non-relevant documents (Salton *et al*., 1975). This situation would fully support the cluster hypothesis. Figure 3.5 shows such an ideal scenario after the application of clustering to a data set. The resulting dendrogram is shown here without the similarity levels at which the clusters are formed. The left branch of the dendrogram contains all the relevant documents, whereas the right branch all the non-relevant ones. In this ideal situation the left branch of the figure would be hierarchically organised in such a way that it would reveal the topical structure of the relevant portion of the database. For IR applications this is the 'known structure' that is present in the original dataset, and which clustering methods should strive to discover.

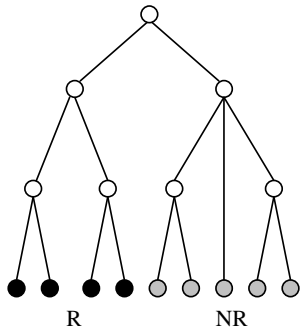


**Figure 3.5.** An ideal scenario: total separation of relevant and non-relevant documents

The quality of the output of various clustering methods can then be measured on the basis of how closely they resemble the ideal situation depicted in Figure 3.5. This measurement can be attained by means of *cluster-based searches* that implement what is known as *cluster-based retrieval* (Jardine & Van Rijsbergen, 1971). Cluster-based retrieval effectiveness is measured using a function of precision and recall, and a number of ways for implementing cluster-based retrieval are presented in Chapter 4. Here it suffices to say that the closer the hierarchic document structure

matches the one of Figure 3.5, the higher the effectiveness will be. Effectiveness decreases as clustering fails to reveal the correct structure of the document space.

Clustering in the context of IR is therefore goal-driven. Its application, as far as effectiveness is concerned, is motivated by the aim to cluster relevant documents together. In the same way, the evaluation of clustering tendency and validity in IR are also goal-driven. What is important for the application of clustering to a document collection for the purposes of IR is different than what may be of importance in any other application area of cluster analysis. However, consideration for the quality of the output of cluster methods in terms of 'meaningful structure' is applicable to IR, and evaluation of the non-randomness of generated structures should be pursued. The methodology followed by Burgin (1995) and Shaw et al. (1997) provides efficient means towards this end.

## 3.7 Recent trends

In this chapter so far I outlined the main aspects of document clustering for IR. The focus was placed on hierarchic clustering methods, because these have been widely applied to IR, and because they will also be applied to the work reported in this thesis. In this section I look into some recent trends that have developed in the area of document clustering.

### 3.7.1 Hypertext and web clustering

With the ever growing popularity of the *World Wide Web* (WWW), it is not surprising that a significant body of recent research has focused on clustering methods for the structuring and organisation of web documents. Clustering of hypertext documents was advocated as early as 1989 by Crouch and his colleagues (Crouch *et al*., 1989). Botafogo (1993), Mukherjea et al. (1994), Weiss et al. (1996), and Johnson and Fotouhi (1996) have also developed methods for clustering hypertext and hypermedia structures. Some of these approaches rely solely on the semantic information embedded in link structures between documents (*link-based* methods, e.g. Botafogo, 1993). Others follow a hybrid approach that combines link and content information in order to calculate interdocument similarities (Weiss *et al*., 1996).

Macskassy et al. (1998) conducted a small scale experiment to investigate the way that humans cluster web documents. Their motivation was to appreciate whether web document clustering implementations can create groupings that are useful and meaningful to users. Their main findings were that users tended to create relatively small clusters, and that any two users had little similarity in the clusters they created. The authors view their findings as "a sobering note on any quest for a single clearly correct clustering method for web pages".

A clustering algorithm designed specifically for web documents has been developed by Zamir and Etzioni (1998) (*Suffix Tree Clustering*, STC). This is a partitioning method that generates overlapping clusters. STC clusters documents based on shared phrases, and also makes use of proximity information between words by treating each document as a string. It uses a suffix tree structure to efficiently organise the initial base clusters that the algorithm subsequently refines.

Modha and Spangler (2000) proposed a hybrid content and link-based algorithm that clusters hypertext documents using words contained in the document, out-links from the document, and in-links to the document. Modha and Spangler use these features to determine the similarities between pairs of documents. The authors also report on a novel method of cluster annotation that is tailored to web documents. They represent each cluster using different *nuggets* of information. Amongst them are the highest weighted keywords of the cluster (*keywords* nugget), the title of the document whose in-link profile is most similar to the in-link profile of the cluster (*breakthrough* nugget), and the title of the document whose out-link profile is most similar to the out-link profile of the cluster (*review* nugget). The highest weighted in and out-links are also used in the representation. No evaluation of the effectiveness of such annotations was reported.

Kumar at al. (1999) describe a link-based clustering method called *trawling*, that combines co-citation and graph analysis to identify clusters of related web documents. The interesting aspect of this approach is its main objective: to automatically identify "emerging web communities". These are clusters of related web pages whose presence on the web is too new, or their topic too fine-grained, to attract the interest of web directory services, such as Yahoo (e.g. the community centred around plane-spotting in U.K. airfields, etc.).

Mukherjea (2000) also uses a hybrid approach in order to organise topic-specific information on the web. He uses a *crawler* to gather web documents specific to a user-supplied topic. To do so, a number of seed documents are chosen, either as selections designated by the user, or as top-ranked documents retrieved from a search engine in response to the specific topic. For each seed the crawler downloads pages that are referenced by the seed, and pages that reference it (these pages are downloaded if their similarity to the seed profile is larger than a specified threshold). The set of pages thus derived is then hierarchically organised into different levels of abstraction. Mukherjea argues that the seeds and the downloaded set of documents constitute a topic-specific set that may reveal useful documents to the user that would have otherwise been missed through a conventional search engine.

Other approaches include a syntactic clustering method that is used to eliminate duplicate web documents (Broder *et al*. 1997), and an algorithm for clustering XML documents that utilises the mark-up language's structured features (Guillaume & Murtagh, 2000).

The WebCluster project (Mechkour et al., 1998; Harper et al., 1999), developed at the Robert Gordon University in Scotland, has proposed a novel approach for document clustering on the web: a mediated access to the web via a clustered collection. The user can initially explore a small, pre-clustered collection that covers a certain, specialised domain of interest (source collection). The clustering can be implemented by any of the hierarchic methods reviewed in section 3.4. The topical structure of the source collection can be revealed to the user through the use of clustering. The user can interact with the structured collection, select clusters and documents of interest. The system subsequently proposes a query based on the selections made by the user. This query is submitted to the target collection, which can be a sub-collection of the web that is indexed by a search engine. This process is expected to assist users with vague information needs, or users who seek information in a domain with which they are unfamiliar.

Comprehensive results from an evaluation of the system have not been published to date. A small scale study that was reported in (Harper *et al*., 1999), revealed that users felt at ease with the idea of mediated access, although users familiar with the source collection would rather be able to formulate their own queries. The issue of cluster representation also came up in the study, as users noted that they would like a more informative representation than the one provided by the system (frequently occurring keywords in the cluster). A potential weakness of Web Cluster that has been identified by Kural (1999), is that the underlying assumption that the small source collection will be representative of a much larger heterogeneous collection, such as the web, is perhaps unrealistic.

#### 3.7.1.1 Other recent trends

Apart from clustering on the web, other emerging applications of clustering in the area of information retrieval include image (Mukherjea *et al*., 1998) and video (Yeung & Yeo, 1998) clustering. Topic detection and tracking is also an area that has attracted much attention recently, and clustering methods have been applied to facilitate the topical and temporal grouping of documents (Yang *et al*., 1998; Hatzivassiloglou *et al*., 2000).

Another body of research has applied clustering in order to detect usage patterns in web-based information systems (Chen & Cooper, 2001). Such approaches cluster user-supplied queries by analysing search engine logs (Beeferman & Berger, 2000; Wen *et al.*, 2001), so that new queries can be matched against similar clusters of past queries. In this way the effectiveness of searches is expected to increase either by expanding user queries with other terms, or by retrieving documents that were relevant to previous similar queries.

In recent years, the emergence of *data mining* has contributed to the introduction of efficient methods for clustering large datasets. Data mining applications impose some special requirements on clustering methods, such as high dimensionality of the feature space, scalability and non-

presumption of canonical data distribution (Agrawal *et al*., 1998). To this end, both partitioning and hierarchic methods have been used (Zhang *et al*., 1996; Guha *et al*., 1998, Karypis *et al*., 1999). Moreover, *density-based* methods that can detect clusters of arbitrary shapes have also been used (Ester *et al*., 1996; Hinneburg & Kleim, 1998), as well as *grid-based* methods that can enhance clustering efficiency even further (Agrawal *et al*., 1998).

## 3.8 Reflections on document clustering research

I will conclude this chapter by reflecting on document clustering research over the past thirty years. The aim of this section is to put forward those aspects of cluster research that are the focal points of the research work in this thesis.

Research on hierarchic document clustering spans arguably for over three decades. During the 1970s research was dominated by the introduction of the cluster hypothesis, the application of the single link method, and the development of search strategies that could potentially increase the effectiveness of document clustering. Van Rijsbergen and Croft carried out most of the work published during this period (Jardine & Van Rijsbergen, 1971; Van Rijsbergen & Sparck Jones, 1973; Van Rijsbergen, 1974b; Van Rijsbergen & Croft, 1975; Croft, 1977, 1978, 1980; Garland, 1982).

The application of other hierarchic methods to IR (e.g. group average, complete link and Ward's methods) was extensively investigated during the 1980s. The majority of the research work was carried out at Cornell University by Ellen Voorhees (1985a), and at Sheffield University by Griffiths et al. (1984, 1986), El-Hamdouchi (1987), El-Hamdouchi and Willett (1987, 1989). Effectiveness was the major consideration in this body of research, and the comparative effectiveness of various hierarchic methods, as well as the comparative effectiveness of clustering and inverted file search, was extensively investigated.

The results of these studies were inconclusive as to whether cluster-based searches or inverted file searches were more effective. Indeed some of the late work in the 1980s by El-Hamdouchi and Willett (1989) suggested that non-clustered searches are to be preferred. In the majority of the work published up to that point, clustering had been applied to entire document collections in a static manner (i.e. once, before querying). Willett (1985) reported on some experiments on post-retrieval clustering, where only documents of the database that match a specific query were clustered. However, no other published research at the time investigated the effectiveness of post-retrieval clustering.

Over the last decade, there has been a considerable shift of focus of clustering research. Effectiveness issues became superseded by considerations for fast, efficient clustering methods

(not necessarily hierarchic) that can support on-line user interaction for browsing document collections (Cutting *et al*., 1992; Allen *et al*., 1993; Silverstein & Pedersen, 1997). Methods for visualising document collections, and search results, and the inter-document relationships that hold in such collections have also been investigated over this period (Dubin, 1996; Leuski & Allan, 1998; Allan *et al*., 2001). Almost ten years after Willett's first experiments on post-retrieval clustering, Hearst and Pedersen (1996) performed some further experiments which reviewed the cluster hypothesis under the light of post-retrieval clustering, and investigated its applicability to a browsing task.

However, this latter period has been dominated by an "efficiency over effectiveness" approach. A potential reason for this might be that no new research has focused on effectiveness issues since perhaps the late 1980s (El-Hamdouchi, 1987). Some effectiveness-oriented research has been published by Shaw (1991, 1993, 1997) and Burgin (1995). However, as I discussed in section 3.2, this work has mainly investigated clustering effectiveness as a function of indexing exhaustivity, and has not put forward a new approach that could enhance clustering effectiveness itself.

In this thesis, I investigate issues pertaining to the effectiveness of cluster-based IR systems. Efficiency issues are not considered. The reasons for this approach are twofold. First, I believe that effectiveness is of primary importance, whereas efficiency is a factor that is heavily dependent on technological advances. Secondly, and more importantly, if one succeeds in improving effectiveness, then one could potentially instigate further development in the field. This development can be materialised in the form of more efficient algorithms and/or hardware that would exploit the improved effectiveness. Alternatively, it can be materialised in the form of new research in areas that are linked with effectiveness and that have been neglected due to the lack of appropriate stimuli. For document clustering, such areas may include new models of cluster-based searches, and new methods of cluster representation.

As mentioned previously in this chapter, the cluster hypothesis is paramount to the issue of effectiveness in hierarchic clustering. In the following chapter I extensively discuss the cluster hypothesis and its implications to cluster-based effectiveness.

## 3.9 Summary

In this chapter I examined the basic steps of the clustering process, and I discussed in detail issues that relate to each of these steps (sections 3.2-3.6). I limited the discussion to hierarchic clustering methods, as this type of clustering is used in this work. The steps of the clustering process which were discussed in this chapter were: the indexing representation of documents, the calculation of interdocument similarities, the application of hierarchic clustering methods, the representation of

cluster contents and the validation of the clustering results. In section 3.7 I also presented some recent trends in document clustering research.

As this work focuses on the effectiveness of hierarchic clustering, I reviewed these issues, where possible, from the perspective of the effect they may have on cluster-based effectiveness. By reviewing previous work which has investigated these issues from an effectiveness point of view, I aimed to justify some decisions that I make in later chapters regarding the implementation of the clustering system used in this work. Such decisions, for example, include the use of the most exhaustive indexing representations for documents (section 3.2.1), the use of *tf-idf* weights for document terms (section 3.2.2) and the use of the cosine coefficient as a measure of interdocument similarity (section 3.3.2).

In section 3.5 I presented a number of issues which relate to the representation of cluster contents (either for cluster-based retrieval, or for the presentation of cluster contents to users). I described the effect that these issues have on the clustering process. I also explained why, based on the current state of research, IR researchers should be cautious when dealing with these issues, and appreciate the effect that these may have on the effectiveness of the clustering process. The discussion in this section relates to the choice of optimal cluster evaluation that is used in this thesis. I further discuss optimal cluster searches in Chapter 4, section 4.3.4.

In section 3.6 I outlined a number of methods for validating the output of clustering methods, and I particularly focused on the way this is performed in IR. In section 3.6.1, I put forward the view that document clustering in IR is a purpose-driven purpose that is characterised by the cluster hypothesis: relevant document should be grouped together, separately from non-relevant ones. Validity studies in IR have therefore been replaced by studies of retrieval effectiveness, since this is a measure of how well a cluster-based system performs at achieving the clustering goal.

I closed this chapter by presenting a short overview of document clustering research, focused on effectiveness issues, that has been carried out over the past three decades (section 3.8). In that section, I emphasised the recent tendency of IR research to not deal with effectiveness issues. In the next chapter, I discuss in detail issues relating to cluster-based effectiveness, and I aim to demonstrate the reasons for which IR researchers seem to have been driven away from pursuing effectiveness-oriented work.

# Chapter 4

# On the Effectiveness of Cluster-Based Information Retrieval

## 4.1 Introduction

In the previous chapter I discussed a number of issues relating to the generation of document hierarchies by means of clustering methods. The issue of effectiveness was mentioned in passing in a number of sections in Chapter 3, and especially in section 3.6.1 where it was related to measuring the goodness of document hierarchies in IR applications. In that chapter the cluster hypothesis was also repeatedly mentioned, as it is upon the hypothesis that the introduction of hierarchic clustering to IR was based. In section 3.6.1 the hypothesis was presented as a measure of the clustering tendency of document collections.

The aim of this chapter is to further expand on issues relating to the cluster hypothesis and the effectiveness of cluster-based IR systems, and through this discussion to bring out the motivation for the experimental work reported in this thesis. First, in section 4.2, I present ways through which the validity of the cluster hypothesis and the clustering tendency of document collections can be measured. The tests that are typically employed in the IR literature are reviewed, and past research that has used these tests is reported. In section 4.3, I discuss methods by which the retrieval of documents from document hierarchies can be implemented. Details of methods for searching document hierarchies are presented, and studies that have investigated the comparative effectiveness of such methods are reviewed.

Hierarchic document clustering was introduced to IR on the grounds of its potential to improve the effectiveness of the IR process. In section 4.4 I examine, based on results that have been published in the literature over the past thirty years, whether clustering has indeed realised this potential. These results mainly reveal a negative picture regarding the success of clustering as a means of effective retrieval. In section 4.5 I outline some reasons for which I think clustering has

failed to fulfil this potential, and I relate these reasons to the motivation behind the work reported in this thesis. Finally, section 4.6 summarises the issues presented in this chapter.

## 4.2 Testing for the validity of the cluster hypothesis

The cluster hypothesis has been used as an indication of the clustering tendency of document collections. Tests for the validity of the cluster hypothesis have been developed as a means of predicting whether the application of clustering to a specific collection would be likely to yield effective retrieval results. To this end, two different tests have been proposed in the document clustering literature. A third test is also presented in this section. Although this third test does not test the validity of the cluster hypothesis, it has been used as a measure of the clustering tendency of document collections and as such it is presented here.

### 4.2.1 Separation of frequency distributions

Van Rijsbergen and his colleagues (Jardine & Van Rijsbergen, 1971; Van Rijsbergen & Sparck Jones, 1973) proposed the *overlap test* that is a natural following of the cluster hypothesis. This test is based on the extent to which documents relevant to the same query are more similar to each other than to non-relevant ones. The procedure for the overlap test is as follows. The associations between all pairs of documents both of which are relevant to the same query, and one of which is relevant and one non-relevant, are first computed. Summing these association values over all the queries of a test collection gives the relative distributions of relevant-relevant (R-R) and relevant-non relevant (R-NR) associations for a test collection. The two distributions can be plotted in the same graph, against the actual association values. If a specific document collection is characterised by the cluster hypothesis, then the separation of the two distributions should be sufficient, and vice versa.

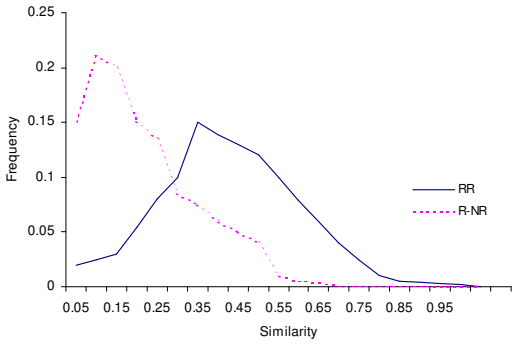


**Figure 4.1.** Separation of frequency distributions

For example, in Figure 4.1 a pair of such distributions is plotted for a hypothetical collection. From this plot, one can note that the collection possesses a reasonably good separation of the two

distributions. Therefore, based on what is postulated by the cluster hypothesis, the application of clustering to the collection of the example is likely to be highly effective.

Griffiths et al. (1986) report a method for obtaining a single numerical value that quantifies the degree of overlap between the two distributions. This measure is defined as the fraction of the two distributions that is common to both, i.e. the extent to which the two distributions overlap each other.

Voorhees (1985a, 1985b) has elicited a problem with the overlap test. She noted that because there are always more R-NR than R-R pairs in a collection, the relative frequency of highly similar R-NR pairs will always be much less than that of highly similar R-R pairs. Voorhees also argued that whether or not the cluster hypothesis holds for a particular collection depends on the absolute number of highly similar R-NR pairs; the overlap test does not provide information at this level of detail.

### 4.2.2 The nearest neighbour test

The *nearest neighbour* (NN) *test* was proposed by Voorhees (1985a, 1985b) in order to address the aforementioned limitation of the overlap test. The *n* nearest neighbours of a given document *d* are the *n* documents that are most similar to *d* using a specific association measure. The NN test examines each of the relevant documents for a specific query in turn, and identifies the number of its *n* nearest neighbours that are also relevant. A single numeric value for the NN test can be obtained for a test collection by calculating the average number of relevant documents that are contained within the *n*-document nearest neighbourhood, when averaged over all the relevant documents for all the queries of a test collection. The higher the average number of relevant documents in the *n*-neighbourhood, the higher the probability that the cluster hypothesis holds for that specific collection.

Voorhees chose to calculate the number of relevant documents contained within a five-document nearest neighbourhood, and she examined the percentage of relevant documents that had 0, 1, 2, 3, 4, and 5 relevant documents in this neighbourhood. El-Hamdouchi and Willett (1987) noted that depending on the chosen size of the nearest neighbourhood different results may be obtained for this test. They also point towards another potential caveat of the NN test: its implicit assumption that the proportion of relevant documents is the same for different test collections. This assumption implies that there is an equal probability of having another relevant document as the nearest neighbour of any specific document for any test collection. Although this may not be an overly realistic assumption, it allows the comparison of the results of the NN test across different test collections.

### 4.2.3 The density test

A third test was proposed by El-Hamdouchi and Willett (1987), and is referred to in the literature as the *density test*. The density test yields a single numeric value for a test collection. This value corresponds to the total number of postings in the collection divided by the product of the number of documents in the collection and the number of terms that have been used for the indexing of those documents. The density of a collection comprising *N* documents, an indexing vocabulary of size V, and *l* terms per document on average, is given by *Nl*/*NV*, i.e. *l*/*V*. The density, as defined by El-Hamdouchi and Willett, is the inverse of the number of clusters resulting from the use of the C3M clustering procedure of Can and Ozkarahan (1990).

The density test associates clustering tendency with the density of the term by document matrix of a document collection. If the matrix is sparse, and each document has only a few terms selected from a large number of possible terms, then most pairs of documents in the collection will have few terms in common, and therefore low association values. If the data matrix is more dense, documents will share a large number of terms in common, and therefore when calculating interdocument associations it will be possible to differentiate between documents that are highly similar to each other and those that are not closely related. El-Hamdouchi and Willett assume that the resulting clustering in the latter case will display the interdocument relationships more accurately than in the former case, where the range of possible similarities is limited. Therefore, a high density value is associated with potentially effective clustering.

It should be noted that the density test does not address the issue of whether the cluster hypothesis holds for a specific document collection. However, it is a measure of clustering tendency, such as tests for the validity of the cluster hypothesis are meant to be. It was therefore presented in the same section with the overlap and NN tests since these three are the most commonly used clustering tendency tests in IR. Other authors have similarly presented these three tests together (El-Hamdouchi & Willett, 1987; Willett, 1988; Rasmussen, 1992). It should also be mentioned that contrary to the other two tests, the density test does not require the existence of relevance assessments for the datasets to which it is applied.

### 4.2.4 Some notes on the tests

Results for the previous three tests, using various IR test collections, have been reported in the literature. Some of the earlier work had focused on the overlap test (Jardine & Van Rijsbergen, 1971; Van Rijsbergen & Sparck Jones, 1973; Van Rijsbergen & Croft, 1975; Croft, 1980; Griffiths *et al*., 1986). These researchers report results of this test using the Cranfield-200 and –1400 collections, as well as the Evans, Harding, Inspec, Keen, LISA, and UKCIS collections. Voorhees (1985a, 1985b) reports results for the overlap and NN tests using the CACM, CISI,

Inspec, and Medline collections. It should be noted that all these collections are of a small size (Table 2.1).

El-Hamdouchi and Willett (1987) provide an extensive list of results for all three tests using the Cranfield-1400, Evans, Harding, Inspec, Keen, LISA, and UKCIS collections by repeating previously published results for the overlap test, and by calculating results for the NN and density tests. The main contribution of this research is that it provided El-Hamdouchi and Willett with the opportunity to compare the success of the three tests in predicting clustering effectiveness. To do so, they used results for the effectiveness of nearest neighbour clusters (NNC) that were published by Griffiths et al. (1986) (a NNC contains just two documents: a document *d* and its nearest neighbour). Based on these results they ranked the seven test collections in decreasing order of effectiveness. They also calculated rankings for each test collection by applying the three tendency tests (i.e. collections were ranked for each test based on its outcome). The authors then calculated the correlation between the rankings obtained by the actual clustering effectiveness (NNC clusters) and the tendency tests. The results demonstrated that the rankings produced by the density test correlated best with the effectiveness rankings, followed by the NN test, and by the overlap test.

It should be mentioned that none of the tests provide an indication of what numerical value is associated with "good clustering tendency" for a specific document collection. For example, what can one infer by the fact that for the overlap test using the LISA collection a value of 0.58 is derived? Does this value imply that clustering the LISA collection is likely to be highly effective, moderately effective, or perhaps not effective at all? Unfortunately, it implies none of the above. The three tests are of value to researchers only as tools for the comparison of the clustering tendency across document collections, and not as indicators of clustering tendency for a single collection, and in this fashion they have been used in the literature.

As a final remark for this section, it is worth noting that both the overlap and the NN tests depend solely on the outcome of the association measure that is used (keeping all other things constant). This outcome can be made to vary if, for example, different levels of indexing exhaustivity, different term-weighting schemes, or different types of association measures are used. If a researcher is interested in measuring the effect of such experimental parameters on the structure of the document collection, then these tests are a useful tool. The NN test in particular, seems to be better suited to such tasks. It provides immediate feedback, in the sense that the researcher can see how the *n* nearest neighbourhood of a relevant document changes in relation to variations in these experimental parameters. R.J. Shaw and Willett (1993) have used this test in this fashion, in order to determine whether actual interdocument associations and randomly generated associations are significantly different to each other.

## 4.3 Cluster-based retrieval

*Cluster-based retrieval* (CBR) was initially proposed by Jardine and Van Rijsbergen (1971) as an alternative to linear associative retrieval. According to CBR, a single cluster is retrieved in response to a query; the documents within the retrieved cluster are not ranked in relation to the query, but rather, the whole cluster is retrieved as an entity. Cluster-based retrieval has as its foundation the cluster hypothesis, since if relevant documents are placed in the same cluster (as the hypothesis postulates), then the effectiveness of a CBR strategy that succeeds in retrieving this specific cluster can be expected to be high. Cluster-based retrieval can be implemented by means of *cluster-based searches*. Before introducing the way that document hierarchies can be searched, I will demonstrate how the effectiveness of CBR can be gauged.

Standard IR evaluation is performed in terms of precision and recall graphs that are calculated based on a ranked document list produced by an IR system (see section 2.4). Cluster-based retrieval strategies, on the other hand, perform a ranking of clusters, instead of individual documents, in response to each request. Even in the case where a ranking of individual documents is performed within the clusters (as for example in (Voorhees, 1985a)), relatively few documents are actually ranked, and therefore different results may be obtained depending on the method that the researcher chooses to generate the precision-recall graphs (Croft, 1978). The generation of precision-recall graphs is thus not possible in such systems, and in order to derive an evaluation measure for clustering systems, the E effectiveness function was proposed by Van Rijsbergen (Jardine & Van Rijsbergen, 1971; Van Rijsbergen, 1974a).

The formula for the measure is given by: $1-\frac{(\beta^2+1)PR}{\beta^2 P+R}$, where P and R correspond to the standard definitions of precision and recall (section 2.4, over the set of documents of a specific cluster), and $\beta$ is a parameter whose values range from 0 to $\infty$; it reflects the relative importance attached to precision and recall. Three values of this parameter are typically used:

$\beta = 1$ attributes equal importance to precision and recall

$\beta = 0.5$ attributes half as much importance to recall as to precision

$\beta = 2$ attributes twice as much importance to recall as to precision

It should be noted that low values of the E measure are associated with higher effectiveness. I will now proceed by presenting the three different types of cluster-based searches that have been proposed in the literature.

### 4.3.1 Top-down search

The difference between the first two types of search lies in the 'direction' in which the document hierarchy is searched for the cluster that best matches the query. A *top-down* strategy enters the hierarchy at the top (*root*), and proceeds in a downward fashion towards the bottom of the tree (Jardine & Van Rijsbergen, 1971). That path is chosen for the downward movement that displays the greater similarity between the query and the cluster centroids. The search then continues moving down the tree until a retrieval criterion is satisfied. Two such criteria have been typically used: either a minimum number of documents needs to be retrieved, or the query-cluster similarity is required to stay sufficiently high. In the former case the search is terminated when a cluster containing the required number of documents is retrieved. In the latter case the search is terminated when the query-document similarity at a specific comparison falls below the similarity attained at the preceding comparison (Jardine & Van Rijsbergen, 1971; Van Rijsbergen, 1974b), or when the similarity falls beyond a user-specified threshold (Willett, 1988; El-Hamdouchi & Willett, 1989).

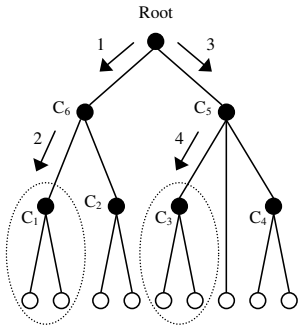


**Figure 4.2.** A sample broad top-down search

The top-down searches can be further distinguished in *broad* and *narrow* (Van Rijsbergen & Croft, 1975). A narrow search proceeds as mentioned previously, that is, by selecting the highest matching cluster at each level, and by expanding it until the stopping criterion is satisfied. It is narrow in the sense that once the decision on the path to be followed has been made it can not be reversed. A broad search, on the other hand, may abandon a path down the tree once the similarities fall below a specific threshold. In such a case the search may backtrack to the cluster that is more similar to the query and that has not been visited before.

In Figure 4.2 an example of a broad top-down search is given. The arrows represent the path that the search follows down the hierarchy, starting from the root of the tree. The numbers on the arrows show the order in which the different paths are followed, depending on the outcome of the query-centroid comparisons. For this example it is assumed that the search will terminate once at least three documents have been retrieved. The search commences by comparing the centroids of clusters $c_5$ and $c_6$ to the query. Cluster $c_6$ is found to be more similar, and hence this path down the

tree is chosen. Since this cluster contains more than three documents, the centroids of clusters $c_1$ and $c_2$ are also compared to the query. Cluster $c_1$ is found to be more similar, and its two documents are retrieved. Since the stopping criterion has not been satisfied, the search backtracks and follows the path down the other branch of the tree towards cluster $c_5$. The search will terminate once the two documents of cluster $c_3$ are retrieved.

A problem with top-down searches is that at the early stage of the search, when the query is matched against the centroids of large clusters close to the root of the hierarchy, it is fairly easy to misdirect the search down the wrong path (clusters $c_5$ and $c_6$ in the previous example). This may happen because of the inherent difficulty with effectively representing large clusters (Willett, 1988). Therefore, the first few cluster-query comparisons can be deemed as arbitrary (Croft, 1978). A remedy to this problem is to start the top-down search not at the root of the hierarchy, but rather at a partition of the hierarchy that can be obtained by applying some similarity thresholding procedure (Van Rijsbergen, 1979; Willett, 1988).

### 4.3.2 Bottom-up search

Bottom-up searches proceed through a document hierarchy in the opposite direction, i.e. the search commences at a document, or cluster, residing at the bottom of the hierarchy, and then moves up towards the root until a retrieval criterion is satisfied (typically until a specified number of documents is retrieved). The issue that has attracted much interest regarding this type of search is the selection of the starting point for the search.

Van Rijsbergen and Croft (1975), in their work on bottom-up searching, suggested that an appropriate starting point is an already known relevant document. If a relevant document is not somehow known *a priori*, then one can rank the documents in decreasing order of similarity to the query, and select the top-ranking document as the starting point (El-Hamdouchi & Willett, 1989). A third alternative, is to utilise the *bottom-level clusters* of the hierarchy (Croft, 1978, 1980). A bottom-level cluster is the cluster that a specific document is assigned to when it first joins the hierarchy (for example, clusters $c_1$, $c_2$, $c_3$, and $c_4$ in Figure 4.3). Croft has demonstrated that it is possible to efficiently access the bottom-level clusters of a document hierarchy by generating an inverted index of these clusters. A starting point for a bottom-up search can then be obtained if the bottom-level clusters are ranked in decreasing order of similarity to the query (similarity is calculated between the centroid of a cluster and the query); the top-ranked cluster is the starting point. If the cluster hypothesis characterises the document collection, then other relevant documents should be located in that same cluster, and hence the search proceeds upwards in this branch of the hierarchy until the stopping criterion is satisfied.

In Figure 4.3 a sample bottom-up search is shown. For this example it is assumed that the search will terminate once at least three documents have been retrieved. The arrows represent the path that the search follows towards the root of the tree. In this example it is also assumed that the starting point of the search is determined by comparing the query against the centroids of the bottom level clusters $c_{1, 2, 3, 4}$. The highest similarity is achieved for cluster $c_3$, and hence the search enters the hierarchy through this cluster. This cluster contains only two documents (less than the stopping criterion requires), and therefore the search proceeds upwards and reaches cluster $c_4$. This cluster contains the desired number of documents, and the search terminates by retrieving cluster $c_4$.

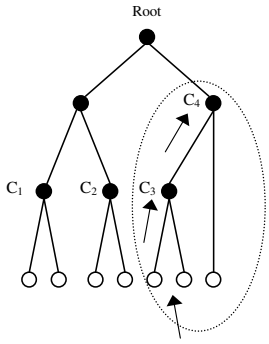


**Figure 4.3.** A sample bottom-up search

It should be noted that the bottom-up search is narrow (Croft, 1978), in the sense that only one path up the tree is pursued. Also, since this type of search commences at the bottom of the hierarchy, it tends to be more efficient than a top-down search in terms of the number of cluster-query comparisons (Croft, 1978).

### 4.3.3 Comparative performance of the two searches

A number of researchers have compared the effectiveness of top-down and bottom-up searches in a variety of experimental settings. In some of the early work (Van Rijsbergen & Croft, 1975; Croft, 1978, 1980) it was suggested that bottom-up searches result in higher retrieval effectiveness. It was also shown that narrow top-down searches are precision-oriented (i.e. result in higher effectiveness values when the parameter $\beta$ of the E measure is set to 0.5), in contrast to broad top-down searches that are recall-oriented (Van Rijsbergen, 1974b; Croft & Van Rijsbergen, 1975). This result is intuitively reasonable, since the former type of search proceeds down only one path of the hierarchy, whereas the latter type pursues many different paths and hence has a better chance of retrieving more relevant documents.

Another result from this early research is that highly effective retrieval can be achieved if one considers only the bottom level clusters of the hierarchy, and ranks those in decreasing order of

similarity to the query (Croft, 1978, 1980). In this case, the best matching bottom level clusters are retrieved until the desired number of documents is reached. As mentioned previously, efficient access to the bottom-level clusters can be obtained by means of an inverted index of these clusters; queries are matched against the cluster centroids. Griffiths et al. (1986) used this type of search to compare the effectiveness of different hierarchic clustering methods, and El-Hamdouchi and Willett (1989) found this type of bottom-up search to be more effective than two other types mentioned in section 4.3.2 (starting with the highest-ranked document of a similarity search, and proceeding upwards from the bottom-level cluster most similar to the query), and almost as effective as a search that commences at a known relevant document.

Based on Croft's work, Griffiths et al. (1986) advocated the use of nearest neighbour clusters (NNC). They noted that since small bottom level clusters seem to result in high effectiveness, then NN clusters (which are the smallest clusters possible, containing only a document and its nearest neighbour) should also result in high effectiveness. They also noted that complete link, group average link, and Ward's methods all produce large numbers of small clusters containing just pairs of documents, most of which contain a single document and its most similar neighbour. Griffiths and his co-workers demonstrated the high effectiveness of NNCs through experiments where NNC effectiveness was superior to that attained using the bottom-level clusters of hierarchies. This result led El-Hamdouchi (1987), El-Hamdouchi and Willett (1989), Croft et al. (1989), and Wilbur and Coffee (1994) to also adopt NNCs in cluster-based retrieval experiments.

Voorhees (1985a) compared a number of different cluster-based searches using the single, complete and group average link methods. Voorhees compared the effectiveness of two different types of top-down and bottom-up searches: those that retrieve entire clusters (as is the case in traditional cluster-based retrieval), and those that retrieve individual documents from each cluster (based on their similarity to the query). The experimental results obtained by Voorhees, in the majority of the cases, indicated that between these two types of search the latter type results in more effective retrieval.

The results also demonstrated that the most effective of all the searches was a top-down search that selects individual documents from a complete link hierarchy. In all other types of hierarchies a bottom-up search was more effective. This led Voorhees to postulate that top-down searches are more affected by hierarchy characteristics than bottom-up ones. This, Voorhees explained, is due to the small size of bottom-level clusters on which bottom-up searches operate. Top-down searches, on the other hand, commence near the root of the hierarchy, were the sizes of clusters are large, and hence the probability of making an erroneous branch selection is high. Regarding the high effectiveness of top-down searches using the complete link hierarchies, Voorhees noted that this method results in a shallow hierarchy where the top-level clusters are smaller compared to those of other hierarchy types. Croft (1978, 1980) had first made the observation that bottom-

up searches involve significantly less amount of uncertainty than top-down ones, and hence are likely to be more effective.

As I already discussed in section 3.5.1, Voorhees also investigated the effect of different cluster centroid lengths on the effectiveness of cluster-based searches, and noted considerable variability in the results. The same variability was noted in measuring retrieval effectiveness for combinations of different clustering methods and different search strategies: for example, a top-down search was more effective than bottom-up searches only for complete link hierarchies. One can therefore note a great degree of variability when actual search strategies are used to compare the effectiveness of different clustering methods, or different clustering strategies. A different type of cluster-based search that rids itself of most of such complications is the optimal cluster search, and is presented in the next section.

### 4.3.4 Optimal cluster search

Optimal cluster searches differ from the other two types in that no actual matching between the query and cluster centroids takes place, and hence no actual need to compute cluster centroids exists. Cluster-based effectiveness is calculated by finding the optimal cluster of a hierarchy, i.e. that cluster, for any given query, that yields the least E value (i.e. highest effectiveness) for that query. Therefore, optimal cluster-based effectiveness represents the maximum effectiveness that is attainable by a cluster-based search strategy that selects a single cluster in response to each query.

Optimal cluster evaluation has been widely employed in the past (Jardine & Van Rijsbergen, 1971; Croft, 1978; Griffiths *et al*., 1984; Shaw, 1991; Burgin, 1995; Aslam *et al*., 1998). The main advantage of optimal cluster search is that it "allows an evaluation of the different hierarchies to be made without the distorting effects of the particular search mechanism adopted" (Griffiths *et al*., 1984, p. 196). Rasmussen (1992) has also noted some inherent problems of making scientific inferences based on the results of actual (i.e. top-down or bottom-up) searches (p. 437): "… there are several ways in which retrieval from a clustered document collection can be performed, making comparisons difficult when using retrieval as an evaluative tool for clustering methods".

Optimal measures, on the other hand, eliminate any bias that may be introduced from sources *external* to the document hierarchy (Shaw, 1991). External sources include the choice of a particular cluster-based search strategy that matches queries to clusters, and the ability of a user during an interactive session to choose the cluster which is most relevant to his information need.

In the case of a cluster-based search strategy, its effectiveness will be determined by a number of parameters that are alien to the document hierarchies. Such parameters have been mentioned in

sections 3.5.1 and 4.3.3, and include the type of search, e.g. bottom-up, top-down, narrow, wide, (Jardine & Van Rijsbergen, 1971; Croft 1980; Voorhees, 1985a; El-Hamdouchi & Willett, 1989), the type and length of the cluster centroid against which queries are matched (Croft, 1978; Voorhees, 1985a), the entry point in the hierarchy in the case of a bottom-up search (Croft, 1978, 1980; El-Hamdouchi & Willett, 1989), etc.

In the case of a user browsing a clustered document collection (e.g. Cutting *et al*., 1992), the cluster considered useful will be influenced by parameters such as the graphical or textual presentation of the clustered space (Hearst & Pedersen, 1996; Leuski & Allan 1998), the way that cluster contents are summarised and displayed (Hearst & Pedersen, 1996; Radev *et al*., 2000; Kural *et al*., 2001), etc. In section 3.5.2 I discussed issues that relate to the effect of cluster representation on the user's perception of relevance of document clusters, and I pinpointed some limitations of the current state of research in this area.

By eliminating such external parameters from the experimental design, one can infer that the variation in effectiveness across experimental conditions is attributed to the different conditions themselves (*internal* parameters, e.g. different similarity measures, clustering methods, etc.), and not to any form of bias that may have been introduced by any of the external parameters. For this reason, I deem optimal cluster evaluation as highly appropriate for cases where the experimenter wishes to vary some parameters of the clustering system and study the effect of this variation on clustering effectiveness. Optimal cluster evaluation is used in this thesis; I will return to this issue in section 5.5.4 where I outline the experimental environment used in this work.

### 4.3.5 On optimal effectiveness measurements

As mentioned previously, the optimal cluster of a hierarchy, for any given query, is the cluster that yields the highest effectiveness for that query, and it may be located at any depth in the document hierarchy. Jardine and Van Rijsbergen (1971) named the effectiveness measure attainable in this way *MK1*. This measure can be used when comparing across different clustering methods, or different clustering strategies. By observing the variation in the effectiveness of the optimal cluster of the hierarchy, a researcher can appreciate the effect that the various experimental conditions have on the effectiveness of the resulting hierarchies, without the confounding effect of parameters external to the hierarchies.

Optimal cluster-based effectiveness can also be compared to non-cluster-based effectiveness. In order to do so, one needs to find appropriate measures to gauge inverted file search (IFS) effectiveness. A first such measure can stem from the MK1 measure: the system finds the optimal cluster in a hierarchy, looks at the size of the cluster (let us assume it contains $k$ documents), and uses this number of top-ranked documents (i.e. $k$) to measure the effectiveness of the ranking $R$ that is produced by the IFS. I will call this measure *MK1-k*. Intuitively, this measure captures the

degree at which IFS effectiveness matches cluster-based effectiveness for the number of documents for which cluster effectiveness is optimal. It should be reminded that, like MK1, the effectiveness of the IFS will also be calculated in terms of the E measure. Hearst and Pedersen (1996), among other researchers, have used MK1-k as a measure of comparison against optimal cluster-based effectiveness.

This comparison of cluster-based and IFS effectiveness, based on MK1 and MK1-k respectively, can be thought of as being unfair on IFS. This is because MK1-k does not take into account IFS optimality, i.e. the rank position for each query for which the set of documents retrieved gives the least value of E. Thus, a second measure for comparing optimal cluster-based effectiveness to IFS effectiveness can be based on the above, i.e. the optimal IFS effectiveness. Jardine and Van Rijsbergen (1971) used this measure, and called it *MK3*. It represents a measurement of the maximum effectiveness that is attainable using an IFS strategy. The effectiveness calculated by MK3 will always be at least as high as that calculated by MK1-k, since the portion of the initial ranking *R* that MK1-k corresponds to is always considered when calculating MK3.

MK3 implicitly assumes that the optimal segment of the initial ranking *R* that is produced by the IFS will always have as its starting point the highest ranked document. However, this may not always be the case. Therefore, a third measure can be used to gauge optimal IFS effectiveness. This measure seeks for that subset of the original ranking *R* that gives the least value of the effectiveness measure E (i.e. the highest effectiveness). Unlike MK3, the starting point of this subset is not required to be the highest ranked document. I will call this new measure *MK4*. It represents the optimal IFS effectiveness that is attainable from any possible segment of the original ranking *R*. MK4 will always yield a highest effectiveness value than MK3 (or, at least, a value equal to MK3), since the portion of *R* that MK3 corresponds to is always considered when calculating MK4. Therefore, MK4 can be seen as a more favourable approximation of optimal IFS effectiveness.

I will illustrate through an example the different effectiveness values obtained, for a specific query and for a specific ranking *R*, by the three different measures MK1-k, MK3 and MK4. For this example I will assume that a document hierarchy yields an optimal value for a cluster that contains six documents, and that the collection to be clustered contains ten documents. Figure 4.4 shows the ranking that is obtained by an inverted file search for this scenario. The first column displays the rank position of the retrieved documents, and the second column shows whether a document is relevant (R) or not relevant (NR) to the query. The total number of relevant documents in this example is assumed to be six.

Since the optimal cluster contains six documents, MK1-k will be calculated for the first six retrieved documents in the example (i.e. $k = 6$). By using the standard definitions for recall (4 relevant documents out of possible 6) and precision (four relevant documents in the six

documents of the set), and by also using the E measure as defined in section 4.3 for $\beta$=1, we derive MK1-k = 0.33.

In order to calculate MK3, the optimal E value that can be obtained from this ranked list by keeping the starting point fixed at rank position one needs to be found. In this way it can be found that the optimal E value can be obtained for the first seven documents of the list, yielding MK3 = 0.23 (recall=5/6 and precision=5/7). Lower E values correspond to higher effectiveness; in this example, MK3 is a much better approximation of optimal IFS effectiveness than MK1-k.

MK4 offers an even more favourable, for the ranked list, effectiveness measurement. In the above example, the most effective segment of the ranking *R* is between rank positions four and seven, yielding a value of 0.2 for measure MK4 (recall=4/6, precision=4/4). It is apparent that, at least in this specific example, MK4 offers a higher effectiveness value for IFS than MK3 does.

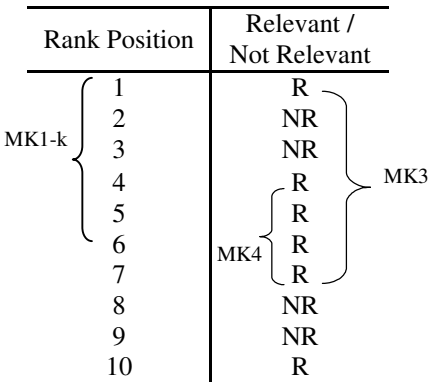


**Figure 4.4.** Example of calculation of optimal effectiveness measures

If the initial retrieval that produces the ranking *R* is highly effective, one can expect the most effective portion of the retrieved list to be located near the top of *R*. Therefore, in this case MK3 is likely to accurately represent the optimal effectiveness attainable by *R*. However, in case of less effective initial retrieval, it is more likely for a large number of relevant documents to be concentrated further away from the top-ranked positions of *R*. In such a scenario MK3 would not give an accurate estimation of optimal IFS effectiveness, since the most effective part of *R* is likely to be located within a segment that does not have rank position one as its starting point. Consequently, one can view MK4 as an attempt to counterbalance the effect of the effectiveness of the initial retrieval when comparing the effectiveness of optimal cluster-based to that of IFS-based retrieval. It can also be argued that the further away from rank position one the optimal MK4 segment of *R* is located, the poorer the initial retrieval has been.

It can furthermore be argued that MK4 is also a more conceptually appropriate measure for comparing optimal IFS effectiveness to optimal cluster-based effectiveness. There is no guarantee that a searcher, or an IR system, will be able to correctly identify the best cluster in a document hierarchy for every query. Similarly, there is no guarantee that a searcher will be able to navigate

his way towards the most effective segment of a ranked list when that segment is located in a lower section of the initial ranking *R*. Therefore, if a comparison between MK1 and MK4 is in favour of the former, one can conjecture that cluster-based effectiveness has indeed the potential to exceed IFS effectiveness.

## 4.4 The effectiveness of hierarchic clustering in IR

Once a document hierarchy has been generated, and once a set of documents has been retrieved from the set of clusters of the hierarchy, the effectiveness of the retrieval can be gauged by means of the E measure presented in section 4.3. In this way, the comparative effectiveness of different clustering strategies can be measured, and also, the comparative effectiveness of cluster-based and non cluster-based strategies can be investigated. In this section I review these two issues: the comparative performance of hierarchic clustering methods is presented in section 4.4.1, and the comparative effectiveness of cluster-based and non cluster-based retrieval is discussed in section 4.4.2

### 4.4.1 Comparisons of hierarchic methods

The hierarchic methods that I focus on in this section are the ones I presented in section 3.4, namely the single link, complete link, group average link and Ward's methods. Apart from the retrieval effectiveness attainable when using each of these four methods, I will also briefly present a comparison of the methods based on their theoretical properties.

#### 4.4.1.1 Theoretical properties

Jardine and Sibson (1968) suggested seven conditions to which any clustering method that transforms a dissimilarity coefficient into a hierarchic dendrogram should adhere. Out of the seven conditions, the authors noted that only three are usually adhered to by most known hierarchic methods. The four conditions that are not adhered to by all methods, state the following:

- A unique result should be obtained from given data
- Small changes in the data should produce small changes in the hierarchy
- The ultrametric similarity coefficient should remain unchanged by the transformation
- The result obtained by the method should impose the minimum distortion upon the similarity coefficient

Jardine and Sibson also noted that the only method to satisfy all seven conditions is the single link method. They placed particular emphasis on the fact that this method preserves the ultrametric

inequality. However, Williams et al. (1971b) noted that although preserving the ultrametric inequality is interesting from a theoretical point of view, there is no known practical utility for such a feature.

Fisher and Van Ness (1971) borrowed the concept of *admissibility* from decision theory, and presented nine properties which one might expect 'reasonable' clustering procedures, or the groups obtained by these, to possess. A method that satisfies all such properties is called admissible; a method that satisfies a specific property *A* is called *A*-admissible. In accordance to Jardine and Sibson's findings, they concluded that the method that displayed a theoretically sound definition, by possessing most (but not all) of the nine properties, was single link. The complete link method was a close second. Whether the theoretical supremacy of the single link method translates to superior effectiveness is examined in the next section.

#### 4.4.1.2 Effectiveness

In the context of IR the single link method was extensively used in early document clustering experiments (Jardine & Van Rijsbergen, 1971; Van Rijsbergen, 1974b; Van Rijsbergen & Croft, 1975; Croft, 1978; Croft, 1980; Garland, 1982). The basis of its application to IR research was its theoretical soundness, as this was outlined in (Jardine & Sibson, 1968), and its computationally attractive implementation (Van Rijsbergen, 1971; Sibson, 1973).

However, a number of studies that were conducted in fields other than IR, mainly in the 1970s and early 1980s, comparing a number of hierarchic methods in their ability to recover true cluster structure (Cunningham & Ogilivie, 1972; Kuiper & Fisher, 1975; Blashfield, 1976; Milligan *et al.*, 1983) suggested that the single linkage method displayed consistently poor performance, rating below the other three hierarchic methods. It should be noted that these were simulation studies that involved the generation of an artificial set whose true clustering structure was known (see section 3.6).

As to the question of which method showed the best performance in these studies, different methods behaved differently under different experimental conditions[15]. Cunningham and Ogilivie (1972) investigated seven agglomerative methods. Their results suggested that the group average was the most effective method, often closely followed by complete link. Blashfield (1976) found that Ward's method formed the solutions that had the greatest accuracy in retrieving true clustering structure, followed by complete linkage. Kuiper and Fisher (1975) examined six hierarchic methods and concluded that Ward's method was better when clusters were of equal size, but that the group average method was superior when cluster sizes varied. Milligan et al.

[15] Griffiths *et al*. (1984) suggested that there are severe methodological problems with the analyses of simulation studies.

(1983) reached conclusions similar to that of Kuiper and Fisher, adding the remark that the complete linkage method was the one most similar to Ward's in terms of the hierarchies produced.

Motivated by the observation that in a number of other fields single link did not seem to be more effective than other methods, a number of researchers applied the complete link, group average link, and Ward's methods, as well as the single link method, to IR laboratory experiments (Griffiths *et al*., 1984; Voorhees, 1985a; Griffiths *et al*., 1986; El-Hamdouchi, 1987; El-Hamdouchi & Willett, 1989, Burgin, 1995). The results of these studies suggested that single link consistently displayed poor performance, confirming results from other disciplines. Explanations for its poor performance were given on the basis of the inherent characteristics of the method (i.e. the chaining effect, section 3.4.1) (Willett, 1988). In a retrieval environment, the single link method leads to a small number of large and loosely defined clusters that seem to perform poorly at recovering the relevance structure of the data (Burgin, 1995).

| *Study* | *Clustering methods* | *Document collections* | *Results* | *Cluster-based searches used* |
|---|---|---|---|---|
| Griffiths at al., 1984 | SL, CL, GA, W | Keen, Cranfield | GA | optimal, bottom-up |
| Voorhees, 1985a | SL, CL, GA | Medline, CACM, CISI, Inspec | CL>GA>SL | bottom-up, top-down |
| Griffiths *et al*., 1986 | SL, CL, GA, W | Keen, Cranfield, Evans, Harding, LISA, Inspec, UKCIS | W | bottom-up |
| El-Hamdouchi, 1987; El-Hamdouchi & Willett, 1989 | SL, CL, GA, W | Keen, Cranfield, Evans, Harding, LISA, Inspec, UKCIS | GA>W>SL>CL | bottom-up |
| Willett, 1988 | SL, CL, GA, W | - | CL | - |
| Burgin, 1995 | SL, CL, GA, W | CF, Medline, Time, Cranfield | GA≈W>CL>SL | optimal |

**Table 4.1.** Studies comparing hierarchic agglomerative methods in IR (CL: complete link, GA: group average, SL: single link, W: Ward)

In Table 4.1 some of the most influential studies that have looked into the comparative effectiveness of different agglomerative methods in the context of IR are summarised. Willett's review article (1988) is included in the table despite that no experimental work was explicitly carried out. However, Willett expressed his personal opinion in that article regarding the most appropriate method for IR applications (p. 592), and it is this opinion that is reported in Table 4.1. In the fourth column of this table, the outcome of the corresponding study as to which of the examined methods was more effective is displayed. In the last column of the table, the type of cluster-based search(es) used in each of the studies is also reported.

There are a few comments regarding the studies mentioned above. In the experiments reported by (El-Hamdouchi, 1987; El-Hamdouchi & Willett, 1989) the CLINK algorithm (Defays, 1977) was used to implement the complete link method. As was mentioned in section 3.4.2, this method does not produce an exact complete link hierarchy. This is attributable for the poor performance of the complete link method in these experiments, something acknowledged by Willett (1988). In their 1989 study, El-Hamdouchi and Willett used the complete link algorithm implemented by Voorhees (1985a, 1986), and repeated their experiments for three collections (Keen, Cranfield, and Evans). The new results suggested that the complete link hierarchies were the most effective.

Griffiths et al. (1984, 1986) used Ward's method with a similarity coefficient (the Dice coefficient). It is known that Ward's method is fully defined only when used with squared Euclidean distances (Willett, 1988). Therefore, the results of Griffiths and his colleagues should be examined with caution. Burgin's study, (1995), aimed mainly at studying the effectiveness of hierarchic clustering methods as a function of indexing exhaustivity (see section 3.2.1). Since in all other experiments reported here complete indexing representations were used, the results of this study should be viewed under the light of Burgin's main experimental aim.

Most experimental studies used either a bottom-up or a top-down search strategy, with the exception of the studies by Griffiths et al. (1984) and Burgin (1995), who employed optimal cluster searches. The reader's attention should be drawn to that, when comparing across clustering methods, the use of actual search strategies introduces a number of factors that may affect the experimental results. This was evident in (Griffiths, *et al*., 1984) where optimal and actual search results were generated for the same methods and collections, and the researchers reached different conclusions for different types of searches. In section 4.3.4 I outlined a number of reasons for which optimal searches are best suited for the comparison of different clustering strategies.

Murtagh (1984b) carried out a comparative study of six hierarchic methods[16] not from an effectiveness point of view, but rather from the perspective of the structure of the clusterings that each of the methods produced. Murtagh defined three different coefficients that quantify the "quality" of hierarchic structures (e.g. how balanced the resulting dendrograms were), and applied these coefficients to the structures generated by the six methods for the Cranfield-200 test collection. The major conclusion of this study was that methods other than the single link should be favoured, and that Ward's and complete linkage methods displayed the most balanced clustering behaviour.

Two of the coefficients that Murtagh (1984b) introduced, were used in the study by Griffiths et al. (1984). Their results confirmed Murtagh's findings, in that the single link method produces a

[16] The methods were singe link, complete link, group average link, Ward's, median, and centroid.

small number of large, unbalanced clusters that suffer from the chaining effect mentioned in section 3.4.1. Complete link and Ward's methods seemed to produce the most balanced hierarchies, again in agreement to Murtagh's findings. However, when Griffiths and his colleagues measured the degree of distortion (see section 3.6 for a discussion on distortion measures) that each of the methods[17] imposed on the initial inter-document similarity matrix, their findings revealed a different picture: the single and the group average link methods seemed to summarise the inter-document similarities more accurately than the complete link method. This led the authors to suggest that methods that impose a small degree of distortion to the similarity matrix may be capable of identifying more natural clusters than methods that result in a cluster structure that is not evident in the original data.

Based on the results presented in the last few pages, what could one conclude, or even hypothesise, on the issue of the most effective clustering method for IR? Not much, it would certainly seem to be the case. Even the inadequacy of the single link method for IR applications seems to be challenged by Griffiths et al.'s (1984) finding that this method is successful at recovering true clustering structure, therefore having potential to effectively recover structure where it is evident. This potential has not been confirmed by the experimental results discussed in this section. As far as the other three methods are concerned, there seems to be little, if any, difference between them. Based on the data of Table 4.1, it would be justifiable to assert that complete link and group average have proved to be the most effective methods so far.

As a closing note for this section, I will quote the following from Cormack (1971): "Some -I am tempted to say most- data are just not classifiable". This note is meant to put the difficulty of the clustering task in perspective, and to serve as a prelude to the next section that compares the effectiveness of cluster-based and non cluster-based retrieval.

### 4.4.2 Cluster-based vs. non cluster-based retrieval

Hierarchic clustering was introduced to IR based on its potential to increase the effectiveness of the IR process. Consequently, a lot of effort has been expanded by researchers in order to investigate whether this potential can indeed be realised by cluster-based retrieval.

The initial studies that were carried out in the 1970s included the use of only the single link method, which, as seen in the previous section, typically displays the lowest retrieval effectiveness amongst the hierarchic methods. Van Rijsbergen and Croft carried out the majority of the work during this period (Jardine & Van Rijsbergen, 1971; Van Rijsbergen, 1974b; Van Rijsbergen & Croft, 1975; Croft, 1978, 1980). When optimal cluster-based retrieval effectiveness

[17] Ward's method was not included in this part of the study.

was compared to IFS effectiveness, the potential of clustering to yield performance improvements was confirmed: in most of the results reported, optimal single link effectiveness was higher than IFS effectiveness. Top-down and bottom-up searches, however, did not often manage to exceed IFS effectiveness. The exception to this was the effectiveness attained by the bottom-up searches reported by Croft (1978, 1980) that searched only the bottom-level clusters of the hierarchy (see section 4.3.2). It also became evident that cluster-based searches compare more favourable to IFS effectiveness when precision is favoured over recall (i.e. when $\beta = 0.5$ in the E measure). It should also be noted that with the exception of Croft's work (1978), the document collections that were used in experiments at that time were of a relatively small size (typically in the order of 1,000-2,000 documents).

Voorhees (1985a) was one of the first researchers to compare the effectiveness of other hierarchic methods to that of IFS. As reported in section 4.3.3, apart from the single link method, Voorhees also used the group average and complete link methods. These two methods often outperformed IFS for certain types of top-down and bottom–up searches (where individual documents were retrieved). Single link, on the contrary, rarely outperformed IFS effectiveness. Other researchers reached similar conclusions regarding the single link method (Griffiths *et al.*, 1986; El-Hamdouchi, 1987; El-Hamdouchi & Willett, 1989). In these studies all four hierarchic methods were investigated (i.e. group average, Ward, complete link and single link methods). The other major conclusion from these last three studies was that nearest neighbour clusters (section 4.3.3) proved to be the most effective type of clustering, and the only strategy used in these studies that consistently yielded higher effectiveness than IFS. The other bottom-up strategies that were investigated failed to exceed IFS effectiveness. In agreement to the early research by Van Rijsbergen and Croft, precision-oriented searches yielded higher effectiveness than recall-oriented ones.

These results led Willett and his co-workers to dismiss the potential of clustering as a means of improving the effectiveness of IR systems. Instead, they proposed that high effectiveness is likely to be achieved through the use of NNCs. This has led a number of other researchers to investigate the effectiveness of NNCs under different experimental environments. Croft and his colleagues (1989), and Wilbur and Coffee (1994) report favourable results from the use of NNCs compared to those obtained from the use of IFS. It should be noted that NNCs had previously been proposed as a method of effective retrieval, for example, Goffman (1969) had advocated the use of chains of NNCs to this end.

As I discussed in section 3.8, research on more effective means of performing clustering seems to have subsided during the past ten years. One reason for this is that the research community seems to have accepted the limitations of clustering, as these were exhibited through the experimental results that I previously mentioned. It is the aim of this thesis to establish a case for the opposite.

## 4.5 What this thesis addresses

The problem that this thesis addresses is that of improving the effectiveness of cluster-based information retrieval. The long standing motivation behind the work reported in this thesis has been the relative failure of cluster-based retrieval to succeed as an effective retrieval mechanism, as this was exhibited by reviewing the research that has been carried out in this area. The belief that CBR effectiveness can indeed be improved, has been based on the view that, because of its intuitively appealing and theoretically sound basis, clustering should indeed be a highly effective information retrieval mechanism. The fact that it has not served as such so far, can be interpreted as a limitation of the way that clustering has been performed to date.

A question that naturally arises, is why has clustering not fulfilled its potential as an effective mechanism for information retrieval. Researchers whose work has demonstrated results not in favour of clustering (e.g. El-Hamdouchi & Willett, 1989), have not offered sufficient insight to the reasons of such failure. Methods for improving the effectiveness of CBR effectiveness, in most cases, are suggested in such studies. For example, Voorhees (1985a) called for more systematic research in cluster representation schemes, El-Hamdouchi and Willett (1989) offer the use of NN clusters as an effective alternative to a clustered file structure, etc.

All the above suggestions are perfectly valid, and indeed, one may also highlight a number of other issues whose addressing may improve CBR effectiveness. However, all these issues are external to the clustering process itself. For example, a cluster representation scheme is put in effect once a document hierarchy has been generated. Different schemes may combine information from documents within clusters in different ways, and it is highly likely that effectiveness improvements can be achieved by researching new representation schemes. The same can be argued for devising more effective cluster-based search strategies, or for using only bottom-level clusters, or NN clusters. The problem with such approaches is that, whatever the representation scheme or the search method, the resulting effectiveness will be constrained by the limitations imposed by the quality of the document hierarchy. Therefore, the view taken in this thesis is that it is in the heart of the clustering process that one has to focus if the effectiveness of the resulting hierarchies is to be improved. To do so, one may need to reconsider some assumptions and practices that implicitly underlie the clustering process.

Shaw and his colleagues (1997) carried out one of the last studies that investigated the effectiveness of CBR. Their findings suggested that the effectiveness attainable by hierarchic clustering methods does not significantly differ from that attainable by random procedures. In an attempt to explain these negative results, Shaw and his co-workers suggested, along with two other possible reasons, the view that “clustering criteria employed to date have failed to reveal the inherent tendency of documents relevant to the same query to be grouped together”. Furthermore,

Shaw et al. also postulated that ”clustering strategies capable of adapting to relevance information may succeed where static clustering techniques have failed”.

This latter comment by Shaw et al. highlights one of the implicit assumptions that underlie the document clustering process: the user is typically left outside the clustering “loop”, i.e. he has no direct input in the way that clustering methods structure the document space. Clustering is thus static, devoid of the ability to adapt to relevance information. I view this as a limitation of document clustering, one that is most likely of the main factors that have contributed to the negative results reported in the literature.

It is this area that the experimental work reported in this thesis addresses: the improvement of the effectiveness of cluster-based retrieval through the generation of document hierarchies that take the user’s search interest into account. The next chapter presents in detail the approaches that are employed in this thesis for enhancing the effectiveness of a cluster-based IR system.

## 4.6 Summary

In this chapter I focused on the effectiveness of cluster-based IR. More specifically, I presented tests for the validity of the cluster hypothesis which are typically used in IR (section 4.2), I discussed the comparative effectiveness of the four hierarchic methods which are used in this thesis (section 4.4.1), and I discussed issues relating to the effectiveness of cluster-based retrieval. The discussion on cluster-based effectiveness had two focal points.

First, the different types of searches which are typically used to search a document hierarchy were presented (top-down, bottom-up and optimal searches), and their comparative effectiveness was examined. I provided justification for the use of optimal cluster searches in this thesis (section 4.3.4), and I also illustrated a number of measures which can be used to measure optimal cluster-based and best-match retrieval effectiveness (section 4.3.5).

The second focal point was the comparative effectiveness of cluster-based and best-match retrieval (section 4.4.2). Studies which have looked into this issue seem to have suggested that document clustering can not act as an effective retrieval mechanism. This in turn has affected IR research, which, as I discussed in Chapter 3 (section 3.8), has not focused on effectiveness issues for a considerable amount of time.

I concluded the chapter in section 4.5, by stating that it is the aim of this thesis to challenge some of the implicit assumptions that characterise the application of document clustering to IR. By doing so, I aim to demonstrate that clustering can indeed act as an effective method for information retrieval, and its failure to do so up to date is attributed to the manner of its

application. The assumptions that this thesis aims to challenge relate to the static manner in which document clustering is applied, and I discuss them in detail in the following chapter.

# Chapter 5

# Query-Based Document Clustering

## 5.1 Introduction

In all the research that I have reviewed so far, clustering has been applied statically, over an entire document collection, prior to querying (i.e. static clustering). Therefore, under static clustering the user has no direct input in the outcome of the clustering process: the generated hierarchies remain the same regardless of what the user's search interest might be. This can be seen as a limitation of document clustering, since a static clustering may not effectively reflect the user's interest (Ottaviani, 1994), especially in large heterogeneous document collections, such as the Internet, with which increasingly larger numbers of users are interacting. If clustering is to act as an effective means of retrieval, it is more likely to do so by adapting to the user's interests. In relation to this issue Ottaviani (1994) argued that static clustering methods "leave the true arbiter of relevance, the searcher, out of the cluster-forming loop… These features result in poor service to the interactive searcher."

The main motivation behind the work reported in this thesis is to investigate the potential effectiveness gains that can be obtained by generating document hierarchies that are based on the user's search request. In this way, document clustering is no longer a static process that does not take the user's information need into account. Instead, clustering is transformed into a dynamic process that adjusts to the user's subject of inquiry. This thesis takes the view that such dynamic methods are more likely to yield effective document structures than a static, *a priori* clustering of the entire document collection. I will call this class of clustering methods *query-based*.

In this chapter I put forward two approaches by which a query can influence document hierarchies generated by clustering methods, and I outline some issues that pertain to the effectiveness of such approaches. I examine the various issues that relate to clustering effectiveness under the view that clustering in IR is a goal-driven process. As I discussed in section 3.6.1, if for each query there exists a perfect separation between relevant and non-relevant documents, then the

effectiveness thereby attained will be the highest possible. This situation represents an ideal case that defines the goal of any cluster-based system: for each search request, to group documents relevant to the request separately from those non-relevant to the request. This goal also holds a strong relation to the cluster hypothesis, since the separation of relevant from non-relevant documents is explicitly postulated in the definition of the hypothesis.

The first way by which the query can influence the output of a hierarchic clustering method, is by clustering documents which have first been retrieved in response to a query (*post-retrieval* clustering). It should be noted that a number of other terms have been used in the literature as an alternative to post-retrieval clustering, as for example *query-specific clustering* (Willett, 1983), *dynamic clustering* (Hearst & Pedersen, 1996; Anick & Vaithyanathan, 1997), *ephemeral clustering* (Maarek *et al*., 2000), etc. The term post-retrieval clustering is used in the rest of this thesis. In section 5.2 I examine issues relating to the effectiveness of post-retrieval clustering, and I also discuss previous research that has looked into this issue. Through the discussion of this past work, limitations of the state of the research in this area are highlighted.

The second approach for influencing document hierarchies towards the user's query is outlined in section 5.3. To elicit this second method, I review the role of the cluster hypothesis in document clustering, and I also challenge the way that interdocument associations are calculated in IR. Previous research that shares similar goals to this proposed approach will also be reviewed.

Having elaborated on the two methods of generating query-based document hierarchies, in section 5.4 I state what the research aims of the experimental work that I report in the following chapters are. Then, in section 5.5 I describe the methodology and the specific details of the environment under which this experimental work is carried out. Section 5.6 summarises the main issues discussed in this chapter.

## 5.2 Post-retrieval clustering

One way of entering the user "in the loop" of the clustering process is by performing post-retrieval clustering, i.e. by clustering documents that have first been retrieved by an IR system in response to a query. In this way a new document hierarchy is generated every time a user inputs a query to an IR system. The generated hierarchies will contain documents that have a greater likelihood of being relevant to the specific query, since they have been highly ranked by the IR system in response to this query. As a consequence, by generating a different hierarchy for each query there seems to be a greater likelihood of reaching the clustering goal than with static clustering, i.e. for each query to achieve a higher degree of separation between relevant and non-relevant documents.

The investigation of post-retrieval clustering, as opposed to that of static clustering, has not been as systematic. Although there can be many viewpoints from which one may examine post-retrieval clustering, I will focus on its effect on the effectiveness of the clustering process, and on its relation to the cluster hypothesis. Preece (1973), Willett (1985) and Hearst and Pedersen (1996) are some of the few researchers who have investigated post-retrieval clustering from similar viewpoints in the context of IR.

S.E. Preece, in 1973, was one of the first researchers to suggest that clustering could be applied as an "output option", to the results of a Boolean or best-match IR system. Preece provides some insight into the potential advantages of post-retrieval clustering, and into its relation to helping achieve the clustering goal. Preece states that (p. 189) "…post-retrieval clustering offers a possible alternative (to inverted file search), since the false drops are likely to be closer to each other than to the relevant documents". This directly relates to the validity of the cluster hypothesis for post-retrieval clustering: by clustering search outputs it will be more likely to capture the relationships between relevant documents, and therefore more likely to cluster relevant documents together, apart from non-relevant ones (false-drops in Preece's terminology).

Preece also argued that "…with pre-retrieval (i.e. static) clustering, each document is attached at full strength to only one cluster. This may mean that documents relevant to a request can not be retrieved by that request". This is a major limitation of static clustering, since the "hard" assignment of documents to clusters for all incoming queries may be problematic in terms of retrieval effectiveness. Documents relevant to a specific query are likely to be dispersed across a few clusters in the static clustering scenario. A cluster-based search strategy is likely to miss such documents that may be placed in clusters with other similar, but irrelevant to the query, documents. In relation to this issue, Preece further adds that post-retrieval clustering can potentially increase retrieval effectiveness, since the initial inverted file search is likely to "filter out" many non-relevant documents that could have otherwise mixed with relevant ones to form clusters.

Given Preece's speculations in 1973, I find surprising that no actual investigation of the effectiveness of post-retrieval clustering took place until 1985[18] (Willett, 1985). The motivation behind Willett's research was not the potential to improve the effectiveness of cluster-based IR systems. Instead, he was motivated by the potential of post-retrieval clustering to address two other limitations of static clustering: efficiency, and updating strategies.

Since small numbers of documents are clustered under post-retrieval clustering, it is possible to avoid the large computational overhead of operating on large data files that is imposed under

[18] Attar and Fraeknel (1977) describe experiments employing post-retrieval keyword, rather than document, clustering.

static clustering (section 3.4). Moreover, post-retrieval clustering alleviates the need to employ any updating mechanisms. Under static clustering, as the composition of the document collection changes over time, methods for updating the static hierarchy, and thus avoiding the need to re-compute the hierarchy, are important. On the other hand, post-retrieval clustering generates a hierarchy that is automatically derived from the current state of the document collection. Documents may be added or deleted from the collection, but the hierarchy is generated based upon the most recent image of the document collection when a query is presented to the system. This advantage could be of great importance if one were to apply clustering to data collections that are highly dynamic by nature (e.g. a collection of web documents stored in an intranet).

Willett experimentally examined whether these two advantages of post-retrieval clustering come at the cost of reduced retrieval effectiveness. He used three test collections (Keen, Evans, and Cranfield) and the single link clustering method. In order to determine the set of documents to be clustered for each query he used a *coordination level search*, with a level of 0 corresponding to the entire collection (static clustering), a level of 1 corresponding to documents that have at least 1 term in common with the query, and so on; levels of 0, 1, 2 and 3 were used. Willett measured the retrieval effectiveness for each of the three levels using a search strategy developed by Croft (1980). The comparison of the results demonstrated that the effectiveness of the dynamic method was inferior to that of static clustering, albeit not substantially inferior.

Willett also examined the effect of variations in the coordination level on the validity of the cluster hypothesis. He used the overlap test (Jardine & Van Rijsbergen, 1971) for all three collections and coordination levels. The degree of the overlap of the frequency distributions (see section 4.2.1) seemed to increase as the coordination level increased, confirming the previous finding that effectiveness deteriorated as the coordination level increased. Willett attributed this behaviour to the increased similarity that documents (both relevant and non-relevant) exhibit amongst each other at higher coordination levels. The similarity, Willett argued, is expected to be higher since documents at high coordination levels will at least share certain terms that characterise the query.

A limitation of Willett's work, that might have affected his experimental results, was the coordination level search that he used, mainly because of the varying indexing exhaustivity of the test collections employed in his experiments. Acknowledging this, Willett reports (p. 30) that "… it would probably be better to rank a document collection in decreasing order of similarity with the query on the basis of some matching function… so as to obtain the desired number of documents". A further limitation of this approach can be found in the use of only one clustering method, namely the single link method, and in the use of only three relatively small test collections.

### 5.2.1 Re-examining the cluster hypothesis for Scatter/Gather

Perhaps the most widely publicised piece of research on post-retrieval clustering is the one reported by Hearst and Pedersen (1996). Hearst and Pedersen re-examined the cluster hypothesis under post-retrieval clustering, by suggesting that, if two documents $D_1$ and $D_2$ are relevant for query $Q_A$, then they need not necessarily be relevant for a different query $Q_B$. In other words, Hearst and Pedersen view the cluster hypothesis on a per-query basis. This is in contrast to the "on-average, across-all queries" basis that Jardine and Van Rijsbergen (1971) originally proposed and viewed the hypothesis.

This dynamic view of the cluster hypothesis is a direct consequence of the clustering of retrieval results, and is similar to the original postulations by Preece (1973). According to the original view of the hypothesis, on average, across all queries, relevant documents tend to be more similar to each other than to non-relevant ones. One can argue that this is a strict assumption that may not be ubiquitously met in realistic retrieval environments. By performing post-retrieval clustering, one relaxes this assumption and reduces it to one that requires, for each query, relevant documents to be more similar to each other than to non-relevant ones. Hearst and Pedersen also postulated that, by clustering retrieval results, clusters have the potential to be more tailored to the characteristics of a specific query than clusters generated by a static clustering.

The authors then tried to experimentally test the validity of their argument using the Scatter/Gather system (Cutting *et al.*, 1992; Pirolli *et al.*, 1996). Scatter/Gather employs partitioning clustering, inheriting from this class of clustering methods the problems mentioned in section 3.1. For example, Scatter/Gather requires the number of clusters to be determined beforehand. One can see this requirement as limiting, especially in the case of dynamic cluster generation, where no prior knowledge of the topical structure of the document set to be clustered is available.

Hearst and Pedersen performed a series of experiments in which they clustered the top-$n$ documents (100, 250, 500, 1000) returned from an inverted file search. In their experiments they used over 2 Gbytes of text from the standard TREC collection (Harman, 1993), as well as 49 queries from TREC-4. For each of the four values of $n$, and for each query, the top-$n$ retrieved documents were clustered into 5 partitions, a value that is arbitrarily chosen. By observing the distribution of the percentage of relevant documents in each of the five partitions, Hearst and Pedersen conjectured that the cluster hypothesis must hold for the Scatter/Gather clustering system, since the best partition (i.e. the one containing the highest percentage of relevant documents) always contains at least 50% of the relevant documents retrieved.

However, their experimental analysis does not provide any information on the varying degree to which the cluster hypothesis may be valid when considering different numbers of top-ranked

documents. Neither does it provide any information on the effect that the transition from static to post-retrieval clustering has on the validity of the cluster hypothesis or on retrieval effectiveness. Moreover, the authors do not provide information about statistics of the partitions generated (e.g. size), and the effect that partition size may have on the number of relevant documents found in the best partition. For example, for large values of *n* one expects the mean size of each partition to increase[19] (since there are always 5 partitions), increasing at the same time the probability to have more relevant documents placed in the same partition.

Given that the main motivation of Hearst and Pedersen's work was to examine the cluster hypothesis under post-retrieval clustering, it may be argued that their experiments do not provide enough evidence towards that end. Therefore, their conclusion (p. 81) that "… clustering (i.e. Scatter/Gather type clustering) does in fact group together the relevant documents, as would follow from the cluster hypothesis" does not appear to be fully supported.

Hearst and Pedersen also compared a ranking of documents in a best cluster, for each of the four values of *n*, to an equivalent cut-off in the original top-ranked documents. It should be noted that this type of evaluation is equivalent to an optimal cluster search (see section 4.3.4), since a best partition is the one that would display the highest effectiveness. Therefore, referring to the evaluation measures reported in section 4.3.5, cluster effectiveness is gauged using the MK1 measure, and IFS effectiveness is gauged using the MK1-*k* measure. Their results showed that if for every query a user was to select the best cluster, then the effectiveness of the clustering would be higher than that of the inverted search for all values of *n*. As I discussed in section 4.3.5, this type of comparison is not fair on inverted search, since it compares a theoretical maximum effectiveness (MK1 is attained if a user was to infallibly select the best partition) against a value that is not optimal (MK1-*k*; there is no guarantee that IFS effectiveness reaches optimality at the cut-off point *k* of the best partition).

Finally, Hearst and Pedersen provide some evidence about the effect of varying numbers of top-ranked documents on clustering effectiveness, however, they do so implicitly when comparing the effectiveness of cluster-based and inverted file searches (p. 82). Their results are obscured by the fact that documents within the best cluster are ranked based on two different methods. When documents are ranked based on their closeness to the query, clustering effectiveness increases as the number of top-ranked documents increases. When, on the other hand, documents are ranked based on their similarity to cluster centroids, then the highest effectiveness seems to be attained when clustering between 250 and 500 top-ranked documents. It should also be noted that Hearst and Pedersen did not test the statistical significance of these results, as this research direction did not fall within their experimental aims.

[19] In fact, the authors mention this in passing in page 81 when referring to a different matter.

### 5.2.2 Some notes on post-retrieval clustering

Despite the limitations of the experimental methodology of Willett (1985), and Hearst and Pedersen (1996), the relation of post-retrieval clustering to retrieval effectiveness has not been examined by other researchers. Most of the reported research that adopts clustering of retrieval results seems to do so rather casually, without consideration for effectiveness issues. For example, (Allen *et al*., 1993; Kirriemuir & Willett, 1995; Leuski & Allan, 1998; Zamir & Etzioni, 1998; Carey *et al*., 2000; Eguchi *et al*., 2001) all arbitrarily select a number *n* of top-ranked documents to cluster, without any investigation into the effect that the choice of *n* may have on retrieval effectiveness.

Kirriemuir and Willett (1995), for example, were interested in examining the effect of hierarchic clustering methods and similarity coefficients on the ability of an IR system to detect duplicate, or near-duplicate, full-text records in a newspaper archive. They did not investigate clustering effectiveness across different numbers of retrieved documents, or the comparative effectiveness of cluster-based and IFS retrieval. Leuski and Allan (1998) examined two and three dimensional visualisations of the top-50 retrieved documents returned from a similarity search. They did not cluster the retrieved document sets, instead they investigated how the visualisation strategies that they proposed affected the spatial proximity of relevant documents (i.e. whether the visualisations placed relevant documents close to each other). Their results demonstrated that relevant documents were visually placed close to each other, providing evidence that the cluster hypothesis holds both for two and three dimensional representations of the document sets.

Similar to the post-retrieval organisation of search results by means of hierarchic clustering is the organisation by means of hierarchic classification, or categorisation, to pre-defined categories (Pratt *et al*., 1999; Chen & Dumais, 2000). This can be an effective and intuitive way to allow users to navigate through retrieved documents, but as I discussed in section 3.1 classification is viewed as a process distinct to that of clustering.

From the discussion of post-retrieval clustering that was presented in the previous paragraphs, three issues need to be further considered. The first issue is the effect that varying numbers of top-ranked documents have on the validity of the cluster hypothesis, and on cluster-based effectiveness. As I previously discussed, Hearst and Pedersen (1996) partially address the effect of varying top-ranked documents on clustering effectiveness, with results depending on the method used to rank the documents within the best clusters, and with no testing for the significance of the results. As far as the effect on the validity of the cluster hypothesis is concerned, Willett (1985) suggested that for static clustering relevant and non-relevant documents seem to be better separated than for post-retrieval clustering. However, his approach was limited by the use of the coordination level search.

The second issue relates to the comparison of the effectiveness of post-retrieval clustering to that of static clustering. Addressing this issue would provide insight into whether it is worthwhile, from an effectiveness point of view, to pursue post-retrieval clustering. Anick and Vaithyanathan (1997) have suggested that by clustering top-ranked documents, one may potentially miss some relevant documents that may have been ranked low by an inverted file search. Such documents might have been picked up by static clusters through interdocument associations, thus enhancing the recall of the system. Post-retrieval clustering simply disregards documents that have been ranked below rank position *n*. The issue of the comparative effectiveness of static and post-retrieval clustering has been left unaddressed by Hearst and Pedersen. Willett (1985) indicated that the effectiveness of static clustering is higher, but this result should be seen tentatively because of the use of the coordination level search.

The third issue pertains to the investigation of the comparative effectiveness of post-retrieval clustering and inverted file search. Willett's study did not address this issue. Hearst and Pedersen's research suggested that, for the Scatter/Gather system, the best cluster was always more effective than an equivalent cut-off of the ranking produced by an inverted file search. However, the use of inappropriate means of comparing effectiveness prohibits the extraction of any definite conclusions. An issue that needs to be emphasised here is that the result of any comparison will depend on the quality of the initial retrieval.

## 5.3 Reviewing the role of the cluster hypothesis

The cluster hypothesis conceptually lies in the heart of the clustering process. If relevant documents are indeed more similar to each other than to non-relevant ones, then the effectiveness of CBR should indeed be high as the likelihood of placing documents relevant to the same requests (*co-relevant* documents) in the same clusters will also be high.

From the definition of the cluster hypothesis it becomes evident that the concept of similarity is central to it: "closely associated documents tend to be relevant to the same requests" (Van Rijsbergen, 1979; p. 45). The tests that are typically used to quantify the degree at which test collections adhere to the cluster hypothesis (see section 4.2) take as input the set of interdocument associations for each collection, and output a numerical value that is treated as an indication of the comparative clustering tendency of these collections (Jardine & Van Rijsbergen, 1971; Voorhees, 1985a, 1985b; El-Hamdouchi & Willett, 1987).

Here I propose an alternative view of the cluster hypothesis. According to this view, the hypothesis should not be seen as a test for an individual collection's clustering tendency. Instead, I argue that the hypothesis should be valid for every collection, and should therefore be seen as an axiom of cluster-based retrieval. I postulate that, for any given query, pairs of relevant documents

will exhibit an inherent similarity which is dictated by the query itself. Under this view, and contrary to the traditional treatment of the hypothesis in the literature so far, failure to validate the hypothesis is not caused by properties of the test collection(s) under examination. Instead, it is caused by failure to structure the document space in such a way that the inherent similarity of documents that are jointly relevant to the same queries can be detected.

Gordon (1991) has also proposed the axiomatic view of the cluster hypothesis. He pursued the clustering of co-relevant documents through altering their indexing representations based on the way that documents are accessed by users[20]. Shaw and his colleagues (1997) have also suggested that the way the cluster hypothesis has so far been treated may be a reason for the failure of cluster-based retrieval to be highly effective. To the best of my knowledge, there is no experimental evidence reported to validate such claims.

The structuring of the document space prior to clustering is implemented through the calculation of the interdocument associations between pairs of documents that are considered for clustering. The outcome of the association calculations dictates the positions of documents relative to each other, and also constitutes the input to a clustering method that may be applied to the database.

### 5.3.1 The static nature of interdocument relationships

Let us consider, for illustration, the cosine coefficient. The formula for it is given by Equation 5.1. Let us assume that a document $D_i$ is a vector of length $n$ comprising binary or weighted entries, and that each entry corresponds to an indexing term: $D_i = \{d_{i1}, d_{i2}, \ldots, d_{in}\}$. The similarity of any two documents $D_i$ and $D_j$ belonging to a document collection $X$ is then given by Equation 5.1.

In a typical document clustering application, interdocument relationships are calculated statically. This means that for any two documents $D_i$ and $D_j$ in a document collection, their similarity $Sim(D_i, D_j)$ will have a value that will be the same under all queries that a user may pose to the IR system. This is clearly demonstrated by Equation 5.1: the similarity between the two objects depends only on the weights of their constituent terms ($d_{ik}$ and $d_{jk}$). Therefore, for a particular document collection $Sim(D_i, D_j)$ will be the same across all requests.

$$Sim(D_i, D_j) = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} \qquad (5.1)$$

Document clustering has also typically been applied statically over an entire collection prior to querying (static clustering). Hence, there has been no practical reason to reconsider static

[20] More information on this class of clustering approaches is given in section 5.3.2.

similarity calculations for a clustering that is itself static. However, even under post-retrieval clustering interdocument similarity is defined in the same static way.

As seen in section 5.2, under post-retrieval clustering a different set of documents is clustered for each query. A consequence of this is that the similarity between any two documents $D_i$ and $D_j$, assuming that they are both included in the document sets to be clustered for different queries, will be different under each query. This difference is introduced implicitly, and is not explicitly defined by the similarity measure used (e.g. Equation 5.1 is typically used for post-retrieval clustering). It is implicitly introduced due to the different documents retrieved in the top-$n$ ranks in response to different queries. Similarity in this case will vary because the term weights of documents ($d_{ik}$ and $d_{jk}$ in Equation 5.1) will also vary depending on other documents that are in the same neighbourhood. However, it should be noted that if binary (presence/absence) term representations are used then similarity will remain static.

Both in the static and in the implicitly variable use of similarity under post-retrieval clustering, interdocument associations are defined through enumeration of common terms, and a mathematical formulation that quantifies this enumeration (e.g. Equation 5.1). According to this view, all dimensions (i.e. terms) are deemed equally relevant at contributing towards the similarity value, and furthermore, the importance of dimensions does not change depending on the query. The use of term weighting schemes for document vectors does not address this issue, firstly because such schemes are not always applied when calculating interobject similarities - binary representations are often used – (Van Rijsbergen, 1979; Willett, 1983; Ellis *et al*., 1993), and secondly because such schemes weight terms according to their indexing importance within a document collection (Van Rijsbergen, 1979), and not according to their value as salient features for the purpose of clustering relevant objects together.

The static calculation of interdocument similarity seems to neglect some potentially important information: the context under which the similarity of the two documents is judged. Evidence by a number of researchers in fields such as those of philosophy, cognition, experimental psychology, and memory based reasoning (MBR) (Goodman, 1972; Tversky, 1977; Nosofsky, 1986; Stanfill & Waltz, 1986) suggest that similarity is a highly dynamic concept that is highly influenced by purpose.

Goodman, (1972), for example, 'accused' similarity of being an insidious and highly volatile concept. He suggested that one can "tie the concept of similarity down" by selecting some important features on which to judge similarity. Tversky, (1977), for the specific task of classification, argued that the salience of features is determined, in part, by their *classificatory significance*, or *diagnostic value*. A feature may acquire diagnostic value, and hence become more salient in a particular context, if it serves as a basis for classification in that particular context. Each class should then contain objects that are similar to each other in the sense that they are

similar in respect to these important features. Nosofsky, (1986), for assessing similarity in a psychological space, and (Stanfill & Waltz, 1986) for determining similarity of cases for MBR, have adopted similar views.

The IR community, on the other hand, has adopted, in a rather casual way, the static nature of interdocument similarity. This seems surprising if one considers that document clustering is a highly goal-driven process: relevant documents should be grouped together, and separately from non-relevant ones. Therefore, a static similarity is unlikely to be able to structure the document space in such a way that the proximity of co-relevant documents is promoted on a per-query basis.

One of the aims of the research that is reported in this thesis, is to devise means by which the document space can be structured in a way suitable to detect the inherent similarity of co-relevant documents. To this end, I propose the use of *query-sensitive similarity measures* (*QSSM*) that bias interdocument relationships towards pairs of documents that jointly possess attributes (i.e. terms) that are expressed in a query. I consider the query terms to be the salient features that define the context under which the similarity of any two documents is judged. This is a novel approach to calculating interdocument relationships, and is motivated by the belief that similarity is a dynamic concept that is highly influenced by purpose. In the context of IR, purpose can be defined as a per-query adherence to the cluster hypothesis as explained in section 5.3. It is this goal that clustering through query-sensitive similarity measures aims to accomplish.

A hierarchic clustering method, like any of the four reviewed in section 3.4, takes the similarity matrix containing all interdocument associations as an input, performs a specific transformation on the matrix, and generates a hierarchic structure. By altering the way that interdocument associations are calculated, one changes the input to the clustering method, and consequently also changes the generated hierarchic structures. The study of hierarchic clustering methods that employ query-sensitive similarity measures for the calculation of interdocument relationships is one of the main aims of this thesis.

### 5.3.2 Related work

The use of query-sensitive similarity measures aims at increasing the similarity of co-relevant documents on a per-query basis, so that the probability that such documents are placed in the same clusters is also increased. A number of approaches that try to ‘force’ co-relevant documents in the same clusters have been developed in the past under the name of *user-oriented*, or *adaptive clustering*. In general, these methods rely on user-supplied feedback in order to determine the degree of association between documents rather than on statistical interpretation of their contents (Gordon, 1991).

The work carried out by Ivie (1966) was pioneering in this field, and it made the first suggestions for the use of an adaptive clustering system. Ivie mentions that (p. 29) "the purpose of each interaction of a user with the system is … a partitioning of the total collection into two disjoint subsets - one containing all documents that are of interest to the user and the other containing those not of interest". Clearly, what Ivie proposes is a per-query adherence to the cluster hypothesis by employing user feedback. He also puts forward the idea of monitoring co-usage patterns of documents in order to determine interdocument similarity (p. 29): "a measure of the relatedness between any two documents based on their usage and co-usage patterns … is to be utilised to facilitate the request-to-answer transformation".

The concepts that Ivie put forward were picked up, almost twenty years later, by a number of other researchers that pursued the adaptive clustering of documents (Yu *et al*., 1985; Deogun & Raghavan, 1986; Gordon, 1991; Bhatia & Deogun, 1993). Typically, such approaches involve the monitoring of the way that users access documents over a period of time. If two documents are jointly accessed by a number of users, then it can be argued that these documents are similar to each other and should be placed in the same clusters. Yu *et al*. (1985), for example, suggested the formation of clusters by adaptively repositioning documents on a real number line that represents the distance between documents. The position on the line is based on user feedback (i.e. jointly relevant documents are pulled closer together), and documents that lie close in this line are assigned to the same clusters. Deogun and Raghavan (1986), developed a method of user-based clustering by partitioning the original document set in such a way that only documents relevant to the same queries are placed in the same clusters. Since this strategy results in many small clusters (even as small as one document), heuristics have to be applied to merge such clusters in larger ones.

Gordon (1991) proposed an adaptive clustering method by redescribing documents (i.e. changing their indexing representations) by means of a genetic algorithm. Each document in a collection is assigned multiple descriptions. The descriptions for a single document compete with each other, with fitter descriptions being those that match relevant queries and do not match non-relevant ones (user-supplied relevance assessments are required). The genetic algorithm causes the set of descriptions associated with a specific document to move over time towards those queries to which they are relevant, and away from those to which they are not. In this way, documents that tend to be relevant to the same requests will have descriptions that move closer to each other. Documents are then clustered based on these descriptions. The goal of Gordon's algorithm is for documents to eventually have closely associated descriptions because they are relevant to the same queries, and consequently, for the validity of the cluster hypothesis to be enforced.

Gordon' s approach is characterised by two rather strong hypotheses. First, it is assumed that queries that are relevant to a specific document will be descriptively similar. If this is not the case,

then the descriptions of documents that are redescribed by the algorithm to match these queries will not group tightly together. The second hypothesis is that documents will be co-relevant to the same queries. If this hypothesis does not hold, then the similarity in the documents' descriptions brought about by their co-relevance to one query, will be reduced as each document separately is relevant to different sets of queries (Gordon, 1991). It must be noted that, to date, there is no experimental evidence to suggest the validity of either of these two hypotheses. Therefore, it can be argued that the applicability of this method is restricted to those environments for which these hypotheses are valid.

Adaptive clustering methods implicitly assume that there are means of monitoring user activities, collecting usage information, and incorporating this information in the clustering system. Moreover, in most of the adaptive approaches it is assumed that the user will perform his searches on the same document collection, since user behaviour over time is monitored to optimise clustering on a specific collection. Most of these assumptions might not be realistic in an operational environment where user searches can be performed on a number of different databases, or where users may not be willing to provide feedback or document usage information. These requirements imposed by such methods have most likely contributed to the limited use of adaptive clustering in widely used IR systems, such as for example web-based search engines.

In contrast to adaptive clustering methods, clustering through the use of query-sensitive similarity measures does not require any form of user feedback, nor does it rely on the user interacting with a single database. Query-sensitive similarity measures assume that the only information available, apart from the collection of documents, is the query posed by the user. In this way it can be argued that the applicability and the utility of the proposed approach is greater than that of the adaptive clustering methods.

Apart from adaptive clustering methods, other approaches that conceptually share the same goal with clustering based on QSSM (i.e. forcing co-relevant documents in the same clusters) can be found in the literature. El-Hamdouchi (1987), for example, proposed a (static) clustering approach that aims at generating clusters with a high probability of containing co-relevant documents. El-Hamdouchi's method uses a probabilistic function that ranks documents, or sets of documents, in relation to a query, and does not challenge the use of static interdocument similarity.

Bartell et al. (1995) propose a method for creating indexing representations of documents based on modelling target (static) interdocument similarity values. The authors partitioned three test collections into a training set and a test set. They used the relevance information contained in the training test to construct the target interdocument similarities (with the similarity of co-relevant documents artificially augmented), and the test set to attempt to model the target similarities. The authors concluded that their method succeeds in enforcing the validity of the cluster hypothesis, as this was demonstrated by results of the overlap test (section 4.2.1). However, the use of the

target similarity matrix is a major limitation of this approach, something that the authors themselves acknowledge. Moreover, the static calculation of interdocument similarity is not addressed in this research.

Wen et al. (2001) applied the idea of usage-based clustering not to documents, but rather to queries issued to a web-based search engine. The authors argue that if two queries result in the user viewing the same documents, then such queries are similar and should be accordingly grouped together. The similarity of two queries is then calculated as a function of their content overlap, and of the sets of documents to which the queries jointly provide access. This work assumes that when a user views a document during a web search session, he implicitly indicates that the document is relevant to the query (implicit relevance feedback). This is a rather strong assumption that has not been fully supported in experiments reported so far (White *et al*., 2002).

## 5.4 Research objectives

In the last two sections (5.2 and 5.3), I outlined two methods through which a user-supplied query can influence the output of hierarchic clustering methods. The main aim of this thesis is to investigate the effectiveness of such query-based document hierarchies.

Through investigating the effectiveness of query-based hierarchic clustering, I also aim to challenge previous findings in the field that have argued for the inappropriateness of clustering as an effective retrieval mechanism, and to argue that such failure can be attributed to the way that clustering has typically been performed. In sections 5.4.1 and 5.4.2 I outline the specific research issues that I will attempt to address through the experimental work that is reported in this thesis.

### 5.4.1 The effectiveness of post-retrieval hierarchic clustering in IR

In section 5.2.2 I raised a number of issues regarding post-retrieval clustering that have been left unaddressed so far. This gap in the state of research is the main motivation for the experimental work that I report in Chapter 6. I believe that the area of post-retrieval clustering merits greater interest than it has received by IR researchers. Also, in contrast to other researchers (e.g. Allen *et al*., 1993; Leuski & Allan, 1998; Allan *et al*., 2001), I do not view post-retrieval clustering merely as a convenient means of presenting and visualising retrieval results to users. Instead, I aim to analytically investigate the viability of post-retrieval clustering based on the grounds of its retrieval effectiveness.

The issues that I aim to investigate are the following:

1. The effect that varying numbers of top-ranked documents that are considered for clustering have on clustering effectiveness. The motivation for pursuing this research

direction is to gain an understanding of the behaviour of post-retrieval clustering under different experimental conditions (e.g. different test collections, clustering methods, etc.)

2. The comparative effectiveness of static and post-retrieval clustering. By comparing the retrieval effectiveness attainable by static clustering to that attainable by clustering variable numbers of top-ranked documents, it should be possible to appreciate whether incorporating query information into the clustering process is worthwhile

3. The effect that different numbers of top-ranked documents have on the validity of the cluster hypothesis. This research direction aims to investigate whether there is a variation in the degree to which the cluster hypothesis is valid, caused by the consideration of different numbers of top-ranked documents

4. The comparative effectiveness of cluster-based (both static and post-retrieval) and IFS retrieval. The primary aim is to examine whether post-retrieval clustering has the potential to act as an effective retrieval mechanism, one that could improve the effectiveness of conventional similarity ranking systems.

It is worth noting that the effectiveness of post-retrieval clustering can also be investigated for other types of clustering methods apart from hierarchic ones. It is not the aim of this thesis to examine such issues. Hierarchic methods form the focus of the work reported here, for reasons that I outlined in section 3.1.

### 5.4.2 Query-sensitive similarity measures for the calculation of interdocument relationships

In section 5.3 I proposed an axiomatic view of the cluster hypothesis that stems from the intuition that documents relevant to the same query exhibit an inherent similarity that constitutes them more similar to each other than to non-relevant documents. In the same section, I also introduced the notion of query-sensitive similarity measures that can be used to detect this inherent similarity. In Chapters 7 and 8 I examine the applicability and effectiveness of QSSM in the context of document clustering.

The main experimental aims of these two chapters are:

1. To propose specific formulas by which QSSM can be defined. In the present chapter I have merely proposed the use of QSSM on a conceptual level. In Chapter 7 I propose specific formulas that can incorporate the query influence in the calculation of interdocument associations

2. To investigate whether the use of QSSM for measuring interdocument relationships succeeds in enforcing the validity of the cluster hypothesis when compared against conventional similarity measures. This issue is also examined in Chapter 7, and is viewed as a test for the applicability of QSSM to IR

3. To investigate whether the application of QSSM to document clustering can improve retrieval effectiveness when compared both to the effectiveness attained by conventional clustering methods (i.e. using a conventional similarity measure), and to the effectiveness obtained by IFS. This issue is examined in Chapter 8.

In addition to these aims that form the focal points of this research, the opportunity to study the comparative effectiveness of four hierarchic clustering methods (those reported in section 3.4) lends itself. Despite that these four methods have been extensively compared in the context of IR (Griffiths *et al*., 1984; Voorhees, 1985a; El-Hamdouchi & Willett, 1989), this has occurred only for static clustering. The comparative effectiveness of these methods under post-retrieval clustering has not been investigated, and neither has their comparative effectiveness using query-sensitive similarity measures.

It should be noted that it is not only hierarchic clustering methods that can employ query-sensitive similarity measures. Other types of clustering methods can also use such measures, since the majority of clustering methods known rely on some form of interdocument association measure. As with post-retrieval clustering, it is not the aim of this thesis to examine the effectiveness of clustering methods other than hierarchic ones.

## 5.5 The experimental environment

In this section I describe the various components that form the experimental environment in which the research objectives of this thesis are investigated. The main components of the experimental system are presented in Figure 5.1. It must be noted that variations in this baseline environment will be made in the following chapters, and specifically in Chapters 7 and 8 where the query-sensitive similarity measures will be investigated. Where such variations occur I will explicitly make appropriate reference to the component of the baseline system in Figure 5.1 that they correspond to.

Documents and queries are taken from standard IR test collections, the details of which are presented in section 5.5.1. Documents and queries are processed, indexed and matched against each other by means of an IR system; the details of this process are outlined in section 5.5.2. Section 5.5.3 presents details of the clustering system that is used, and section 5.5.4 outlines the method used for obtaining measurements of retrieval effectiveness.

### 5.5.1 Test collections

Six document collections are used in the experimental work. Four of them (CACM, CISI, LISA, and Medline) have been used by other researchers for experimentation with hierarchic clustering methods (Voorhees, 1985a; Griffiths *et al*., 1986; El Hamdouchi & Willett, 1989; Burgin, 1995), and the remaining two are part of the TREC standard collections (Harman, 1993).

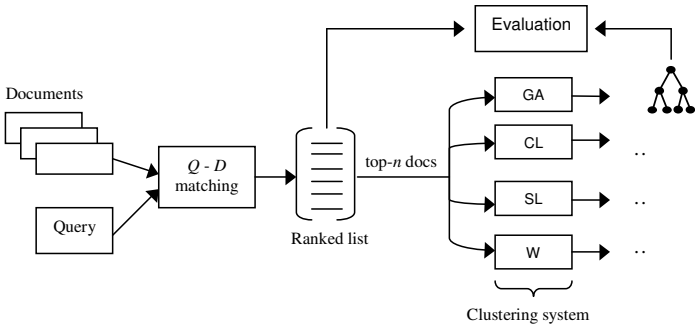


**Figure 5.1.** The experimental system

The part of the Medline database that is used in the experiments is widely known as Medlars. This database is atypical as Bartell et al. (1995) note, in the sense that the 30 queries of this collection partition it into disjoint sets. This means that each document is relevant to either one or none of the 30 queries. There are therefore 31 disjoint classes of documents: one class of relevant documents per query, and one "garbage" class that contains documents that are relevant to no query. One can therefore argue that Medlars displays a structure that is well suited to clustering. If a clustering method succeeds at grouping relevant documents separately from non-relevant ones, then this structure may be sufficient for all queries of this test collection, i.e. a static clustering of this collection may be as effective as a dynamic one. Voorhees (1985a) showed that documents relevant to a query can be distinguished from other documents in the dataset because they appear in a (static) cluster of topically-related documents associated with the query.

Statistics for the six document collections are presented in Table 5.1. It should be noted that the four smallest collections (CACM, CISI, LISA, and Medline) are homogeneous, treating one major subject area (e.g. Library and Information Science, Biomedicine, etc.), and such topical homogeneity may effect the experimental results. The AP and WSJ collections, on the other hand, cover in their documents a wide variety of topics, providing two collections with different characteristics. For these two collections, TREC topics (i.e. queries) 1-50 were randomly chosen and used in the experiments.

<num> Number: 003
<dom> Domain: International Economics
<title> Topic: Joint Ventures
<desc> Description:
Document will announce a new joint venture involving a Japanese company.
<narr> Narrative:
A relevant document will announce a new joint venture and will identify the partners ...
<con> Concept(s):
1. joint venture, tie up
2. partner, cooperation, joint management, …
...
</top>

**Figure 5.2.** A sample TREC topic

TREC topics comprise many fields, and can be long and detailed (Figure 5.2). This raises the issue of what part of the topic to use for retrieval purposes. The *'title'* section of the topics is felt to be typical of the queries that users might enter in an actual IR system, while the other sections are regarded as a much more detailed description of the information need that is unlikely to be used by an actual user. The 'title' sections of the 50 topics that were used contain on average 3.2 terms. Since this average is rather low, and given that no special consideration is taken to cater for very short queries in this experimental environment, it was felt that the effectiveness of the initial retrieval may consequently suffer for these two collections.

To this end, a number of manually selected terms from the '*concepts'* field are added to the terms of the 'title' section of the topics. The 'concepts' field usually lists terms and phrases that the creator of the query thinks are related to it (Harman, 1993). On average 4.4 terms per query are added from the concepts field, yielding an average of 7.6 terms per query for the AP and WSJ collections (Table 5.1).

| | *AP* | *CACM* | *CISI* | *LISA* | *MED* | *WSJ* |
|---|---|---|---|---|---|---|
| Number of docs. | 79,919 | 3204 | 1460 | 6004 | 1033 | 74,520 |
| Mean terms per doc. | 370 | 22.5 | 43.9 | 39.7 | 51.6 | 377 |
| Number of queries | 50 | 52 | 35 | 35 | 30 | 50 |
| Mean terms per query | 7.6 | 13 | 7.6 | 19.4 | 9.9 | 7.6 |
| Mean relevant docs per query | 42.4 | 15.3 | 49.8 | 10.8 | 23.2 | 71.4 |
| Total relevant docs. | 2122 | 796 | 1742 | 379 | 696 | 3572 |

**Table 5.1.** Collection statistics

Apart from the topical differences among the six test collections, one can also note a significant degree of variability in their statistics. For example, CACM and LISA both have few relevant documents per query, whereas AP, CISI, and especially WSJ have a large number of relevant documents per query. The average length of the documents belonging to the various collections is also considerably variable. The two TREC collections, for example, have a much larger average

number of terms per document compared to any of the other four collections. CACM stands at the other end of the spectrum, with the smallest number of index terms assigned per document.

The variability in the characteristics of the test collections used is a desirable property of the experimental environment. It allows a researcher to investigate the research aims in a variety of settings, and hence to demonstrate that any results obtained are generally valid (Voorhees, 1985a).

### 5.5.2 IR system

In order to cluster a set of documents that have been retrieved in response to a query, one needs to define the environment under which the initial retrieval takes place. The SMART experimental information retrieval system (Salton, 1971) is used to this end. Documents and queries in SMART are represented as vectors in a multidimensional space (sections 2.2, 2.3).

$$w_k = \frac{(\ln(tf_k)+1)\cdot\log\frac{N}{n}}{\sqrt{\sum_{vector}(\ln(tf_k)+1)\cdot\log\frac{N}{n})^2}} \qquad (5.2)$$

A *tf-idf* weighting scheme (section 2.2) is used in the experiments, both for document and query terms, that involves cosine length normalisation - SMART's *ltc* scheme (Salton & Buckley, 1988). According to this scheme, the weight $w_{dk}$ of term $k$ in document $d$ (or query $q$) is given by Equation 5.2, where $tf_k$ is the term frequency of the term in document $d$ (or query $q$), $N$ is the total number of documents of the collection, and $n$ the number of documents in which term $k$ occurs. The default SMART stoplist and stemming algorithm are used for processing the documents and queries of all test collections.

SMART performs comparisons between documents and queries, and outputs a ranking of documents in decreasing order of their computed similarity to the queries. The matching function employed by SMART is given by the cosine formula in Equation 5.3 (assuming that length-normalised term weights are used) (Salton & Buckley, 1988). In this formula $w_{qk}$ and $w_{dk}$ represent the weights of terms that belong to the query $Q$ and the document $D$ respectively.

$$Sim(Q,D) = \frac{\sum_{k=1}^{t} w_{qk}\cdot w_{dk}}{\sqrt{\sum_{k=1}^{t}(w_{qk})^2\cdot\sum_{k=1}^{t}(w_{dk})^2}} \qquad (5.3)$$

Table 5.2 gives the average precision values for 11 recall points, and 3 recall points (0.2, 0.5, 0.8) for the initial retrieval for all six collections. As I mentioned in section 5.5.1, the documents of the two TREC collections (AP and WSJ) are much longer than those of the other four collections. Singhal *et al*. (1996) have suggested that the cosine coefficient, used by the SMART system to

match documents and queries, can be affected by document length. Regarding this issue, I feel that the retrieval effectiveness for these two collections is sufficient for the aims of the experimental work.

| | AP | CACM | CISI | LISA | MED | WSJ |
|---|---|---|---|---|---|---|
| 11 pt. Avg. | 0.2568 | 0.3778 | 0.1945 | 0.3115 | 0.5699 | 0.2546 |
| 3 pt. Avg. | 0.2361 | 0.365 | 0.1678 | 0.2991 | 0.5836 | 0.2259 |

**Table 5.2.** Initial retrieval evaluation

After the initial retrieval, and for each query of each test collection, the top-*n* ranked documents are used to create the collections that are clustered. Seven different values of *n* are used: 100, 200, 350, 500, 750, 1000, and full collection size (*n* = collection size). The value of 1000 is not used in the CISI and Medline collections because their sizes are 1460 and 1033 documents respectively. The full AP and WSJ collections (79,919 and 74,520 documents respectively) are not clustered for practical reasons.

### 5.5.3 Clustering methods

Four hierarchic methods are used in the experiments, namely the single link, complete link, group average, and Ward's methods. The main reason behind the choice of these four methods is that they have been extensively used and examined in the context of IR (e.g. Van Rijsbergen & Croft, 1975; Griffiths *et al*., 1984; Voorhees, 1985a; El Hamdouchi & Willett, 1989). The methods are implemented in ANSI C based on the algorithms that are given in (Späth, 1980).

Apart from Ward's method which requires a specific form of distance measure that minimises the within group variance (Wishart, 1969), the association measure used for the other three methods is the cosine coefficient. Experiments with the normalised Euclidean distance and the Dice coefficient did not produce significantly different results, something which is in agreement with previous suggestions and findings (Van Rijsbergen, 1979; Willett, 1983; Ellis *et al*., 1993; also see section 3.3.2).

The document sets to be clustered comprise either all the documents of a test collection, or the *n* top-ranked documents of a test collection (*n* = 100, 200, 350, 500, 750, 1000). Document terms belonging to these sets are weighted using the same weighting scheme as for the initial retrieval (*ltc*). After initial experimentation with different vector weighting schemes (binary, term frequency weights) for clustering, no significant differences were found – again in agreement with previous findings (Willett, 1983; also see section 3.2.2).

Document terms are weighted locally within the retrieved sets prior to clustering, as opposed to globally over an entire data set. Korpimies and Ukkonen (1998) suggested that local term

weighting for the purpose of document clustering is more beneficial than global weighting. Whether there are effectiveness gains by local document term weighting was not investigated in this thesis. Also, the most exhaustive indexing representations are used for documents, as the variation of indexing exhaustivity was not considered to be a relevant experimental parameter in the present study (see section 3.2.1).

| | *AP* | | | | *WSJ* | | | |
|---|---|---|---|---|---|---|---|---|
| *n* | *Group Average* | *Ward* | *Complete Link* | *Single Link* | *Group Average* | *Ward* | *Complete Link* | *Single Link* |
| 100 | 11.7 | 10.1 | 8 | 24.8 | 12.6 | 8.8 | 8 | 28.3 |
| 200 | 15.4 | 12.3 | 9.3 | 46.9 | 16.7 | 10.4 | 9.4 | 54.1 |
| 350 | 18.6 | 13.9 | 10.5 | 79.3 | 21 | 11.9 | 10.6 | 91.2 |
| 500 | 21.4 | 15 | 11.2 | 112 | 24.3 | 12.7 | 11.6 | 129.7 |
| 750 | 24.6 | 16.2 | 12.6 | 167.4 | 28.6 | 13.6 | 13 | 196.7 |
| 1000 | 26.6 | 17 | 13.6 | 221.4 | 31.8 | 14.5 | 14.5 | 263.4 |

**Table 5.3.** Average cluster sizes for the four methods using the AP and WSJ collections

In Table 5.3 the average cluster sizes for the two TREC collections (AP and WSJ) are presented, using the hierarchies generated by the four clustering methods. From the data presented one can note that the only method for which average cluster size significantly increases as the number *n* of top-ranked documents increases, is single link. The other three methods produce hierarchies that are little affected by the increase in the number of documents clustered. This is especially true for the complete link and Ward's methods, which tend to produce small, compact clusters whose size does not significantly vary based on the number of documents clustered (Milligan *et al*., 1983). This behaviour is typical of the four methods used (Murtagh, 1984b), and is consistent across the six document collections.

Hierarchic agglomerative methods usually have a time complexity of $O(N^3)$, something that makes them an inefficient solution for the clustering of large data sets. A dynamic, post-retrieval clustering method should have efficiency as a high priority (Zamir & Etzioni, 1998). However, efficiency issues are not tackled in this thesis for a number of reasons. I provided two such reasons in section 3.8. A further reason is that for dynamic, post-retrieval clustering small numbers of documents are clustered (Willett, 1985). For small values of *n* (e.g. 100, 200) hierarchic methods have acceptable performance for on-line clustering. Moreover, improvements in the time efficiency of the hierarchic methods can be achieved by using efficient algorithms for their implementation, such as the ones by (Van Rijsbergen, 1971; Sibson, 1973; Defays, 1977; Voorhees, 1985a, 1986). Further improvements can be achieved by using efficient methods for the calculation of the similarity matrix, such as the ones proposed by Croft (1977) and Willett (1981). Such improvements are not considered in this thesis, since the present research is focused solely on issues of effectiveness.

### 5.5.4 Measuring retrieval effectiveness

Two separate issues need to be considered with regard to the evaluation of retrieval effectiveness. First, how cluster-based effectiveness is to be measured, and secondly, how IFS effectiveness is to be gauged and compared to cluster-based effectiveness.

Regarding the first issue, optimal cluster-based searches are used to gauge the effectiveness of the various clustering strategies used. In Chapters 3 and 4 (sections 3.5.1, 3.5.2, 4.3.4) I argued on the appropriateness of optimal cluster searches when a researcher is comparing the effectiveness of different clustering strategies. This type of comparison is in agreement with the research objectives of this thesis, which mainly involve measuring cluster-based effectiveness using different strategies. Comparisons are primarily made across static clustering, post-retrieval clustering that considers varying numbers of top-ranked documents, and clustering that employs query-sensitive similarity measures. In addition, comparisons are also made among the effectiveness obtained by different clustering methods. For the reasons that I have mentioned in Chapters 3 and 4, I consider optimal cluster search to be better suited to these experimental aims than other cluster-based searches. Consequently, the MK1 measure (section 4.3.5) is used to quantify cluster-based effectiveness.

As far as the issue of the comparison of cluster-based and IFS effectiveness is concerned, the three measures that were presented in section 4.3.5 are used to gauge IFS effectiveness (MK1-k, MK3, MK4), and to compare it to cluster-based effectiveness (MK1). The benefit of using three distinct measures is that one can examine the degree at which cluster-based effectiveness exceeds (if at all) IFS effectiveness. For example, a clustering strategy $C_1$ may fail to exceed IFS effectiveness at the MK4 level, but it may succeed to do so at the MK3 level. By noting how optimal cluster effectiveness compares to different levels of IFS effectiveness, one may be able to extract useful conclusions about the potential of the former to exceed the latter.

The comparison of optimal cluster effectiveness to IFS effectiveness may raise some criticism, since the former does not correspond to actual effectiveness values obtained by an operational search strategy. Regarding this issue, I feel that given the research objectives of this thesis, the use of optimal cluster effectiveness is warranted. It is the aim of this thesis to examine the effectiveness improvements that can be introduced in the clustering process by utilising information from the query. By using optimal searches, one demonstrates that clustering has the potential to achieve a certain level of effectiveness. Whether this level will actually be achieved, is an issue that should concern researchers interested in the development of, for example, more effective search strategies or more effective cluster representation schemes.

Moreover, a specific search strategy may favour a certain hierarchy type, or may be more effective when a certain cluster representation scheme is used as Voorhees (1985a) suggested.

Given that it is not within the objectives of this thesis to study such behaviour, I feel that the choice of optimal cluster searching is justified. The dependence of cluster-based effectiveness on cluster-based search strategies and representation schemes are separate research topics in their own right, and not ones that I aim to address in this thesis. It may also be the case that the research and the results reported here will instigate further research towards addressing such issues.

I also view the use of variable effectiveness measures for gauging IFS effectiveness, and especially measure MK4, as providing further credibility to the experimental results. One can argue that measure MK4 is as optimal as MK1 is, since there is no guarantee in either case that the calculated effectiveness value will be reached in an operational environment by a search or a user. Therefore, if a cluster-based strategy exceeds an IFS strategy at the MK4 level, it can be argued that the former has the potential to exceed the latter in an operational environment.

A final issue that needs to be examined is that of the statistical comparison of sets of experimental results. To this end, the Wilcoxon signed-ranks test (Siegel & Castellan, 1988) is used in this thesis. This test utilises information about the direction of differences between pairs of values, as well as the relative magnitude of the difference. The Wilcoxon test has been employed in a similar experimental environment in the past by Croft (1978) and El-Hamdouchi (1987). These researchers considered it as a powerful statistical tool that makes reasonable assumptions about the distributions of the values it is comparing. In the experiments reported in this thesis the sets of values that the test compares are values of the E measure, and the only assumption made by the Wilcoxon test is that such values come from the same family of distributions.

The procedure for the Wilcoxon signed ranks test is as follows (Siegel & Castellan, 1988):

- For any matched pair of E values let $d_i$ be the signed difference between the values. Ignore pairs where $d_i = 0$.
- Rank these $d_i$ values without respect to sign, i.e. give rank 1 to the smallest $d_i$, etc. When a tie occurs, assign the average of the tied ranks.
- To each rank affix the sign of the difference that the corresponding $d_i$ represents.
- Determine $T$ as the smaller of the sums of the like-signed ranks, and $N$ as the total number of $d_i$ having a sign.
- Calculate $z = \dfrac{T - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}$
- Calculate the probability $p$ of the value $z$ under the null hypothesis $H_0$ (i.e. that the difference in the E values is not significant). If $p$ is equal to, or less than, some level of

significance (typically .05), then reject the null hypothesis (i.e. the difference in the E values is significant).

## 5.6 Summary

In this chapter I outlined two methods by which a query can influence the output of hierarchic clustering methods. I called the class of clustering approaches that utilise query information *query-based*, and I established that the main aim of this thesis is to investigate the effectiveness of this type of clustering in the context of IR. I examined the two proposed methods under the view that clustering is a goal-driven process that aims, for each query, to group relevant documents together and separately from non-relevant ones.

Post-retrieval clustering was presented as one way of generating query-based hierarchies. Related work in this area was reviewed, and through this process it was established that there are a number of issues regarding the effectiveness of post-retrieval clustering that have been left unaddressed. These issues are experimentally investigated in Chapter 6, where I aim to systematically study the effectiveness of post-retrieval clustering.

The second approach for generating query-based document hierarchies uses query-sensitive similarity measures. To illustrate the motivation of this approach, the role of the cluster hypothesis was challenged in section 5.3, and an axiomatic view of the hypothesis was proposed. This view was based on the argument that co-relevant documents exhibit, on a per-query basis, an inherent similarity that is influenced by the context under which similarity is judged. By placing the calculation of interdocument associations in the context of the cluster hypothesis, I criticised their traditionally static nature. I proposed a different class of similarity measures, query-sensitive similarity measures (QSSM) which aim to detect the inherent similarity of co-relevant documents.

In this chapter I merely proposed the use of QSSM in document clustering based on their potential to increase the similarity of co-relevant documents. In Chapter 7 I propose specific formulas that can define this class of measures, and in Chapter 8 I examine the effectiveness of clustering methods that employ query-sensitive similarity measures for the calculation of interdocument relationships.

# Chapter 6

# The Effectiveness of Hierarchic Post-Retrieval Clustering

## 6.1 Introduction

This chapter aims to investigate the effectiveness of hierarchic post-retrieval document clustering. In the previous chapter (section 5.4.1) I outlined the research objectives that are pursued in this chapter, and I also described the experimental environment under which the research is carried out (section 5.5).

Under post-retrieval clustering varying numbers of documents may be considered as an input to the clustering system. The effect that the number of top-ranked documents has on the effectiveness of the clustering process is examined from two perspectives. First, in section 6.2, I examine the effect of the number of top-ranked documents on the structure of the document space in terms of the proximity of co-relevant documents. In this way it is possible to appreciate how effectively similarities between documents are calculated; the closer co-relevant documents are placed, the more likely cluster-based retrieval is to be effective. Subsequently, in section 6.3.1, I investigate the effect of the varying number of top-ranked documents on the optimal effectiveness of the document hierarchies. The comparative effectiveness of static and post-retrieval clustering is also examined in this same section.

Post-retrieval clustering challenges one of the implicit assumptions that have long determined the application of document clustering to IR: its static nature. The static nature of document clustering was criticised in sections 4.5 and 5.2 as being responsible for the relative failure of clustering to act as an effective alternative to conventional similarity search. In order to examine whether post-retrieval clustering can act as such an effective alternative, in section 6.3.2 I compare its effectiveness to that attained by inverted file search.

A by-product of the experimental procedure, since four clustering methods are employed, is a comparison of the effectiveness of these methods (section 6.4). This is in addition to the main focus of this chapter, since the comparative effectiveness of these four hierarchic methods has been extensively studied in the past (e.g. Griffiths *et al*., 1984; Voorhees, 1985a; El-Hamdouchi & Willett, 1989; Burgin, 1995), albeit under different experimental settings (i.e. static clustering). Finally, in section 6.5 a summary of the aims and findings of this chapter is presented.

## 6.2 Clustering tendency

The experiments reported in this section aim to examine the degree at which the clustering tendency of the six document collections is affected by the different values of $n$ used for the clustering of the top-$n$ ranked documents. The clustering tendency of the resulting collections is examined in relation to the validity of the cluster hypothesis. By examining whether co-relevant documents are more similar to each other than to non-relevant ones, it is also possible to see how the structure of the document space changes in terms of the proximity of co-relevant documents. Variations in this structure can therefore be examined when considering different numbers of top-ranked documents, and when considering all the documents in a collection. The closer pairs of co-relevant documents are within a document collection, the higher the likelihood that clustering will be effective when using this collection.

In section 4.2 I presented two methods that are typically used to test the validity of the cluster hypothesis in IR test collections. These were the test proposed by Jardine and Van Rijsbergen for the separation of the distributions of pairs of relevant-relevant and relevant-non relevant documents (section 4.2.1), and the nearest neighbour (NN) test proposed by Voorhees (1985a, 1985b). The nearest neighbour test (section 4.2.2), is used here. This test consists of finding the N nearest neighbours (i.e. most similar documents) for each relevant document for a specific query, and of counting the number of relevant documents in this neighbourhood. The size of the NN neighbourhood used in the experiments is 5, the same as Voorhees used in her study. The higher the number of relevant documents in the NN neighbourhood, the higher the probability that the cluster hypothesis holds for the collection. For each of the six collections used, and for each value of $n$, a single value is calculated that corresponds to the number of relevant documents contained in the NN set when averaged over all of the relevant documents for all the queries in a collection.

The reason for using the NN test is that it fits better with the specific experimental objective of this section, which is to examine how the structure of the document space changes with regard to the proximity of pairs of co-relevant documents (section 4.2.4). The results of the NN test provide a direct measure of the extent to which pairs of co-relevant documents are close to each other in

terms of their corresponding content similarity, and also provide an opportunity to study variations of the results as the number of top-ranked documents changes.

The results for the NN test are displayed in Table 6.1, where the highest value in each column is displayed in bold. These results suggest that the number of highly similar co-relevant documents for each collection tends to decrease for increasing values of *n*. Statistical analysis of the results showed significant differences for the AP (all combinations of *n*, except 100-200, 100-350, 500-750), CACM (all combinations between *n*=100 and the rest, and *n*=200 and the rest), Medline (between *n*=100 and the rest), CISI (all combinations of *n*), and WSJ (all combinations of *n*) collections. No significance was found using the LISA dataset.

It should also be noted that the results when using the Medline collection display comparatively the smallest degradation as *n* increases, and also the highest values for the NN test among the six collections used. This can be attributed to the atypical nature of this database (section 5.5.1): each document of this collection is relevant to at most one query. Because of the way this dataset is constructed, subject descriptions of documents associated with one query are not likely to be related to the representations of documents associated with other queries (Shaw *et al*., 1997). It may therefore be easier to detect the similarity of co-relevant documents for this dataset, since there is already a "relevance" structure imposed on its constituent documents.

| *n* | *AP* | *CACM* | *CISI* | *LISA* | *MED* | *WSJ* |
|---|---|---|---|---|---|---|
| 100 | **2.447** | **1.621** | **1.53** | **0.896** | **3.143** | **2.122** |
| 200 | 2.184 | 1.511 | 1.37 | 0.845 | 3.022 | 2.051 |
| 350 | 2.111 | 1.415 | 1.253 | 0.784 | 3.023 | 1.909 |
| 500 | 2.085 | 1.393 | 1.203 | 0.783 | 3.003 | 1.863 |
| 750 | 2.041 | 1.376 | 1.14 | 0.776 | 3.004 | 1.734 |
| 1000 | 2.010 | 1.35 | - | 0.768 | - | 1.711 |
| full | - | 1.366 | 1.119 | 0.859 | 3.016 | - |

**Table 6.1.** Results of the NN test. Highest values in bold

One explanation for the results presented in Table 6.1, is that by increasing the number of top-ranked documents, larger numbers of non-relevant than relevant documents are introduced. As the number of top-ranked documents increases, the number of non-relevant documents increases as well, and so does the probability of a relevant document having a non-relevant one in its N-document neighbourhood.

Moreover, for increasing values of *n*, the new relevant documents that are introduced in the sets are more likely to either have fewer query terms, less of the important terms, or less agreement on the relevance of the documents (e.g. in the TREC collections). These documents can therefore be seen as more "fuzzy" in respect to relevance, and may introduce a confounding effect. This behaviour (i.e. the results of the NN test to decrease as *n* increases) is also more evident for

smaller values of *n* (e.g. 100 or 200), something which is also displayed by the statistical significance of the results presented. In Table 6.2 the average number of relevant documents for each number *n* of top-ranked documents is presented. The figures in this table demonstrate that as *n* increases larger numbers of non-relevant than relevant documents are introduced in the datasets.

In Table 6.1, for the four test collections that data for the full number of documents are available (CACM, CISI, LISA, Medline), only in one (CISI) the results for the full collection are the lowest among all values of *n*. For CACM, the results obtained when using all the documents in the collection are better than the ones for *n*=1000, for LISA they are better than the ones for *n*=200, 350, 500, 750 and 1000, and for the Medline collection they are better than *n*=500, 750. However, no statistical significance is obtained for these results. On the other hand, the results obtained when using the full collections are significantly lower than those obtained using *n*=100 and 200 for the CACM collection, all other values of *n* using the CISI collection, and *n*=100 using the Medline collection.

| *n* | *AP* | *CACM* | *CISI* | *LISA* | *MED* | *WSJ* |
|---|---|---|---|---|---|---|
| 100 | 14.35 | 10.46 | 16.31 | 7.12 | 18.97 | 16.63 |
| 200 | 19.44 | 11.62 | 24.71 | 8.62 | 20.37 | 24.02 |
| 350 | 24.21 | 12.69 | 32.31 | 9.17 | 21.03 | 31.88 |
| 500 | 27.67 | 13.21 | 37.06 | 9.83 | 21.13 | 37 |
| 750 | 31.5 | 13.58 | 42.34 | 10.2 | 21.3 | 43.54 |
| 1000 | 34.25 | 13.83 | - | 10.34 | - | 47.75 |
| full | - | 15.31 | 49.77 | 10.83 | 23.2 | - |

**Table 6.2.** Average number of relevant documents for different numbers of top-ranked documents

An interpretation of the results presented in this section is that post-retrieval clustering does not effectively manage to re-structure the document space for each query. When one considers only the top 100 ranked documents, the probability of co-relevant documents being in the same NN neighbourhood is relatively large, especially for test collections with a large number of relevant document per query such as AP, CISI and WSJ. As the number *n* of top-ranked documents increases, one can view the increasing numbers of non-relevant documents introduced in the sets as noise, and the relevant documents as the ones we wish to separate from the noise.

The results in Table 6.1 suggest that noise is not effectively filtered out as *n* increases, and consequently for larger numbers of top-ranked documents the average number of highly similar co-relevant documents significantly decreases. Therefore, it can be argued that for larger numbers of top-ranked documents, pairs of co-relevant documents are not effectively placed in close proximity to each other. If the document space was effectively structured, one would expect similar, or at least not significantly less, numbers of pairs of co-relevant documents in close

proximity to each other for increasing numbers of *n*. This is not supported by the data of Table 6.1.

The pattern of the results of the NN test prompted some further investigation. Intuitively, the results presented in Table 6.1 seem to display patterns of the kind that could be obtained by chance. This is not to suggest that the actual interdocument similarities are not significantly different to ones generated randomly – R.J. Shaw and Willett (1993) have provided experimental evidence to support the non-randomness of interdocument associations. Instead, what I aim to investigate next is whether the pattern of the results of Table 6.1 is similar to randomly generated results. To do so, the procedure followed is similar to the one described by Burgin (1995):

- Random interdocument associations are produced by means of a random number generator (Matsumoto & Nishimura, 1998)
- Thirty matrices per query per test collection are generated for each number *n* of top-ranked documents
- The NN test is then performed for each of the randomly generated matrices, generating thirty results per query; a single value is then calculated per query by averaging these thirty results.

The random number generator used in the experiments generates pseudorandom real numbers that are uniformly distributed on the [0,1] interval. Although it could be the case that random numbers generated by means of any automatic procedure may not exhibit a perfectly random behaviour, it is felt that the use of this procedure to approximate randomly generated similarity values is sufficient. The use of random number generators in similar experimental conditions has also been used by other IR researchers in the past (Shaw & Willett, 1993; Burgin, 1995; Shaw *et al.*, 1997).

| *n* | *AP* | *CACM* | *CISI* | *LISA* | *MED* | *WSJ* |
|---|---|---|---|---|---|---|
| 100 | 0.775 | 0.515 | 0.824 | 0.522 | 1.002 | 0.79 |
| 200 | 0.56 | 0.291 | 0.626 | 0.202 | 0.639 | 0.571 |
| 350 | 0.431 | 0.184 | 0.459 | 0.134 | 0.496 | 0.443 |
| 500 | 0.351 | 0.125 | 0.374 | 0.105 | 0.452 | 0.36 |
| 750 | 0.274 | 0.097 | 0.286 | 0.069 | 0.434 | 0.285 |
| 1000 | 0.228 | 0.071 | - | 0.052 | | 0.234 |

**Table 6.3.** Results for the NN test generated by random similarity values

The outcome of this process is displayed in Table 6.3. Random values for the full collection size of the CACM, CISI, LISA and Medline collections are not generated, as the aim is to study the pattern of results for other values of *n*. It should also be noted that for all experimental conditions the results obtained with actual interdocument association values are significantly higher than the ones obtained by random means. By observing the results in Table 6.3 it becomes apparent that, as expected, fewer relevant documents are in the neighbourhood of a given relevant document for

increasing numbers *n* of top-ranked documents. To gain an understanding of how the patterns of the results obtained by actual and random interdocument similarities compare, for each collection the results obtained by each method were plotted against each other.

In Figure 6.1 such a plot is displayed for the WSJ collection. The horizontal axis represents the various numbers *n* of top-ranked documents, and the vertical axis represents values corresponding to the result of the NN test. The dotted line corresponds to the results obtained by random similarity values. Comparing the two plots in Figure 6.1, one can see that the behaviour (and not the absolute values) of the plot corresponding to the results using actual association values is highly similar to the one using random values. In fact, this is the case for five out of the six collections used in the experiments, LISA being the only exception.

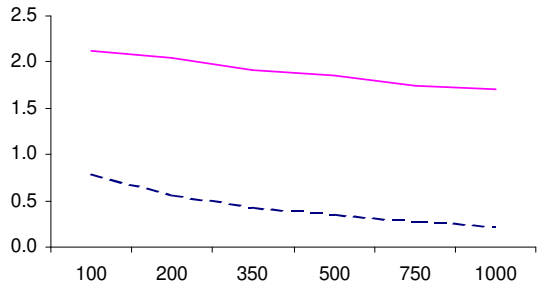


**Figure 6.1.** Random vs. actual values for the NN test using the WSJ collection

This specific observation raises questions regarding the effective structuring of the document space prior to clustering. The proximity of co-relevant documents seems to be affected by the irrelevant documents that are introduced in the datasets as the number of top-ranked documents increases. The numbers of highly similar co-relevant documents seems to decrease in a way that resembles the behaviour of randomly generated interdocument similarities. In the next section, I examine how the effectiveness of hierarchic document clustering is affected by variations in the numbers of top-ranked documents clustered.

## 6.3 Cluster-based effectiveness

Optimal effectiveness values based on the MK1 measure for cluster-based effectiveness, and MK1-k, MK3 and MK4 measures for IFS effectiveness, are presented in this section. These results allow the examination of the behaviour of optimal cluster effectiveness when the number of top-ranked documents varies, and the examination of how optimal cluster effectiveness compares to IFS effectiveness. Table 6.4 presents the results (E values) for the group average method using all six document collections, for $\beta$=0.5 and 2. The full results for all four clustering

methods are presented in Appendix B (Tables B1-B4). Full results for the MK4 measure are presented in Appendix B, Tables B5-B7.

| | *β=0.5* | | | | *β=2* | | | |
|---|---|---|---|---|---|---|---|---|
| *AP* | *MK1* | *MK1-k* | *MK3* | *MK4* | *MK1* | *MK1-k* | *MK3* | *MK4* |
| top100 | 0.511 | **0.752** | **0.663** | 0.550 | 0.619 | 0.749 | 0.667 | 0.628 |
| top200 | 0.514 | 0.778 | 0.685 | **0.543** | 0.604 | 0.741 | 0.663 | 0.613 |
| top350 | 0.507 | 0.790 | 0.695 | 0.552 | 0.576 | 0.739 | 0.663 | 0.611 |
| top500 | 0.508 | 0.792 | 0.692 | 0.552 | 0.560 | **0.735** | **0.652** | 0.614 |
| top750 | 0.488 | 0.785 | 0.699 | 0.548 | 0.562 | 0.745 | 0.657 | 0.605 |
| top1000 | **0.482** | 0.798 | 0.699 | 0.548 | **0.550** | 0.736 | 0.656 | **0.604** |
| *CACM* | *MK1* | *MK1-k* | *MK3* | *MK4* | *MK1* | *MK1-k* | *MK3* | *MK4* |
| top100 | **0.438** | 0.660 | 0.503 | 0.448 | **0.502** | **0.642** | 0.503 | 0.500 |
| top200 | 0.476 | **0.646** | **0.498** | 0.448 | 0.512 | 0.651 | **0.501** | 0.497 |
| top350 | 0.469 | 0.647 | 0.503 | **0.444** | 0.520 | 0.667 | 0.501 | **0.492** |
| top500 | 0.461 | 0.660 | 0.503 | 0.444 | 0.540 | 0.667 | 0.501 | 0.492 |
| top750 | 0.465 | 0.658 | 0.503 | 0.444 | 0.537 | 0.667 | 0.501 | 0.492 |
| top1000 | 0.463 | 0.652 | 0.503 | 0.444 | 0.537 | 0.662 | 0.501 | 0.492 |
| full | 0.641 | 0.713 | 0.503 | 0.444 | 0.782 | 0.806 | 0.501 | 0.492 |
| *CISI* | *MK1* | *MK1-k* | *MK3* | *MK4* | *MK1* | *MK1-k* | *MK3* | *MK4* |
| top100 | 0.630 | 0.827 | 0.727 | 0.651 | 0.702 | 0.777 | 0.738 | 0.717 |
| top200 | 0.609 | 0.820 | 0.729 | 0.651 | 0.658 | **0.741** | 0.699 | 0.674 |
| top350 | 0.589 | **0.811** | **0.726** | 0.639 | 0.655 | 0.753 | 0.680 | 0.653 |
| top500 | 0.593 | 0.815 | 0.726 | 0.639 | 0.656 | 0.765 | **0.676** | 0.649 |
| top750 | **0.567** | 0.818 | 0.726 | **0.638** | **0.649** | 0.776 | 0.676 | **0.648** |
| full | 0.790 | 0.873 | 0.726 | 0.638 | 0.798 | 0.824 | 0.676 | 0.648 |
| *LISA* | *MK1* | *MK1-k* | *MK3* | *MK4* | *MK1* | *MK1-k* | *MK3* | *MK4* |
| top100 | 0.517 | 0.699 | **0.577** | 0.438 | 0.576 | 0.677 | 0.584 | 0.570 |
| top200 | 0.504 | 0.695 | 0.577 | 0.420 | 0.559 | **0.672** | **0.580** | 0.549 |
| top350 | 0.493 | **0.693** | 0.577 | **0.400** | 0.553 | 0.698 | 0.580 | **0.529** |
| top500 | 0.487 | 0.717 | 0.577 | 0.400 | 0.568 | 0.721 | 0.580 | 0.529 |
| top750 | 0.489 | 0.700 | 0.577 | 0.400 | 0.571 | 0.705 | 0.580 | 0.529 |
| top1000 | **0.475** | 0.707 | 0.577 | 0.400 | **0.549** | 0.725 | 0.580 | 0.529 |
| full | 0.643 | 0.736 | 0.577 | 0.400 | 0.716 | 0.739 | 0.580 | 0.529 |
| *MED* | *MK1* | *MK1-k* | *MK3* | *MK4* | *MK1* | *MK1-k* | *MK3* | *MK4* |
| top100 | 0.300 | **0.456** | **0.354** | **0.327** | 0.308 | **0.399** | **0.333** | **0.331** |
| top200 | 0.281 | 0.468 | 0.354 | 0.327 | 0.294 | 0.413 | 0.333 | 0.331 |
| top350 | 0.281 | 0.462 | 0.354 | 0.327 | **0.271** | 0.404 | 0.333 | 0.331 |
| top500 | 0.279 | 0.471 | 0.354 | 0.327 | 0.273 | 0.399 | 0.333 | 0.331 |
| top750 | **0.276** | 0.462 | 0.354 | 0.327 | 0.272 | 0.400 | 0.333 | 0.331 |
| full | 0.682 | 0.596 | 0.354 | 0.327 | 0.711 | 0.403 | 0.333 | 0.331 |
| *WSJ* | *MK1* | *MK1-k* | *MK3* | *MK4* | *MK1* | *MK1-k* | *MK3* | *MK4* |
| top100 | 0.608 | 0.767 | 0.693 | 0.645 | 0.696 | 0.779 | 0.719 | 0.712 |
| top200 | 0.604 | 0.762 | 0.69 | 0.640 | 0.661 | 0.741 | 0.686 | 0.679 |
| top350 | 0.603 | **0.760** | **0.689** | 0.638 | 0.65 | 0.742 | 0.666 | 0.659 |
| top500 | **0.585** | 0.774 | 0.689 | 0.636 | 0.642 | 0.731 | 0.659 | 0.651 |
| top750 | 0.585 | 0.775 | 0.689 | 0.633 | **0.64** | **0.729** | 0.655 | 0.647 |
| top1000 | 0.586 | 0.776 | 0.689 | **0.633** | 0.641 | 0.732 | **0.654** | **0.646** |

**Table 6.4.** Results using the group average method. Highest effectiveness (lowest E value) for each column appears in bold

The values in Table 6.4 have been calculated based on the total number of relevant documents for each query, and not on the number of relevant and retrieved documents. Initially, evaluation was

performed using the relevant and retrieved documents to calculate recall, and results showed a consistent and significant drop in effectiveness for increasing values of *n*.

However, as it was demonstrated in section 5.5.3 (Table 5.3), average cluster size does not always increase in proportion to the number of documents clustered. Therefore, if recall is defined by using the relevant and retrieved documents, the comparison is not fair for collections resulting from large values of *n*: the number of relevant and retrieved documents increases, but the average cluster size does not always increase in proportion for increasing values of *n*, resulting in a decrease in recall which in turn translates into lower effectiveness. For example, in Table 6.5, the effectiveness values obtained when using the relevant and retrieved documents are displayed. These results have been generated by the four clustering methods using the WSJ collection and $\beta$=0.5. From these results it can be seen that effectiveness decreases as the number of top-ranked documents increases, and it does so in a significant way: the difference between the effectiveness at 100 retrieved documents and 350 retrieved documents is approximately 25-30% in favour of the former.

| *n* | *Group Average* | *Ward* | *Complete Link* | *Single Link* |
|---|---|---|---|---|
| 100 | **0.368** | **0.388** | **0.37** | **0.417** |
| 200 | 0.424 | 0.437 | 0.439 | 0.479 |
| 350 | 0.469 | 0.48 | 0.485 | 0.527 |
| 500 | 0.472 | 0.505 | 0.507 | 0.555 |
| 750 | 0.495 | 0.521 | 0.53 | 0.57 |
| 1000 | 0.51 | 0.535 | 0.548 | 0.579 |

**Table 6.5.** Results obtained using the relevant and retrieved documents to calculate recall, for the WSJ collection and $\beta$=0.5. Highest values in bold

The data in Table 6.5, as well as the rest of the results calculated in the same way, are in agreement with the results presented in Table 6.1 regarding the clustering tendency of the collections resulting by considering different numbers of top-ranked documents. The results of Table 6.1 can be interpreted as indicative of the clustering tendency of the various collections (Voorhees, 1985a, 1985b; Griffiths *et al.*, 1986; El-Hamdouchi and Willett 1987). These results suggest that the clustering tendency of the collections decreases as the number of top-ranked documents considered increases. The data in Table 6.5 verify this tendency in terms of the optimal effectiveness of the generated hierarchies.

Single link is the only of the four clustering methods whose average cluster size significantly increases with the increase of the number of top-ranked documents (see Table 5.3). Consequently, one may argue that by basing the evaluation process on a specific characteristic of the other three methods, the effectiveness of the single link method may suffer comparatively to these other methods. The experimental data, however, do not support this view, as this is exhibited in Table

6.5. As I will further discuss in section 6.4, and as is evident from the data in Tables B1-B4, single link is the least effective of the four methods; this result is not affected by the way effectiveness values are calculated. By comparing across the rows of Table 6.5 it is evident that the single link method is the least effective. This is the case in all other experimental conditions as well.

Moreover, clustering that uses large numbers of top-ranked documents (and especially static clustering) can be seen as favoured by taking into account all relevant documents for a query when evaluating the effectiveness of the various hierarchies. This is especially evident when one compares the effectiveness of static clustering to that obtained by considering relatively small numbers of top-ranked documents, e.g. 100 or 200. For example, from the data of Table 6.2 one can note that for the CISI collection when $n$=200 there are approximately 25 relevant documents per query "available" to be clustered, whereas for the full collection there are almost twice as many relevant documents per query. Effectiveness for $n$=200 will be calculated taking into account relevant documents that are not "available" for clustering, whereas for the full collection all relevant documents are available. Therefore, by considering all relevant documents per query to calculate recall, the evaluation can be seen as being favourable for static clustering (when $n$=full), and more strict on clusterings generated by other values of $n$.

Based on these observations, all subsequent discussion is based on effectiveness values that are calculated by taking into account the total number of relevant documents per query. This type of evaluation can be viewed as an attempt to normalise the results over the various experimental conditions.

In the remainder of this section, I examine various aspects of the effectiveness of post-retrieval clustering More specifically, in section 6.3.1 I examine the cluster-based effectiveness obtained for different numbers of top-ranked documents, in section 6.3.2 I compare the effectiveness of cluster-based to IFS retrieval and in section 6.3.3 I compare the effectiveness of actual and random cluster-based retrieval. In sections 6.3.4 and 6.3.5 I also examine some characteristics of optimal clusters.

### 6.3.1 Effectiveness for different numbers of top-ranked documents

One of the research objectives of this thesis, is to investigate how clustering effectiveness varies when different numbers of top-ranked documents are clustered by different clustering methods, and how this effectiveness compares to that attainable by static clustering. This issue is examined in this section.

Based on the data in Table 6.4 and in Tables B1-B4 in Appendix B, there seems to be a small degradation of effectiveness for decreasing values of $n$. Also, static clustering effectiveness (i.e.

*n*=full) seems to be significantly lower than that obtained at any other value of *n*. Statistical analysis of these results leads to two conclusions: first that for the majority of experimental conditions there is no significant degradation of effectiveness for decreasing values of *n*, and secondly that static clustering is significantly inferior to any level of post-retrieval clustering.

The effectiveness for different values of *n* (across rows for the MK1 column of Table 6.4) appears to increase as *n* increases, but the gains in effectiveness do not always prove to be statistically significant. In fact, the majority of the statistically significant differences are noted when MK1 at *n*=100 is compared against MK1 at other values of *n*. Few statistically significant differences exist for other combinations of values of *n* (for example, for *n*=200 and 350 when using the AP collection). Table 6.6 gives a summary of the cases in which significance is achieved when using the group average method for $\beta$=1 (values represent one-tailed probabilities for the Wilcoxon signed-ranks test). It should be noted that for the CACM collection the values showed improved effectiveness for *n*=100 against *n*=200, 500, and 750. Also, no statistical significance is observed when using LISA.

| *n* | *AP* | *CACM* | *CISI* | *MED* | *WSJ* |
|---|---|---|---|---|---|
| 100 - 200 | - | 0.05 | - | - | <0.0001 |
| 100 - 350 | 0.03 | - | 0.03 | 0.001 | <0.0001 |
| 100 - 500 | - | 0.05 | - | 0.002 | 0.0002 |
| 100 - 750 | - | 0.04 | 0.009 | 0.003 | 0.0003 |
| 100 - 1000 | 0.03 | - | N/A | N/A | 0.001 |

**Table 6.6.** Significance levels for comparisons across values of *n*. Results are for the group average method and $\beta$=1

These results suggest that, with the exception of the smallest value of *n* (i.e. 100), there is no significant increase in effectiveness when considering larger numbers of top-ranked documents. This is further strengthened by the fact that the effectiveness values are based on the total number of relevant documents for each query: one would expect hierarchies generated from larger numbers of documents to display significantly higher effectiveness. Based on these results, if one were to choose a unique value for *n*, one would also have to consider practical issues. It may be advantageous, from an efficiency point of view, to cluster the top-200 or top-350 documents returned from a search rather than, for example, the top-1000 documents. Moreover, it can be argued that if the resulting cluster structure is to be presented to a user in an interactive task environment, then a reduced document space may be advantageous (e.g. allowing the user to easily and quickly find a few relevant documents which could start a relevance feedback iteration or satisfy the user's information need).

As far as the second conclusion of this section is concerned, the effectiveness of static clustering is significantly lower than any level of post-retrieval clustering, for all clustering methods, all

collections and all values of $\beta$. In fact, statistical tests gave significance at level < 0.0001 for all experimental conditions. This is despite that the effectiveness values have been calculated based on the total number of relevant documents for each query, something that, as I explained in the previous section, should have favoured static clustering effectiveness.

It should be noted that the poor effectiveness of static clustering is not in agreement with the data that was presented in Table 6.1 regarding the clustering tendency of the collections resulting from various numbers of top-ranked documents. Those results had demonstrated that, when using the CACM, LISA and Medline collections, considering all the documents in the set can lead to higher clustering tendency than considering other numbers *n* of top-ranked documents. However, these differences were not statistically significant.

The results in this section have demonstrated that the effectiveness of post-retrieval clustering is significantly higher to that of static clustering for all experimental conditions studied. It seems that dynamically re-arranging the document space on a per-query basis customises the document space to the request, increasing the chance of relevant documents being placed in the same clusters (Hearst & Pedersen, 1996). Although the results reported in Table 6.1 demonstrated that there may be some problems with the structuring of the document space under post-retrieval clustering, the results presented in this section leave little doubt as to whether post-retrieval clustering is an affective means of performing clustering.

In the next section I examine the comparative effectiveness of both static and post-retrieval clustering effectiveness to that attained by an inverted file search.

### 6.3.2 Cluster-based vs. inverted file search effectiveness

Document clustering has been criticised in the past by IR researchers on the grounds of its failure to provide an effective alternative to conventional similarity search (El-Hamdouchi & Willett, 1989). In this section I examine whether post-retrieval clustering can act as such an effective alternative. The data in Table 6.4, and in Tables B1-B7 in Appendix B, allow a comparison of cluster-based and IFS effectiveness by considering effectiveness values across the rows of the MK1, MK1-k, MK3, and MK4 columns.

By observing the data in Table 6.4, where the group average method is used, for cases other than static clustering, one can see that the general trend of the results is for MK1 to outperform IFS at the MK4 level for the majority of the experimental conditions. Exceptions to this are noted when using the CACM and LISA datasets, where MK1 outperforms IFS at the MK3 level, and when performing recall-oriented searches on the CISI collection where again MK1 exceeds IFS effectiveness at the MK3 level (for *n*=350, 500, 750).

For static clustering, on the other hand, the picture is quite different. For the four collections that data for static clustering are available (CACM, CISI, LISA, Medline), cluster-based effectiveness manages to outperform IFS effectiveness only at the MK1-k level. Exception to this is noted when using the Medline collection, where static cluster-based effectiveness is lower than all levels of IFS effectiveness. This latter result is rather surprising taking into account the properties of the documents of the Medline database that were mentioned earlier (also, see section 5.5.1). Conventional similarity search seems to be successful at separating relevant from non-relevant documents (hence the high effectiveness values for MK4), but static clustering performs poorly at the same task.

| | *β=0.5* | | *β = 2* | |
|---|---|---|---|---|
| *AP* | *MK1* | *IFS* | *MK1* | *IFS* |
| top100 | 0.511 | 0.550 (*MK4*) | 0.619 | 0.667 (*MK3*) |
| top200 | 0.514 | 0.543 (*MK4*) | 0.604 | 0.663 (*MK3*) |
| top350 | 0.507 | 0.552 (*MK4*) | 0.576 | 0.611 (*MK4*) |
| top500 | 0.508 | 0.552 (*MK4*) | 0.560 | 0.614 (*MK4*) |
| top750 | 0.488 | 0.548 (*MK4*) | 0.562 | 0.605 (*MK4*) |
| top1000 | 0.482 | 0.548 (*MK4*) | 0.550 | 0604 (*MK4*) |
| *CACM* | *MK1* | *IFS* | *MK1* | *IFS* |
| top100 | 0.438 | 0.503 (*MK3*) | 0.502 | 0.642 (*MK1-k*) |
| top200 | 0.476 | 0.646 (*MK1-k*) | 0.512 | 0.651 (*MK1-k*) |
| top350 | 0.469 | 0.467 (*MK1-k*) | 0.520 | 0.667 (*MK1-k*) |
| top500 | 0.461 | 0.66 (*MK1-k*) | 0.540 | 0.667 (*MK1-k*) |
| top750 | 0.465 | 0.658 (*MK1-k*) | 0.537 | 0.667 (*MK1-k*) |
| top1000 | 0.463 | 0.652 (*MK1-k*) | 0.537 | 0.662 (*MK1-k*) |
| full | 0.641 | 0.713 (*MK1-k*) | 0.782 | - |
| *WSJ* | *MK1* | *IFS* | *MK1* | *IFS* |
| top100 | 0.608 | 0.645 (*MK4*) | 0.696 | 0.719 (*MK3*) |
| top200 | 0.604 | 0.64 (*MK4*) | 0.661 | 0.686 (*MK3*) |
| top350 | 0.603 | 0.638 (*MK4*) | 0.650 | 0.742 (*MK1-k*) |
| top500 | 0.585 | 0.636 (*MK4*) | 0.642 | 0.731 (*MK1-k*) |
| top750 | 0.585 | 0.633 (*MK4*) | 0.640 | 0.729 (*MK1-k*) |
| top1000 | 0.586 | 0.633 (*MK4*) | 0.641 | 0.732 (*MK1-k*) |

**Table 6.7.** Comparative effectiveness of cluster-based and inverted file searches using the group average method for $\beta$=0.5 and 2

As far as the effectiveness of the other three hierarchic methods is concerned, from the data in Tables B2-B4 one can note that Ward's and complete link methods manage to outperform IFS effectiveness at the MK4 level for a large number of experimental conditions. In fact, the results obtained when using Ward's method follow highly similar patterns to the ones obtained when using the group average method. The effectiveness of the complete linkage hierarchies on the other hand, seems to be less successful at competing with IFS effectiveness. The effectiveness of

single link hierarchies does not manage to exceed IFS effectiveness at the MK4 level in any experimental condition. In fact, in most cases single link effectiveness outperforms IFS effectiveness only at the MK1-k level.

In Table 6.7 a view of the data in Table 6.4 focused on the comparative effectiveness of the two searches is presented (using the group average method for $\beta$=0.5 and 2). Data for three test collections are presented in this table (AP, CACM, and WSJ). The first column displays the number $n$ of documents clustered for each test collection. The second column shows the optimal cluster-based effectiveness as calculated by the MK1 measure for $\beta$=0.5. In the next column, the effectiveness value of the IFS measure that the corresponding cluster-based effectiveness significantly outperforms (as calculated by the Wilcoxon signed-ranks test, for significance level $p$<0.05) is displayed, along with the name of the IFS measure in brackets. For example, when using the CACM collection for $\beta$=0.5, the MK1 measure is significantly higher than the MK3 measure for $n$=100, and higher than the MK1-k measure for $n$=200, 350, 500, 750, 1000, and for $n$=full (i.e. static clustering). Columns four and five display similar information for recall-oriented searches (i.e. $\beta$=2).

It should be noted that there are cases where in Table 6.4 the numeric value of MK1 is higher than the corresponding value of MK4, but for which the respective cell in Table 6.7 does not display this result. For example, when using the AP collection for $\beta$=2, from Table 6.4 it follows that MK1 (0.619) is more effective than MK4 (0.628). However, in the corresponding cell of Table 6.7 this result is not displayed, simply because the difference between MK4 and MK1 in this case is not statistically significant.

The results in Tables 6.4, 6.7 and B1-B7 display some interesting patterns. The first such pattern is that precision-oriented searches ($\beta$=*0.5*), in general, compare favourably to IFS effectiveness, and do so more than recall-oriented searches ($\beta$=2). An example of this behaviour can be seen in Table 6.7 for the WSJ collection. Cluster-based effectiveness for precision-oriented searches outperforms IFS effectiveness at the MK4 level for all values of $n$, whereas for recall-oriented searches it manages to exceed IFS effectiveness either at the MK3 or Mk1-k level. The better performance of precision-oriented searches is in general agreement with findings of previous research in document clustering (Croft, 1978; Voorhees, 1985a; Griffiths *et al*., 1986) that have suggested that clustering can be used as a precision-enhancing retrieval method.

Another observation is that for certain collections (e.g. CACM, LISA) cluster-based effectiveness does not compare well to IFS effectiveness. For example, when using the CACM collection and the group average method, cluster-based effectiveness manages to exceed IFS effectiveness only at the MK1-k level for the majority of the experimental conditions (Table 6.7). Apart from these two collections, when using CISI for recall-oriented searches, MK1 is significantly more effective than IFS only at the MK3 level.

If one looks at the effectiveness of the initial retrieval for these collections (section 5.5.2, Table 5.2), then one will note that the average precision for LISA and CACM is high compared to that of the other test collections. Consequently, it may be argued that the reason for the poor comparative effectiveness of cluster-based searches is the high effectiveness of the initial similarity search for these collections. Such a suggestion, however, is challenged by the results obtained when using the Medline collection. From Table 5.2 it follows that the average precision for the similarity search of the Medline collection is the highest amongst the six test collections used. It would therefore be reasonable to expect IFS effectiveness to compare well to cluster-based effectiveness for this collection.

However, the data in Table 6.4 provide evidence for the contrary: cluster-based effectiveness significantly outperforms IFS effectiveness at the MK4 level for the majority of the experimental conditions. Despite the atypical nature of this database (section 5.5.1), the case of the Medline collection demonstrates that although the effectiveness of the initial retrieval is a major issue when comparing IFS to cluster-based effectiveness, it is not the issue that will determine the outcome of the comparison. The use of the MK4 measure to gauge IFS effectiveness can further be seen as counterbalancing the effect of the initial retrieval on optimal IFS effectiveness (section 4.3.5).

The results reported in this section also demonstrate that using MK1-k to gauge IFS effectiveness is not as fair an approximation as when using other measures that take IFS optimality into account. If MK1-k is used to gauge IFS effectiveness, then all four clustering methods significantly outperform conventional similarity search, for all values of $\beta$, and for all values of $n$ (including static clustering, although not always significantly). A comparison based on MK1-k can thus lead researchers to conclusions that may not hold as strongly when comparisons are made using more favourable, for IFS effectiveness, measures. However, MK1-k should not be completely dismissed, since it offers a comparison demonstrating that IFS does not do as well as optimal clusters under the conditions that 'define' cluster optimality.

| | *AP* | | | *MED* | | |
|---|---|---|---|---|---|---|
| $n$ | $\beta$=1 | $\beta$=2 | $\beta$=0.5 | $\beta$=1 | $\beta$=2 | $\beta$=0.5 |
| 100 | 12.8 | 10 | 21.4 | 1 | 0.5 | 2.9 |
| 200 | 25.8 | 22.8 | 33.4 | 1 | 0.5 | 2.9 |
| 350 | 37.9 | 33.7 | 53.5 | 1 | 0.5 | 2.9 |
| 500 | 37.9 | 33.7 | 56.7 | 1 | 0.5 | 2.9 |
| 750 | 59 | 46.8 | 76.9 | 1 | 0.5 | 2.9 |
| 1000 | 59 | 46.8 | 76.9 | - | - | - |
| full | - | - | - | 1 | 0.5 | 2.9 |

**Table 6.8.** Average offsets for the MK4 measure, for the AP and MED collections

Regarding the use of measures MK3 and MK4, from the data of Table 6.4 it follows that for precision-oriented searches MK4 tends to yield significantly higher effectiveness values than MK3 does, something which indicates that in most cases the most effective portion of the ranked list may be located at a starting point other than the top of the list. For recall-oriented searches the difference between the two measures tends to be much smaller. For the Medline collection, for which the effectiveness of the initial retrieval is particularly high, one can note in Table 6.4 that the difference between the two measures is, in most cases, negligible. This behaviour for the Medline collection is displayed in Table 6.8, where the average offsets (from rank position 1) of the optimal portions of the ranked list, as calculated by the MK4 measure for each of the $\beta$ values, are presented in columns 5-7.

If one compares these results to the ones displayed in columns 2-4 of the same table for the AP collection, the differences are remarkable. First, the offsets for the AP collection are much larger than the ones for Medline, suggesting that the optimal segments of the ranked list for the AP collection are located significantly lower than rank position one. Moreover, the offsets for AP stabilise only at $n$=750, whereas for Medline the initial optimal segment at $n$=100 does not improve as the number of top-ranked documents increases. By comparing the results for the two collections one can therefore conclude that initial IFS retrieval for Medline is much higher than for AP. A consequence of the extremely high IFS retrieval effectiveness for Medline, is that the difference between the MK3 and MK4 measures is insignificant, since both measures practically locate the same segment of the ranked list.

It should be noted that for the results reported in this section, different effectiveness values can be obtained depending on the definition of recall used. Over the two possible definitions of recall that can be used (section 6.3), the results obtained when using all relevant documents for a query (i.e. the ones reported in this section) are the strictest on cluster-based effectiveness, and therefore represent a “pessimistic” form of evaluation. When recall is defined over only relevant and retrieved documents, results mainly follow the same patterns as the ones presented here. In this case however, and especially for recall-oriented searches, cluster-based effectiveness compares more favourably to IFS effectiveness. By presenting results in this section using the more strict evaluation scenario for cluster-based effectiveness, the aim is to further strengthen the validity of any conclusions that may be drawn based on these results.

For example, in Table 6.9 the effect of the way that effectiveness values are calculated is presented. In the second and third columns of this table the values for the MK1 and MK4 measures are displayed (using the WSJ collection, the group average method, and $\beta$=2) when recall is calculated over relevant and retrieved documents, and in the next column the percentile difference between the two measures is displayed. Columns 5-7 display similar information when recall is calculated using all relevant documents for a query. By observing the percentile

differences in columns 4 and 7, one can note that when recall is defined over only relevant and retrieved documents MK1 compares more favourably to MK4. Moreover, the difference between MK1 and MK4 in columns 2 and 3 is statistically significant (except when $n$=1000), something which is not the case when recall is calculated over all relevant documents (Table 6.7).

| | *Recall over relevant & retrieved* | | | *Recall over all relevant* | | |
|---|---|---|---|---|---|---|
| *n* | *MK1* | *MK4* | *% difference* | *MK1* | *MK4* | *% difference* |
| 100 | 0.36 | 0.413 | 14.7 | 0.696 | 0.712 | 2.3 |
| 200 | 0.425 | 0.49 | 15.4 | 0.661 | 0.679 | 2.6 |
| 350 | 0.465 | 0.538 | 15.5 | 0.65 | 0.659 | 1.3 |
| 500 | 0.505 | 0.564 | 11.8 | 0.642 | 0.651 | 1.4 |
| 750 | 0.538 | 0.584 | 8.5 | 0.64 | 0.647 | 1.1 |
| 1000 | 0.563 | 0.597 | 6.2 | 0.641 | 0.646 | 0.7 |

**Table 6.9.** Comparative MK1 and MK4 effectiveness when using the two different definitions of recall

From the results presented in this section, it follows that in the experimental environment used in this thesis, optimal post-retrieval cluster-based effectiveness significantly outperforms optimal IFS effectiveness for a large number of combinations of clustering methods (especially when using the group average method), numbers $n$ of top-ranked documents, and values of $\beta$, the exceptions being recall-oriented searches and the single link method. Precision-oriented searches tend to show much better results, something which has been suggested in previous research (e.g. Croft, 1978; Griffiths *et al*., 1986).

Regarding the issue of selecting a specific number of top-ranked documents to cluster, the results in Table 6.4, 6.7 and Tables B1-B4 in Appendix B, suggest that considering small numbers of top-ranked documents (i.e. 100-350) is an effective option if one considers how cluster-based effectiveness compares to IFS effectiveness at various values of $n$. If this result is seen in conjunction with the suggestions made in section 6.3.1 regarding the issue of the choice of a specific value of $n$, then a number of documents in the order of 200-350 should be considered, since section 6.3.1 suggested significant effectiveness losses when effectiveness at $n$=100 is compared to effectiveness at other values of $n$.

Moreover, the results demonstrate that static clustering effectiveness only manages to exceed IFS effectiveness for a few experimental conditions, only at the MK1-k level, and mainly for precision-oriented searches. Viewing this result in addition to the findings of section 6.3.1, which suggested that static clustering effectiveness is significantly lower than that attained by any level of post-retrieval clustering, it seems justifiable to conclude that static clustering is not an effective means of organising a document collection. This also leads to the suggestion that the use of static clustering in previous research has been a major reason for the failure of clustering to act as an effective retrieval mechanism.

However, there are some negative results regarding the effectiveness of post-retrieval clustering. These results come in the form of the poor comparative effectiveness against conventional similarity search in three document collections (CACM, CISI and LISA), where cluster-based effectiveness does not generally manage to exceed IFS effectiveness at the MK4 level. These results suggest that although post-retrieval clustering is a major improvement over static clustering, it still fails to consistently act as an effective alternative to conventional inverted file searches.

Based on these results, it can be concluded that for most of the experimental conditions there exists an optimal cluster in a document hierarchy that is more effective than an optimal document set retrieved by an IFS. Highly effective clusters can prove useful in an operational environment by, for example, triggering a relevance feedback process (Buckley *et al*., 2000; Iwayama, 2000), or by providing a selection of browsing points for path-based ostensive browsing (Campbell, 2000).

The optimal cluster in a document hierarchy is determined by the clustering scheme used. The issue of whether this optimal cluster will be retrieved by a search strategy, or chosen by a user in a browsing session, depends on a number of parameters that I mentioned in sections 3.6.1 and 3.6.2 (e.g. type of search strategy, cluster visualisation, cluster summarisation method, etc.). These parameters are external to the document hierarchies, and form separate research issues in their own right. The same can be said on whether a user using an IFS system will be able to benefit from its optimal MK4 effectiveness in an operational environment. The way that document contents are presented to the user as relevance clues, for example, highly determines the actual effectiveness of the IR process (Tombros & Sanderson, 1998).

I believe that this study, as well as that of other researchers that investigate effectiveness issues, can motivate research into areas that are related to such external parameters. Two such areas, that have long been neglected in IR, are cluster-based search strategies and cluster representation schemes. It has also been acknowledged by Kural et al. (2001) that users have difficulty in recognising ‘good’ clusters based on conventional representations of cluster contents. Therefore, as I mentioned in section 4.6, more research is warranted in these areas in order to investigate more effective cluster representations.

As I shall point out in section 6.4, the group average method proved to be the most effective of the four clustering methods. The pattern of the results is similar for Ward’s and the complete link methods: optimal cluster-based effectiveness is generally significantly higher than IFS at the MK4 level, except when using the CACM and LISA collections where significance is mostly reported at the MK1-k level. Single link, on the other hand, rarely outperforms IFS in levels other than MK1-k. Finally, for these three methods precision-oriented searches, in general, compare more favourably to IFS effectiveness than recall-oriented searches.

### 6.3.3 Random cluster-based effectiveness

It has been suggested in the past that the effectiveness obtained by static hierarchic clustering does not significantly differ from that obtained by random structures (Shaw *et al*., 1997). A number of other studies have also investigated whether cluster-based effectiveness is significantly higher than random effectiveness (Shaw & Willett, 1993; Burgin, 1995). In order to investigate this issue when post-retrieval clustering is used, the effectiveness obtained by random means was studied, and the results obtained are reported in this section.

The procedure used to generate the random hierarchies is similar to the one reported by Burgin (1995), and stems from the procedure used in section 6.2:

- Random interdocument associations are produced by means of a random number generator (Matsumoto & Nishimura, 1998). The values produced are the same ones used for the experiments reported in section 6.2
- When static clustering is used, thirty random similarity matrices per test collection are produced. When post-retrieval clustering is considered, thirty matrices per query per test collection are generated
- The randomly generated matrices are clustered by each of the four clustering methods
- Retrieval is then performed on the random hierarchies in the same way as for non-random clustering (i.e. the optimal cluster is retrieved). The reported random effectiveness values are obtained by averaging the results over the thirty iterations.

The results that are obtained by this procedure, using each of the four clustering methods, are presented in Appendix B, Tables B8-B11. For ease of reference, actual cluster-based effectiveness values (MK1) are presented next to the randomly obtained values in these tables. By comparing the values of the MK1 and random columns, one can notice that for post-retrieval clustering, in general, actual optimal effectiveness is much higher than random effectiveness. Statistical testing confirms the significance of these results in the vast majority of the experimental conditions.

There are two exceptions to this. First, when using the CISI collection for recall-oriented searches with any of the four clustering methods, the difference in the effectiveness of random and actual hierarchies is small, and in one case (single link method, $\beta$=2, $n$=100, Table 6.10) it is in favour of random cluster-based effectiveness. Few of the differences between actual and random effectiveness are statistically significant when using this collection with $\beta$=2, especially for small values of $n$ (100 or 200). The other exception is noted when using the LISA collection and the single link method for recall-oriented searches (Table B8). In this case the difference in

effectiveness between actual and random effectiveness is small and not statistically significant. In Table 6.10 the results using the CISI collection and the single link method are presented.

Another observation that can be made from the results in Tables B8-B11 is that random cluster-based effectiveness consistently decreases as the number *n* of top-ranked documents increases. This behaviour is in agreement with the data presented in Table 6.3 of section 6.2 about the results of the NN test when randomly generated similarities are used. The only exceptions to this behaviour are when using the CISI and LISA collections, and comparing effectiveness between small values of *n* (i.e. 100-200 in CISI and 200-350 in LISA). An example of this behaviour can be seen in Table 6.10. As mentioned previously, it is in these cases that random cluster-based effectiveness is closer to actual effectiveness.

| | *β=1* | | *β=0.5* | | *β=2* | |
|---|---|---|---|---|---|---|
| *n* | *MK1* | *Random* | *MK1* | *Random* | *MK1* | *Random* |
| 100 | 0.749 | 0.762 | 0.677 | 0.740 | 0.733 | 0.723 |
| 200 | 0.719 | 0.764 | 0.657 | 0.751 | 0.669 | 0.692 |
| 350 | 0.723 | 0.789 | 0.661 | 0.769 | 0.666 | 0.700 |
| 500 | 0.728 | 0.809 | 0.660 | 0.780 | 0.677 | 0.720 |
| 750 | 0.735 | 0.832 | 0.659 | 0.795 | 0.685 | 0.753 |
| full | 0.876 | 0.884 | 0.821 | 0.843 | 0.825 | 0.831 |

**Table 6.10.** Random vs. actual effectiveness values for the CISI collection using single link

Based on these results, it follows that post-retrieval clustering is significantly more effective than random clustering for most clustering methods and document collections. To the best of my knowledge, there is no previous research that has investigated this issue, and therefore the results obtained here can not be compared to that of other researchers. However, the fact that in a number of cases, mostly when using the CISI and LISA collections, random and actual effectiveness values are close should raise some questions about the effectiveness of all clustering methods in these conditions.

When comparing random and actual cluster-based effectiveness for static clustering, the former is always lower than the latter. However, in most of the experimental conditions, the difference between the two is small. Statistical testing mostly confirms the significance of these results for precision-oriented searches. Similar to the results for post-retrieval clustering, when using the CISI collection with any of the four clustering methods, the difference between actual and random static cluster-based effectiveness is much smaller comparatively to other collections. It should be reminded that CISI is one of the databases that display poor cluster-based effectiveness when compared to that of IFS (section 6.3.2).

It should also be noted that the random cluster-based effectiveness that is obtained when using the single link method (either post-retrieval or static) is, in the majority of the cases, lower than the random effectiveness obtained by the other three methods (typically in the order of 8-12%). No

statistical testing was performed to examine the significance of these differences as this was not within the aims of this chapter. This behaviour of random structures for the single link method is in agreement with previous findings that have been reported by Burgin (1995).

From the results presented about static clustering it follows that optimal static clustering effectiveness is higher than optimal random effectiveness, but not always statistically significant. This result seems to be in general agreement with Shaw et al.'s study (1997) that concluded that cluster-based effectiveness can mostly be explained on the basis of chance. In that study operational cluster-based effectiveness was compared against optimal random effectiveness, and in many cases the latter was higher than the former. In the study reported here, optimal static cluster-based effectiveness is always higher than random effectiveness, however, not always significantly higher. In addition, the percentile differences between the two are not *material* (Keen, 1992), i.e. they do not exceed 10%, which suggests that the differences may not be important enough. However, actual effectiveness consistently outperforms random effectiveness, and the consistency of this behaviour may well be important (Keen, 1992).

### 6.3.4 Optimal cluster characteristics

This section aims to provide details about characteristics of optimal clusters. In Table 6.11, columns 3-6, the average number of documents and the average number of relevant documents that are contained in optimal clusters for the LISA and WSJ collections (MK1 measure) are presented. The optimal clusters in Table 6.11 have been generated by the group average method for $\beta$=0.5 and $\beta$=2. Column 2 of the table contains the average number of relevant documents per query for each value of $n$ for each of the two collections. Columns 7-10 contain the average number of documents and the average number of relevant documents that comprise the optimal set (for the same values of $\beta$ as for the optimal clusters) returned by a conventional similarity search (MK4 measure). The optimal clusters and IFS sets in Table 6.11 correspond to the E values presented in Table 6.4.

Data for these two collections (LISA & WSJ) are chosen so as to better demonstrate the dependence of optimal cluster size on the average number of relevant documents per query. The WSJ collection has the largest number of relevant documents per query between the six collections used, whereas LISA the least.

By definition, an optimal cluster (or an optimal set returned by an IFS) is the one which best combines precision and recall. For a collection with a small number of relevant documents per query (such as LISA) one expects the average size of optimal clusters to be small. On the other hand, for a collection with a large number of relevant documents per query (such as WSJ), the size of optimal clusters is expected to be large. This is confirmed by the data presented in Table 6.11. For the LISA collection, for all numbers of top-ranked documents, optimal cluster and

optimal IFS sizes are significantly smaller than for the WSJ collection. What is also apparent from Table 6.11, is that optimal cluster and optimal IFS size depends on the value of the parameter $\beta$ that are investigated. Precision-oriented searches ($\beta$=0.5) lead to smaller sizes, both for clusters and IFS sets, than recall-oriented ($\beta$=2) searches.

| *LISA* | | *MK1* $\beta$=0.5 | | *MK1* $\beta$=2 | | *MK4* $\beta$=0.5 | | *MK4* $\beta$=2 | |
|---|---|---|---|---|---|---|---|---|---|
| *n* | *Mean rel. per query* | *Avg. size* | *Avg. rel.* | *Avg. size* | *Avg. rel.* | *Avg. size* | *Avg. rel.* | *Avg. size* | *Avg. rel.* |
| 100 | 7.1 | 4.7 | 2.9 | 23.5 | 5.3 | 4.3 | 3.1 | 28.2 | 6.3 |
| 200 | 8.6 | 4 | 2.6 | 30.7 | 5.8 | 4.4 | 3.2 | 32.3 | 6.9 |
| 350 | 9.2 | 4.3 | 2.7 | 27.4 | 5.2 | 3.7 | 3.1 | 37.1 | 7.4 |
| 500 | 9.8 | 3.5 | 2.4 | 37.7 | 5.4 | 3.7 | 3.1 | 37.1 | 7.4 |
| 750 | 10.2 | 3.9 | 2.7 | 38.3 | 5.3 | 3.7 | 3.1 | 37.1 | 7.4 |
| 1000 | 10.3 | 4.1 | 2.8 | 20.3 | 4.7 | 3.7 | 3.1 | 37.1 | 7.4 |
| full | 10.8 | 2.2 | 1.2 | 17.7 | 1.9 | 3.7 | 3.1 | 37.1 | 7.4 |
| *WSJ* | | *MK1* $\beta$=0.5 | | *MK1* $\beta$=2 | | *MK4* $\beta$=0.5 | | *MK4* $\beta$=2 | |
| *n* | *Mean rel. per query* | *Avg. size* | *Avg. rel.* | *Avg. size* | *Avg. rel.* | *Avg. size* | *Avg. rel.* | *Avg. size* | *Avg. rel.* |
| 100 | 16.6 | 25.1 | 12.3 | 55.8 | 16.1 | 24.8 | 11.2 | 65.5 | 16 |
| 200 | 24 | 32.5 | 15.1 | 98.3 | 22.5 | 31.2 | 13.4 | 119 | 22.6 |
| 350 | 31.9 | 48.2 | 19.1 | 129.3 | 28.1 | 35.5 | 14.3 | 178.5 | 28.9 |
| 500 | 37 | 37.2 | 15.6 | 165.1 | 32.1 | 36.3 | 14.4 | 215.5 | 32.5 |
| 750 | 43.5 | 25 | 13.5 | 224.5 | 36.3 | 40.4 | 14.8 | 250.6 | 35.4 |
| 1000 | 47.7 | 22.6 | 12.5 | 245.4 | 37.9 | 38.5 | 14.7 | 266.3 | 36.6 |

**Table 6.11.** Average size and average number of relevant documents for optimal clusters using the group average method (MK1), and for optimal IFS sets (MK4)

Regarding the comparative characteristics of optimal clusters retrieved by the four clustering methods, in general, the size of the clusters retrieved by the group average, Ward and complete link methods is comparable. Single link, on the other hand, tends to retrieve much larger clusters in the majority of the experimental conditions. This is illustrated in Figure 6.2, where the sizes of optimal clusters are plotted for the LISA collection and $\beta$=2. The horizontal axis represents the number *n* of top-ranked documents, and the vertical axis the average size, in documents, of optimal clusters. This figure displays the highly similar average size of optimal clusters produced by the other three methods, and the consistent increase of the average size of optimal single link clusters.

Burgin (1995) listed a number of factors imposed by experimental test collections that may affect the level of performance of cluster-based systems. Such factors include the number of relevant documents per query, the mean number of index terms assigned per document, etc. The results presented in this section support this view, since optimal characteristics vary depending on the

characteristics of test collections. This in turn indicates that experimental results should be examined in the context of the specific environment that generated them.

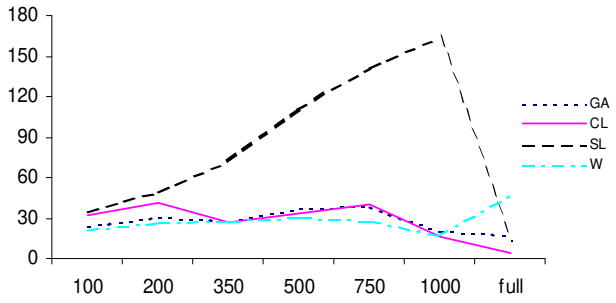


**Figure 6.2.** Average size of optimal clusters for the LISA collection and $\beta$=2

## 6.3.5 Bottom-level optimal clusters

In the results presented in the previous sections no restrictions have been imposed on the characteristics of optimal clusters. Croft (1978, 1980), Griffiths *et al*. (1986), and El Hamdouchi and Willett (1989), among others, suggested that if a bottom-up search considers only the *bottom-level clusters* of a document hierarchy, then its effectiveness exceeds all other types of cluster searches (see section 4.4.3).

Table 6.12 presents some statistics about the size of the bottom-level clusters of the full LISA hierarchy (6003 clusters in total) that were generated by the four clustering methods. The second and third columns of this table display the total number of bottom level clusters available in the hierarchy, and their average size respectively. Columns 4-9 display the percentage of the bottom-level clusters whose size falls within a specified limit, e.g. 73.85% of the group average bottom-level clusters have a size between 2 and 3 documents. It should be noted that for the group average, Ward, and complete link methods, the average size of bottom-level clusters remains fairly constant for all values of *n* used. These three methods tend to produce a large number of small bottom-level clusters (Murtagh, 1984b, Voorhees, 1985a). Results for the other 5 collections are similar and not presented for brevity.

| *LISA full* | *bt. level clusters* | *avg. size* | *2 - 3* | *4 - 10* | *11 - 20* | *21 - 30* | *31 - 40* | *> 40* |
|---|---|---|---|---|---|---|---|---|
| Group average | 3989 | 3.6 | 73.85 | 22.59 | 2.61 | 0.83 | 0.08 | 0.05 |
| Ward | 3561 | 2.4 | 92.31 | 7.69 | 0 | 0 | 0 | 0 |
| Complete link | 3722 | 2.6 | 87.26 | 12.44 | 0.30 | 0 | 0 | 0 |
| Single link | 4915 | 2076.2 | 30.97 | 10.68 | 2.22 | 1.18 | 0.39 | 54.57 |

**Table 6.12.** Bottom-level cluster size statistics for LISA hierarchies

Based on the suggestions by Croft (1978, 1980), Griffiths et al. (1986), and El Hamdouchi and Willett (1989), I considered limiting the definition of an optimal-cluster to a bottom-level cluster. However, a study into the characteristics of optimal clusters suggested that such a constraint would not be beneficial for effectiveness.

Table 6.13 presents the percentage of optimal clusters that are bottom-level for the WSJ and LISA collections. Results for two values of $\beta$ (0.5, 2) are presented. When using the WSJ collection, the percentage of optimal bottom-level clusters is low for the group average, Ward, and complete link methods. This can be explained by the large number of relevant documents per query: the optimal cluster is the one that best combines precision and recall (given the different values of $\beta$). For these three methods, bottom-level clusters have a consistently small size (Table 6.12), so for a collection with a large number of relevant documents per query, such as WSJ, they would not be ideal candidates for optimality. In the case where they are chosen as optimal clusters, it happens for queries that have a small number of relevant documents. For LISA, on the other hand, percentages are significantly higher due to the small number of relevant documents per query.

| *LISA* | *Group Average* | | *Ward* | | *Complete Link* | | *Single Link* | |
|---|---|---|---|---|---|---|---|---|
| *n* | $\beta$=2 | $\beta$=0.5 | $\beta$=2 | $\beta$=0.5 | $\beta$=2 | $\beta$=0.5 | $\beta$=2 | $\beta$=0.5 |
| 100 | 32.3 | 67.7 | 33.3 | 80 | 29 | 71 | 55.2 | 89.7 |
| 200 | 25.8 | 77.4 | 29 | 80.6 | 38.7 | 90.3 | 71 | 93.5 |
| 350 | 33.3 | 76.7 | 35.5 | 71 | 41.9 | 67.7 | 83.9 | 96.8 |
| 500 | 36.7 | 76.7 | 29 | 67.7 | 48.4 | 83.9 | 77.4 | 100 |
| 750 | 67.7 | 67.7 | 29 | 74.2 | 38.7 | 74.2 | 77.4 | 93.5 |
| 1000 | 41.9 | 77.4 | 35.5 | 77.4 | 48.4 | 71 | 80.6 | 96.8 |
| full | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| *WSJ* | *Group Average* | | *Ward* | | *Complete Link* | | *Single Link* | |
| *n* | $\beta$=2 | $\beta$=0.5 | $\beta$=2 | $\beta$=0.5 | $\beta$=2 | $\beta$=0.5 | $\beta$=2 | $\beta$=0.5 |
| 100 | 20.8 | 35.4 | 6.3 | 31.3 | 8.3 | 20.8 | 62.5 | 72.9 |
| 200 | 6.3 | 25 | 0 | 22.9 | 2.1 | 20.8 | 64.6 | 62.5 |
| 350 | 12.5 | 25 | 2.1 | 25 | 4.2 | 25 | 60.4 | 58.3 |
| 500 | 10.4 | 25 | 2.1 | 20.8 | 8.3 | 25 | 66.7 | 54.2 |
| 750 | 14.6 | 35.4 | 4.2 | 22.9 | 10.4 | 31.3 | 58.3 | 56.3 |
| 1000 | 12.5 | 39.6 | 6.3 | 20.8 | 6.3 | 29.2 | 54.2 | 54.2 |

**Table 6.13.** Percentage of optimal bottom-level clusters

The most notable result from Table 6.13 is that for the full (static) LISA hierarchy all optimal clusters are bottom-level (the same happens for the other 3 collections for which the full number of documents is clustered). From these results it follows that optimal clusters for static clustering are always bottom level, and for non-recall oriented searches (i.e. $\beta \neq 2$) have an average size that significantly deviates from that obtained by other values of *n*. The size of optimal static clusters is generally either much larger or much smaller than that at various post-retrieval levels. Therefore,

optimality for static clustering is reached only in extreme cases where too many or too few documents are contained in clusters. This happens because the quality of the clustering is not good enough to allow a different behaviour. For example, average sizes of optimal clusters for the CACM collection ($\beta$=1) range from 8 to 27 documents for all four methods and values of $n$, whereas static optimal clusters have an average size of 3.6, 7.2, 3.7 and 3.3 documents for group average, complete link, single link and Ward methods respectively.

It therefore seems that previous research that suggested the benefits of bottom-level clusters for retrieval is restricted to the case of static clustering. Post-retrieval clustering, on the other hand, tends to reach optimality in much more practical settings.

## 6.4 Comparative effectiveness of the four clustering methods

In addition to the main objective of this chapter, which is to study the effectiveness of post-retrieval clustering under different experimental conditions, the opportunity to study the comparative effectiveness of the four clustering methods used also presents itself. As I discussed in Chapter 4 (section 4.4.1), the effectiveness of these four methods has been extensively investigated in the past, albeit under static clustering. The results presented in the previous sections offer a pool of data through which one can examine whether the conclusions of past research regarding the effectiveness of these methods (Griffiths *et al*., 1984, 1986; Voorhees, 1985a; El-Hamdouchi & Willett, 1989) are valid under post-retrieval clustering. The opportunity to study the optimal behaviour of these four methods under static clustering also lends itself.

In the next paragraphs, I first examine the comparative effectiveness of the four methods under post-retrieval clustering (section 6.4.1) and then under static clustering (section 6.4.2).

### 6.4.1 Effectiveness under post-retrieval clustering

Out of the four hierarchic methods used, the group average method proved to be the most effective in the majority of the experimental conditions. Ward's method ranked second in most of the cases, followed closely by the complete link method. The single link method, in the majority of the conditions, is the least effective of the four.

More specifically, the group average method outperforms the other three methods in the vast majority of the experimental conditions investigated. Group average is almost always significantly more effective than the single link method, in many conditions significantly more effective than complete link, and in comparatively fewer cases significantly more effective than Ward's method. No method is significantly more effective than group average.

Ward's method is more effective than the complete link method in a large number of experimental conditions. The differences between Ward and complete link methods are rarely statistically significant. In fact, out of the 102 total experimental conditions (for each collection: 3 values of $\beta$ x number of values of $n$), only in 19 cases Ward's method is significantly more effective than the complete link method.

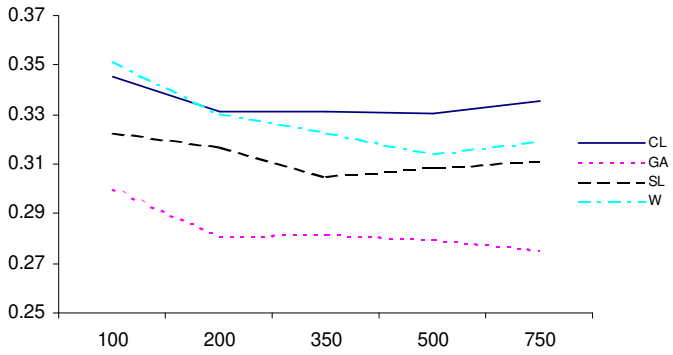


**Figure 6.3.** Comparative effectiveness of the four methods using the Medline collection for β=0.5

In agreement with findings of previous research (Griffiths *et al*., 1984,1986; Voorhees, 1985a; El-Hamdouchi & Willett, 1989; Burgin, 1995), the single link method is the least effective of the four methods used in the experiments. As mentioned previously, group average is almost always significantly more effective than single link, and so is Ward's method (for 62 out of possible 102 conditions), whereas complete link is significantly more effective than single link in 32 out of 102 conditions. It should be noted that when using the CACM and Medline collections, complete link is never significantly more effective than the single link method. In fact, when using Medline single link outperforms complete link and Ward's methods in a number of experimental conditions, but never significantly. Figure 6.3 displays the comparative effectiveness of the four methods when using Medline, for $\beta$=0.5 (the horizontal axis corresponds to the number $n$ of top-ranked documents).

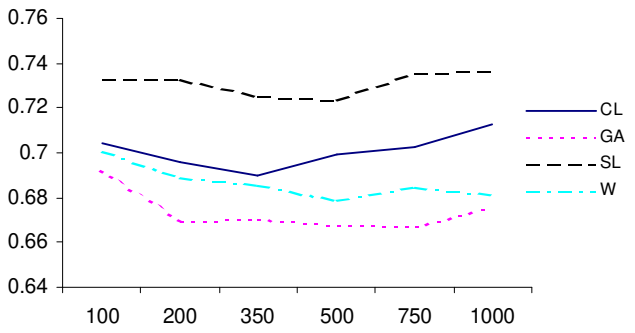


**Figure 6.4.** Comparative effectiveness of the four methods using the WSJ collection for β=1

Although not much can be inferred from the behaviour of the four clustering methods in one small and topically homogeneous collection such as Medline, this collection offers an interesting environment from a clustering perspective because of the property of its documents to be relevant to at most one query (section 5.5.1). This property can be seen as denoting an *a-priori* structure that clustering methods are asked to recover. The experimental results suggest that the group average and single link methods are the ones that most successfully recover this structure. Given that for the single link method Medline is the only collection for which it outperforms the other two methods (complete link and Ward), it may be suggested that single link is successful at recovering clustering structure where this is evident. However, this suggestion should be seen tentatively, especially given the poor effectiveness of the method in the other experimental conditions.

This generally poor effectiveness of the single link method is attributed to its tendency to generate large clusters which are characterised by the chaining effect mentioned in section 3.4.1 (Griffiths *et al*., 1984; Voorhees, 1985a). The complete link and Ward's methods tend to generate hierarchies which display comparable characteristics: their average size is small and relatively unaffected by the increase in the numbers of documents clustered, and their bottom-level clusters are small, typically containing a pair of documents (Murtagh, 1984b; Voorhees, 1985a). These similar characteristics can explain their comparable behaviour in terms of optimal cluster-based effectiveness.

The group average method, generates clusters whose characteristics are similar to those of the two other methods (Ward and complete link). However, the average size of clusters of this method tends to be slightly larger than complete link and Ward clusters, and also tends to slightly increase as the number of documents clustered increases (Table 5.3, section 5.5.3). Burgin (1995) suggested that clustering methods for which the mean cluster size is closest to the mean number of relevant documents per query are likely to display good retrieval performance. This can explain the higher effectiveness of the group-average method, especially for increasing values of $n$: its mean cluster size increases slightly for increasing $n$, and in this way it better adapts to the increasing average number of relevant documents per query for increasing values of $n$.

Recently, Leuski (2001) examined the comparative effectiveness of hierarchic methods under post-retrieval clustering. Leuski did not study the effectiveness at different numbers of top-ranked documents, instead he only used the top 50 retrieved documents to compare the effectiveness of the methods. Leuski's results are in agreement with the findings of this section, in that the group average method was the most effective, followed by Ward's and the complete link methods. The single link method was the least effective. Leuski also suggested that the differences between the group average and Ward's methods are generally insignificant, something which is not the case in this experimental environment.

### 6.4.2 Effectiveness under static clustering

Previous research has suggested that, under static clustering, the complete link and group average are the most effective methods (Voorhees, 1985a; Willett, 1988; El-Hamdouchi & Willett, 1989). The results reported in this chapter however do not fully support these findings, at least as far as the optimal effectiveness of these methods is concerned. Complete link, group average and Ward's methods, under different conditions, all outperform each other.

There are 12 total static clustering conditions in the experimental environment, defined by the 4 collections for which static clustering is performed (CACM, CISI, LISA, Medline) and by the 3 values of $\beta$ used. Out of these 12 conditions, complete link is the most effective method in 7 of them, and Ward's and group average methods in 3 and 2 conditions respectively. Moreover, results seem to vary within the same test collection for different values of $\beta$. For example, when using the Medline collection, group average is the most effective method for recall-oriented searches ($\beta$=2), whereas Ward's method is the most effective for precision-oriented searches ($\beta$=0.5). However, the statistical significance of these results is rarely confirmed. For example, the complete link method is significantly more effective than the other two methods only when using the LISA collection for $\beta$=0.5, and when using the CACM collection for $\beta$=2.

On the other hand, the differences between these three methods and single link are, in most cases, statistically significant. This result suggests that the single link method is the least effective of the four in terms of optimal static clustering effectiveness, and is in agreement with previous research (Voorhees, 1985a; Griffiths *et al*., 1984,1986; El-Hamdouchi & Willett, 1989).

Although the results seem inconclusive, it is possible to obtain a ranking of the four methods under static clustering based on the results presented here. This can be achieved by assigning a number (1-4) to each method, for each of the 12 experimental conditions, corresponding to the relative rank of the method (1 being the most effective, etc.). By averaging the ranks over the 12 experimental conditions, it follows that complete link is the most effective method (average rank 1.58), followed by Ward's method (2), and group average (2.42). Single link has a rank of 4 for all experimental conditions. It should again be emphasised that statistical significance rarely confirms the differences between the three most effective methods, and hence any conclusions drawn regarding their comparative effectiveness can only be tentative.

Another interesting finding under static clustering is that for precision-oriented searches (i.e. $\beta$=0.5), the differences between the three most effective methods seem to become smaller, and with the exception of the LISA collection, statistically insignificant. An explanation for this result can be given in terms of the characteristics of optimal clusters under static clustering. As it was mentioned in section 6.3.5 and displayed in Table 6.13, optimal clusters for static clustering are bottom level clusters that typically contain few documents (2 or 3). This is especially so when

precision is more important than recall. In such cases it is more likely for effectiveness across methods to display less variation.

## 6.5 Summary

The research reported in this chapter investigated the effectiveness of hierarchic post-retrieval document clustering. Four hierarchic clustering methods and six document collections were used in the experiments. Four main research objectives were pursued in this chapter.

The first objective was to investigate the structure of document spaces that result by considering varying numbers of top-ranked documents. The structure of document spaces was examined under the specific focus of the proximity of documents which are relevant to the same query (co-relevant documents) (section 6.2). Second, the comparative effectiveness of document hierarchies generated by varying numbers of top-ranked documents was examined in section 6.3.1. The third objective was to compare the effectiveness of post-retrieval clustering to that of static clustering; this issue was also examined in section 6.3.1. The fourth issue was that of the comparative effectiveness of document clustering (both post-retrieval and static) and conventional similarity search (section 6.3.2).

In addition to these objectives, a number of other issues were also investigated in this chapter. Such issues include the comparative effectiveness of actual and randomly generated hierarchies (section 6.3.3), the study of characteristics of the optimal clusters of document hierarchies (sections 6.3.4 and 6.3.5), and the examination of the comparative effectiveness of the four clustering methods used (section 6.4).

The main finding of this chapter can be summarised as follows:

- The number of highly similar co-relevant documents seem to significantly decrease as the number of documents considered increased. This behaviour was consistent across the six document collections, and resembled the behaviour of documents whose similarities have been randomly generated.

- Optimal cluster-based effectiveness did not generally significantly decrease as the number of documents clustered increased. An exception to this was noted when comparing the effectiveness at $n$=100 to that obtained at other values of $n$.

- Optimal post-retrieval cluster-based effectiveness was always significantly more effective than optimal static cluster-based effectiveness.

- Post-retrieval effectiveness exceeds IFS effectiveness at the MK4 level for a large number of experimental conditions. This is mainly noted when using the group average method, and

when performing precision-oriented searches. However, for certain document collections (CISI, CACM, LISA) cluster-based effectiveness only manages to exceed IFS effectiveness at the MK1-k level. Static cluster-based effectiveness only manages to exceed IFS effectiveness at the MK1-k level.

- Post-retrieval effectiveness is generally significantly higher than the effectiveness obtained by random means. An exception to this is when using the LISA and CISI collections with small numbers of top-ranked documents. Static effectiveness is consistently higher than random effectiveness, however not significantly higher.

- The group average method was the most effective of the four methods used for post-retrieval clustering. Ward's method ranked second. Single link was the least effective of the four, apart from when using the Medline collection. For static clustering, results were not as clear, but there are indications to suggest that, in agreement with previous research, the complete link method is the most effective, followed by Ward's method. Single link was again the least effective.

The main implication of the results obtained in this chapter, is that they provide evidence that static clustering is not an effective means of organising a document collection. Its effectiveness is significantly lower than that of any level of post-retrieval clustering, and also does not compare well to IFS effectiveness. Moreover, the results provide evidence that the application of post-retrieval clustering to IR bears significant effectiveness improvements compared both to static clustering and to best-match searches. These results also provide evidence to support the view that the static application of document clustering has been a major reason for its failure to act as an effective mechanism for IR.

However, there were also a number of shortcomings noted, mainly regarding the almost random patterns noted at the structure of document spaces resulting from increasing numbers of top-ranked documents (section 6.2), and the consistently poor comparative effectiveness of cluster-based to IFS retrieval when using the CACM and LISA collections, and to a lesser extent the CISI collection. Also, the close-to-random behaviour in a number of experimental conditions when using the CISI and LISA collections provides further evidence to suggest that, although post-retrieval clustering is a significant improvement over static clustering, it is not acting as an ideal solution for organising document collections.

In the next two chapters, I demonstrate how the effectiveness of post-retrieval clustering can be enhanced by challenging an assumption which has traditionally characterised the way document clustering has been applied: the static nature of interdocument similarity calculations. This experimental work expands on the issues that I discussed in Chapter 5 (section 5.3) regarding the use of query-sensitive similarity measures.

# Chapter 7

# Query-Sensitive Similarity Measures

## 7.1 Introduction

The results presented in the previous chapter demonstrated some shortcomings regarding the effective structuring of the document space prior to clustering. In section 6.2 specifically, by isolating the effect of interdocument associations, I showed that the number of highly similar co-relevant documents tends to decrease as the number of retrieved documents increases, and it does so in a pattern that is highly similar to that displayed by random similarity values. This result should be seen in relation to the argument that was made in section 5.3 regarding the static nature of interdocument relationships. In that section I had argued that this static nature is a limitation of the way that document clustering is applied to IR, since it does not take into account the context (i.e. the query) under which the similarity of any two documents is judged.

This chapter aims to build upon the argument made in section 5.3 for the application of query-sensitive similarity measures to the calculation of interdocument relationships. It aims to do so by specifying means of implementing query-sensitive measures, and by investigating their effectiveness at structuring the document space prior to clustering. The study of query-sensitive measures in this chapter will be performed under the view that document clustering in IR is a goal-driven process that aims to group relevant documents together on a per-query basis. Consequently, if the document space is structured effectively (with regards to the proximity of co-relevant documents), then the effectiveness of the clustering process may also increase. It should also be mentioned that although there are other perspectives through which query-sensitive measures could be investigated (e.g. their applicability to visualising the interdocument relationships in a dataset), it is only in relation to document clustering that they will be examined in this thesis.

This chapter is structured as follows. First, in section 7.2 I propose specific formulas that can incorporate the influence of the query in the calculation of similarity measures, mention some of their limitations, and review some related work. Subsequently, in section 7.3 I describe some modifications to the basic experimental environment that was outlined in section 5.5. The results of the evaluation of the effectiveness of the query-sensitive measures are reported in section 7.4, and in section 7.5 I conclude this chapter by summarising its main findings.

# 7.2 Query-sensitive similarity measures

In section 3.3 I discussed a number of issues pertaining to the use of association measures in document clustering, and in Appendix A I present a number of such measures that are commonly used in IR. Regarding the use of a single measure for document clustering, through the discussion in section 3.3.2, it transpired that the only evidence offered by previous research so far is that measures should be normalised by the length of the documents which they compare. No other significant evidence exists to suggest that the use of one measure instead of another may significantly influence the effectiveness of the clustering process.

$$Sim(D_i, D_j) = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} \qquad (7.1)$$

In this section I present means by which query-sensitive similarity measures can be defined. Taking the previous suggestions into account, any discussion about query-sensitive similarity measures (QSSM) hereafter will be based on the cosine coefficient, which is presented again in Equation 7.1 for ease of reference. However, any formulas and arguments can easily be applied to other measures that are typically used for document clustering, such as the Dice coefficient, or Euclidean distances. The choice of the cosine coefficient in this chapter is based on this measure's widespread application to document clustering. As Ellis et al. (1993) noted, there does not seem to be any reason for the IR community to revise the historical attachment to association coefficients provided by (among others) the cosine formula.

In this section I first define formulas for the calculation of query-sensitive measures in section 7.2.1, present an example to illustrate their use in section 7.2.2, mention some limitations in section 7.2.3, and discuss related work in section 7.2.4.

## 7.2.1 Defining query-sensitive similarity measures

Before proceeding with the definition of QSSM, it is worth reiterating that documents and queries in this thesis are represnted as vectors in a *n*-dimensional space (section 2.2), where *n* is the size

of the indexing vocabulary. In this space it is assumed that a document $D_i$ can be represented as: $D_i = \{d_{i1}, d_{i2}, ..., d_{in}\}$, where $d_{ij}$ is the weight assigned to the $j$-th term of $D_i$.

The rationale behind query-sensitive measures is that similarity is a dynamic, purpose (and context)-sensitive concept. In the specific case of interdocument similarity calculations in IR, purpose is defined by the query under which the associations of documents are examined. According to this view, documents that are jointly relevant to a query display an inherent similarity that is dictated by the query itself. Query-sensitive similarity measures aim to detect this inherent similarity. This can be achieved by viewing the query terms as salient features that define the context under which similarity is examined.

Query-sensitive measures can be defined as a function of two components. The first one corresponds to the conventional similarity between two documents $D_i$ and $D_j$, and is given by Equation 7.1. The second component corresponds to the common similarity of all three objects: the pair of documents $D_i$, $D_j$ and the query $Q$, and I will represent this component by $Sim(D_i, D_j, Q)$. This is the variable component of the similarity measure. The query-sensitive similarity $Sim(D_i, D_j \mid Q)$ can therefore be defined as:

$$Sim(D_i, D_j \mid Q) = f(Sim(D_i, D_j), Sim(D_i, D_j, Q)) \qquad (7.2)$$

The similarity given by the variable component $Sim(D_i, D_j, Q)$ can be defined by finding all common terms between documents $D_i$ and $D_j$, and seeing which of these common terms are also terms that appear in the query $Q$. The similarity between pairs of documents that have a large number of common terms that are query terms should then be accordingly augmented. This idea can be defined in terms of the cosine coefficient in Equation 7.3. In this equation $Q=\{q_1, q_2, \ldots, q_n\}$ is the query vector, $D_i$ and $D_j$ are the two document vectors, and $C = D_i \cap D_j = \{c_1, c_2, \ldots, c_k, \ldots, c_n\}$ is a vector which contains the common terms of documents $D_i$ and $D_j$.

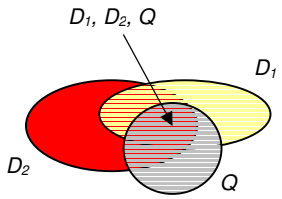


**Figure 7.1.** The variable similarity $Sim(D_1, D_2, Q)$

In order to visualise the concept of the common similarity between the documents and the query, an example is presented in Figure 7.1. Documents $D_1$ and $D_2$, and query $Q$, are represented as sets of their constituent terms, and therefore overlaps between the sets denote common terms. The set of terms that is common to the two documents and to query $Q$ corresponds to the area that is common to all three sets, and is the one that defines the similarity $Sim(D_1, D_2, Q)$. When a

different query *Q'* is posed to the IR system, the area of overlap between the two documents and the query will accordingly change, and so will the similarity between documents $D_1$ and $D_2$. It should be noted that if the two documents do not share any terms at all (i.e. Equation 7.1 gives a similarity value of zero), then the variable similarity $Sim(D_i, D_j, Q)$ will also equal zero.

The terms of the common vector *C* can be represented by $c_k = (d_{io} + d_{jp}) / 2$, where $d_{io}$, and $d_{jp}$ are the weights of each of the common terms in $D_i$ and $D_j$ respectively. Vector *C* then contains the set of common terms of the two documents, and each term of *C* is weighted by the average of the weights of the common terms. Other representations of $c_k$ were also investigated ($min(d_{io}, d_{jp})$, $max(d_{io}, d_{jp})$, $(d_{io} \cdot d_{jp})$), but no significant differences were found. I report this specific form which proved to be consistently the most effective.

$$Sim(D_i, D_j, Q) = \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \qquad (7.3)$$

Having established ways to define the two components of Equation 7.2, what remains is to define the function that combines these two sources of evidence. One way to do so is by using a linear combination of the two sources: $\mathrm{Sim}(D_i, D_j \mid Q) = \vartheta_1 Sim(D_i, D_j) + \vartheta_2 Sim(D_i, D_j, Q)$, where $\vartheta_1 + \vartheta_2 = 1$. By substituting Equations 7.1 and 7.3 in the above, we derive Equation 7.4 which gives the query-sensitive similarity between $D_i$ and $D_j$. I will call this measure *M3*. It should be noted that a linear combination of sources of evidence is commonly used in IR applications. For example, Weiss et al. (1996) use a linear combination of hyperlink and content evidence to define the similarity between hypertext documents, and Wen et al. (2001) follow a similar approach in order to define the similarity between queries posed to a search engine.

$$Sim(D_i, D_j \mid Q) = \vartheta_1 \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} + \vartheta_2 \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \qquad (7.4)$$

Intuitively, if one bases the calculation of interdocument similarities on measure M3, then for a specific query, pairs of documents that have more terms in common with the query than other pairs will be assigned higher similarity values (assuming that they have the same number of non-query terms in common). This reflects the belief that under the context defined by the query, query terms possess greater salience when determining interdocument relationships. The relative importance of each of the two components of Equation 7.4 can be determined by assigning appropriate values to the two parameters $\vartheta_1$ and $\vartheta_2$.

More specifically, the first parameter ($\vartheta_1$) determines the importance assigned to the conventional, static similarity of the documents under comparison, while the other parameter determines the importance assigned to the varying component of Equation 7.4. If $\vartheta_2$ is set equal to zero, then the similarity given by Equation 7.4 is simply the cosine coefficient between the two documents $D_i$ and $D_j$ adjusted by the parameter $\vartheta_1$. The same effect can be achieved when none of the common terms between the two documents is a query term; in this case Equation 7.3 will give a similarity value of zero.

On the other hand, if parameter $\vartheta_1$ is set equal to zero, then the query-sensitive similarity between the two documents becomes equivalent to the one given by Equation 7.3. In this case, the effect of the static similarity is totally ignored, and the resulting formula can be seen as the most extreme form of query-biasing. I will call this measure *M2*. Measure M2 only takes into account common terms between the two documents that are also query terms. Unlike the measure defined by Equation 7.4, M2 will equal zero if none of the common terms between the documents is a query term. Also unlike Equation 7.4, the overall similarity between $D_i$ and $D_j$ does not take into account the co-occurrence of other terms (apart from query terms) in the two documents. The effectiveness attained with M2 can be seen as a lower limit of the effectiveness of query-sensitive measures.

A note that should be made regarding the value of these two parameters is that their absolute value is of no practical significance. Instead, it is the ratio of one parameter over the other that is of importance. The reason for this, is that it is not the absolute value of interdocument similarities that affects the clustering process, but rather the relative ranking of these similarities (Van Rijsbergen, 1979). For example, setting $\vartheta_1$=0.5 and $\vartheta_2$=0.5 is equivalent to setting $\vartheta_1$=2 and $\vartheta_2$=2, since the actual similarity values in the latter case will be four times larger than in the former case, but the relative ranking of the similarity values will remain the same. The constraint set earlier ($\vartheta_1+\vartheta_2=1$) reflects this.

To summarise, so far two measures have been proposed for the calculation of query-sensitive similarities. These are the measures given by Equation 7.4 (measure M3), and by Equation 7.3 (measure M2). Moreover, varying forms of M3 can be obtained by varying the relative ratio of parameters $\vartheta_1$ and $\vartheta_2$.

One more measure will be defined in this section, its definition being highly similar to the one of M3. This third measure differs in the way that it combines the two sources of evidence given by Equations 7.1 and 7.3. Instead of a linear combination of the two components (Equation 7.4), the new measure is defined as the product of the two sources of information. This is presented in

Equation 7.5; I will call this measure *M1*. The rationale behind measure M1 is exactly the same as for M3, i.e. for a specific query, pairs of documents that have more terms in common with the query than other pairs will be assigned higher similarity values.

$$Sim(D_i, D_j \mid Q) = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} \cdot \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \qquad (7.5)$$

However, there is one significant difference between the two measures. When using M1, if none of the common terms between the two documents is a query term (i.e. $Sim(D_i, D_j, Q) = 0$), then the overall similarity $Sim(D_i, D_j \mid Q)$ will equal zero. This is in contrast to when using M3, where $Sim(D_i, D_j \mid Q)$ will be equal to the conventional similarity of the two documents (adjusted by the parameter $\vartheta_1$). The aim of query-sensitive measures is to increase, on a per-query basis, the similarity of documents that are likely to be co-relevant. Measure M1 attempts to do so in a rather "greedy" way, by setting the similarity of pairs of documents that do not possess any query terms in common to zero. This choice for M1 reflects the assumption that the presence of query terms is required for a document to be relevant.

This is verified by the behaviour of the test collections used in this experimental environment. Table 7.1 presents in the first row the percentage of relevant documents which contain at least one query term for each of the six collections[21], and in the second row the average number of query terms contained in a relevant document. The figures in the first row of this table all exceed 91%, an exceptionally high value that verifies the highly topical and algorithmic nature of relevance that is employed in standard IR evaluation (Ingwersen, 1994). The implication of this for the query-sensitive measures presented here, and especially for M1, is that the likelihood for pairs of co-relevant documents to contain at least one query term in common is high.

| | *AP* | *CACM* | *CISI* | *LISA* | *MED* | *WSJ* |
|---|---|---|---|---|---|---|
| % | 96.32 | 93.22 | 92.31 | 100 | 91.81 | 97.06 |
| Avg. q.terms per doc. | 3.2 | 3 | 2.4 | 4.5 | 2.8 | 3.5 |

**Table 7.1.** Query term statistics for the six test collections

Therefore, by disregarding pairs of documents that contain no query terms in common (and hence have a low likelihood of being jointly relevant to a query), M1 can be seen as adapting to the topical nature of relevance in typical IR test collections. However, it may cause the similarity

[21] For the AP and WSJ collections the figures have been calculated using 7.6 terms on average per query. The procedure for deriving these terms was explained in section 5.5.1. Also, calculations are based on the stemmed forms of terms that the SMART IR system uses to match documents and queries and to calculate interdocument similarities.

between documents that share a large number of (non-query) terms to equal zero. In the specific case where one of these documents is relevant and contains a number of query terms, it is likely to "miss" a relevant document which contains no query terms, but is highly associated to that relevant document. It should be reminded that this exact potential of document clustering (to discover relevant documents by association with other relevant ones) has been put forward in the past for the application of document clustering to IR (Jardine & Van Rijsbergen, 1971; Croft, 1978).

It should also be mentioned that if the pair of documents under comparison contains non-overlapping sets of query terms, this will not be taken into account as an indication of co-relevance by any of the similarity measures presented here. Although the presence of query terms in both documents can be seen as a source of evidence of their co-relevance, this is not incorporated by the query-sensitive similarity measures. The main reason for this decision is that if two documents contain non-overlapping sets of query terms, this may be an indication that the documents are discussing these terms under different topics.

Consider for example the query: "aviation accident reports in the United Kingdom". If two documents $D_1$ and $D_2$ contain the query terms "United Kingdom aviation" and "Accident reports" respectively, it is highly likely that each document discusses the query terms in different contexts. The first document is more likely to discuss general issues concerning aviation in the U.K., whereas the second document may be focused on accident reports. If we assume that the second document is relevant, it would not seem appropriate to associate the first document with it through the set of non-overlapping query terms, since these terms discuss a different topic.

For measures M1 and M2, $0 \leq Sim(D_i, D_j \mid Q) \leq 1$. For measure M3 this property can be retained by appropriate selection of parameters $\vartheta_1$ and $\vartheta_2$ (e.g. by constraining the parameters so that $\vartheta_1+\vartheta_2=1$). To preserve the reflexivity of the measures defined by M1, M2 and M3 (i.e. $Sim(D_i, D_i)=1$), the similarity of a document with itself is defined to be equal to 1. This does not follow as a result of either Equations 7.3, 7.4, or 7.5, but can be introduced by definition. Finally, for all three measures $Sim(D_i, D_j \mid Q) = Sim(D_j, D_i \mid Q)$ (i.e. query-sensitive similarity is symmetric). These properties are in accordance with those of conventional similarity measures (Van Rijsbergen, 1979).

### 7.2.2 An example

To better illustrate the concept of query-sensitive similarity, and to demonstrate the way that this is calculated by the three measures introduced, I present a specific example in this section. I will consider a sample query posed to an IR system by a user interested in finding out information about engine specifications for aircraft manufactured by Boeing. The terms input to the IR system

in this scenario are assumed to be: "*Boeing aircraft engine specification*". I will also assume that in this simplified example only four documents are retrieved in response to this query. These four documents are displayed below, and are represented by sets of their alphabetically sorted terms (query terms are displayed in italics). No lexical processing is assumed to have been performed on document terms (e.g. stemming). Although many aspects of this specific example are not realistic, it is sufficient to illustrate the application of query-sensitive measures.

$D_1$: {*aircraft*, *boeing*, commercial, company, *engine*, model, *specification*}

$D_2$: {*boeing*, company, employment, management, products, sales}

$D_3$: {*aircraft*, commercial, company, leasing, model, products, sales, services}

$D_4$: {*engine*, model, general-electric, products, rolls-royce, *specification*}

In this example, documents $D_1$ and $D_4$ are assumed to be relevant to the query. The other two documents discuss issues relating to the management section of the Boeing company ($D_2$), and to products and services offered by a commercial aircraft leasing company ($D_3$).

It is possible to construct a (symmetric) similarity matrix for these four documents. The matrix is given in Figure 7.2, where instead of displaying a numeric value in each cell of the matrix corresponding to the similarity between pairs of documents, the set of common terms between pairs of documents is presented (query terms are displayed in italics). For example, documents $D_1$ and $D_3$ have four terms in common: aircraft, commercial, company, and model; the first of these is also a query term.

| | $D_1$ | $D_2$ | $D_3$ |
|---|---|---|---|
| $D_2$ | *boeing*<br>company | - | - |
| $D_3$ | *aircraft*<br>commercial<br>company<br>model | company<br>products<br>sales | - |
| $D_4$ | *engine*<br>model<br>*specification* | products | model<br>products |

**Figure 7.2.** The similarity matrix for the example

Let us now examine a given relevant document, $D_1$, in relation to the other documents in this dataset. If one ranks the other three documents in decreasing order of the number of terms in common with $D_1$, then the ranking would be (the number of common terms is given in brackets): $D_3$ (4), $D_4$ (3), $D_2$ (2). For simplicity, document vectors are assumed to have a binary representation (presence/absence of index terms), interdocument similarity is calculated using the simple matching coefficient (Appendix A), and therefore similarity values correspond to the

number of terms in common between documents. Consequently, the static similarity of $D_1$ to the rest of the documents is given by the values in brackets in the above ranking. Based on this ranking, the most similar document to $D_1$ is $D_3$, a document that is not relevant to the query.

By using the query-sensitive measures, a re-ranking of the rest of the documents based on their similarities to $D_1$ will occur. For this specific example, the variable component of the similarity (Equation 7.3) for each of the other three documents to $D_1$ corresponds to the number of common terms between the pair of documents and the query, and therefore $Sim(D_1,D_2,Q)=1$, $Sim(D_1,D_3,Q)=1$, $Sim(D_1,D_4,Q)=2$.

From this it becomes apparent that if measure M2 (Equation 7.3) is used to gauge similarity, then the most similar document to $D_1$ is $D_4$, a document that is also relevant to the query. If M1 is used (Equation 7.5), then:

$$Sim(D_1,D_2 \mid Q)=2\cdot1,\ Sim(D_1,D_3 \mid Q)=4\cdot1,\ Sim(D_1,D_4 \mid Q)=3\cdot2$$

This will also give $D_4$ as the nearest neighbour of $D_1$. If M3 is used (Equation 7.4), then:

$$Sim(D_1,D_2 \mid Q)=\vartheta_1\cdot2+\vartheta_2\cdot1,\ Sim(D_1,D_3 \mid Q)=\vartheta_1\cdot4+\vartheta_2\cdot1,\ Sim(D_1,D_4 \mid Q)=\vartheta_1\cdot3+\vartheta_2\cdot2$$

In the above, $D_4$ will become the most similar document to $D_1$ for any ratio $\vartheta_1:\vartheta_2$ that assigns at least twice as much importance to $\vartheta_2$ as to $\vartheta_1$.

### 7.2.3 Limitations

The assumption that query terms are sufficient indicators of document relevance is made for all three measures defined in the previous section, and especially for measures M1 and M2. Therefore, implicitly the notion of *topicality* (Saracevic, 1970) is adopted for relevance. It is well established in IR research that relevance is a multidimensional concept, and that topicality is only one such aspect (Schamber *et al.*, 1990). Research into the concept of relevance has indicated that topicality plays a significant role in the determination of relevance (Saracevic, 1975), although it does not automatically result in relevance for users (Barry, 1994).

Apart from the topical view of relevance taken, query-sensitive measures only take one instance of the user's information need into account (i.e. the set of query terms posed by the user to the IR system). Due to this treatment, contextual and temporal factors that may affect the user's perception of relevance are not incorporated.

Ottaviani (1994) and Ingwersen (1994), for example, argue that information needs evolve and develop during the course of a search session. Campbell (2000) also suggested that there is a temporal aspect to the notion of relevance, and this temporal aspect should be incorporated in the

retrieval model. In the same way, one can argue that the similarity between two objects may change over time due to new evidence presented, or due to the contextual effect of other objects (Tversky, 1977). Contextual factors are not considered by the measures presented in the previous section. In other words, the system decides, based on the query, how the similarity between objects in the retrieved set should change, but it does not take into account other factors that may influence interobject similarities.

As far as the temporal aspects are concerned, these are not explicitly incorporated into the query-sensitive measures. These measures take into account the current instance of the user's query. If the user's information need (and thus the query) changes during the course of a search session, then the modified query will be incorporated into the calculation of interdocument similarity by the query-sensitive measures. Therefore, it can be argued that dealing with temporal aspects of information needs follows logically from this work. However, this is not examined experimentally in this thesis.

These limitations are not unique to the approach proposed in this thesis. The majority of IR research to date has focused on the topical aspect of relevance, taking the view that query terms offer the only evidence about the user's information need. As far as this thesis is concerned, the choice not to consider factors such as the ones mentioned previously was taken on the basis that in a non-interactive laboratory-based environment it is difficult to model such factors. The importance of such factors in IR research is fully acknowledged; however, it is not within the aims of this thesis to investigate them.

A further limitation relates to the problem of short queries, the type usually encountered in web search engines, averaging about 2-3 terms per query (Jansen *et al.*, 2000). The three measures defined previously, regard query terms as the dimensions that acquire significant discriminatory power. If only 2 or 3 such terms are supplied by the user, it is doubtful whether these measures (especially M2) will have enough information to effectively bias similarity. This is a well-known research problem in IR, and methods that have been used to tackle it previously (Van Rijsbergen *et al.*, 1981; Voorhees, 1994; Xu & Croft, 1996) could also be applied here. The effect of query length on the effectiveness of these measures is investigated in section 7.4.5.

### 7.2.4 Related work

In section 5.3.2 I presented research work that is conceptually similar to the ideas put forward by the axiomatic view of the cluster hypothesis, and to the use of QSSM as an attempt to increase the similarity of co-relevant documents on a per-query basis. In this section, I discuss research work that has attempted to generate clusterings of documents focused on the query. I first summarise such research work, and then discuss the differences between such approaches and the research reported in this thesis.

Iwayama (2000) defined query-biased clusters by modifying a probabilistic clustering algorithm that assigns documents *d* to clusters *c* based on the probability $P(c \mid d)$. The modified probability takes the query *q* into account, and hence becomes $P(c \mid d,q)$. The conditional part is calculated by adding the term weights for *q* to the term weights for *d*. According to Iwayama, this modified conditional probability "raises the importance of terms occurring in a query" when forming the clusters.

Chang and Hsu (1997) suggested an approach which incorporates the information learnt from a series of queries into the computation of interdocument similarities. For a new query *q*, the similarity measure between clusters (or documents) *c*(*i*) and *c*(*j*) is given by:

$$R(c(i), c(j)) = \sum_{t \in C_i, C_j} TF(i,t) \cdot TF(j,t) \cdot \frac{DF(t,q)}{QF(t)},$$

where TF(*i*,*t*) is the relative frequency of term *t* in cluster *c*(*i*), DF(*t*,*q*) is the document frequency of term *t* in the whole collection of query *q*, and QF(*t*) is the frequency of term *t* in queries that have appeared so far. The values of DF(*t*,*q*) and QF(*t*) are updated after user feedback so as to increase the weights of terms that are in topics selected by the user. It should be noted that the effectiveness of this method was not evaluated by the authors.

Eguchi et al. (2001) proposed a query-biased similarity measure which takes into account a series of incrementally expanded queries. His main research focus was the investigation of incremental query expansion, and the effectiveness of the expanded query as a viewpoint for clustering. The weight $w_t^{d_i}$ of a term *t* in document $d_i$ that matches a query term is modified by adding a quantity equal to $\xi \cdot w_t^q$ to it, where $\xi$ is a coefficient that has a positive value, and $w_t^q$ is the weight of the same term *t* belonging to query *q*. In this approach, if $\xi$=0, then a conventional (static) similarity measure is derived.

Eguchi reported an evaluation of this method using 10,000 HTML documents in Japanese and 10 queries. The clustering method used was of a partitioning type, and the top-200 documents retrieved by an initial similarity search were clustered. As mentioned previously, the main aim of Eguchi's research was to investigate the effectiveness of clustering based on the similarity calculated by incrementally expanded queries (as opposed to the initial query posed by the user). Consequently, most of the results reported investigate expanded forms of the initial query.

However, Eguchi also reports some comparisons of cluster-based effectiveness using a conventional similarity measure (i.e. when $\xi$=0) to that using only the initial unexpanded query posed by the user. The results reported by Eguchi demonstrate an improvement in the average

precision of the best cluster[22], when using the query-biased similarity measure, that is in the order of 2%-4%. However, the results also displayed a decrease in the percentage of relevant documents which were included in the best cluster when only the initial query is used (pp. 71-72), compared to the percentage observed when the query is not taken into account (i.e. when $\xi$=0). For unexpanded queries, the results demonstrated that the best value of the parameter $\xi$ is 0.5.

Regarding the relationship of these approaches to the query-sensitive measures defined in this chapter, the following should be noted:

- The focus of all these approaches is different to the one pursued in this thesis. Cluster-based effectiveness is not the main focus *per se* of these approaches, but rather the investigation of other issues, like incremental query expansion (Eguchi *et al.*, 2001), or the effect of few relevance judgements on retrieval performance (Iwayama, 2000). As a consequence, these methods do not focus on the issue of the static nature of interdocument associations, nor do they focus on the implications of the similarity measures on the structure of the document space in terms of the proximity of co-relevant documents

- Limited experimental results are presented by these researchers (or, in the case of Chang and Hsu, no results at all). Also, in cases where results that compare the effectiveness of the modified similarity measure to that of conventional measures (Eguchi *et al.*, 2001) are available, the differences seem insignificant, and in some cases in favour of conventional measures. The results obtained by using the QSSM proposed in this thesis, although applied to different sets of documents, introduce significant effectiveness improvements

- The formulas presented by these authors are of different forms from the ones presented in this thesis

- Finally, the research reported in this thesis has been developed independently from these approaches, and its contributions to the field are significantly different.

## 7.3 Experimental environment

The use of QSSM is advocated in this thesis based on the premise that they can "force", on a per query basis, documents that are likely to be co-relevant to be highly similar to each other. This in turn relates to the axiomatic view of the cluster hypothesis that I proposed in section 5.3: QSSM aim to capture the inherent similarity that co-relevant documents exhibit, a similarity that is dictated by the query itself. Moreover, if query-sensitive measures are effective in placing co-relevant documents close to each other, then their application to document clustering can also be

[22] Documents within clusters are ranked in decreasing order of their similarity to the query.

expected to prove effective. In the remainder of this chapter I examine whether the use of QSSM introduces improvements compared to the use of conventional, static similarity measures with regards to the proximity of co-relevant documents.

A slightly modified version of the experimental environment that was defined in section 5.5 is used in the rest of this chapter. The modification relates to the use of QSSM instead of a conventional similarity measure (cosine coefficient). The effectiveness of QSSM with regards to increasing the similarity of co-relevant documents is measured by using the NN test, i.e. by using the same experimental procedure as in section 6.2. Following the same experimental procedure will also facilitate an immediate comparison of the results obtained when using QSSM to those obtained when using a conventional similarity measure (section 6.2). The comparison of the results will indicate whether the use of QSSM increases the similarity of co-relevant documents, and consequently, whether query-sensitive measures succeed at achieving a greater degree of adherence to the cluster hypothesis.

It should be clarified that when using any of the query-sensitive measures and all the documents in a test collection (i.e. when *n*=full), then the interdocument associations are dynamic, i.e. the relationships between all the documents in the dataset to each other change on a per-query basis as a result of the use of QSSM. Recall from Chapter 6 that using *n*=full was associated with a static organisation of the document space (either in terms of interdocument similarities, or in the case of hierarchic clustering).

The three query-sensitive measures defined in section 7.2.1 are studied. Their effectiveness is compared both to that of the cosine coefficient, and to that of each other. Moreover, for measure M3 varying results can be attained depending on the values assigned to parameters $\vartheta_1$ and $\vartheta_2$. The issue of the effect of the ratio of these parameters on the effectiveness of M3 is also examined.

In section 7.2.3, when discussing the potential limitations of QSSM, I highlighted the issue of short queries, and the effect that query length may have on the effectiveness of QSSM. This issue is also investigated, and the modification of query length for these purposes is another alteration to the basic experimental environment of section 5.5.

The NN test used in section 6.2 does not give information about the relevance status of the immediate NN (i.e. most similar) document of a relevant document. A number of researchers have suggested that, for the purposes of clustering, it may be worth considering clusters containing only a document along with its nearest neighbour (e.g. Griffiths *et al*., 1986; El-Hamdouchi, 1987). Therefore, in addition to the NN test proposed by Voorhees, the percentage of relevant documents whose most similar neighbour is also relevant will be calculated. In order to distinguish the two tests, in the remaining of this chapter Voorhees's test will be called 5NN

(since it is a 5 document neighbourhood it is using), and the test examining only the nearest neighbour will be called 1NN.

As a final note, it should be mentioned that document and query terms are weighted based on the formula given by Equation 5.2 in section 5.5.2. When considering varying numbers of top-ranked documents, document-term weighting is performed locally, that is, by using evidence from within the retrieved document set only. This approach to document term weighting has been known to produce effective results when calculating interdocument relationships (Korpimies and Ukkonen, 1998) (see section 5.5.3). However, query-sensitive measures, apart from document terms, also employ query terms (and consequently their weights) in the calculation of similarities. The effect that global or local query-term weighting has on the effectiveness of the query-sensitive measures is examined in section 7.4.1.

# 7.4 Experimental results

In this section I report and analyse experimental results that are obtained for the query-sensitive measures presented in this chapter. The presentation of the results consists of five parts. First, in section 7.4.1 I examine the effect of local and global query-term weighting on the effectiveness of query-sensitive measures. Then, in section 7.4.2 I examine how the effectiveness of measure M3 varies as a function of the two parameters ($\vartheta_1$ and $\vartheta_2$). Subsequently, in section 7.4.3 I investigate the comparative effectiveness of the QSSM and the cosine coefficient, in section 7.4.4 I study the comparative effectiveness of the three query-sensitive measures, and in section 7.4.5 I consider the effect of the query length on M1, M2 and M3.

## 7.4.1 Global vs. local query-term weighting

The weighting formula which is used to assign weights to query terms (Equation 5.2, section 5.5.2) is based on the *tf-idf* measure. Consequently, query terms that appear frequently in the document set are assigned lower weights than terms that do not appear as frequently (given that most query terms appear once in a query). In the local document sets, query terms will be of the most frequently occurring terms: all documents in the top-*n* set will contain some of the query terms, otherwise these documents would not have been retrieved in the first place. If, on the other hand, information from the whole dataset is used (global weighting), then the weights assigned to query terms will depend upon their pattern of occurrence across the entire collection, and not within a localised set that has been retrieved in response to these terms.

Previous research has demonstrated the benefits of local information for tasks such as automatic thesaurus construction (Attar & Fraeknel, 1977) and query expansion (Xu & Croft, 1996). The use of local evidence has been suggested as effective, since the information contained in local sets

is more focused on the query (Attar & Fraeknel, 1977). Croft and Harper (1979), used information from the top-ranked documents returned from a query to re-estimate the probabilities of term occurrence within the relevant set for a query. In this work, Croft and Harper used the local information to modify the weights of query terms, and not to select candidate terms for query expansion.

These findings suggest that the use of local information for weighting the evidence offered by query terms is likely to be more effective than the use of global information. To examine whether this suggestion is valid in the experimental environment used, the 5NN test was performed using the AP, Medline, and WSJ collections and all three QSSM. A ratio of 1:4 was selected for the two parameters of M3 for AP and WSJ (i.e. $\vartheta_2$ is weighted four times more heavily than $\vartheta_1$), and a ratio of 1:7 for Medline. Document terms are weighted using local information in all conditions (i.e. even when query terms are weighted using global evidence).

The rationale behind this experiment is not to investigate factors that may influence the effectiveness of local or global information. Instead, the purpose of this experiment is to examine whether, in this specific environment, there is a difference in the effectiveness of the query-sensitive measures depending on the method of query-term weighting.

The results of the 5NN test are presented in Table 7.2 for the WSJ collection (highest value in each column appears in bold). The results obtained for the other collections display similar patterns, and are not presented for brevity. The results clearly indicate that local query-term weighting is more effective than global weighting. Local query-term weighting results in significantly higher values for the 5NN test for all experimental conditions. All differences are significant at levels <0.001.

| *n* | *M1 global* | *M1 local* | *M2 global* | *M2 local* | *M3 global* | *M3 local* |
|---|---|---|---|---|---|---|
| 100 | **1.741** | 2.357 | **1.113** | **1.872** | **2.172** | 2.354 |
| 200 | 1.698 | 2.446 | 1.018 | 1.827 | 2.081 | **2.443** |
| 350 | 1.699 | **2.468** | 0.904 | 1.832 | 2.021 | 2.389 |
| 500 | 1.682 | 2.463 | 0.897 | 1.856 | 1.966 | 2.377 |
| 750 | 1.587 | 2.421 | 0.874 | 1.838 | 1.867 | 2.300 |
| 1000 | 1.576 | 2.416 | 0.865 | 1.799 | 1.895 | 2.269 |

**Table 7.2.** Global vs. local query-term weighting for WSJ

The results of Table 7.2 do not reveal different behaviour for any of the three QSSM, i.e. all three measures significantly benefit from the use of local weighting. M3 is less (but still significantly) affected by the use of local weighting than M1 and M2. This can be explained on the basis that M3 uses a linear combination of common terms and common query terms between documents, and it is therefore less reliant on query terms. The other two measures, and especially M2, rely

more heavily on characteristics of query terms. Confirming this, the decrease in effectiveness for M2 is the largest among the three measures.

Query-sensitive measures, especially measure M2, use query terms as discriminators between relevant and non-relevant documents. From this point of view, it seems logical that local weighting of the evidence is more effective, as previous research has suggested the effectiveness of this approach in other environments. Local weighting is dependent on the set of documents that is retrieved in response to the query, and can reflect the relationships that hold between terms in this environment more effectively than when using global weighting (Xu & Croft, 1996). Global weighting, on the other hand, is independent of the query, and reflects the relationships that hold between terms in a static manner.

As mentioned previously, it is not the aim of this section to examine the factors that may influence the comparative effect of local and global weighting on the effectiveness of query-sensitive measures. Based on the results presented here however, the use of local query-term weighting is warranted in this specific experimental environment. This method of weighting for query terms is used in the remainder of this chapter, as well as in the experiments reported in Chapter 8.

### 7.4.2 Selecting parameters for M3

In this section I examine the selection of appropriate values for the parameters $\vartheta_1$ and $\vartheta_2$ of M3 (section 7.2.1, Equation 7.4). As I explained in that section, it is not the absolute values of these parameters that is of interest, but rather, their ratio. By varying the ratio of these parameters, one can investigate the effect of assigning different importance to the two components of Equation 7.4. More specifically, $\vartheta_1$ determines how much importance is associated to the static similarity of the two documents, whereas $\vartheta_2$ how much importance is associated to the common similarity of the two documents and the query. It should also be reminded that for $\vartheta_1=1$ and $\vartheta_2=0$ M3 becomes equivalent to the cosine coefficient (Equation 7.1), and also that for $\vartheta_1=0$ and $\vartheta_2=1$ M3 becomes equivalent to M2 (Equation 7.3). The former case will not be dealt with in this section, as the comparative effectiveness of query-sensitive measures and the cosine coefficient is presented in section 7.4.3.

Intuitively, one would expect the results of the 5NN test to resemble those attained by the cosine coefficient when the values of the parameter $\vartheta_1$ are much higher than those of $\vartheta_2$. Then, by decreasing the difference in the values of the two parameters (and hence their ratio), the results should start to differ to those obtained by the cosine. This is evident in Table 7.3, where the results of the 5NN test are presented for four different ratios of the two parameters (9:1, 4:1, 2:1,

1:1) when using LISA (highest value for each ratio is in bold). The results presented in Table 7.3 are representative of the results obtained using the other five test collections. Full results for all six collections for varying ratios of the two parameters are presented in Appendix C, Tables C1-C6. For comparison, the results for this collection using the cosine coefficient are reported in Table 7.6.

| *n* | *9:1* | *4:1* | *2:1* | *1:1* |
|---|---|---|---|---|
| 100 | **0.946** | **0.99** | 1.055 | 1.206 |
| 200 | 0.926 | 0.972 | 1.027 | 1.195 |
| 350 | 0.821 | 0.93 | 1.029 | 1.199 |
| 500 | 0.849 | 0.938 | 1.037 | **1.237** |
| 750 | 0.859 | 0.94 | 1.041 | 1.208 |
| 1000 | 0.83 | 0.91 | **1.076** | 1.204 |
| full | 0.913 | 0.946 | 1.028 | 1.177 |

**Table 7.3.** Results of the 5NN test for the LISA collection by varying the $\vartheta_1$: $\vartheta_2$ ratio in favour of $\vartheta_1$

By observing the results, it becomes evident that significant improvements are introduced by reducing the importance of the component of Equation 7.4 that corresponds to the static similarity between the two documents and the query terms (i.e. by increasing the importance of $\vartheta_2$). If one compares, for instance, the results obtained when the static component of Equation 7.4 is nine times more important than the variable component (i.e. $\vartheta_1$:$\vartheta_2$=9:1) to the results obtained when both components are weighted equally, the differences range between 27.5 and 46% in favour of the latter ratio. The differences in the majority of cases are statistically significant, especially as the relative importance assigned to $\vartheta_1$ is reduced.

Having established that the results obtained by the 5NN test significantly increase when the ratio of the two parameters increases in favour of $\vartheta_2$, what remains to be established is whether there is a specific ratio for each collection that displays the highest effectiveness. In Table 7.4, the results of the 5NN test are presented when using the LISA collection, and when the ratio of the two parameters is varied in favour of $\vartheta_2$. The last column of this table contains the results obtained when using $\vartheta_1$=0 and $\vartheta_2$=1; as I mentioned earlier this corresponds to the M2 measure (Equation 7.3). The results of this table demonstrate that, in general, the effectiveness of M3 tends to increase as the weight assigned to $\vartheta_2$ increases. When M3 becomes equivalent to M2 (last column of the table), there seems to be a rather significant drop in the effectiveness of the measure[23]. For the specific case of the LISA collection, the peak in effectiveness seems to occur between the

[23] The comparative effectiveness of QSSM is examined in section 7.4.4.

ratios of 1:7 and 1:10. However, the differences in effectiveness at this region are not statistically significant.

| *n* | *1:2* | *1:4* | *1:7* | *1:9* | *1:10* | *M2* |
|---|---|---|---|---|---|---|
| 100 | 1.352 | 1.383 | 1.402 | 1.392 | 1.404 | **1.395** |
| 200 | 1.311 | 1.327 | 1.390 | 1.391 | 1.389 | 1.269 |
| 350 | 1.335 | 1.418 | **1.429** | **1.42** | **1.428** | 1.315 |
| 500 | **1.374** | **1.415** | 1.423 | 1.403 | 1.406 | 1.317 |
| 750 | 1.358 | 1.395 | 1.421 | 1.413 | 1.392 | 1.287 |
| 1000 | 1.341 | 1.384 | 1.393 | 1.385 | 1.380 | 1.303 |
| full | 1.303 | 1.332 | 1.376 | 1.354 | 1.341 | 1.269 |

**Table 7.4.** Results of the 5NN test for the LISA collection by varying the $\vartheta_1$: $\vartheta_2$ ratio in favour of $\vartheta_2$

It should also be noted that the behaviour when using LISA with increasing importance assigned to $\vartheta_2$ is not typical of the two larger collections (AP and WSJ). In general, when using LISA, as the data in Table 7.4 demonstrate, when the importance attributed to the common similarity between the documents and the query increases it does not seem to significantly impair the effectiveness of the measure, at least not until M3 becomes equivalent to M2 (i.e. the difference in effectiveness when using ratios 1:7, 1:9, 1:10 are small). This is especially evident for small values of *n* (i.e. 100, 200, 350).

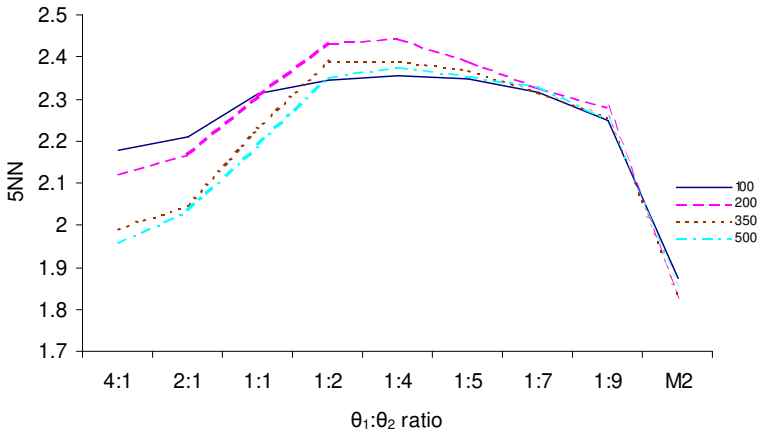


**Figure 7.3.** The effectiveness of M3 as a function of $\vartheta_1$ and $\vartheta_2$ for the WSJ collection

Figure 7.3 demonstrates the variation in the effectiveness of M3 for varying ratios of the two parameters for *n*=100, 200, 350 and 500 when using the WSJ collection. The pattern of the results for the WSJ collection is for the effectiveness of M3 to peak when the ratio between the two parameters is in the region of 1:4. The results display a consistent decrease past this point as the weight assigned to $\vartheta_2$ increases (i.e. ratios 1:7, 1:9 yield lower results).

A reason for the rather different behaviour of the two databases can be given in terms of their characteristics. Documents of the LISA collection are rather short, with 39.7 terms on average per document. The length of the queries for this collection is large (almost 20 terms per query on average, Table 5.1), almost half the average document size. Moreover, as it was mentioned in section 7.2.1, relevant documents in this collection contain on average 4.5 query terms (Table 7.1). Taking these characteristics into account, it can be appreciated why query influence in this database is strong: the combination of short documents, long queries and relatively large number of query terms per relevant document increases the likelihood of pairs of co-relevant documents to be assigned high similarity by M3.

The WSJ collection on the other hand, is characterised by long documents (377 terms on average per document), and shorter queries than LISA (7.6 terms on average). In this collection, as the weight assigned to the static similarity of documents is decreased and calculations are increasingly biased towards common query terms between documents, the effectiveness of the measure seems to be obscured by the length of the documents and the relatively few query terms (especially comparatively to document length). In addition, documents of the WSJ collection are more topically diverse than those of the smaller collections, and therefore query terms can be used under a varying number of contexts in the bodies of such documents. M3, in such an environment, is more likely to reach a higher effectiveness when the importance assigned to common query terms and common "content" terms is more balanced (but still in favour of the former) than in more topically homogeneous collections. The other TREC collection (AP) displays a similar behaviour (Appendix C, Table C1).

As far as the other three collections are concerned (CACM, CISI and Medline), the effectiveness of M3 seems to peak when the ratio of the two parameters is set to around 1:7 (Appendix C, Tables C2, C3 and C5). This behaviour is similar to the one noted for LISA. These four collections are topically homogeneous, treating mainly a single subject area (e.g. library and information science for LISA).

Another interesting finding from the study of the variation of the effectiveness of M3 by adjusting the two parameters, is that especially when using CISI, Medline and LISA, as the number $n$ of top-ranked documents increases, the effectiveness of M3 tends to be higher when less weight is attributed to $\vartheta_2$. For example, when using CISI (Table C3), for $n$=100 and 200 the most effective ratio of the two parameters is 1:7, for $n$=350 and 500 the most effective ratio becomes 1:5, and for $n$=750 the most effective ratio is 1:4. Also, when using LISA (Table 7.3), for $n$=100, 200 and 350 the highest effectiveness is noted for ratios 1:9 and 1:10, and for the remaining values of $n$ the highest effectiveness is noted for a ratio of 1:7.

An explanation for this behaviour can be given in terms of what happens when *n* increases. As the number of top-ranked documents increases, so does the number of non-relevant documents. However, non-relevant documents will also contain some of the query terms, otherwise they would not have been retrieved in the top-*n* set in the first place. This has as a consequence that some of the relevant documents share a number of query terms with non-relevant documents. Therefore, in order to counterbalance this effect to retain high effectiveness, the importance assigned to the variable part of the similarity defined by M3 seems to decrease for increasing values of *n*. In other words, more weight needs to be assigned to the other common terms between documents, so as to define the context under which the query terms are used for larger values of *n*.

As a conclusion regarding the selection of parameters for M3, the data obtained support the view that this is heavily dependent on the characteristics of the test collection under investigation. What was noted for all six collections was that the effectiveness of M3 for the 5NN test increases as the relative importance of $\vartheta_2$ over $\vartheta_1$ increases, and it reaches its peak when the ratio between the two parameters is considerably in favour of $\vartheta_2$. The effectiveness of the measure then tends to drop past this point, and when $\vartheta_1$ becomes equal to zero M3 generally displays its lowest effectiveness.

It should also be emphasised that as the ratio of the two parameters increases in favour of $\vartheta_2$, the differences in the effectiveness of M3 are generally not statistically significant. For example, in Table 7.3 all the differences across the various ratios are significant (apart from *n*=100 when comparing ratios 4:1 and 2:1), whereas in Table 7.4 there are few statistically significant differences between the results for ratios 1:4, 1:7, 1:9, 1:10. For the two TREC collections there are significant differences as the ratios of the values move past the peak point (1:4), i.e. the differences between the ratios of 1:4 and 1:7, 1:9 are significant in favour of the former.

In the next section I examine how the effectiveness of the three QSSM compares to that of a static measure - the cosine coefficient.

### 7.4.3 Comparative effectiveness of the query-sensitive measures and the cosine coefficient

In this section I examine three issues. First I compare the effectiveness of M1, M2 and M3 to the cosine coefficient for the 5NN test, then in section 7.4.3.1 I examine the effectiveness for different numbers of top-ranked documents, and in section 7.4.3.2 I present results for the 1NN test.

In Tables 7.5-7.7 the results of the 5NN test for each of the six test collections and each of the three QSSM are presented. Each table comprises five columns[24]. In the first column the different values of *n* are given for which results are calculated. Columns 2-5 contain the results obtained for the 5NN test with the cosine coefficient, and measures M1, M2 and M3 respectively. In columns 3-5 the percentage difference between the results for M1-cosine, M2-cosine and M3-cosine, respectively, are also calculated. The differences are displayed in brackets. For each of the four columns (2-5), the highest value for the 5NN test across all values of *n* is displayed in bold.

| *AP* | *Cosine 5NN* | *M1 5NN* | *M2 5NN* | *M3 5NN* | *WSJ* | *Cosine 5NN* | *M1 5NN* | *M2 5NN* | *M3 5NN* |
|---|---|---|---|---|---|---|---|---|---|
| top100 | **2.447** | **2.619** (7.02%) | **2.079** (-15.06%) | **2.652** (8.35%) | top100 | **2.122** | 2.357 (11.1%) | **1.872** (-11.74%) | 2.354 (10.95%) |
| top200 | 2.184 | 2.406 (10.18%) | 1.834 (-16.02%) | 2.404 (10.07%) | top200 | 2.051 | 2.446 (19.29%) | 1.827 (-10.88%) | **2.443** (19.15%) |
| top350 | 2.111 | 2.39 (13.22%) | 1.671 (-20.84%) | 2.349 (11.26%) | top350 | 1.909 | **2.468** (29.29%) | 1.832 (-4.01%) | 2.389 (25.15%) |
| top500 | 2.085 | 2.442 (17.1%) | 1.663 (-20.25%) | 2.387 (14.49%) | top500 | 1.863 | 2.463 (32.19%) | 1.856 (-0.39%) | 2.377 (27.61%) |
| top750 | 2.11 | 2.457 (16.41%) | 1.605 (-23.93%) | 2.431 (15.18%) | top750 | 1.734 | 2.421 (39.62%) | 1.838 (6.01%) | 2.3 (32.63%) |
| top1000 | 2.01 | 2.37 (17.95%) | 1.517 (-24.52%) | 2.337 (16.28%) | top1000 | 1.711 | 2.416 (41.23%) | 1.799 (5.17%) | 2.269 (32.6%) |

**Table 7.5.** AP and WSJ results

Document terms are weighted locally within the top-*n* document sets. This is applied to both experimental conditions, i.e. both when the cosine and the query-sensitive measures are used. In addition, as I mentioned in section 7.4.1, query terms are also weighted using local evidence from within the top-*n* document sets.

As far as measure M3 is concerned, the values presented here are the ones resulting from a single setting of the ratio of the two parameters $\vartheta_1$ and $\vartheta_2$ for each test collection. The ratio selected is the one that displayed the highest effectiveness for each collection across values of *n* based on the results reported in the previous section. For the four smaller collections (CACM, CISI, LISA and Medline) the ratio selected is that of 1:7, whereas the ratio selected for the two TREC collections is 1:4. In cases where there is not a clear best ratio for all values of *n*, the ratio that displays the best average rank among all ratios is selected.

An alternative procedure for reporting results for M3 would have been to select, for each value of *n*, that ratio that gives the highest effectiveness. This strategy would have resulted in the best possible values for M3. However, it was deemed as more realistic to select values from a single ratio for all values of *n*, rather than to do so selectively from the best ratio for each value of *n*.

[24] Each table contains results for two collections, so in fact each table contains ten columns. In each table I consider the data corresponding to each collection as a separate table. The arrangement of Tables 7.4-7.6 in this way is purely for organisational reasons.

Moreover, it was mentioned in the previous section that the differences in the effectiveness of M3 for the ratios that give the highest values are not significantly different. Consequently, there should not be any distortion of the results by presenting the single best ratio for each test collection.

| *CACM* | *Cosine 5NN* | *M1 5NN* | *M2 5NN* | *M3 5NN* | *LISA* | *Cosine 5NN* | *M1 5NN* | *M2 5NN* | *M3 5NN* |
|---|---|---|---|---|---|---|---|---|---|
| top100 | **1.621** | 1.924 (18.72%) | 1.754 (8.24%) | 1.911 (17.93%) | top100 | **0.896** | 1.362 (52.05%) | **1.395** (55.74%) | 1.402 (56.5%) |
| top200 | 1.511 | 1.981 (31.17%) | **1.902** (25.89%) | 2.04 (35.03%) | top200 | 0.845 | 1.376 (62.84%) | 1.269 (50.13%) | 1.39 (64.53%) |
| top350 | 1.415 | 2.028 (43.27%) | 1.875 (32.45%) | **2.073** (46.45%) | top350 | 0.784 | **1.449** (84.8%) | 1.315 (67.75%) | **1.429** (82.21%) |
| top500 | 1.393 | 2.039 (46.37%) | 1.85 (32.87%) | 2.051 (47.24%) | top500 | 0.783 | 1.425 (81.92%) | 1.317 (68.17%) | 1.423 (81.64%) |
| top750 | 1.376 | **2.045** (48.67%) | 1.761 (28.04%) | 2.006 (45.82%) | top750 | 0.776 | 1.41 (81.68%) | 1.287 (65.81%) | 1.421 (83.09%) |
| top1000 | 1.35 | 2.017 (49.45%) | 1.731 (28.25%) | 1.987 (47.26%) | top1000 | 0.768 | 1.391 (81.18%) | 1.303 (69.71%) | 1.393 (81.49%) |
| full | 1.366 | 1.859 (36.08%) | 1.655 (21.2%) | 1.873 (37.11%) | full | 0.859 | 1.381 (60.73%) | 1.289 (49.97%) | 1.388 (61.5%) |

**Table 7.6.** CACM and LISA results

The results obtained for the 5NN test across all test collections show that query-sensitive measures, in the vast majority of experimental conditions, are more effective than the cosine coefficient at placing co-relevant documents in the same "neighbourhood". The only exception to this is noted when using the M2 measure in the two TREC collections (AP and WSJ), where M2 is less effective than the cosine for all values of $n$ when using the AP collection, and for $n \leq 500$ when using the WSJ (Table 7.5).

| *CISI* | *Cosine 5NN* | *M1 5NN* | *M2 5NN* | *M3 5NN* | *MED* | *Cosine 5NN* | *M1 5NN* | *M2 5NN* | *M3 5NN* |
|---|---|---|---|---|---|---|---|---|---|
| top100 | **1.53** | **1.728** (12.96%) | 1.703 (11.34%) | 1.761 (15.13%) | top100 | **3.143** | **3.569** (13.57%) | 3.361 (6.94%) | **3.576** (13.79%) |
| top200 | 1.37 | 1.652 (20.62%) | **1.733** (26.49%) | **1.789** (30.61%) | top200 | 3.022 | 3.54 (17.13%) | **3.367** (11.4%) | 3.532 (16.86%) |
| top350 | 1.253 | 1.66 (32.51%) | 1.555 (24.13%) | 1.692 (35.09%) | top350 | 3.023 | 3.501 (15.8%) | 3.31 (9.5%) | 3.476 (14.98%) |
| top500 | 1.203 | 1.625 (35.09%) | 1.436 (19.38%) | 1.652 (37.36%) | top500 | 3.003 | 3.475 (15.71%) | 3.305 (10.06%) | 3.436 (14.2%) |
| top750 | 1.14 | 1.55 (35.84%) | 1.357 (19.01%) | 1.575 (38.12%) | top750 | 3.004 | 3.466 (15.4%) | 3.285 (9.37%) | 3.431 (14.23%) |
| full | 1.119 | 1.433 (28.06%) | 1.328 (18.69%) | 1.442 (28.87%) | full | 3.016 | 3.235 (7.26%) | 3.124 (3.57%) | 3.216 (6.63%) |

**Table 7.7.** CISI and Medline results

Statistical tests of the results reveal significant improvements of M1 and M3 over the cosine (significance level <0.001 for the majority of cases) for all experimental conditions except for the CISI collection when $n$=100. Measure M2 is significantly more effective than the cosine for the CACM (except for $n$=100), LISA (all values of $n$), and Medline (except for $n$=100, 750, full) collections. It is also significantly more effective than the cosine when using the WSJ collection

for $n$=750, 1000. Significance levels for M2 are not as low as the ones for M1 and M3, but they are still lower than 0.04 for all significant cases.

The gains in effectiveness introduced by using QSSM are in most cases "material", i.e. over 10%, which confirms the significance of the results (Keen, 1992). The largest differences occur when using the LISA collection, where all three query-sensitive measures are over 50% more effective than the cosine in all experimental conditions. Even M2, which relies only on common terms between documents that are query terms, introduces improvements of that magnitude. This behaviour for LISA can be explained on the basis of its characteristics: on average, queries contain as much as half the number of terms that documents do, and also relevant documents for this collection are strongly characterised by the presence of query terms. CACM, that possesses similar properties, also displays high effectiveness gains for all three QSSM.

Regarding the two TREC collections, it is perhaps not surprising that the use of M2 does not introduce effectiveness gains. The documents of the two TREC collections are large (370 and 377 terms on average per document for AP and WSJ respectively), and the queries relatively short (7.6 terms per query). Moreover, as mentioned previously, these two collections are topically diverse, and therefore terms that appear in queries are likely to be used in documents under many different contexts, not necessarily under the ones dictated by the query. M2 does not use any further contextual information (i.e. the rest of the content overlap between documents), and hence the topical diversity of these collections may mislead the similarity calculations. For example, a relevant document can be deemed as highly similar to another document with which it shares some query terms, but which treats these query terms under a context that is unrelated to the one dictated by the query. By not combining content and query-term overlap, M2 will not capture that these two documents are in fact discussing the query terms under totally different contexts. In such a setting it would seem unlikely that the use of only common query terms between documents can improve the effectiveness of the cosine coefficient.

As far as the AP collection is concerned, this is verified: the use of M2 is always significantly lower than that of the cosine. However, when using the WSJ collection, for $n$=750 and 1000, M2 is significantly more effective than the cosine coefficient (although the differences in effectiveness are not material). Despite that relevant documents of the WSJ collection are strongly characterised by the presence of query terms (Table 7.1), this result is rather surprising. This is especially so, given that for large numbers of top-ranked documents one would expect the confounding effect of non-relevant documents that contain query terms to be stronger on the effectiveness of M2. As this result is not confirmed when using the other TREC collection, it should be seen with caution since it is more likely to be attributed to particular characteristics of the WSJ documents rather than to the actual effectiveness of M2.

### 7.4.3.1 Effectiveness for different numbers of top-ranked documents

In section 6.2, when examining the results of the 5NN test using the cosine coefficient, I demonstrated how, for the majority of the experimental conditions, the cosine coefficient displays the highest result for $n$=100. It was also noted that values past $n$=100 follow a decreasing pattern for increasing values of $n$, a decrease that for a large number of cases is also statistically significant. In section 6.2 I also demonstrated how this behaviour of the cosine coefficient is similar to the one displayed when interdocument associations are calculated randomly.

By observing the data in Tables 7.5-7.7, it seems that measures M1, M2 and M3 (across rows of the tables for columns 3-5) seem to be less affected by the increasing numbers of non-relevant, and "not-as-clearly" relevant, documents (see section 6.2) that are introduced as the value of $n$ increases. M1 for the CACM, LISA and WSJ collections shows the highest scores for $n$=750, 350, and 350 respectively. M2 for the CACM, CISI and Medline collections displays the highest scores for $n$=200. M3 displays the highest effectiveness when using CISI and WSJ for $n$=200, and when using CACM and LISA for $n$=350.

Another observation that can be made from the data in these three tables, is that values for M1, M2 and M3, in the majority of the cases, are more "balanced" across different numbers of top-ranked documents. In fact, the only cases where the effectiveness of a QSSM consistently drops as $n$ increases is noted for the M2 measure using the AP collection (Table 7.5), and for the M1 and M3 measures using the Medline collection (Table 7.7). Moreover, for all experimental conditions using query-sensitive measures, the results obtained using the full document collection are the lowest among all values of $n$; recall that this was not the case when using the cosine coefficient (section 6.2).

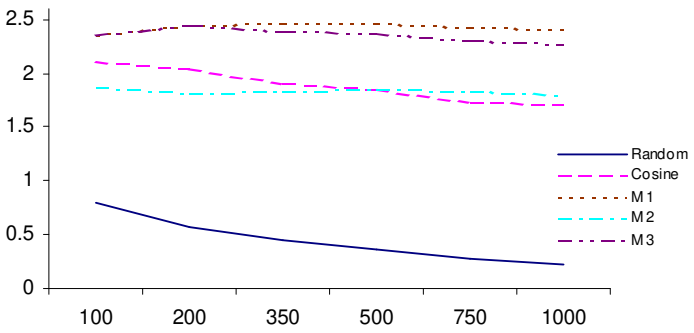


**Figure 7.4.** Random vs. actual values for the 5NN test using the WSJ collection

The statistical significance of the results confirms these observations. As reported in section 6.2, when using the cosine coefficient, the effectiveness at $n$=100 was significantly higher than that at other values of $n$ for the majority of experimental conditions, and also the results were significantly decreasing across values of $n$. This behaviour is not noted when using QSSM. The

only cases where such a behaviour is evident is when using Medline and measures M1 and M3, and when using CISI and M2. In these cases the effectiveness at $n$=100, 200 and 350 is significantly higher than that attained at larger values of $n$. It should also be noted that the effectiveness when using the full number of documents for the CACM, CISI, LISA, and Medline collections is significantly lower than that attained at most other values of $n$.

Figure 7.4 displays the results of the 5NN test (vertical axis) across different values of top-ranked documents (horizontal axis) when using the WSJ collection. The results have been obtained by the three QSSM, the cosine coefficient, and the randomly generated values reported in section 6.2. This figure visually demonstrates that the patterns of results obtained by the QSSM do not follow the patterns noted using the cosine coefficient and the randomly generated similarity values.

Based on the results obtained by the cosine coefficient and the QSSM for varying numbers of top-ranked documents, it seems that the latter cope better with the increasing numbers of non-relevant and "not-as-clearly relevant" documents (section 6.2) introduced in the sets for increasing values of $n$. The pattern of consistently decreasing results for increasing values of $n$ is not noted when using measures M1, M2 and M3, and consequently, numbers of highly similar co-relevant documents do not significantly decrease.

Query-sensitive measures, as opposed to the cosine coefficient, are influenced by the presence of query-terms, and consequently they are likely to be more effective at dealing with the not-as-clearly relevant documents introduced at higher values of $n$. Such documents may only have few query terms, or some of the not important or not common query terms. By biasing the similarity towards query terms, QSSM may be more effective than the cosine coefficient at increasing the similarity of such documents to other relevant documents. Moreover, the use of local query-term weighting (section 7.4.1), may also contribute to this. Not common query terms will likely have high weights when weighted locally within the retrieved sets, as such terms are likely to be infrequent within the retrieved sets. The "not-as-clearly" relevant documents will contain query terms of this type, and the high weight attributed to such terms by the QSSM may contribute to increasing the similarity between such documents and other relevant ones.

The results presented for the query-sensitive measures support this view to an extent. In general, there are no statistical differences between results for the smaller values of $n$ (i.e. between $n$=100, 200 and 350). In fact, most of the differences occur between these values and those of much larger $n$ (e.g. 750, 1000, and full). At that point, the influence of non-relevant documents (or the influence of the "not-as-clearly relevant documents") seems to increase so much as to significantly decrease the effectiveness of the similarity measures in a number of cases. This is especially so for M2, since this measure is more susceptible to erroneous similarity judgements based only on query terms that may be used in different contexts within documents.

### 7.4.3.2 Co-relevant nearest neighbours

In section 7.3 I mentioned that the 5NN test does not provide any information on the number of immediate co-relevant nearest neighbours. To provide information at this level of detail, a variation of the 5NN test (the 1NN test) is performed. The results for this test using the CISI and WSJ collections are presented in Table 7.8. In columns 2-5 (7-10 for WSJ) the percentage of documents whose nearest neighbour is also relevant is displayed when using the cosine coefficient, and when using each of the three QSSM. For M3 the same best ratio $\vartheta_1:\vartheta_2$ for each collection is used as for the calculations in Tables 7.5-7.7. The results of these two collections are representative of the results obtained using the other four collections. The results for all six test collections are presented in Appendix C, Tables C7-C9.

These results reveal a similar pattern to those obtained for the 5NN test. M1 and M3 are significantly more effective than the cosine coefficient for all test collections and values of *n* (all significance levels < 0.02). M2 is significantly more effective than the cosine for the CACM (except for *n*=100), LISA, and Medline (except for *n*=full) collections (significance levels < 0.03). It is worth noting that similar to the 5NN test, for the two TREC collections measure M2 performs significantly worse than the cosine for most values of *n*.

| | *CISI* | | | | | *WSJ* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *n* | *Cosine NN (%)* | *M1 NN (%)* | *M2 NN (%)* | *M3 NN (%)* | *n* | *Cosine NN (%)* | *M1 NN (%)* | *M2 NN (%)* | *M3 NN (%)* |
| 100 | **45.44** | **52.11** | 55.79 | **55.79** | 100 | **64.41** | **67.42** | **56.02** | **65.16** |
| 200 | 39.98 | 49.25 | **56.2** | 55.39 | 200 | 57.24 | 62.1 | 49.7 | 61.67 |
| 350 | 35.75 | 47.88 | 54.34 | 52.92 | 350 | 54.05 | 63.73 | 50 | 60.72 |
| 500 | 33.87 | 46.53 | 50.85 | 51.08 | 500 | 52.65 | 62.9 | 48.64 | 58.83 |
| 750 | 32.82 | 45.1 | 44.77 | 48.21 | 750 | 49.19 | 61.82 | 48.18 | 57.32 |
| 1000 | - | - | - | - | 1000 | 47.6 | 60.43 | 47.73 | 55.76 |
| full | 32.85 | 41.3 | 37.05 | 42.79 | full | - | - | - | - |

**Table 7.8.** Results of the 1NN test when using CISI and WSJ

Based on the results for the 1NN test, it seems that all three QSSM (and especially M1 and M3) manage to increase the proportion of co-relevant nearest neighbours for all test collections. Consequently, such measures are likely to increase the effectiveness of a clustering system that employs nearest neighbour clusters (NNC), such as those proposed by Griffiths and his colleagues (1986) (also see section 4.3.3).

The results of both the 5NN and 1NN tests suggest that measures M1 and M3 are significantly more effective than the cosine at placing co-relevant documents closer to each other. In this way, the likelihood of a more effective clustering of the document space is increased. Augmenting term co-occurrence similarity with query-term co-occurrence information in a pair of documents, is shown to be an effective way of detecting the similarity of co-relevant documents.

The results obtained with measure M2, as I discussed in section 7.2.1, can be seen as a lower limit for the effectiveness of query-sensitive measures. However, despite the extreme form of query biasing that M2 employs, it manages to introduce significant improvements over the cosine in a large number of cases. This result can be seen as providing further evidence for the applicability of query-sensitive measures to IR.

In the following section I examine the comparative effectiveness of the three query-sensitive measures M1, M2 and M3.

### 7.4.4 Comparative effectiveness of M1, M2 and M3

The results of the 5NN test in Tables 7.5-7.7 (columns 3-5) show that measures M1 and M3 achieve higher scores than M2 for the majority of experimental conditions. The only two exceptions are noted when using CISI for $n$=200, and when using LISA for $n$=100; in both cases M2 is more effective than M1 (though not significantly more effective). Statistical testing showed that M1 and M3 are significantly more effective than M2 for all values of $n \neq 200$ when using CACM, for $n$>200 when using CISI, and only for $n$=750 and full when using LISA. For the two TREC collections and Medline, all differences are significant.

The results regarding the comparatively lower effectiveness of M2 are not surprising, given that this measure uses less information than the other two measures. Especially when using the topically diverse TREC collections, the lower effectiveness of M2 compared to M1 and M3 is attributed to its reliance only on common query terms between documents. M2 ignores other common terms between documents that may define the context under which query terms are used within documents.

The other issue to be examined here is the comparative effectiveness of M1 and M3. The results in Tables 7.5-7.7 reveal that the effectiveness of these measures is comparable in most experimental conditions. When using CACM, CISI, LISA or Medline, the differences between the two measures are generally negligible, and never statistically significant. Moreover, none of the two measures consistently outperforms the other in these collections so as to offer an indication of superior effectiveness. For example, when using CACM M3 is more effective than M1 in 4 out of 7 possible values of $n$; there is no pattern to relate smaller values of $n$ with superior effectiveness of one measure over the other. The only consistent behaviour noted is when using CISI where M3 is always more effective than M1, and when using Medline where M1 is always more effective than M3.

The only indication of superior performance comes when using the two TREC collections. When using AP, M1 is more effective than M3 for all but one ($n$=100) values of $n$, and when using WSJ it is more effective than M3 for all values of $n$. Significant differences occur for $n$=750 when

using AP, and for $n$>200 when using WSJ. These differences are confirmed even when the best ratio $\vartheta_1$:$\vartheta_2$ is selected for M3 for each value of $n$ for these two collections. This occurs only once in each collection: for $n$=500 when using AP the result for a ratio of 1:5 is better than the one used (1:4), and when using WSJ for $n$=350 a ratio of 1:2 gives a value higher than the one of the ratio used (1:4).

To appreciate why any significant differences in performance occur between these two measures, one has to look at the way they use information from the query to augment interdocument similarity values. Both measures use information from the content overlap and from the query-term overlap between documents. Consequently, when query terms are common between documents, both measures will augment the content similarity value between those documents by a factor that is incorporated differently for each measure (product for M1, linear combination for M3).

More important than the way similarity values are augmented, is the behaviour of the two measures when no common terms between the two documents are query terms (i.e. when Equation 7.3 outputs zero): M1 sets the similarity of the two documents to zero, whereas M3 sets it equal to a value corresponding to the static similarity between the two documents, adjusted by the parameter $\vartheta_1$[25].

Let us consider the case of a relevant document $D_i$ that contains a few query terms. According to the static component of the similarity, this document will be similar to other documents with which it shares a large number of content terms (not necessarily including query terms). M1 and M3 will re-order this initial similarity ranking in such a way so as to promote documents that share a large number of content terms and query terms with document $D_i$. The re-ordering generated by M1 will remove documents with no query-term overlap with $D_i$ from the top of the list in a rather crude way, by setting their similarities to $D_i$ to zero. The reordering generated by M3 will promote documents with query-term overlap with $D_i$, but may not promote such documents sufficiently to "force" them to obtain a similarity to $D_i$ higher than documents with no query-term overlap (but significant content term overlap) may have. This is also more likely to occur for TREC documents because of their length: it is more likely to have documents with a strong (non-query term) content overlap than it is for documents of shorter lengths, as those of the other four collections.

The results for the 1NN test (Table 7.8 and Tables C7-C9) offer a slightly different view regarding the comparative effectiveness of these three measures. The main results of this test are

[25] Whether the returned similarity value is adjusted by $\vartheta_1$ or not makes no difference to the results of the tests. This was proven experimentally when using the 5NN test. These results are not considered significant enough to be reported.

first that M1 is always more effective than M3 for the two TREC collections (and in most cases significantly so), second that M3 is always more effective than M1 when using the other four collections (and in the majority of these cases significantly), and third that M2 is more effective than the other two measures (especially than M1) in a large number of experimental conditions when using CACM, CISI, LISA or Medline.

More specifically, M2 is the most effective measure when using CISI for $n$=100, 200 and 350, when using LISA for $n$=100, and when using Medline for $n$=200, 350 and 500. In addition to these cases, it also exceeds M1 when using CACM for $n$=200,350 and 500, when using CISI for $n$=500, when using LISA for $n$=750 and 1000, and when using Medline for $n$=750. The differences when using CISI, LISA and Medline are statistically significant. As far as the two TREC collections are concerned, in agreement with the results of the 5NN test, M2 is significantly less effective than both M1 and M3.

It is worth noting from Table C8, where results of the 1NN test are presented when using LISA, that, for all values of $n$ and all similarity measures used, less than half the relevant documents have another relevant as their most similar neighbour. This result is especially surprising given the success of the three query-sensitive measures when using this collection (as the results of Table 7.6 demonstrated). This result demonstrates that there are aspects of the similarity of co-relevant documents that are not captured by these measures. However, it should be emphasised that all three query-sensitive measures introduce significant effectiveness improvements in this collection compared to the cosine coefficient.

Based on the results presented in this section, it is valid to state that M1 and M3 are both more effective than M2 at placing co-relevant documents at close proximity to each other. This is especially evident when using short queries, since M2 relies only on the information supplied by the query terms. In the following section the effect that query length has on query sensitive measures is examined.

### 7.4.5 Effect of query length on the query-sensitive measures

In the results for the 5NN test in Tables 7.5-7.7 (columns 3-5), M2 was more effective than the cosine for the CACM, LISA and Medline collections, where the average query length is relatively large (on average, 13 terms for CACM, 19.4 for LISA, and 10 for Medline, compared to 7.6 for AP, CISI and WSJ). This is a consequence of the strong dependence of M2 on query terms.

In order to investigate the effect of query length on the effectiveness of all three measures, an expanded and a shorter version of the 50 TREC topics for the AP and WSJ collections were used. For the expanded version, terms from the *Title*, *Description*, and *Concepts* fields of each topic were used (see section 5.5.1), yielding on average 23.4 terms per query (compared to 7.6 terms

initially). For the shorter version of the queries only the *Title* field was used, with an average of 3.2 terms per query.

In relation to the data presented in Table 7.1, when using short queries in the WSJ collection 89.7% of relevant documents contain at least one query term, with an average of 2.9 query terms per relevant document. When using the AP collection and short queries 88.6% of relevant documents contain at least one query term, with an average of 3 query terms per relevant document. Not surprisingly, when the longer form of the topics is used, all relevant documents for both collections contain at least one query term, with an average of 8.5 query terms for AP, and 8 terms for WSJ.

The expansion terms for the TREC topics are not generated algorithmically, and this can perhaps be seen as a point of criticism. For example, a query expansion algorithm might have selected terms that are better discriminators than the ones selected manually, by analysing distribution patterns over an entire document corpus, or locally over a set of retrieved documents (Xu & Croft, 1996). However, it is felt that the experimental procedure followed in this section is sufficient to demonstrate the behaviour of the query-sensitive measures when variations in query length occur, as any research relating to query-expansion issues is not pursued in this thesis.

| *n* | *M1 expanded* | *M2 expanded* | *M3 expanded* | *M1 short* | *M2 short* | *M3 short* |
|---|---|---|---|---|---|---|
| 100 | **2.67** (1.95%) | **2.364** (13.75%) | **2.687** (1.34%) | **2.459** (-6.45%) | **1.616** (-22.25%) | **2.541** (-4.18%%) |
| 200 | 2.39 (-0.67%) | 2.128 (16.05%) | 2.402 (-0.05%) | 2.095 (-12.92%) | 1.313 (-28.39%) | 2.254 (-6.23%%) |
| 350 | 2.408 (0.75%) | 2.1 (25.65%) | 2.384 (1.5%) | 2.081 (-12.92%) | 1.21 (-27.59%) | 2.216 (-5.66%) |
| 500 | 2.422 (-0.83%) | 2.124 (27.71%) | 2.401 (0.57%) | 2.08 (-14.81%) | 1.199 (-27.88%) | 2.228 (-6.68%) |
| 750 | 2.494 (1.52%) | 2.191 (36.5%) | 2.464 (1.38%) | 2.137 (-13%) | 1.192 (-25.72%) | 2.237 (-7.95%) |
| 1000 | 2.428 (2.41%) | 2.129 (40.39%) | 2.387 (2.14%) | 2.05 (-13.51%) | 1.127 (-25.7%) | 2.167 (-7.25%) |

**Table 7.9.** The effect of query length for AP: results of the 5NN test

The 5NN test was repeated for both the expanded and shorter versions of the queries, on the same sets of documents as for the original queries[26], for each value of *n*. For measure M3 the best ratio (1:4) of parameters $\vartheta_1$ and $\vartheta_2$ was used for both collections so as to allow these results to be compared to the results reported in Table 7.5. Other ratios were tried in order to examine whether query length would change the most effective ratio for these collections, but there were no significant deviations from the pattern of the results presented in Tables C1 and C6. The results using the modified queries for the AP and WSJ collections are presented in Tables 7.9 and 7.10

[26] This choice was made so as to be able to compare the results between the modified and the original queries.

respectively, where the highest values for each column are displayed in bold. For columns 2-7 the percentage differences between the reported values and those obtained with the standard queries (Table 7.5, columns 3-5) are displayed in brackets.

The results in Tables 7.9 and 7.10 confirm the strong dependence of M2 on query length. M2 with the expanded queries (column 3) is significantly more effective than using the initial queries for all values of *n* (significance levels <0.001). Moreover, when using WSJ, M2 is significantly more effective than the cosine coefficient for all values of *n* (significance levels <0.03), and is not significantly worse than M1 or M3 (either with expanded or initial queries). It is also more effective that M3 for *n*=750 and 1000, but not significantly so.

| *n* | *M1 expanded* | *M2 expanded* | *M3 expanded* | *M1 short* | *M2 short* | *M3 short* |
|---|---|---|---|---|---|---|
| top 100 | 2.457 (4.22%) | 2.372 (26.67%) | 2.414 (2.54%) | **2.32** (-1.59%) | **1.672** (-12%) | **2.295** (-2.52%) |
| top 200 | 2.535 (3.63%) | 2.37 (29.69%) | **2.474** (1.25%) | 2.271 (-7.7%) | 1.631 (-12.04%) | 2.236 (-8.49%) |
| top 350 | **2.54** (2.91%) | 2.415 (31.82%) | 2.448 (2.46%) | 2.241 (-10.14%) | 1.536 (-19.31%) | 2.159 (-9.52%) |
| top 500 | 2.54 (3.14%) | **2.425** (30.71%) | 2.44 (2.65%) | 2.195 (-12.22%) | 1.525 (-21.67%) | 2.173 (-8.59%) |
| top 750 | 2.441 (0.83%) | 2.407 (30.93%) | 2.344 (1.9%) | 2.101 (-15.24%) | 1.434 (-28.17%) | 2.05 (-10.88%) |
| top 1000 | 2.437 (0.85%) | 2.399 (33.35%) | 2.325 (2.47%) | 2.064 (-17.09%) | 1.435 (-25.36%) | 2.022 (-10.88%) |

**Table 7.10.** The effect of query length for WSJ: results of the 5NN test

When using the AP collection, M2 exceeds the cosine for some values of *n* (500, 750 and 1000) but not significantly, and it is also not significantly worse than the cosine for the other values of *n*. In contrast to when using WSJ, M2 with the expanded queries is still significantly worse than both M1 and M3 for all values of *n*.

The behaviour of M2 for expanded queries can be explained on the basis of the role that the added query terms play for this measure. Because M2 relies only on common query terms between documents, it lacks the contextual information provided by other common terms between documents. The addition of terms to the query provides more information to M2 to effectively assess the likelihood of two documents to be jointly relevant to the same query. When using WSJ, this addition of extra terms seems to be enough for M2 to be almost as effective as the other two query-sensitive measures (that use more information to assess similarity), and more effective than the cosine. In practical terms, this implies that these extra terms that M2 uses are almost as good a source of information as all the common terms between documents. When using the AP collection however, this does not seem to be the case.

Column 6 of Tables 7.9 and 7.10 shows a significant decrease in effectiveness for M2 when average query length is decreased to 3.2 terms. The decrease in effectiveness is sizeable if one

considers that the difference in query length between the initial and the short queries is on average just 4.4. terms. This result, in combination with the result for M2 when using expanded queries, indicates that there is a range in the number of query terms within which M2 can perform reasonably effectively. The addition of more query terms past the upper limit of this range, for example, is unlikely to prove more effective for M2, whereas the removal of query terms, as demonstrated here, has significantly negative effects on the measure's performance.

Measures M1 and M3, on the other hand, are less affected by the increase in query length from 7.6 terms per query (initial queries) to 23.4 (expanded). None of the differences in effectiveness reported in Tables 7.9 and 7.10 (columns 2 and 4) between the expanded and the initial form of the queries are significant. In some cases when using AP, there is even a minor decrease in the effectiveness of the measure when expanded queries are used ($n$=200 and 500 for M1 and $n$=200 for M3).

When short queries are used (columns 5 and 7 of Tables 7.9 and 7.10), both measures (M1 and M3) display a significant decrease in effectiveness. The decrease is smaller in scale than that reported for M2, but significant (significance levels <0.03) for both collections and all values of $n$, except for $n$=100. Despite this decrease, M1 and M3 using the short queries are still significantly more effective than the cosine when using the WSJ collection (Table 7.5, column 2, significance levels <0.003). When using AP, M1 is more effective than the cosine for $n$=100, 750 and 1000, and more effective than M3 for all values of $n$. However, no significant differences between these measures and the cosine are noted.

Comparing the effectiveness of M1 and M3 when expanded queries are used for the WSJ collection (columns 2 and 4 of Table 7.10), the effectiveness of the former is always higher than that of the latter, and for $n$>100 the differences are statistically significant (significance levels <0.02). The comparative effectiveness of the two measures does not seem to change when queries are augmented (recall from section 7.4.4 that M1 was more effective than M3 for all values of $n$ when using WSJ). When using AP a similar behaviour is noted (for $n$>200 M1 is more effective than M3). However, any differences noted between the two measures are not significant.

When short queries are used, (columns 5 and 7 of Tables 7.9 and 7.10), the comparative performance of M1 and M3 differs in the two collections (i.e. when using AP M3 is more effective than M1, and vice versa when using WSJ). However none of the differences between the two measures are significant. These results suggests that the comparative effectiveness of the two measures does not significantly change depending on the length of the query, something that is not surprising given that both measures take the same amount of information into account when assessing interdocument similarities.

A further observation from the results presented in Tables 7.9 and 7.10 is that M1 seems to be more affected by the reduction of query length than M3. By observing the decrease in the performance of the two measures when short queries are used (columns 5 and 7 of Tables 7.9 and 7.10), it follows that the relative differences are in general much larger when using M1 (i.e. the effectiveness using short queries compared to that using the standard queries decreases more when using M1 than M3, but the absolute effectiveness of the two measures is comparable).

The results presented here suggest that M2 is highly affected by query length, and it would therefore not seem suitable to be applied to environments where very short queries are usually input by users, unless effective ways to expand the query could be used. Assessing the likelihood of two documents to be jointly relevant to a query based on the amount of information provided by approximately 3 terms on average is not likely to be effective.

The other two measures, perhaps not surprisingly, do not seem that much affected by variations in query length. This is due to that they combine contextual information (the whole set of terms between documents) with increased weight assigned to query information. In this way, M1 and M3 are more likely to cope well when query length is decreased: the contextual information may be a good indicator of whether the few query terms are used in the same topic between documents. It is for this same reason that the effectiveness of the two measures does not significantly benefit from the addition of terms to the query.

This behaviour of measures M1 and M3 might appear useful in an operational environment, like a web search engine for example, where user queries comprise only few terms (Jansen *et al*., 2000). In the specific experimental environment used in this thesis, M1 and M3 outperformed the cosine coefficient in a large number of cases when short queries were used. It remains to be seen whether such improvements would occur in operational environments.

It should also be mentioned that the results reported in this section regarding the effect of query length, may have been affected by the way that the expanded forms of the queries were obtained. If the expanded terms were chosen in a different way, then a different picture regarding the effectiveness of the measures for varying query lengths might have been obtained. If, for example, expansion terms were obtained algorithmically, then the effectiveness of M2 compared to M1 and M3 may improve. Query terms added algorithmically may be better at discriminating between relevant and non-relevant documents than the ones used here.

If one assumes that the retrieval effectiveness of query terms is an indication of their effectiveness at discriminating between relevant and non-relevant documents, then some useful insight is provided from a per-query analysis of the results of the 5NN test. An analysis of the queries for which M2 is consistently more effective than the other measures (M1, M3 and cosine) for the WSJ collection, demonstrated that a large number of these queries display high retrieval

effectiveness. For example, three of the topics for which (in their original, unexpanded form) M2 is consistently more effective are TREC topics 31, 32 and 34. The 11-point average precision for these three topics is 0.31, 0.47 and 0.39 respectively, which is significantly higher than the average for the WSJ collection which is 0.25 (Table 5.2, section 5.5.2). The length of these three topics in their original form is 7, 7 and 10 terms respectively, near the 7.6 average of the collection. It should however be noted that this correlation is not general, i.e. not all of the most effective queries, in terms of retrieval, are also the ones for which M2 is the most effective measure.

The observation, however, that there is some correlation between the effectiveness of measure M2 and the retrieval effectiveness of the queries, suggests that further research would be needed to appreciate the dependence of query-sensitive measures on query length, as well as on the "quality" of the terms that the query contains.

The way that the expanded forms of queries were obtained may also be a contributing factor for the different effect of query length when using the two TREC collections. Tables 7.9 and 7.10 show that the results of the 5NN test in each of the two collections are differently affected by variations in query length. The discriminating power of the query terms in each query form examined is likely to be different for each collection, and therefore likely to have a different effect on the effectiveness of the query-sensitive measures.

The way that TREC relevance assessments are constructed may also be a factor contributing to the results reported in this section. In Chapter 2 (section 2.4) I mentioned the pooling technique which is used to generate the relevance assessments for the TREC collections (Harman, 1993). In brief, the top 200 documents retrieved in response to each topic by each of the IR systems participating in TREC were retrieved (25 systems in total), and it was only for these documents that relevance assessments were made. Any documents not retrieved in the top 200 were assumed to be non-relevant. Each of the systems which participated in TREC used different fields of the TREC topics to retrieve documents (e.g. *title*, *concepts*, etc). This naturally affects the type of documents which were retrieved by these systems, and consequently, the type of documents for which relevance assessments were made. For the expanded version of the queries used here, the *title*, *description* and *concepts* fields were used. The discriminating power of query terms contained in these fields may be influenced by the method that has been used to assess the relevance of the TREC documents.

## 7.5 Summary

In this chapter I introduced means by which query-sensitive similarity measures can be defined. Query-sensitive measures bias similarity towards pairs of documents that jointly possess terms

that are expressed in a query. This is based on the view that similarity is a dynamic and purpose-sensitive notion, and that query-sensitive measures have the potential to capture the dynamics of similarity for the calculation of interdocument relationships.

I presented three such measures. Two of them take into account all common terms between a pair of documents, but bias the similarity measure towards those common terms that are also query terms (measures M1 and M3). Each of these two measures uses a different function to combine static and variable similarity (M1 uses a product of the two sources, where M3 uses a linear combination). The third measure only takes into account common terms between documents that are query terms (measure M2).

Four main issues were experimentally investigated in this chapter. The first issue was the effectiveness of the query-sensitive measure M3 as a function of the ratio of two parameters that assign importance to the static and variable similarity between two documents. The results across the six document collections were consistent, in that significantly higher effectiveness occurred when the ratio of the parameters was set so as to assign greater importance to the variable part of the similarity. The actual setting depends on the characteristics of the test collection under investigation. For the test collections that are examined in this thesis, a setting between 1:4 and 1:7 proved to be the most effective.

The second issue was the comparative effectiveness of the three QSSM and the cosine coefficient. The results demonstrated that measures M1 and M3 are always significantly more effective than the cosine at placing co-relevant documents close to each other. M2 outperformed the cosine for a large number of experimental conditions, mainly for small homogeneous collections. It was also demonstrated that the effectiveness of query-sensitive measures does not follow the same pattern of the cosine coefficient, and of randomly generated similarities, i.e. to consistently decrease as the number of documents increases.

The third issue related to the comparative effectiveness of the three query sensitive measures. M2 was less effective than the other two for the vast majority of experimental conditions, and especially for long, topically diverse documents. Measures M1 and M3, in general, performed comparably, with M1 proving more effective at adjusting to the topical nature of relevance typically employed in IR research.

The fourth issue investigated in this chapter, was the effect of query length on the effectiveness of the three measures. Again, M1 and M3 displayed a similar behaviour, and although influenced by short query length still managed to outperform the cosine coefficient for a large number of experimental conditions. Measure M2, on the other hand, proved highly sensitive to variations of query length.

The main conclusion from this chapter is that the use of query-sensitive measures for the calculation of interdocument relationships is highly effective. Regarding the motivation behind the introduction of QSSM to IR, the per-query adherence to the cluster hypothesis (section 5.3), the results presented in this chapter demonstrate that, compared to static measures, query-sensitive measures achieve a significantly higher adherence to the hypothesis. A perfect per-query adherence is not achieved, and it would seem unlikely that considering only topical aspects of relevance would achieve this.

The results presented in this chapter demonstrate the applicability of query-sensitive measures to IR. A more thorough evaluation of such measures can be performed if one integrates them in a wider application area. This is the aim of the following chapter, where query-sensitive measures are applied to hierarchic document clustering.

# Chapter 8

# Hierarchic Document Clustering Using Query-Sensitive Similarity Measures

## 8.1 Introduction

In this chapter I investigate the effectiveness of the second form of query-based clustering that is considered in this thesis: the generation of document hierarchies by using query-sensitive similarity measures.

In Chapter 6 I examined the effectiveness of query-based clustering in the form of post-retrieval clustering. The results that I presented in Chapter 6 suggested significant improvements compared to static clustering. The results also demonstrated that post-retrieval clustering has the potential to exceed the effectiveness of inverted file searches. However, a number of shortcomings regarding the effectiveness of post-retrieval clustering were noted in that chapter. These mainly involved the unfavourable comparative effectiveness to inverted file searches at the MK4 level for a large number of cases (section 6.3.2), and the close-to-random effectiveness for a number of cases when using the LISA and CISI databases (section 6.3.3).

The form of hierarchic clustering that is investigated in this chapter can be seen as introducing a further level of query influence on top of post-retrieval clustering. In addition to clustering retrieval results, the query is also taken into account when calculating interdocument associations. This is done by using the query-sensitive measures proposed in Chapter 7 (section 7.2.1). The effectiveness of these measures in structuring the document space in terms of the proximity of co-relevant documents was demonstrated in Chapter 7.

I investigate the effectiveness of query-based clustering which uses query-sensitive similarity measures by comparing its effectiveness to that of clustering which uses conventional static

similarity measures. Using query-sensitive similarity measures with any clustering method that makes use of a similarity matrix (such as for example the four hierarchic methods used in Chapter 6) is a straightforward process: the only step of the clustering process that is affected is that of the generation of the interdocument similarity matrix (section 3.3.1).

This in turn means that the comparison of the effectiveness of document clustering using query-sensitive measures to document clustering using static measures is feasible. Keeping all other experimental settings constant (i.e. indexing exhaustivity, term weighting schemes, etc.) and varying only a single experimental parameter (i.e. the similarity measure), it is possible to attribute any variations in retrieval effectiveness to that single parameter that is varied. In this chapter, the two different "values" of the experimental parameter under consideration are defined by the use of different types of similarity measures, i.e. static (cosine coefficient), and query-sensitive (measures M1, M2 and M3 as defined in section 7.2.1).

Moreover, by employing optimal evaluation to measure cluster-based retrieval effectiveness, it is possible to attribute any differences in effectiveness across experimental conditions to the variation of the experimental conditions themselves (internal factors), and not to other (external) factors that may influence the outcome of the evaluation. Such external factors were outlined in previous chapters (mainly in sections 3.5.1, 3.5.2 and 4.3.3).

The study of the effectiveness of the application of query-sensitive measures to hierarchic document clustering is organised in four parts, each part presented in a section of this chapter. At the end of each section I also present a discussion that summarises the main findings of each part.

In the first part, in section 8.2, I examine the effectiveness of hierarchic clustering using query-sensitive similarity measures, and I compare it to three aspects of the effectiveness of clustering using static similarity measures presented in Chapter 6. More specifically, in section 8.2.1 I compare the effectiveness of hierarchic clustering using a conventional static similarity measure (i.e. the cosine coefficient), to the effectiveness using the three QSSM defined in the previous chapter. In section 8.2.2 I then examine how the effectiveness attained with query-sensitive measures compares to that obtained by an inverted file search, and in section 8.2.3 I compare the query-sensitive effectiveness to that of random clustering.

The second part of the study, in section 8.3, involves the investigation of the characteristics of hierarchies generated by using QSSM. In section 8.3.1 I examine whether the use of QSSM to generate document hierarchies alters the behaviour of the clustering methods in terms of the characteristics of the hierarchies that they generate. Then, in section 8.3.2 I present statistics about the optimal clusters generated using the QSSM, in terms of average size and number of relevant documents they contain (section 8.3.2.1), and in terms of the levels of the hierarchy in which

optimal clusters occur (section 8.3.2.2). I also compare these data to when using static measures (i.e. the data presented in Chapter 6).

The third part of the study involves the comparison of the effectiveness of hierarchies generated by each of the three query-sensitive measures (section 8.4). The aim of this comparison is to examine whether there is a single measure that tends to consistently yield the highest effectiveness in the experimental environment used. In this section, I also examine the effect that query length has on the effectiveness of the query-sensitive measures (section 8.4.2), and the effectiveness of the QSSM across different numbers of top-ranked documents (section 8.4.3).

The last part of the study looks into the comparative effectiveness of the four clustering methods. In Chapter 6 (section 6.4) I examined the comparative effectiveness of these methods under post-retrieval clustering (and also under static clustering). In section 8.5 I present results that compare these methods when QSSM are used for the calculation of interdocument relationships. I finish this chapter by summarising its main findings in section 8.6.

## 8.2 The effectiveness of hierarchic clustering using query-sensitive similarity measures

In this section I report on results that are obtained by the application of query-sensitive measures to hierarchic clustering methods. The results are generated by applying each of the three query-sensitive measures defined in Chapter 7 (section 7.2.1) to the four clustering methods used (group average, Ward, complete link and single link). As mentioned previously, the effectiveness of the hierarchies is evaluated by using optimal cluster evaluation, and by calculating effectiveness for the three values of the parameter $\beta$ (0.5, 1, 2) of the E measure (section 4.3).

Evaluation in this chapter is performed in the same way as in Chapter 6, that is, by considering all relevant documents for a query for all numbers of top-ranked documents. As I discussed in section 6.3, this type of evaluation does not distort the results presented and the conclusions extracted. However, this type of evaluation can explain some of the results that are reported in this section, as well as in later sections of this chapter. When this happens, I explicitly report this effect in the respective section.

The results in this section are examined under three different viewpoints. First, in section 8.2.1, the results are examined in comparison to the results reported in section 6.3 (Table 6.4) and in Appendix B (Tables B1-B4). These results correspond to cluster-based retrieval effectiveness obtained by the use of a static similarity measure (the cosine coefficient). Then, in section 8.2.2, the results are examined in comparison to the effectiveness obtained by an inverted file search (IFS). Results for the effectiveness of IFS are reported in Appendix B (Tables B1-B7), and were

also reported in section 6.3 (Table 6.4). In section 8.2.3, the results are viewed in comparison to cluster-based effectiveness obtained from random structures. Results for random cluster-based effectiveness were reported in section 6.3.3, and are presented in Appendix B (Tables B8-B11).

### 8.2.1 Comparatively to clustering using static similarity measures

In Table 8.1 optimal cluster-based retrieval results are presented for the six test collections used. The results are obtained from document hierarchies generated by the group average method. The results in this table correspond to E values, and therefore the lower the values the higher the retrieval effectiveness. The E values presented in this table are calculated for $\beta$=0.5 (precision-oriented searches) and for $\beta$=2 (recall-oriented searches). The measure used to gauge optimal cluster effectiveness is the MK1 measure (section 4.3.5), i.e. the same that was used in Chapter 6. The results for all four clustering methods are presented in Appendix D, Tables D1-D4.

For each value of $\beta$ there are four columns in Table 8.1. Each column corresponds to the E value obtained when using a different similarity measure. The values under the "Standard" column correspond to the ones obtained when using the cosine coefficient, and they are the ones reported in Chapter 6 (Table 6.4); they are also presented here for ease of reference. The other three values correspond to E values obtained with each of the three QSSM: M1, M2 and M3. Throughout this chapter I mostly refer to cluster-based retrieval effectiveness obtained using the cosine coefficient as "*standard effectiveness*", and to effectiveness obtained using any of the query-sensitive measures as "*query-sensitive effectiveness*".

The results using measure M3 are obtained by the ratio that proved to be the most effective in the previous chapter (section 7.4.2). Experiments were carried out using other ratios of the two parameters in the calculation of M3 ($\vartheta_1$, $\vartheta_2$), however no significant deviations from the results presented in 7.4.2 were noted in relation to the comparative effectiveness obtained by the various ratios. These data do not alter the pattern of the results presented in this chapter. Therefore, for the remaining of this chapter, all results reported for measure M3 have been obtained by using the single most effective setting of the two parameters that was described in section 7.4.2.

The results presented in Table 8.1, and those presented in Tables D1-D4, demonstrate that, in the majority of the experimental conditions, the use of query-sensitive measures in document clustering is more effective than the use of a static measure. Effectiveness improvements are consistent and significant. The extent of the improvements that each of the query-sensitive measures introduces depends on the number of top-ranked documents clustered, on the clustering method used, on the type of search performed (i.e. recall or precision-oriented), and on the document collection clustered. I discuss each of these issues in the following paragraphs.

| | | $\beta$=0.5 | | | | $\beta$=2 | | |
|---|---|---|---|---|---|---|---|---|
| *AP* | *Standard* | *M1* | *M2* | *M3* | *Standard* | *M1* | *M2* | *M3* |
| top100 | 0.511 | 0.527 | 0.573 | 0.520 | 0.619 | 0.613 | 0.643 | 0.618 |
| top200 | 0.514 | 0.509 | 0.567 | 0.511 | 0.604 | 0.576 | 0.603 | 0.580 |
| top350 | 0.507 | 0.494 | **0.557** | 0.497 | 0.576 | 0.545 | 0.580 | 0.551 |
| top500 | 0.508 | 0.477 | 0.562 | 0.486 | 0.560 | 0.522 | 0.575 | 0.536 |
| top750 | 0.488 | 0.452 | 0.564 | 0.477 | 0.562 | **0.505** | 0.569 | 0.518 |
| top1000 | **0.482** | **0.448** | 0.561 | **0.467** | **0.550** | 0.513 | **0.565** | **0.512** |
| *CACM* | *Standard* | *M1* | *M2* | *M3* | *Standard* | *M1* | *M2* | *M3* |
| top100 | **0.438** | 0.435 | 0.468 | 0.438 | **0.502** | 0.472 | 0.509 | 0.490 |
| top200 | 0.476 | 0.426 | 0.417 | 0.418 | 0.512 | 0.476 | 0.480 | 0.484 |
| top350 | 0.469 | 0.412 | 0.423 | 0.408 | 0.520 | 0.450 | 0.480 | 0.475 |
| top500 | 0.461 | 0.404 | **0.412** | 0.417 | 0.540 | 0.440 | **0.478** | 0.468 |
| top750 | 0.465 | **0.400** | 0.427 | **0.407** | 0.537 | **0.442** | 0.483 | 0.470 |
| top1000 | 0.463 | 0.405 | 0.417 | 0.412 | 0.537 | 0.445 | 0.479 | **0.462** |
| full | 0.641 | 0.639 | 0.642 | 0.641 | 0.782 | 0.787 | 0.788 | 0.787 |
| *CISI* | *Standard* | *M1* | *M2* | *M3* | *Standard* | *M1* | *M2* | *M3* |
| top100 | 0.630 | 0.635 | 0.667 | 0.642 | 0.702 | 0.702 | 0.710 | 0.708 |
| top200 | 0.609 | 0.592 | 0.648 | 0.604 | 0.658 | 0.649 | 0.671 | 0.649 |
| top350 | 0.589 | 0.578 | **0.641** | 0.598 | 0.655 | 0.614 | 0.642 | 0.623 |
| top500 | 0.593 | 0.570 | 0.648 | 0.593 | 0.656 | 0.615 | **0.639** | **0.615** |
| top750 | **0.567** | **0.561** | 0.643 | **0.577** | **0.649** | **0.609** | 0.642 | 0.615 |
| full | 0.790 | 0.787 | 0.787 | 0.787 | 0.798 | 0.797 | 0.796 | 0.797 |
| *LISA* | *Standard* | *M1* | *M2* | *M3* | *Standard* | *M1* | *M2* | *M3* |
| top100 | 0.517 | 0.463 | 0.524 | 0.492 | 0.576 | 0.524 | 0.584 | 0.550 |
| top200 | 0.504 | 0.423 | 0.478 | 0.438 | 0.559 | 0.503 | 0.550 | 0.507 |
| top350 | 0.493 | **0.411** | **0.451** | **0.432** | 0.553 | **0.490** | **0.496** | 0.477 |
| top500 | 0.487 | 0.425 | 0.465 | 0.458 | 0.568 | 0.497 | 0.503 | 0.507 |
| top750 | 0.489 | 0.446 | 0.475 | 0.449 | 0.571 | 0.490 | 0.523 | 0.500 |
| top1000 | **0.475** | 0.450 | 0.466 | 0.444 | **0.549** | 0.496 | 0.520 | **0.490** |
| full | 0.643 | 0.641 | 0.644 | 0.642 | 0.716 | 0.713 | 0.716 | 0.714 |
| *MED* | *Standard* | *M1* | *M2* | *M3* | *Standard* | *M1* | *M2* | *M3* |
| top100 | 0.300 | 0.264 | 0.324 | 0.294 | 0.308 | 0.296 | 0.335 | 0.319 |
| top200 | 0.281 | 0.264 | 0.300 | 0.287 | 0.294 | 0.277 | **0.319** | 0.300 |
| top350 | 0.281 | **0.254** | **0.288** | **0.274** | **0.271** | **0.269** | 0.321 | 0.294 |
| top500 | 0.279 | 0.258 | 0.291 | 0.275 | 0.273 | 0.270 | 0.322 | **0.291** |
| top750 | **0.276** | 0.259 | 0.292 | 0.278 | 0.272 | 0.271 | 0.324 | 0.295 |
| full | 0.682 | 0.682 | 0.687 | 0.684 | 0.711 | 0.734 | 0.740 | 0.736 |
| *WSJ* | *Standard* | *M1* | *M2* | *M3* | *Standard* | *M1* | *M2* | *M3* |
| top100 | 0.608 | 0.585 | 0.609 | 0.586 | 0.696 | 0.679 | 0.682 | 0.682 |
| top200 | 0.604 | 0.560 | 0.594 | 0.558 | 0.661 | 0.620 | 0.633 | 0.629 |
| top350 | 0.603 | 0.541 | 0.591 | 0.546 | 0.650 | 0.582 | 0.590 | 0.583 |
| top500 | **0.585** | **0.534** | 0.581 | **0.535** | 0.642 | 0.568 | 0.575 | 0.574 |
| top750 | 0.585 | 0.537 | **0.569** | 0.537 | **0.640** | 0.563 | 0.573 | 0.567 |
| top1000 | 0.586 | 0.541 | 0.572 | 0.536 | 0.641 | **0.559** | **0.569** | **0.557** |

**Table 8.1.** Optimal cluster-based effectiveness using the group average method. Highest effectiveness (lowest E value) for each column appears in bold

The study of retrieval effectiveness using each of the three query-sensitive measures when different numbers of top-ranked documents are clustered is presented in section 8.4.3. Instead, what is of interest here, is that there seems to be a general pattern across experimental conditions for query-sensitive effectiveness to increase its improvement over standard effectiveness as the number of top-ranked documents increases.

For example, in Table 8.1, when using the CISI collection and measure M1, the effectiveness obtained at $n$=100 is slightly lower than that obtained using the cosine coefficient (though not significantly so). When more documents are clustered, query-sensitive effectiveness with M1 becomes higher than the standard effectiveness, and for $n \geq 350$ for recall-oriented searches it becomes significantly more effective. Another such example can be found by looking at the results for the WSJ collection. When using M1 and recall-oriented searches ($\beta$=2), the difference in effectiveness between the *M1* (column 7) and *Standard* (column 6) columns of the table range from 5.6% for $n$=100, to 23% for $n$=1000 in favour of M1, with the differences consistently increasing in between.

The only deviation from this general trend is noted when comparing query-sensitive and standard effectiveness for $n$=full. It should be reminded that $n$=full corresponds to a static clustering when using the cosine coefficient, whereas it corresponds to a dynamic clustering which changes on a per-query basis when using any of the query-sensitive measures. The data presented in Table 8.1, and in Tables D2-D4 in Appendix D, show that there are no significant changes between the effectiveness of standard and query-sensitive clustering. This is the case for all four test collections (CACM, CISI, LISA, and Medline), all values of $\beta$, and all query-sensitive measures.

Regarding the effectiveness gains across different clustering methods, the application of query-sensitive measures introduces significant improvements to all four methods, but to a different extent. Of the four clustering methods used, Ward's method (Table D2) in general displays the smallest difference between query-sensitive and standard effectiveness. This difference is statistically significant in many of the experimental conditions, and especially for values of $n$>100. An explanation for this behaviour of the Ward method is presented in section 8.5.4.

The single link method, on the other hand, generally displays the largest improvement when comparing query-sensitive and standard effectiveness. Recall from section 6.4 that single link was the least effective of the four methods in all experimental conditions (with the exception of the Medline collection). The comparative effectiveness of the four methods when query-sensitive measures are used is investigated in section 8.5.

Another observation from the results, is that recall-oriented searches generally introduce larger improvements than precision-oriented searches. The significance of the results in favour of query-sensitive effectiveness using recall-oriented searches is much more consistent than that of

precision-oriented searches. In fact, there are a number of cases where query-sensitive effectiveness significantly outperforms standard effectiveness for $\beta$=2 but fails to do so for $\beta$=0.5. Such an example in Table 8.1 is found when using the CISI collection: for $n \geq 350$ query-sensitive recall-oriented effectiveness is significantly more effective than standard effectiveness, but precision-oriented effectiveness is not.

It should however be emphasised that the results for precision-oriented searches are also, in the majority of the experimental conditions, improving the effectiveness of standard-based retrieval. It is just that they do not achieve statistical significance in as many cases as recall-oriented searches. However, the consistency of the improvements should be taken in consideration (Keen, 1992).

When examining the results on a per-collection basis, the trends seen in Table 8.1, using the group average method, are representative of the behaviour of the other three clustering methods. More specifically, for three test collections (CACM, LISA, and WSJ) the use of any query-sensitive measure improves the effectiveness of standard cluster-based retrieval. These three collections also show the largest differences when comparing query-sensitive and standard effectiveness. Most of the differences in these cases are material (i.e. over 10%), and in some cases they exceed 20%. The significance of the results varies for M1, M2 and M3, but in general for these three collections it is justifiable to say that all three QSSM improve standard cluster-based effectiveness.

When using AP and Medline, the use of measures M1 and M3 introduces improvements over the use of the cosine coefficient for the majority of clustering methods and types of searches. In most cases the improvement in effectiveness is significant, and in many cases the difference in effectiveness exceeds 30% (e.g. single link method, Medline collection, recall-oriented searches, measures M1, M2 and M3 in Table D4). There are some exceptions to this behaviour that are mainly noted when using Ward's method with either of these two collections and either of M1 or M3. When using this method, query-sensitive effectiveness is not consistently higher than standard effectiveness, and in some cases (recall-oriented searches using Medline and M1) it is consistently worse.

When applying measure M2 to the AP collection effectiveness also generally decreases. The decrease in effectiveness is significantly higher for precision-oriented searches than for recall-oriented ones. In the case of recall-oriented searches, the decrease in effectiveness is generally not statistically significant. The only case where the use of M2 consistently introduces improvements over the use of the cosine coefficient for this collection, is for recall-oriented searches using the single link method (Table D4). When using this measure (M2) with Medline, improvements occur for the other three clustering methods, but not for the group average method.

The database that displays the most variable behaviour is CISI. Measures M1 and M3 display a comparable behaviour in this database, with mainly significant improvements for larger values of $n$ (typically $\geq$350), except for precision-oriented searches with the single link method where effectiveness consistently (but not significantly) decreases. The use of M2 has generally a significant negative effect on precision-oriented searches, and an insignificant effect on recall-oriented searches where differences in effectiveness occur in favour of either query-sensitive or standard clustering.

To sum up the findings of this section, the use of query-sensitive measures in hierarchic clustering introduces significant effectiveness improvements compared to using a static similarity measure. The improvements in effectiveness vary depending on the clustering method used, the number of documents clustered, and the type of search performed. The only cases where the use query-sensitive measures consistently results in lower effectiveness than standard clustering are noted when using the M2 measure with the AP and CISI databases.

In the following section I examine how the effectiveness of query-based clustering using query-sensitive measures compares to that of a conventional inverted file search.

### 8.2.2 Comparatively to IFS effectiveness

In this section, the effectiveness of cluster-based retrieval using query-sensitive measures is compared to the effectiveness of inverted file searches (IFS). Results for IFS effectiveness calculated through measures MK3 and MK4 (section 4.3.5) for the six collections are presented in full in Appendix B (Tables B1-B7). These measures do not depend on the outcome of the clustering process, and therefore the values that were calculated using these two measures in Chapter 6 apply here as well. Measure MK1-k depends on the outcome of the clustering process, since it calculates effectiveness based on the number of documents that are contained within an optimal cluster. Therefore, the values that were presented for this measure in Chapter 6 do not apply here. However, IFS results at the MK1-k level are not presented since they do not compare well with query-based effectiveness and do not form part of the discussion in this section.

In section 6.3.2, when I examined the comparative effectiveness of post-retrieval clustering to that of IFS, it was demonstrated that cluster-based effectiveness significantly outperformed IFS effectiveness for a large number of experimental conditions, and especially using the group average method. It was also demonstrated that precision-oriented searches compared more favourably to IFS effectiveness than recall-oriented searches, and that the single link method only managed to outperform IFS effectiveness at the MK1-k level for most experimental conditions. Moreover, the effectiveness obtained by static clustering (i.e. $n$=full when using a static similarity measure) did not compare favourably to IFS effectiveness, which it only managed to exceed at the MK1-k level for a small number of experimental conditions.

In the previous section (8.2.1) I demonstrated how the use of query-sensitive measures for the calculation of interdocument relationships improves the effectiveness of hierarchic clustering methods for the majority of cases that were investigated. In the remaining of this section, I examine whether these improvements translate into more favourable cluster-based effectiveness comparatively to IFS effectiveness.

| | $\beta$=0.5 | | $\beta$=2 | |
|---|---|---|---|---|
| *AP* | *M1* | *IFS* | *M1* | *IFS* |
| top100 | 0.527 | 0.550 (*MK4*) | 0.613 | 0.628 (*MK4*) |
| top200 | 0.509 | 0.543 (*MK4*) | 0.576 | 0.613 (*MK4*) |
| top350 | 0.494 | 0.552 (*MK4*) | 0.545 | 0.611 (*MK4*) |
| top500 | 0.477 | 0.552 (*MK4*) | 0.522 | 0.614 (*MK4*) |
| top750 | 0.452 | 0.548 (*MK4*) | 0.505 | 0.605 (*MK4*) |
| top1000 | 0.448 | 0.548 (*MK4*) | 0.513 | 0604 (*MK4*) |
| *CACM* | *M1* | *IFS* | *M1* | *IFS* |
| top100 | 0.435 | 0.503 (*MK3*) | 0.472 | 0.503 (*MK3*) |
| top200 | 0.426 | 0.498 (*MK3*) | 0.476 | 0.501(*MK3*) |
| top350 | 0.412 | 0.444 (*MK4*) | 0.450 | 0.492 (*MK4*) |
| top500 | 0.404 | 0.444 (*MK4*) | 0.440 | 0.492 (*MK4*) |
| top750 | 0.400 | 0.444 (*MK4*) | 0.442 | 0.492 (*MK4*) |
| top1000 | 0.405 | 0.444 (*MK4*) | 0.445 | 0.492 (*MK4*) |
| full | 0.639 | 0.713 (*MK1-k*) | 0.787 | - |
| *WSJ* | *M1* | *IFS* | *M1* | *IFS* |
| top100 | 0.585 | 0.645 (MK4) | 0.679 | 0.712 (MK4) |
| top200 | 0.560 | 0.640 (MK4) | 0.620 | 0.679 (MK4) |
| top350 | 0.541 | 0.638 (MK4) | 0.582 | 0.659 (MK4) |
| top500 | 0.534 | 0.636 (MK4) | 0.568 | 0.651 (MK4) |
| top750 | 0.537 | 0.633 (MK4) | 0.563 | 0.647 (MK4) |
| top1000 | 0.541 | 0.633 (MK4) | 0.559 | 0.646 (MK4) |

**Table 8.2.** Comparative effectiveness of cluster-based and inverted-file searches using the group average method

In Table 8.2, a view of the data in Table 8.1 focused on the comparative effectiveness of the two searches is presented (using the group average method for $\beta$=0.5 and 2). Data for three test collections are presented in this table (AP, CACM, and WSJ). These are the same collections that were presented in Table 6.7 comparing standard cluster-based effectiveness to IFS effectiveness. The first column of this table displays the number $n$ of documents clustered for each test collection. The second column shows the optimal cluster-based effectiveness as calculated by the MK1 measure for $\beta$=0.5. This column contains values obtained when the M1 measure is used to calculate similarities. In the next column, the effectiveness of the IFS measure that the corresponding cluster-based effectiveness significantly outperforms (as calculated by the Wilcoxon signed-ranks test, for significance level $p$<0.05) is displayed, along with the name of the IFS measure in brackets. It should be reminded that the "ranking" in decreasing order of

importance of the three measures which are used to calculate IFS effectiveness is MK4, MK3 and MK1-k.

For example, when using the CACM collection for $\beta$=0.5, cluster-based effectiveness is significantly higher than the MK3 measure for $n$=100 and 200, and higher than the MK4 measure for $n$=200, 350, 500, 750, 1000. Columns four and five display similar information for recall-oriented searches (i.e. $\beta$=2).

It should also be noted that the results for query-sensitive effectiveness in Table 8.2 have been calculated using the M1 measure. Comparatively to IFS effectiveness, measures M1 and M3 displayed comparable effectiveness in the majority of cases (Tables D1-D4). As a consequence, by presenting the results based on M1 here there is no major distortion of the experimental results. The comparative effectiveness of the three query-sensitive measures is presented in section 8.4.

By comparing the levels at which cluster-based effectiveness is significantly higher than IFS effectiveness between Table 8.2 (query-sensitive effectiveness) and Table 6.7 (standard effectiveness), one can note that when QSSM are used (Table 8.2) the levels improve. For example, when the M1 measure is applied to WSJ, the resulting effectiveness is significantly higher than IFS effectiveness at the MK4 level for all values of $n$ and both types of searches. In Table 6.7 recall-oriented searches managed to significantly outperform IFS effectiveness only at the MK3 (for $n$=100 and 200), and MK1-k levels (for the rest values of $n$).

This observation is not only valid for the data presented in Table 8.2, but for the majority of the experimental conditions. As I mentioned in the previous section, in general, the improvements introduced by the QSSM are greater for recall-oriented searches. This has as a consequence for recall-oriented searches to perform more favourably when compared to IFS effectiveness, and indeed to manage to exceed it significantly at the MK4 level (especially for $n$>100). The only cases where cluster-based effectiveness using QSSM fails to exceed IFS effectiveness at the MK4 level for recall-oriented searches, is when using CACM and Medline and the complete link method, CISI and the single link method, and Medline and Ward's method.

Precision-oriented searches are also improved, and this translates into even higher effectiveness compared to that of IFS. This is also displayed in the data of Table 8.2 in the case of the CACM collection. In Table 6.7 it was shown that the standard effectiveness of the group average method significantly outperforms IFS at the MK3 level for $n$=100, and at the MK1-k level for the rest values of $n$. In Table 8.2, however, cluster-based effectiveness (using QSSM) exceeds IFS effectiveness at the MK3 level for $n$=100, 200 and full, and at the MK4 level for all other values of $n$. This behaviour is typical for the other collections and clustering methods.

The case of the single link method should also be emphasised. The results of section 6.3.2 had demonstrated that the effectiveness of this method did not compare favourably to IFS

effectiveness for most experimental conditions. However, when using QSSM (and especially measures M1 and M3), the effectiveness of the single link method significantly increases, and this results into a much better comparative performance to IFS.

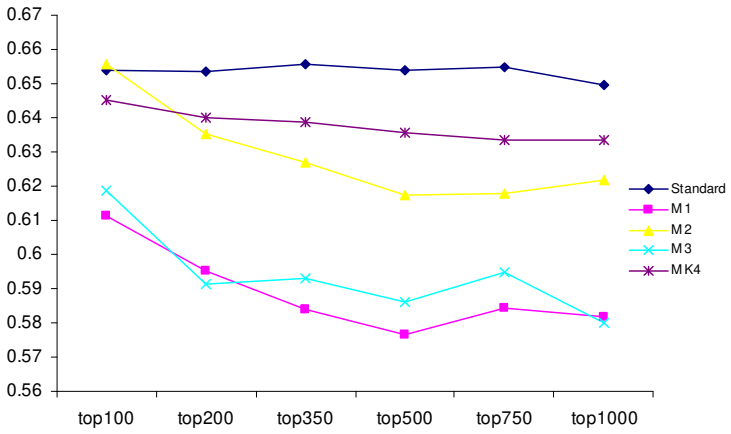


**Figure 8.1.** Precision-oriented effectiveness using the single link method and the WSJ collection

For example, in Figures 8.1 and 8.2 the precision and recall-oriented effectiveness obtained using the single link method and the WSJ collection is displayed. In these two figures the effectiveness (vertical axis) obtained under three different conditions is displayed: standard (using the cosine coefficient), query-sensitive (using M1, M2 and M3) and IFS (calculated by measure MK4). Effectiveness values are plotted against the number $n$ of top-ranked documents (horizontal axis).

In both these figures, the effectiveness using the cosine coefficient is lower (and for $n \geq 350$ significantly lower) than IFS effectiveness at the MK4 level. When using QSSM however, cluster-based effectiveness becomes significantly higher than IFS effectiveness. For recall-oriented searches (Figure 8.2) all three QSSM consistently outperform IFS (and also significantly for $n>100$). For precision-oriented searches (Figure 8.1), measures M1 and M3 result in significantly more effective retrieval than IFS for all values of $n$. Clustering based on M2 is more effective than IFS for $n>100$, and significantly more effective for $n=500$ and 750.

The only test collection for which cluster-based effectiveness still does not compare favourably to IFS effectiveness is LISA. Despite that the use of all three QSSM introduces significant improvements over standard effectiveness for this collection (section 8.2.1), this is not enough for cluster-based effectiveness to consistently exceed IFS effectiveness at the MK4 level. More specifically, no precision-oriented search using the group average, complete link and Ward's methods, and no search of either type using the single link method, manage to exceed IFS at the MK4 level.

However, in all other cases using LISA, query-sensitive measures improve the levels at which cluster-based effectiveness exceeds IFS effectiveness. For example, when using the complete link method to perform recall-oriented searches in this collection, standard cluster-based effectiveness is significantly worse than IFS effectiveness at the MK4 level for most values of *n* (Table B3). When using query-sensitive measures, measures M1 and M3 consistently outperform IFS at the MK4 level (and significantly for many cases).

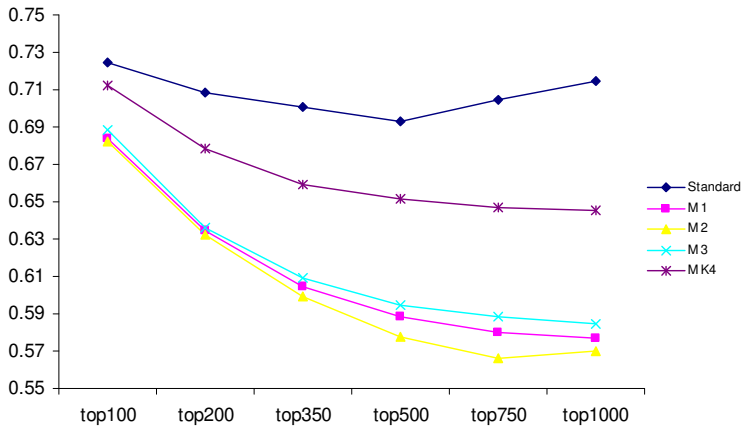


**Figure 8.2.** Recall-oriented effectiveness using the single link method and the WSJ collection

The results presented in this section further strengthen the findings of the previous section in relation to the effectiveness of hierarchic clustering using QSSM. Not only does the use of QSSM improve effectiveness compared to standard clustering, it also improves the effectiveness of cluster-based retrieval comparatively to IFS. As I demonstrated, it does so in a consistent and significant way, so that in the majority of the cases cluster-based effectiveness exceeds IFS effectiveness at the MK4 level.

### 8.2.3 Comparatively to random cluster-based effectiveness

The last issue that I examine in section 8.2 is the comparative effectiveness of hierarchies constructed by using similarity matrices generated by query-sensitive measures, and by using similarity matrices generated by random means.

In Chapter 6 (section 6.3.3), I examined the comparative effectiveness of standard and random document hierarchies for post-retrieval and static (i.e. *n*=full when using a static similarity measure) clustering. The main conclusion of that section was that post-retrieval clustering effectiveness is significantly higher than random effectiveness, with the exception of the CISI and LISA databases. More specifically, when using CISI all four clustering methods for recall-oriented searches and small values of *n* displayed close-to-random effectiveness, and in the specific case of the single link method, random effectiveness was slightly higher than actual

effectiveness for $n$=100. When using LISA, it was only the single link method for recall-oriented searches that displayed effectiveness values not significantly higher than that of random structures.

The comparative investigation of query-sensitive and random effectiveness is limited to these two cases. In all other cases, query-sensitive cluster-based effectiveness is significantly higher than random effectiveness (since query-sensitive effectiveness is in general higher than standard effectiveness). Even in cases where the use of QSSM results in lower effectiveness than that of standard clustering (e.g. the use of measure M2 with the AP collection), it is still significantly higher than that of random clustering.

The data in Table 8.1 and Tables D2-D4 demonstrate that, when using the LISA collection, all three QSSM result in effectiveness that is significantly higher than standard effectiveness. Query-sensitive effectiveness is much higher than standard effectiveness when using this database, especially for recall-oriented searches of all clustering methods (section 8.2.1). This has as a consequence that query-sensitive cluster-based effectiveness is significantly higher than random effectiveness (Tables B8-B11) in all cases using this database, including the "problematic", under standard clustering, case of the single link method for recall-oriented searches.

When using the CISI collection and the group average, complete link and Ward's methods, in general, the use of M1 and M3 improves cluster-based effectiveness (for the case of recall-oriented searches for $n$=100, 200) comparatively to random effectiveness, but not always in a statistically significant manner. The use of M2 with these three clustering methods does not significantly differ from standard and random effectiveness for these cases, and therefore does not change the comparative effectiveness of actual and random clustering.

| *n* | *Standard* | *Random* | *M1* | *M2* | *M3* |
|---|---|---|---|---|---|
| 100 | 0.733 | 0.723 | 0.725 | 0.719 | 0.704 |
| 200 | 0.669 | 0.692 | 0.676 | 0.691 | 0.679 |
| 350 | 0.666 | 0.700 | 0.663 | 0.676 | 0.659 |
| 500 | 0.677 | 0.720 | 0.656 | 0.668 | 0.652 |
| 750 | 0.685 | 0.753 | 0.646 | 0.662 | 0.646 |
| full | 0.825 | 0.831 | 0.824 | 0.824 | 0.823 |

**Table 8.3.** Random vs. actual effectiveness for the CISI collection using the single link method for $\beta$=2

When using the single link method, recall-oriented effectiveness for small values of $n$ remains problematic. Table 8.3 displays the effectiveness for this method using the cosine coefficient (standard), random similarities and the three query-sensitive measures. By observing the results for small values of $n$, one can note that for $n$=100 and 200 all actual cluster-based effectiveness values (i.e. standard, M1, M2 and M3) are close to random values. Any significant differences in favour of actual clustering occur for $n \geq 350$.

Therefore, as a general comment for the behaviour of CISI with any of the clustering methods used, for small numbers of top-ranked documents (i.e. 100 and 200) and recall-oriented searches, it is valid to conclude that cluster-based effectiveness does not significantly differ from that of random clustering. The only indication of superior performance of query-sensitive clustering is that it consistently outperforms random effectiveness for all values of *n* (e.g. in Table 8.3, M2 and M3 are always more effective than random clustering).

In section 8.2.1 I mentioned that when an entire document collection (i.e. *n*=full) is considered for clustering with any of the query-sensitive measures, there are no significant differences in effectiveness compared to standard clustering (Tables D1-D4). Consequently, the behaviour of query-sensitive actual clustering comparatively to random clustering does not change, and therefore any conclusions that were drawn in section 6.3.3 also apply here. In brief, the results in that section had suggested that when clustering an entire collection, the resulting effectiveness is consistently higher than random effectiveness, but not significantly so.

### 8.2.4 Discussion

The results that I presented in this section demonstrate that the use of query-sensitive measures in hierarchic document clustering significantly improves the effectiveness of hierarchic clustering that uses static similarity measures. The improvements are in general large, consistent and significant. All clustering methods benefit from the use of query-sensitive measures, albeit each to a different extent. Providing further evidence for the utility of query-sensitive measures, is that their use significantly improves optimal cluster-based effectiveness when using the two topically heterogeneous TREC databases (AP and WSJ). The only significant exception is noted when using measure M2 with the AP collection. The effectiveness of QSSM using the TREC databases is emphasised, as topically diverse datasets are more likely to be used in actual, operational environments.

Query-sensitive measures further improve cluster-based effectiveness comparatively to inverted file search effectiveness. The results presented in section 8.2.2 demonstrate that query-sensitive hierarchic clustering significantly outperforms IFS effectiveness at the MK4 level for the majority of cases. In this way, it significantly improves the comparative effectiveness of cluster-based to IFS effectiveness. Moreover, the use of query-sensitive measures improves the comparative effectiveness of actual to random clustering, by removing a number of cases where static similarity measures resulted in close-to-random effectiveness.

The results presented in this section further strengthen the findings of Chapter 7 regarding the effectiveness of query-sensitive measures in IR. By measuring the similarity between documents in a way that takes the query into account, document hierarchies are more influenced by the information contained in the query, and relevant documents are grouped more effectively than

using a static similarity measure. These results also suggest that the use of static similarity measures in document clustering has been a limiting factor for cluster-based effectiveness.

## 8.3 Hierarchy characteristics

In this section I examine some characteristics of hierarchies that are generated by using query-sensitive similarity measures, and I compare them to characteristics of hierarchies generated by the cosine coefficient. I first examine the variation of the average size of clusters when using different similarity measures in section 8.3.1. I then focus on the characteristics of optimal clusters (section 8.3.2) in terms of their size and number of relevant documents they contain, and also in terms of the levels in the document hierarchies in which they occur.

### 8.3.1 Size of clusters

In section 5.5.3 (Table 5.3) I had presented the average size of clusters generated by the four clustering methods for the AP and WSJ collections. These statistics were calculated when the hierarchies were generated by the cosine coefficient as a measure of interdocument similarity.

It was then demonstrated that the complete link and Ward's methods produce clusters of a small size that does not significantly increase as the number *n* of top-ranked documents increases. This is a consequence of the way that these methods operate, employing a stringent clustering criterion that leads to the formation of small, tightly bound clusters (Milligan *et al*., 1983; Murtagh, 1984b; Voorhees, 1985a). Moreover, the behaviour of these methods is consistent across all test collections.

The single link method, on the other hand, tends to produce clusters whose average size increases consistently as the number *n* of top-ranked documents increases. The clustering criterion employed by the single link method is not as stringent as that employed by the other two methods, leading in large clusters that are characterised by the chaining effect (section 3.4.1) (Jardine & Sibson, 1971).

The group average method lies somewhere in between the cases defined by the extremes of the single link method on one hand, and the complete link and Ward's methods on the other hand. The size of clusters generated by this method increase as the number *n* of top-ranked documents increases, but do not do so in a significant manner.

The output of these four clustering methods depends on the composition of the similarity matrix, since the clustering criterion which they employ performs a transformation on the matrix. Keeping all other parameters that may affect similarity calculations constant (e.g. indexing exhaustivity), the characteristics of document hierarchies will be determined by the properties of

the similarity matrix generated by the measure in use. In this chapter, the three query-sensitive measures have been used to calculate interdocument associations. Whereas measures M1 and M3 result in structures of comparable characteristics, measure M2 generates structures that in some cases differ from those generated by the other two measures.

| | *Group Average* | | *Ward* | | *Complete Link* | | *Single Link* | |
|---|---|---|---|---|---|---|---|---|
| *n* | *Cosine* | *M2* | *Cosine* | *M2* | *Cosine* | *M2* | *Cosine* | *M2* |
| 100 | 12.6 | 16.3 | 8.8 | 11.4 | 8 | 11.9 | 28.4 | 45.6 |
| 200 | 16.7 | 25.9 | 10.4 | 17.2 | 9.4 | 17.8 | 54.1 | 91.4 |
| 350 | 21 | 38.6 | 11.9 | 23.7 | 10.6 | 24.9 | 91.1 | 163.2 |
| 500 | 24.3 | 50.9 | 12.7 | 30.6 | 11.6 | 31.8 | 129.7 | 234.7 |
| 750 | 28.6 | 69.9 | 13.6 | 41.2 | 13 | 42.6 | 196.7 | 353.6 |
| 1000 | 31.8 | 90.6 | 14.5 | 52 | 14.5 | 53.3 | 263.4 | 474.9 |

**Table 8.4.** Average size of clusters generated using the cosine coefficient and measure M2 for the WSJ collection

More specifically, the average size of clusters generated using the M2 measure, in some test collections, is significantly larger compared to that of clusters generated using the cosine coefficient. Moreover, in these cases for all clustering methods using measure M2, the average size of clusters increases as *n* (number of documents) increases. This happens even for Ward's and the complete link methods, which as mentioned previously are typically characterised by small variations in size across different numbers of documents. Table 8.4 displays this for the WSJ collection, where the average size of clusters generated using the cosine coefficient and measure M2 are presented for all four clustering methods.

| | | *Ward* | | *Complete Link* | | |
|---|---|---|---|---|---|---|
| *n* | *Cosine* | *M1* | *M3* | *Cosine* | *M1* | *M3* |
| 100 | 8.8 | 9 | 8.2 | 8 | 8.2 | 8 |
| 200 | 10.4 | 10.8 | 9.7 | 9.4 | 9.6 | 9.5 |
| 350 | 11.9 | 12.3 | 10.9 | 10.6 | 10.8 | 10.6 |
| 500 | 12.7 | 13.2 | 11.7 | 11.6 | 11.4 | 11.4 |
| 750 | 13.6 | 14.3 | 12.8 | 13 | 12.3 | 12.5 |
| 1000 | 14.5 | 15.2 | 13.5 | 14.5 | 12.8 | 13.4 |

**Table 8.5.** Average size of clusters generated using the cosine coefficient and measures M1 and M3 for the WSJ collection (Ward and complete link methods)

This increase in size does not occur when using measure M1 or M3. In these cases, the clustering methods display their typical behaviour, approximating the characteristics of the hierarchies generated using the cosine coefficient. This is demonstrated in Table 8.5, where the average size of Ward and complete link hierarchies are presented using the WSJ collection and measures M1 and M3. Data using the cosine are also presented for comparison. The data in the table clearly present that when using these two measures, the average cluster size remains similar to when using the cosine coefficient.

An explanation for this behaviour of hierarchies generated using M2 can be given in terms of the definition of M2 itself. More specifically, when measure M2 is used to calculate interdocument similarities, only common query terms between documents contribute towards the calculation of similarity values (section 7.2.1). Especially in cases where queries are short, the evidence used to calculate similarity values is limited, and as a consequence, the range of the generated similarity values will also be limited.

Another way to view this behaviour is in terms of the reduction of the dimensions of the space based on which the similarity of two documents is judged. M2 collapses the space in which documents are represented in such a way that only the dimensions corresponding to query terms are used to represent (and therefore discriminate between) documents. When fewer dimensions are used, poorer discrimination between documents occurs, a direct consequence of the reduction of the dimensionality of the document space. The range of the similarity values of interdocument associations will be limited because only a limited number of terms (document and query terms) will determine the outcome of the calculations. The lack of any other evidence (i.e. other common terms between documents) also contributes to this.

This in turn, forces the clustering methods to increase the cluster size so as to accommodate the poorer separation between documents. As far as these method are concerned, more documents seem to be (almost) equally similar to each other. As a consequence, the stage of the clustering process where pairs of documents (or clusters) are joined will be affected. In order to reflect the limited range of similarity values, documents will be more likely to join other clusters of documents, creating in this way larger bottom level clusters (the cluster that a document joins when it enters the hierarchy for a first time).

This effect will become more pronounced as more documents are clustered. The range in the similarity values will not change, since the dimensions based on which similarity is calculated remain constant, but the number of documents that are likely to be poorly discriminated increases (because $n$ increases). This explains the atypical increasing cluster size for increasing values of $n$ for clustering methods such as Ward's and the complete link.

The large difference in the composition of the bottom level clusters in the hierarchies can be seen in Table 8.6. Hierarchies of the WSJ collection using the complete link method are used in this example, generated by using the cosine coefficient and measure M2. The first column of this table contains the number $n$ of documents clustered, and the second column the average size of the bottom level clusters. Columns 3-8 display the number of bottom level clusters which contain a number of documents that falls within the range of its corresponding heading (e.g. using the cosine coefficient and $n$=100, there are 9.7 bottom level clusters on average whose size is between 4 and 10 documents).

| *Cosine* | | | | | | | |
|---|---|---|---|---|---|---|---|
| *n* | *Avg. size* | *1 - 3* | *4 - 10* | *11 - 20* | *21 - 30* | *31 - 40* | *> 40* |
| 100 | 2.8 | 53.5 | 9.7 | 0.3 | 0.1 | 0 | 0 |
| 200 | 2.8 | 106.4 | 18.5 | 0.7 | 0.1 | 0.1 | 0 |
| 350 | 2.8 | 186.1 | 32.3 | 1.5 | 0.2 | 0.1 | 0 |
| 500 | 2.8 | 266.6 | 45.8 | 1.9 | 0.3 | 0 | 0.1 |
| 750 | 2.8 | 399.9 | 71.0 | 2.6 | 0.4 | 0.1 | 0.3 |
| 1000 | 2.8 | 534.5 | 93.2 | 3.8 | 0.3 | 0.2 | 0.4 |
| *M2* | | | | | | | |
| *n* | *Avg. size* | *1 - 3* | *4 - 10* | *11 - 20* | *21 - 30* | *31 - 40* | *> 40* |
| 100 | 8.7 | 26.7 | 34.8 | 14.5 | 4.7 | 2.5 | 1.2 |
| 200 | 14.2 | 41.1 | 59.5 | 35.4 | 16.7 | 8.5 | 14.7 |
| 350 | 20.9 | 60.7 | 90.2 | 59.4 | 34.6 | 19.4 | 50.1 |
| 500 | 27.8 | 75.1 | 115.3 | 80.3 | 49 | 33.3 | 103.1 |
| 750 | 38.6 | 95.5 | 149.6 | 111.9 | 71.1 | 51.6 | 214.4 |
| 1000 | 49.3 | 115.3 | 178.5 | 136.8 | 91.6 | 64.7 | 345 |

**Table 8.6.** Composition of bottom-level clusters for the complete link method using the WSJ collection

The difference in the data of Table 8.6 is quite remarkable. The typical behaviour of the complete link method is to create a large number of small bottom level clusters, which are mainly formed either by pairs of documents or by a document that joins a single cluster consisting of a pair of documents (Voorhees, 1985a). However, when using measure M2 the number of large bottom level clusters increases significantly compared to using the cosine coefficient, and as a result the average size of the bottom level clusters increases accordingly. This is a direct consequence of the effect that was mentioned previously, and is also noted for all other clustering methods. M1 and M3, on the other hand, generate hierarchies which are much like those of the cosine coefficient in this respect.

| | *Group Average* | | *Ward* | | *Complete Link* | | *Single Link* | |
|---|---|---|---|---|---|---|---|---|
| *n* | *Cosine* | *M2* | *Cosine* | *M2* | *Cosine* | *M2* | *Cosine* | *M2* |
| 100 | 10.7 | 10.4 | 8.5 | 8.3 | 8.3 | 9.3 | 32.7 | 40 |
| 200 | 13.5 | 13.2 | 10.1 | 10 | 11.6 | 11.8 | 63.8 | 81.6 |
| 350 | 16.3 | 16.8 | 11.4 | 11.7 | 13.1 | 13.9 | 112.2 | 147.5 |
| 500 | 19.4 | 20.3 | 12.1 | 13.6 | 14.2 | 15.5 | 160 | 216.5 |
| 750 | 22.2 | 26.4 | 13.2 | 16.3 | 15.4 | 17.9 | 246 | 339.7 |
| 1000 | 25.3 | 33.1 | 14.2 | 19.3 | 16 | 20.7 | 328.9 | 465.8 |

**Table 8.7.** Average size of clusters generated using the cosine coefficient and measure M2 for the LISA collection

The larger size of clusters generated when using measure M2 in the previous example was demonstrated by using the WSJ collection, which is characterised by short queries. I mentioned that the increase in average cluster size is attributed to the limited range of similarities that is a consequence of the few terms upon which M2 is based. It would therefore seem reasonable to expect a more "typical" behaviour of the clustering methods (in terms of the size of the clusters that they produce), when using test collections with larger query length (e.g. CACM, LISA).

This is verified by the data presented in Table 8.7 using the LISA collection, which has, on average, almost 20 terms per query. The difference in size between clusters generated using the cosine coefficient and using measure M2 is much smaller than that in Table 8.4. In fact, for the group average, Ward and complete link methods, cluster size is not significantly larger when using M2. The single link method is more sensitive to such effects, and there is an increase in the size of the clusters, however, the increase is much smaller than that displayed in Table 8.4.

## 8.3.2 Optimal cluster characteristics

The study of the characteristics of optimal clusters is divided into two parts. First, in section 8.3.2.1 I examine the composition of optimal clusters in terms of documents and relevant documents contained within optimal clusters. The second part is reported in section 8.3.2.2, where I examine the hierarchy levels at which optimal clusters occur.

### 8.3.2.1 Average size and numbers of relevant documents

In section 6.3.4 I had presented details about characteristics of optimal clusters using the cosine coefficient as a measure of similarity. The variation of optimal cluster size as a function of test collection characteristics, and especially as a function of the average number of relevant documents per query, was demonstrated in that section. Also, the effect of the type of search performed was also emphasised, i.e. optimal clusters for recall-oriented searches tend to be of much larger size.

In this section I examine the composition of optimal clusters of hierarchies generated using the query-sensitive measures. I will not focus on the same issues as in section 6.3.4 (i.e. parameters that affect optimal cluster size), as such issues do not depend on the use of different types of measures (i.e. the findings of 6.3.4 apply to the case of optimal clusters generated using query-sensitive measures). Instead, I will focus on the composition of optimal clusters in terms of the number of relevant documents they contain, and in terms of their size. The aim is to investigate whether optimal clusters generated by query-sensitive measures result in more “useful” clusters, in terms of their composition, compared to those generated by the cosine coefficient.

In section 8.2.1 I demonstrated that query-sensitive cluster-based effectiveness is, in general, significantly higher than standard effectiveness. An interpretation of this, is that the hierarchies generated using query-sensitive measures contain an optimal cluster which yields a higher effectiveness (E) value than hierarchies generated using the cosine coefficient. This behaviour, as it was explained in section 8.2.1, is consistent across clustering methods and test collections, and although variations are noted, in general query-sensitive effectiveness is higher than standard effectiveness.

| | β=2 | | | | β=0.5 | | | |
|---|---|---|---|---|---|---|---|---|
| *n* | *Cosine* | *M1* | *M2* | *M3* | *Cosine* | *M1* | *M2* | *M3* |
| 100 | 95.47 | 98.44 | 96.03 | 96.21 | 68.78 | 71.56 | 72.49 | 71.93 |
| 200 | 92.10 | 93.35 | 93.10 | 94.73 | 56.09 | 61.10 | 59.35 | 62.74 |
| 350 | 90.81 | 91.88 | 94.12 | 91.01 | 50.46 | 48.81 | 54.45 | 52.79 |
| 500 | 88.33 | 89.93 | 90.77 | 88.08 | 46.61 | 44.84 | 38.02 | 48.46 |
| 750 | 84.25 | 87.55 | 87.47 | 88.13 | 42.12 | 38.90 | 35.97 | 42.86 |
| 1000 | 79.84 | 86.73 | 79.77 | 84.97 | 34.58 | 37.49 | 35.39 | 41.95 |

**Table 8.8.** Percentage of relevant and retrieved documents contained in an optimal cluster using the group average method and the AP collection

By examining the characteristics of optimal clusters generated by each of the similarity measures, the findings of section 8.2.1 are further strengthened. Not only do query-sensitive measures result in higher effectiveness, but they also tend to generate more compact optimal clusters, that contain a higher proportion of relevant documents.

In Table 8.8 I present the percentage of relevant and retrieved documents that are contained within optimal clusters that are generated using each of the similarity measures (cosine, M1, M2 and M3). Results are presented for recall and precision-oriented searches, using the group average method and the AP collection. It should be noted that the use of measure M2 in this case results in lower effectiveness than the use of the cosine coefficient (Table 8.1). The other two measures improve standard effectiveness with the exception of $n$=100 for precision-oriented searches, where there is an insignificant decrease in effectiveness.

By observing the data in Table 8.8 it follows that, in general, optimal clusters generated using query-sensitive measures contain higher percentages of all the relevant documents "available" at each value of $n$. It should however be noted that a higher percentage of relevant and retrieved documents contained within an optimal cluster does not necessarily imply higher effectiveness. For example, using measure M2 in Table 8.8, the percentage of relevant and retrieved documents contained within an optimal cluster is higher than using the cosine, but the effectiveness achieved by measure M2 in this case is much lower than standard effectiveness. This can be explained on the basis of the average size of the optimal clusters generated.

In Figure 8.3, the average size (in documents, vertical axis) of optimal clusters is plotted against the number $n$ of top-ranked documents (horizontal axis) for recall-oriented searches of the AP collection using the group average method. The average size of optimal clusters using M2 is much larger than that using the other measures, and this can explain the poor effectiveness of the measure in this case. As I mentioned in section 8.3.1, average cluster size using this measure tends to be much larger than that using the other three measures, and this characteristic can have a negative effect on the effectiveness attainable with this measure.

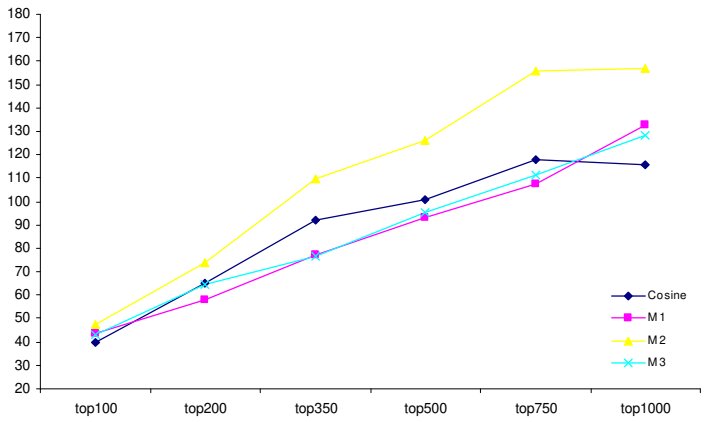


**Figure 8.3.** Average size of optimal clusters using the group average method and the AP collection for recall-oriented searches

Measures M1 and M3, on the other hand, tend to produce optimal clusters of much smaller sizes than M2, and in most cases of smaller sizes than using the cosine coefficient. This is noted in the majority of experimental conditions, and is in agreement with the higher effectiveness that these two measures display compared to standard effectiveness. This characteristic of M1 and M3, in combination with their higher effectiveness and their better composition in terms of relevant documents than the cosine, further strengthens the utility of such measures in hierarchic document clustering.

Regarding the comparison of characteristics of optimal clusters to optimal sets of a best-match search, query-sensitive measures generate optimal sets that display better characteristics. Query-sensitive effectiveness is, in general, higher than IFS effectiveness at the MK4 level (section 8.2.2), and this translates into better characteristics of the optimal sets returned by clustering which uses query-sensitive measures.

| | *MK4* | | *M1* | | *M2* | | *M3* | |
|---|---|---|---|---|---|---|---|---|
| *n* | *#docs* | *#rel* | *#docs* | *#rel* | *#docs* | *#rel* | *#docs* | *#rel* |
| 100 | 24.75 | 11.21 | 22.56 | 12.35 | 23.71 | 11.92 | 22.81 | 12.08 |
| 200 | 31.23 | 13.40 | 35.13 | 16.58 | 34.10 | 15.71 | 31.52 | 15.69 |
| 350 | 35.46 | 14.31 | 43.79 | 19.40 | 39.35 | 17.08 | 38.13 | 18.00 |
| 500 | 36.25 | 14.44 | 33.00 | 18.13 | 29.81 | 15.10 | 33.90 | 18.13 |
| 750 | 40.38 | 14.83 | 38.00 | 20.31 | 34.40 | 17.88 | 36.71 | 18.58 |
| 1000 | 38.54 | 14.71 | 35.67 | 19.79 | 39.00 | 19.08 | 33.38 | 18.69 |

**Table 8.9.** Average size and average number of relevant documents in optimal sets, using the WSJ collection for precision-oriented searches

In Table 8.9, the average number of documents and relevant documents contained within optimal sets returned by an IFS (MK4) and by cluster-based searches using each of the QSSM (M1, M2 and M3) are presented. The clustering method used in this case is the group average method. Results are generated using the WSJ collection and precision-oriented searches. M1 and M3 in

this case tend to generate optimal clusters that are smaller and that contain more relevant documents than the sets returned by an optimal IFS search. M2 in some cases tends to generate larger clusters due to its tendency to generate large clusters using this dataset (section 8.3.1). The difference in the characteristics of the optimal sets is more pronounced (in favour of clustering) for recall-oriented searches. The results presented here are representative of the other experimental conditions.

#### 8.3.2.2 Hierarchy levels

An additional source of experimental evidence which confirms the potentially higher utility of optimal clusters generated by query-sensitive measures, comes from examining the levels within the document hierarchies in which such clusters occur. Before examining this issue, it is worth restating the notion of levels within document hierarchies. I also discussed this issue in section 3.4.

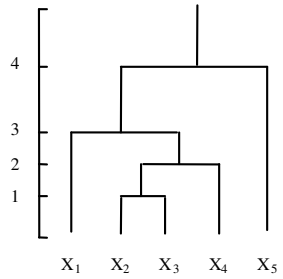


**Figure 8.4.** An example document hierarchy

In Figure 8.4 an example of a document hierarchy is presented. In this example only the levels at which pairs of documents (or clusters) merge are shown in the vertical axis, and not the actual similarity values at which the merges occur. For a hierarchy comprising *n* documents there will be *n*-1 levels in the hierarchy. The actual number of documents contained within clusters at each level is a characteristic of a particular clustering method. For example, the complete link method produces a large number of small clusters, and therefore at each level of the hierarchy few documents (usually a pair) typically merge. The single link method displays an opposite behaviour.

If two hierarchies are of comparable types (e.g. are generated by the same clustering method by varying the similarity measure used), then it can be argued that the lower the level at which optimal clusters are formed, the more useful these clusters may be. Lower levels in the hierarchy are usually associated with more cohesive clusters (since lower levels also represent higher similarities at which the merge occurs). If the hierarchies are different to each other in terms of their characteristics, then such a comparison does not allow the extraction of useful conclusions.

In section 8.3.1 I demonstrated that the use of measures M1 and M3 with each of the four clustering methods used, results in hierarchies which display characteristics similar to the hierarchies generated by the cosine coefficient. The use of measure M2 on the other hand, results in hierarchies whose characteristics differ widely to those generated by the cosine coefficient (and by measures M1 and M3), especially for test collections for which query length is short. Therefore, a comparison of the levels at which optimal clusters are formed between hierarchies generated using the cosine coefficient, and measures M1 and M3, can lead to useful conclusions regarding the utility of the hierarchies.

Such a comparison is presented in Table 8.10, where the average levels at which optimal clusters are formed for the WSJ collection are presented. The results have been compiled using the group average, complete link and single link methods. The cosine coefficient, and measures M1, M2 and M3 have also been used. Optimal clusters in this table are based on precision-oriented searches. The results for Ward's method are similar to those for the complete link method.

| | *Group Average* | | | | *Complete Link* | | | | *Single Link* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *n* | *Cosine* | *M1* | *M2* | *M3* | *Cosine* | *M1* | *M2* | *M3* | *Cosine* | *M1* | *M2* | *M3* |
| 100 | 68 | 48 | 43 | 55 | 72 | 58 | 48 | 63 | 54 | 37 | 36 | 43 |
| 200 | 138 | 88 | 70 | 99 | 143 | 116 | 80 | 116 | 103 | 58 | 53 | 62 |
| 350 | 250 | 150 | 104 | 180 | 242 | 192 | 116 | 186 | 184 | 82 | 78 | 100 |
| 500 | 339 | 190 | 112 | 237 | 338 | 260 | 145 | 259 | 239 | 87 | 94 | 121 |
| 750 | 476 | 268 | 140 | 321 | 483 | 340 | 204 | 352 | 335 | 105 | 101 | 146 |
| 1000 | 611 | 339 | 158 | 387 | 647 | 438 | 233 | 465 | 410 | 130 | 123 | 181 |

**Table 8.10.** Hierarchy levels at which optimal clusters are formed. Using the WSJ collection and precision-oriented searches

For each of the three clustering methods presented in this table, a simple comparison of the levels at which optimal clusters are formed reveals that when using query-sensitive measures, optimal clusters tend to form at significantly lower levels in the hierarchy. The data displayed here for the WSJ collection are also typical of the other test collections. As it was explained previously, comparisons between the data for hierarchies generated using the cosine coefficient on one hand, and measures M1 and M3 on the other, are appropriate since such hierarchies have similar characteristics.

The outcome of this comparison can be interpreted as indicating that when using measures M1 and M3, relevant documents tend to merge into clusters earlier than when using the cosine coefficient. In fact, in most cases more relevant documents tend to merge into clusters of smaller (or approximately equal) size at an earlier stage using these two measures than when using the cosine coefficient. That optimal clusters generated by query-sensitive measures typically contain more relevant documents, and are of the same or smaller size than those generated by using the cosine coefficient, was demonstrated in section 8.3.2.

Measure M2, for reasons that were illustrated in section 8.3.1, tends to from large bottom-level clusters in a number of cases. Optimal clusters using this measure tend to occur lower in the hierarchies than when using any other measure, but this comes at the cost of a chaining effect that can also lead to poor effectiveness (e.g. when using the AP collection).

### 8.3.3 Discussion

In this section I examined some characteristics of hierarchies generated by the query-sensitive measures, and I also compared such characteristics to those of hierarchies generated using the cosine coefficient. It was demonstrated that hierarchies generated using measures M1 and M3 tend to have characteristics that are highly similar to those of hierarchies generated by the cosine coefficient. M2, on the other hand, tends to differ significantly, especially when using collections with short queries.

More specifically, it was demonstrated that the average size of clusters generated by measure M2 increases significantly compared to the size of clusters generated by the other similarity measures. This increase is mainly attributed to the increase in the size of the bottom level clusters of the hierarchies. Croft (1978), Voorhees (1985a), Griffiths et al. (1986) and El-Hamdouchi (1987), among others, have suggested that it is beneficial for a clustering method to produce a hierarchy which is characterised by small bottom level clusters. This is because if a bottom-up search strategy (section 4.3.2) is used to search the hierarchy, it is more effective to produce cluster representatives for small rather than for large bottom-level clusters. The hierarchy can then be searched by using an inverted file structure of the bottom level clusters (section 4.3.3) (Croft, 1978, 1980). Therefore, the characteristic of hierarchies produced by measure M2 can be seen as a shortcoming that may cause actual cluster-based searches not to achieve effectiveness close to that of the optimal clusters of the hierarchies.

The other point of investigation of this section involved the characteristics of optimal clusters produced by using different similarity measures. In the majority of cases, the use of query-sensitive measures results in more effective optimal clusters than using the cosine coefficient, and consequently the “query-sensitive” optimal clusters tend to contain a higher proportion of relevant documents than the “standard” ones. Moreover, in section 8.3.2.2 I demonstrated that query-sensitive optimal clusters tend to occur in lower levels of the hierarchies.

The combination of these three results (higher effectiveness, higher proportion of relevant documents and lower levels in the hierarchy), suggests that query-sensitive measures tend to produce optimal clusters that may be of higher utility in an interactive environment. The characteristics of such optimal clusters may prove beneficial to users in a browsing task for example, where the concentration of more relevant documents in a smaller part of the hierarchy may lead users to useful information easier and quicker. Moreover, the characteristic of optimal

clusters to form at lower levels may also have consequences on efficiency aspects of clustering. By exploiting this characteristic, it is possible to disregard clusters that join at higher levels of the hierarchy, or even, to stop the clustering process at an appropriately low level.

The interpretation of the findings of this chapter for interactive cluster-based retrieval can be highly subjective. Since I do not tackle such issues experimentally, the validity of the claims that I make in this section are not proven in this thesis. It should however be an interesting subject for further work to investigate what criteria users would attribute greater importance to in an interactive cluster-based environment, and what compromises regarding these criteria users would be willing to make. Such criteria would include the effectiveness of the clusters, the size of the clusters, the proportion of relevant documents they contain, the levels at which useful clusters merge, etc.

## 8.4 Comparison of the query-sensitive measures

The data collected during the experiments reported in this chapter offer the opportunity to compare the effectiveness of the three query-sensitive measures. This study is reported in this section. First I compare the optimal effectiveness of the three measures in section 8.4.1, then I examine the effect of query length on each of the three measures (section 8.4.2), and in section 8.4.3 I examine the effectiveness of the three measures for different numbers of top-ranked documents.

### 8.4.1 Comparative effectiveness of M1, M2 and M3

The results obtained in this chapter suggest that there is a difference in the comparative effectiveness of the three QSSM depending on the clustering method that is used. More specifically, when using the group average and the single link methods the rank order of the three measures based on their effectiveness is, in the majority of cases, M1, M3 and M2. When using the other two methods however, M1 does not perform as well, and measure M3 produces the most effective clusterings in the majority of cases.

It should however be noted that the differences between measures M1 and M3 do not tend to be statistically significant. There are relatively few experimental conditions in which one of the two measures significantly outperforms the other. These tend to occur more when using the group average and single link methods, where measure M1 is more effective than measure M3. More specifically, M1 is significantly more effective than M3 for a greater variety of clustering methods (group average, single link, and once using the complete link method), types of searches and test collections (in total, in 23 experimental conditions). M3, on the other hand, significantly

outperforms M1 only using Ward's method for various values of $n$ when using the CACM, Medline and WSJ collections (in total, in 11 experimental conditions).

Furthermore, when M1 outperforms M3 it does so in a more consistent fashion across all (or most) values of $n$ for a specific test collection and type of search. This is evident, for example, when using the CACM ($\beta$=1, 2, 0.5), Medline ($\beta$=1, 2, 0.5) and WSJ ($\beta$=1, 2) collections with the group average and single link methods. This more consistent behaviour, in combination with the relatively larger number of cases in which M1 significantly outperforms M3, suggest that in the experimental environment used in this thesis, measure M1 results in more effective clusterings than M3.

M2 is, in the majority of experimental conditions, the measure that produces the least effective clusterings. The hierarchies produced by the other two measures are in the majority of cases consistently more effective than the ones produced by M2. They are also significantly more effective for a large number of experimental conditions, and especially for precision-oriented searches and for collections with short queries. The only case where M2 manages to generate significantly more effective hierarchies than any of the other two measures is when using the Medline collection and Ward's method for recall-oriented searches. There are also a number of cases where M2 produces the most effective clustering. Such cases only occur when using either the complete link or Ward's methods.

The comparatively poorer effectiveness of M2 in relation to that of M1 and M3 seems to correlate to the results presented in section 7.4.4 when comparing the effectiveness of the three measures using the 5NN test. The results in that section had demonstrated that, in general, M2 was less effective than the other two methods at placing co-relevant documents close to each other. This was more strongly evident when using the two TREC collections (AP and WSJ), and especially when using AP. This is also the case for the optimal cluster-based effectiveness of this measure.

### 8.4.2 The effect of query length

A number of experiments were also conducted to determine the effect of the query length on the effectiveness of the resulting hierarchies. The experimental procedure is similar to the one reported in section 7.4.5. More specifically, the two TREC collections are used (AP and WSJ) with the standard queries (7.6 terms per query), short queries (3.2 terms per query) and expanded queries (23.4 terms per query). The four clustering methods are applied to each set of queries using each of the three query-sensitive measures. The effectiveness of the resulting hierarchies is gauged using the MK1 measure for optimal cluster-based evaluation.

When expanded queries are used, in general there are consistent but not significant improvements (compared to using the standard queries) using measures M1 and M3 in either collection.

Improvements tend to be slightly larger when using the AP collection. When using measure M2, the improvements are larger in both collections, however they are rarely statistically significant. Precision-oriented searches tend to benefit more. When significant differences occur, they mainly do using either the group average or the single link method.

The larger improvements using the expanded form of the queries with measure M2 can be explained by viewing the added terms as added dimensions (or evidence) based on which to discriminate between the documents. M2 does not take into account any other information when measuring similarities, apart from common query terms between documents. A consequence of the better discrimination between documents is that the average cluster size reduces, and the behaviour of the clustering methods tends to resemble more that using the cosine coefficient (and also M1 and M3) (section 8.3.1). The reduction in the average size of the clusters produced has as a consequence that optimal clusters are more positively affected for precision-oriented than for recall-oriented searches.

| *n* | *Cosine* | *M2* | *M2 expanded* |
|---|---|---|---|
| 100 | 8.8 | 11.4 | 9.2 |
| 200 | 10.4 | 17.2 | 11.7 |
| 350 | 11.9 | 23.7 | 14.1 |
| 500 | 12.7 | 30.6 | 16.6 |
| 750 | 13.6 | 41.2 | 20.7 |
| 1000 | 14.5 | 52 | 24.7 |

**Table 8.11.** Average cluster size for Ward's method, using expanded queries and the WSJ collection

Table 8.11 displays this reduction in size for Ward's method, using the WSJ collection. Measure M2 using the expanded queries generates clusters whose average size is significantly smaller than using the shorter queries. The average size using the expanded queries however, is still larger than the cosine, especially for larger values of *n*.

When the length of the queries is reduced to 3.2 terms on average, the effectiveness of all clustering methods in both collections is negatively influenced. When using the AP collection effectiveness always decreases compared to the original queries, and it does so to a greater extent than when using WSJ. When using this collection (WSJ), in few cases the effectiveness improves slightly when using short queries (Table 8.12). The query-sensitive measure that is more strongly influenced is M2. Another observation is that precision-oriented searches are more affected by the reduction of query length, yielding effectiveness that is typically significantly lower than using the original queries.

The data in Table 8.1 and in Tables D2-D4, showed that when using the WSJ collection and measure M2 (with the original queries), query-sensitive effectiveness is higher than standard effectiveness for all but two experimental conditions (*n*=100, precision-oriented searches using

the group average and single link methods). Also, in most cases it is significantly more effective. When using short queries, the effectiveness of M2, as discussed previously, is negatively affected. However, despite this decrease in effectiveness, it is still higher than standard effectiveness in the majority of experimental conditions (but not significantly so).

| | *Group Average* | | *Ward* | | *Complete Link* | | *Single Link* | |
|---|---|---|---|---|---|---|---|---|
| *n* | *β=2* | *β=0.5* | *β=2* | *β=0.5* | *β=2* | *β=0.5* | *β=2* | *β=0.5* |
| 100 | -0.4 | -1.4 | 0.7 | -1.2 | 0.2 | -1 | 0.1 | 1.3 |
| 200 | -1.1 | -3.2 | 0.4 | -2.4 | -0.6 | -3.4 | -1.3 | -2 |
| 350 | -3.3 | -7.4 | -1.3 | -6.1 | -0.8 | -7.9 | -2.8 | -5.4 |
| 500 | -4.3 | -10.4 | -1.4 | -7.3 | -2 | -8.3 | -5.8 | -7.4 |
| 750 | -4.9 | -13 | -0.4 | -10 | -1 | -11.2 | -7.5 | -7.4 |
| 1000 | -3.9 | -11.4 | -0.7 | -10.8 | -0.3 | -11.6 | -6.6 | -7.1 |

**Table 8.12.** Percentage difference in effectiveness between short and original queries using the WSJ collection and measure M2

In general, the effect of variations in query length on cluster-based effectiveness reported here, is in agreement with the effect of query length on the effectiveness of the three query-sensitive measures reported in Chapter 7 (section 7.4.5). Also, the same limitations that were reported in that section regarding the expanded form of the queries used apply here as well. Further research is warranted to investigate the comparative effectiveness of the measures when algorithmically selected terms are added to the queries.

### 8.4.3 Effectiveness for different numbers of top-ranked documents

In section 6.3.1 I examined the variation of standard cluster-based effectiveness for different numbers of top-ranked documents. The results had demonstrated that cluster-based effectiveness tends to increase as the number of documents increases, but it does not always do so in a significant manner. In the majority of cases, the only significant variations occurred when comparing the effectiveness at $n$=100 to that at other values of $n$. Also, it was demonstrated that the effectiveness of static clustering (i.e. $n$=full when using the cosine coefficient) was always significantly inferior to that attained by considering any number of top-ranked documents. In the present section I examine these issues when query-sensitive measures are used.

The data in Table 8.1 and Tables D2-D4 in Appendix D show that query-sensitive cluster-based effectiveness increases as the number $n$ of top-ranked documents increases. The data also show that the effectiveness for $n$=full is always lower than that at any other value of $n$. A closer look at the results, also reveals that the increase in effectiveness for increasing values of $n$ is, in general, more evident for recall-oriented searches. Also, all three query-sensitive measures display similar patterns in their effectiveness across numbers of top-ranked documents. The effectiveness of

clusters generated by M2 tends to increase in some cases more as a function of *n* than using either M1 or M3.

In the majority of cases, the decrease when comparing the effectiveness at *n*=100 and 200 to that at other values of *n* is statistically significant. Exceptions to this are noted when using the Medline collection. Effectiveness for the Medline collection does not normally vary significantly as a function of the number of documents (especially when using measure M1), and when it does so, it is only the effectiveness at *n*=100 that is significantly lower than that at other values of *n*.

Effectiveness for values of $n \geq 350$ is rarely significantly lower than that at higher values of *n*. The only test collection that consistently displays such a behaviour is the AP collection. This is displayed in Figure 8.5, where effectiveness is plotted against the number of top-ranked documents using measure M3, the AP collection and precision-oriented searches. This behaviour of AP was also evident when examining standard effectiveness (Tables B1-B4), although to a lesser extent.

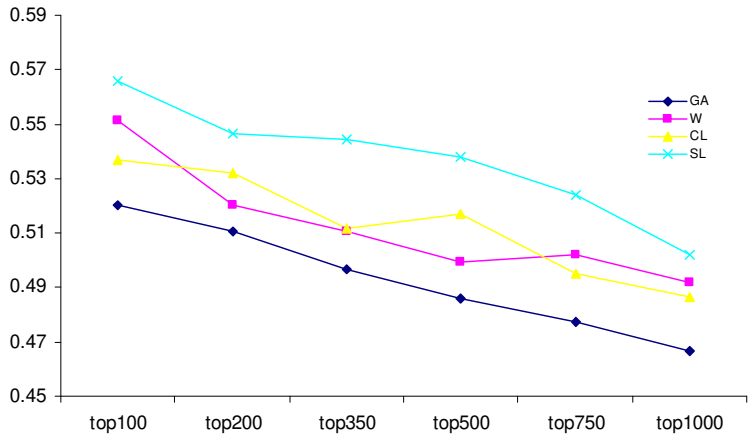


**Figure 8.5.** Effectiveness across numbers of top-ranked documents using the AP collection, M3 and precision-oriented searches

As it was mentioned previously, the effectiveness of recall-oriented searches seems to vary more with the increase in the number of documents than precision-oriented searches. It also tends to vary more compared to the effectiveness of recall-oriented searches using the cosine coefficient. This can be explained by recalling that evaluation of effectiveness is performed by taking into account the total number of relevant documents per query (as opposed to the retrieved and relevant). As the number *n* of documents increases, more relevant documents are available to be clustered. Therefore, optimal clusters at larger values of *n* have a better chance of achieving higher effectiveness since there are more actual relevant documents to be clustered together. This is especially so for recall-oriented searches, where the number of relevant documents contained within optimal clusters is more important. That this behaviour is more evident when using the

query-sensitive measures is attributed to the higher effectiveness achieved with these measures: optimal clusters contain more relevant documents, and this leads to an increase in the effectiveness of recall-oriented searches, especially for larger values of *n*.

The same explanation can be given for the tendency of M2 in some cases to produce optimal clusters whose effectiveness increases more with the increase in the value of *n*. As I discussed in section 8.3.1, the average size of clusters generated by M2 is much larger than that using either M1 or M3, and therefore when the number of documents to be clustered increases, there are more relevant documents available to be placed in the clusters. The increasing cluster size for increasing values of *n*, increases the likelihood of more relevant documents to be contained within optimal clusters generated using the M2 measure. This is especially so given the way that recall is calculated in this experimental environment.

Based on the results presented in this section, if one were to select a single number of documents to be clustered, then a number in the order of 350 documents is likely to prove effective when using query-sensitive measures. The effectiveness past this number does not increase significantly in general. Moreover, as I discussed in section 8.2.1, query-sensitive effectiveness generally becomes significantly higher than standard effectiveness for values of $n \geq 350$. It should however be noted that the choice of a single number of documents to cluster depends on the way that effectiveness is evaluated. The results that I report here have been generated using all relevant documents for a query. If only relevant and retrieved documents are used, then the highest query-sensitive effectiveness is attained for $n=100$, and at this value of *n* differences between query-sensitive and standard effectiveness are also, in general, statistically significant.

### 8.4.4 Discussion

The results presented in this section suggest that, in the experimental environment used in this thesis, measure M1 is the most effective of the three QSSM used in terms of optimal cluster-based effectiveness. Measure M3 displays an effectiveness that is comparable to that of M1 for a large number of experimental condition. Moreover, these two measures seem to be equally affected by the variations in query length that were investigated in section 8.4.2, and in general there seems to be little reason not to prefer M3. However, as I mentioned in section 8.4.1, M1 tends to be significantly more effective than M3 in a larger and more varied number of experimental conditions. Moreover, measure M1 displays significantly and consistently higher effectiveness than M3 when using the group average method, which as I discuss in the next section, is the most effective clustering method.

Measure M2 does comparatively poorer than these other two measures, but not always significantly so. M2 manages to outperform M1 and M3 in a number of cases, mainly when using Ward and the complete link methods (section 8.4.1). Also, as I discussed in section 8.2.1, query-

sensitive effectiveness using M2 manages to significantly outperform standard effectiveness for a large number of experimental conditions. This success of M2 can be seen as surprising if one considers the limited evidence this measure uses to measure the similarity between any two documents.

The effectiveness improvements that M2 introduces compared to standard effectiveness are largely attributed to how this measure treats pairs of documents that do not have any query terms in common. Such documents, in the experimental environment used in this thesis, are not likely to be both relevant to the same query. M2 sets the similarity of such pairs of documents to zero. Consequently, documents that have zero similarity to most others documents in the set will be filtered out from the low levels of the generated hierarchies (i.e. will not merge with other documents at high similarity values). Such documents are more likely to join the hierarchy at a high level (low similarity), separately from documents that have a higher likelihood of being relevant to the query. It should be noted that the same applies to M1, with the difference that M1 uses additional information to judge the similarity between documents.

For the two TREC collections, in which only few dimensions (query terms) are used to discriminate between documents, M2 is highly effective using the WSJ collection (significantly more effective than standard clustering for the majority of cases), and ineffective when using the AP collection (less effective than standard clustering for all cases except recall-oriented searches of the single link hierarchies). One potential reason for the different behaviour of M2 with these two collections may be the relative effectiveness of query terms to discriminate between relevant and non-relevant documents in each collection. All experimental evidence collected in this chapter, and in Chapter 7 for the effectiveness of M2 when using the two TREC collections at the 5NN and 1NN tests, demonstrate that M2 displays a more effective behaviour when using the WSJ collection. The two TREC collections are highly comparable in their characteristics (Table 5.1), and in the way relevance judgements have been constructed (Harman, 1993). The different behaviour of M2 can then be attributed to the different discriminating power of query terms for these two collections.

Measure M2 also tends to produce much larger clusters than the other measures in experimental conditions where there are limited dimensions on which to discriminate between documents (e.g. the two TREC collections). Moreover, it is more affected by variations in query length as it was demonstrated in section 8.4.2. By combining all the experimental evidence for M2, it is justifiable to conclude that the use of this measure seems appropriate in cases where longer than typical queries are expected, or in cases where highly discriminating query terms are provided either by the user himself, or by means of a query expansion procedure.

# 8.5 Comparative effectiveness of the four clustering methods

The data obtained in this chapter allow the comparison of the effectiveness of each of the four hierarchic clustering methods used in the experiments. In section 6.4 I presented a comparison of the four methods when the cosine coefficient was used to measure interdocument similarities. The comparative effectiveness of the four methods was studied under both post-retrieval and static clustering (i.e. *n*=full when using a static similarity measure).

Regarding the comparative performance under post-retrieval clustering, the group-average was the most effective of the methods, followed by Ward's method, the complete link method, and finally the single link method. The superiority of the group average method was consistent, and in most of the cases statistically significant. The poor effectiveness of the single link method was equally consistent and significant. The only exception was noted when using the Medline collection, where the single link method outperformed both the Ward and the complete link methods.

The effectiveness of the four methods under static clustering, on the other hand, revealed few and not consistent differences between the three methods (group average, Ward's and complete link). Single link was again the least effective of the four methods.

Taking into account that there are no significant variations in cluster-based effectiveness using the query-sensitive measures and using the cosine coefficient for *n*=full, there does not seem to be a significant reason to re-examine the effectiveness of the four methods for *n*=full. This is further supported by the finding of section 6.4 that suggested that, with the exception of the poor performance of the single link method, few differences existed among the other three methods. Consequently, in the rest of this section I consider the comparative effectiveness of the four clustering methods for other values of *n*.

In the next paragraphs, I discuss the effectiveness of the group average method in section 8.5.1, the effectiveness of the Ward and complete link methods in section 8.5.2, and the effectiveness of the single link method in section 8.5.3. In section 8.5.4 I discuss the findings of this section.

## 8.5.1 The group average method

The group average method is the most effective method in the majority of the experimental conditions, using any of the query-sensitive measures. This result is in agreement with the behaviour of this method using the cosine coefficient (section 6.4). The differences between the other three methods and group average are in a large number of conditions statistically significant.

On the other hand, none of the other methods manage to significantly outperform the group average method.

The differences in favour of the group average method are generally less pronounced for precision-oriented searches, where the behaviour of all four methods becomes more comparable. Small differences between group average, and Ward's and complete link methods also occur for small values of $n$ using the CISI collection. Recall that at small values of $n$ using this collection, the effectiveness of all clustering methods is similar to that of random clustering (section 8.2.3). Figure 8.6 displays the superior effectiveness of the group average method over the other three methods using measure M1, the AP collection and $\beta$=1.

### 8.5.2 Complete link and Ward's methods

Complete link and Ward's methods display comparable effectiveness in most experimental conditions, and the statistical significance of any differences between these two methods are few and inconsistent. In section 6.4 it was demonstrated Ward's method was significantly more effective than the complete link method in a large number of experimental conditions. This is not the case in the results presented in this chapter. Complete link, using any of the query-sensitive measures, performs more comparably to Ward's method, and manages to significantly outperform it in a few experimental conditions (e.g. using CACM, measure M2, and recall-oriented searches for $n \geq 500$).

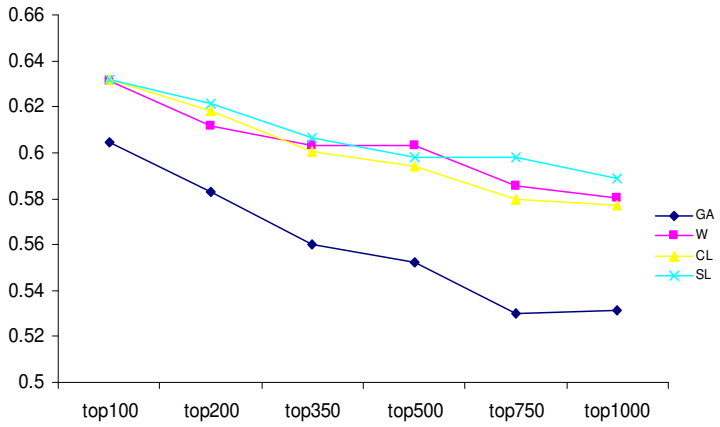


**Figure 8.6.** Comparative effectiveness of the four methods using M1, the AP collection and $\beta$=1

However, in the majority of the cases, the pattern displayed in Figure 8.6 is characteristic of the comparative effectiveness of the two methods. The differences in effectiveness are small, and the comparative effectiveness of the two methods varies across values of $n$ for the same test collection, same similarity measure and same type of search. The highly comparable effectiveness of these two methods is not surprising, given their tendency to produce hierarchies with similar

properties (Murtagh, 1984b; Voorhees, 1985a). Therefore, in contrast to section 6.4 where I concluded that Ward's method was more effective than the complete link method using the cosine coefficient, there seems no significant reason to prefer one method over another when using query-sensitive measures.

### 8.5.3 The single link method

Figure 8.6 displays another interesting result, which is perhaps the most important finding of the comparison of the four methods when using query-sensitive measures. This is that the comparative effectiveness of the single-link method improves significantly in a large number of experimental conditions. In fact, it improves in such a manner as to significantly outperform Ward's and complete link methods in a large number of conditions, and to outperform the group average method (though not significantly) when using the CACM and (especially) Medline collections. In the experimental condition that corresponds to Figure 8.6, when using the cosine coefficient in Chapter 6, the single link method was significantly less effective than all other methods for all values of *n*. When using M1 however (Figure 8.6) there are no differences between single link, complete link and Ward's methods.

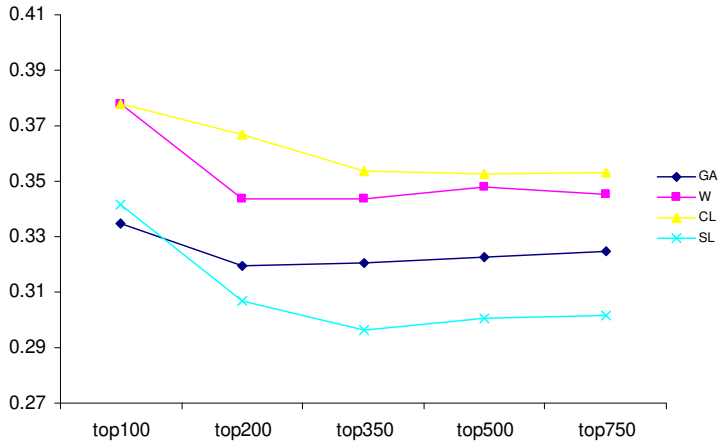


**Figure 8.7.** Comparative effectiveness of the four methods using M2, the Medline collection and $\beta$=2

In Figure 8.7 the comparative effectiveness of the three methods when using the Medline collection, measure M2 and recall-oriented searches is presented. In this case single link is significantly more effective than both the complete link and Ward's methods, and outperforms (though not significantly) the group average method. In fact, when using this collection the single link method is significantly more effective than the complete link and Ward's methods for all experimental conditions using measure M1, and all conditions using measures M2 and M3 except for precision-oriented searches for *n*=100, and 200.

Apart from when using the Medline collection, the single link method outperforms the complete link and Ward's methods using the CACM collection (most conditions using measures M1, M3), and for the majority of conditions using the WSJ collection. Furthermore, it is not significantly worse than these two methods using the AP collection and measures M1 and M3. The only two collections for which single link is consistently and significantly outperformed by all three methods are CISI and LISA.

The effectiveness of the single link method generally compares better to the other methods for non-precision oriented searches (exceptions are noted as discussed previously, especially in the case of the Medline collection). This can be attributed to the characteristics of the hierarchies that this method produces. In general, single link produces clusters whose size is considerably larger than that of the other three methods. Especially when using collections with a relatively small average number of relevant documents per query (e.g. LISA, CACM), the effectiveness of this method for precision-oriented searches is likely to be affected by the large average size of the clusters. It is therefore more likely for this type of hierarchies to display high recall-oriented effectiveness, especially given the way that recall is calculated in this experimental environment (i.e. over all relevant documents for a query).

The comparative effectiveness of precision-oriented searches of the single link hierarchies to those of the other three clustering methods becomes worse using measure M2 than when using measures M1 or M3. This has as a consequence that even in cases where precision-oriented searches of single link hierarchies are more effective than other methods using M1 or M3, they are significantly inferior to these other methods when using M2. For example, using the CACM collection and measure M1, precision-oriented searches using the single link method are more effective than those using either the Ward or the complete link methods (except for $n$=500). When using M2 however, the effectiveness of single link becomes significantly worse than that of all other three methods for all values of $n$.

As I discussed in section 8.3.1, the average size of single link clusters using measure M2 tends to significantly increase compared to that using either of the other two query-sensitive measures. This significant increase in cluster size has as a consequence that precision-oriented searches of the single link hierarchies are more affected by the large size of the clusters generated by this method.

It should also be mentioned that the better comparative performance of the single link method does not depend on the definition of recall (over all relevant documents for a query) used in the experiments. Results obtained by calculating the E measure over the retrieved and relevant documents for the two TREC collections (AP and WSJ) and for the LISA collection, demonstrated the same pattern of results as the ones reported here.

### 8.5.4 Discussion

Based on the results presented in the previous sections about the comparative effectiveness of the four clustering methods, it is justifiable to conclude that in the experimental environment used in this thesis, the group average method is the method to be preferred. It consistently and significantly outperforms the other three methods using any of the three query-sensitive measures. Moreover, the group average method is the most effective also when using the cosine coefficient (section 6.4). The group average method was not significantly outperformed by any of the other three methods in any experimental condition.

Regarding the use of the other three methods, there does not seem to be a significant reason to prefer Ward's method over the complete link method, and vice versa. Single link, manages to significantly improve its comparative effectiveness to the other three methods when using the query-sensitive similarity measures. In a large number of cases, this method significantly outperforms Ward's and the complete link methods, and is also more effective than the group average method. However, the effectiveness of this method seems to be hampered by the large size of the clusters that it produces, and therefore for precision-oriented searches it would not seem an effective choice.

An interpretation of the behaviour of the group-average and singe link methods using query-sensitive measures, can be given by noting that these two methods generally preserve the relationships defined in the similarity matrix when generating the hierarchies. Griffiths et al. (1984) had noted that these two methods introduce the smallest amount of distortion on the similarity matrix when generating document hierarchies. Other researchers (e.g. Farris, 1969; Jardine & Sibson, 1971) have suggested the same for fields other than information retrieval. In Chapter 3 (section 3.6) I discussed issues relating to the measurement of distortion introduced by clustering methods.

If one views the results of Chapter 7 as suggesting that the application of query-sensitive measures results in the generation of a similarity matrix that more closely approximates the relevance structure of the document space than using the cosine coefficient, then one can also suggest that it is a desirable property for a clustering method not to introduce a large degree of distortion on this similarity matrix. The group average and single link methods have the tendency to do so more than the complete link and Ward's methods.

It should be mentioned that the results obtained using Ward's method in this chapter should be viewed with caution. This method has been explicitly defined when squared Euclidean distances are used (Lance & Williams, 1967; Wishart, 1969). Griffiths et al. (1984, 1986) implemented this method using the Dice coefficient, something which prompted Willett (1988) to note that the

method resulting is not Ward's method *per se*. When implementing this method using query-sensitive measures, squared Euclidean distances were not used. This consequently means that the results obtained by Ward's method in this chapter may not be the results of an actual implementation of the algorithm given by Wishart (1969).

The comparison of the "query-sensitive" results for Ward's method to the "standard" results for this method (using squared Euclidean distances, Chapter 5, section 5.5.3) is permissible. This is because experiments had demonstrated no significant differences in the effectiveness of the clustering methods using the cosine coefficient, squared Euclidean distances and the Dice coefficient (section 5.5.3). In those experiments Ward's method was also implemented using the cosine and Dice coefficients, and the effectiveness of the method was not affected compared to using Euclidean distances. This suggests that there may not be a significant difference between the results reported in this chapter using Ward's method, and an actual implementation based on squared Euclidean distances. This however, was not examined in this work, and therefore the results for Ward's method in this chapter should examined with caution.

## 8.6 Summary

In this chapter I investigated the effectiveness of the application of query-sensitive measures to hierarchic document clustering. I presented experimental evidence which suggests that the use of such measures results in higher optimal cluster-based effectiveness than the use of conventional static similarity measures (section 8.2.1). The results presented in this chapter complement, and further strengthen, the results presented in Chapter 7 regarding the effectiveness of query-sensitive measures in IR. This method of query-based clustering proves to be more effective than post-retrieval clustering using a static similarity measure, and it also shows greater potential to offer an effective alternative to conventional best-match retrieval (section 8.2.2).

All three query-sensitive measures improve the effectiveness of standard clustering. I provided evidence to support the view that measure M1 is to be preferred, on the grounds of its more consistently high effectiveness (section 8.4.1). I also showed how the characteristics of hierarchies generated using measure M2 tend to differ from those generated by other similarity measures (section 8.3.1). This was especially evident in cases where few query terms were used as the only evidence to gauge the similarity between documents.

The characteristics of optimal clusters generated by the query-sensitive measures were also examined. Optimal clusters generated by query-sensitive measures tend to contain a higher proportion of relevant documents than those generated using static similarity measures (section 8.3.2.1), and also tend to form at lower levels of the hierarchies (section 8.3.2.2). These characteristics provide further evidence towards the utility of hierarchic clustering based on

query-sensitive measures, as they may have significant extensions to issues relating to the efficiency of the clustering process, and to the utility of QSSM for cluster-based interactive retrieval.

The experimental evidence presented in this chapter suggests that the traditional use of static similarity measures for document clustering has been a limiting factor to cluster-based effectiveness. It was demonstrated that by incorporating information from the query into the calculation of interdocument similarities, the generated hierarchies are more effectively tailored to the query.

The findings of previous research which have dismissed the potential of clustering as an effective alternative to best-match search, (El-Hamdouchi & Willett, 1989) for example, should be re-examined. In this chapter I showed that query-sensitive cluster-based effectiveness has the potential to significantly outperform best-match effectiveness. Whether this potential will be materialised remains to be investigated. It is one of the aims of this thesis to instigate further research in issues that have long been neglected in cluster-based research. Such issues relate to the development of effective cluster-based strategies to search query-based document hierarchies, and to the development of novel models of cluster summarisation for the presentation of the improved clustering structure to users in an interactive environment.

# Chapter 9

# Contributions and Future Work

## 9.1 Contributions and conclusions

In this section I list the contributions that this thesis has made, and outline its main conclusions.

### 9.1.1 Contributions

This thesis investigated the effectiveness of query-based hierarchic clustering of documents for the purpose of information retrieval. I outlined and investigated two approaches for query-based clustering from the perspective of retrieval effectiveness. In the following paragraphs I list the contributions that this work has made. I first present the overall contribution that I believe has been achieved, when the work of this thesis is taken as a whole. I then list in more detail some of the individual contributions that this work has made.

#### 9.1.1.1 Document clustering can be effective

It has been the long standing motivation for this work to challenge the assumptions that have characterised clustering research in IR. Such assumptions include the static application of document clustering prior to querying, and the static calculation of interdocument associations. A further motivation has been to challenge the results of previous research which had dismissed clustering as an effective method for information retrieval. This was based on the view that document clustering has the potential to act as an effective retrieval mechanism if its static application is reviewed. The experimental evidence that was presented in this thesis demonstrated that clustering can indeed act as a highly effective method for information retrieval.

This work focused purely on effectiveness issues. This is in contrast to recent work on document clustering, which, having accepted limitations in the effectiveness of the clustering process, has focused on other aspects (e.g. efficiency). I believe that the work reported in this thesis provides

leverage for the further advancement of research in the area of cluster-based retrieval effectiveness, and I outline some possible areas for future research in section 9.2.

In the following paragraphs I outline specific contributions that this work made regarding the effectiveness of document clustering.

#### 9.1.1.2 Investigating the effectiveness of post-retrieval hierarchic clustering

In Chapter 5 I described post-retrieval clustering as the first approach for generating query-based document hierarchies. I also outlined the implications of post-retrieval clustering for the effectiveness of cluster-based retrieval. By reviewing previous work on the effectiveness of post-retrieval clustering, I presented a number of issues that had not been addressed. I investigated these issues in Chapter 6 in a study that consisted of three parts.

In the first part of the investigation I examined the effect that different numbers of top-ranked documents have on the structure of the document space, in terms of the proximity of pairs of documents which are relevant to the same query (co-relevant documents). In the second part I examined the effectiveness of four hierarchic clustering methods (group average, Ward, complete link and single link) under four different viewpoints: the variation of effectiveness across different numbers of top-ranked documents, the comparative effectiveness of post-retrieval and static clustering, the comparative effectiveness of cluster-based and best-match retrieval, and the comparative effectiveness of actual and random clustering. In the third part of this study I investigated the comparative effectiveness of the four clustering methods under both post-retrieval and static clustering.

The results demonstrated that the effectiveness of post-retrieval clustering is significantly higher than that of static clustering. It was also demonstrated that cluster-based retrieval through post-retrieval clustering has the potential to exceed the effectiveness of a best-match IR system. However, the experiments also demonstrated a number of shortcomings regarding the effectiveness of post-retrieval clustering. These shortcomings mainly came in the form of poor comparative effectiveness to best-match retrieval for some experimental conditions, and also in the form of close-to-random effectiveness in a number of cases. The use of query-sensitive measures for document clustering aimed to address these shortcomings.

#### 9.1.1.3 Challenging the static nature of interdocument similarity

In Chapter 5 I proposed an axiomatic view of the cluster hypothesis. In agreement with the hypothesis, I argued that co-relevant documents are more similar to each other than to other documents. However, in contrast to the traditional treatment of the hypothesis, I argued that this similarity is inherent, and is dictated by the query itself (i.e. all pairs of documents which are relevant to the same query should exhibit this inherent similarity). According to this view, if

documents fail to display this inherent similarity, then this is attributed to the way the similarity between documents is measured, and not to the properties of the document collection to which these documents belong. As part of the same argument, I viewed document clustering for information retrieval as a goal-driven process. I argued that the purpose of document clustering for IR is to group relevant documents together, separately from non-relevant ones.

Based on these arguments, I proposed the use of query-sensitive similarity measures (QSSM). These measures view the similarity between documents as a dynamic concept which changes depending on the purpose for which it is measured. By accepting purpose to be the separation between relevant and non-relevant documents, and by accepting that in the context of IR relevance is dictated, among other factors, by the presence of query terms in documents, I postulated that query terms should acquire greater salience in defining the inherent similarity between co-relevant documents.

#### 9.1.1.4 Proposing measures for the calculation of query-sensitive similarity

In Chapter 7 I expanded on the notion of query-sensitive similarity, and I proposed specific measures for its calculation. I proposed three query-sensitive measures, and I illustrated the way that they incorporate information from the query into the calculation of interdocument similarity.

The working of these measures is dictated by the view of query-sensitive similarity that was proposed in Chapter 5. These measures assign higher similarity values to pairs of documents that possess a larger number of query terms in common than other pairs, with the aim to force co-relevant documents to become more similar to each other, and therefore to achieve a greater per-query adherence to the cluster hypothesis.

#### 9.1.1.5 Investigating the effectiveness of QSSM at structuring the document space

In Chapter 7 I experimentally investigated the effectiveness of query-sensitive measures at structuring the document space. I evaluated the measures based on their effectiveness at "forcing" documents relevant to the same query to be more similar to each other. I did so by examining how many relevant documents are contained within a five-document neighbourhood of any given relevant document. The investigation consisted of three parts. First I examined the comparative effectiveness of query-sensitive measures and static measures (I used the cosine coefficient as a static measure). I then studied the comparative effectiveness of the three measures to each other, and I also investigated the effect that query length has on the effectiveness of each of the three query-sensitive measures.

The experiments demonstrated that query-sensitive measures are significantly more effective than static measures at increasing the similarity of pairs of co-relevant documents. In this way, query-

sensitive measures achieve a greater per-query adherence to the cluster hypothesis, and are therefore more likely to result in high cluster-based retrieval effectiveness.

#### 9.1.1.6 Investigating the effectiveness of hierarchic document clustering using QSSM

In Chapter 8 I investigated the second form of query-based clustering, which uses query-sensitive measures for the calculation of interdocument similarities. The investigation consisted of four parts.

In the first part I examined the comparative effectiveness of clustering using query-sensitive measures and of clustering using static measures, and I also considered whether the use of query-sensitive measures improves the effectiveness of cluster-based retrieval compared to best-match retrieval. The second part included the study of the characteristics of hierarchies generated using query-sensitive measures, the study of the characteristics of optimal clusters generated by query-sensitive measures, and the comparison of these characteristics to those of hierarchies and optimal clusters generated using static similarity measures. The third part of the study looked into the comparative effectiveness of the three query-sensitive measures. The effect that query length has on the effectiveness of each of the three measures was also investigated in this part, together with the effectiveness of the three measures across different numbers of top-ranked documents. In the last part of the study I examined the comparative effectiveness of the four clustering methods under document clustering which uses query-sensitive measures.

The experiments demonstrated that hierarchic document clustering which uses query-sensitive similarity measures is significantly more effective than clustering which uses static measures. The significant effectiveness improvements also translated into more favourable comparison to the effectiveness of best-match retrieval. Moreover, the characteristics of clusters generated using the query-sensitive measures proved more useful than those of clusters generated using static measures. Such characteristics include the average size, average number of relevant documents, and the levels of the document hierarchies in which they occur.

### 9.1.2 Conclusions

The results of this thesis demonstrated that by incorporating information from the query into the clustering process (query-based clustering), the effectiveness of the clustering process is enhanced. The experiments investigated two approaches for query-based clustering, and demonstrated that:

- Hierarchic post-retrieval clustering using static similarity measures is significantly more effective than static clustering, and also has the potential to significantly exceed the effectiveness of best-match retrieval.

- Query-sensitive similarity measures are significantly more effective than static measures at increasing the similarity of pairs of co-relevant documents. In this way, query-sensitive measures achieve a better per-query adherence to the cluster hypothesis.
- Hierarchic clustering using query-sensitive similarity measures is significantly more effective than hierarchic clustering using static measures. It also significantly increases the potential of cluster-based retrieval to be more effective than best-match retrieval.

## 9.2 Future Work

In the following paragraphs I outline a number of areas for possible future work. These areas either describe aspects of the work of this thesis that might be worthy of further investigation, or that stem as a consequence of the findings of this thesis.

### 9.2.1 Cluster-based search strategies

The results presented in this thesis demonstrated that using query-sensitive measures for the calculation of interdocument similarity enhances the effectiveness of the clustering process. An area of cluster-based research that is directly affected by this finding is that of cluster-based searches.

Research into models of cluster-based searches, and also into models of cluster representatives, has been rather limited over the past fifteen years. One reason for this is that IR researchers seem to have accepted the limitations of the effectiveness of document clustering, demonstrated by research carried out at that time (El-Hamdouchi & Willett, 1989). The results presented in this work provided evidence that the effectiveness of document clustering can be enhanced. Therefore, further work would be needed to investigate whether the effective characteristics of query-based hierarchies can be exploited by cluster-based search models. A particular type of cluster-based search that may benefit from the use of query-sensitive similarity measures is that of nearest-neighbour clusters (NNC) (Griffiths *et al*., 1986, El-Hamdouchi, 1987).

### 9.2.2 Summarisation of cluster contents

The results presented in Chapter 8 suggested that, compared to using a static similarity measure, optimal clusters generated using query-sensitive measures contain a higher proportion of relevant documents, are more compact in size, and tend to occur at significantly lower levels in the hierarchy.

An important, yet relatively unexplored problem of cluster-based information retrieval, is the representation of the contents of document clusters for the specific purpose of providing relevance

clues to users. It would be worthwhile to investigate models of cluster summarisation that would exploit these characteristics of query-sensitive clusters. I would expect query-biased summarisation models (Tombros & Sanderson, 1998) to be better suited to this task. It may prove to be the case that specific characteristics of clusters generated using query-sensitive measures are better tailored to summarisation models. One such characteristic is the distribution of query terms in clusters. The issue of cluster summarisation is closely related (and yet distinct at the same time) to that of the development of models of cluster representatives for the purpose of cluster-based searches that was discussed in the previous section.

### 9.2.3 Interactive retrieval

Following naturally from the previous point, it is possible to investigate implications of this work that may lead to further experimentation in interactive IR.

One point for future work is to examine whether the effectiveness improvements presented in this work would also translate into effectiveness improvements in an interactive, task-based retrieval environment. It may be the case that the improved structure of document hierarchies generated by query-sensitive measures, in terms of the characteristics that I mentioned in 9.2.2, could eventually lead into document hierarchies which are more useful for users in an interactive environment. By increasing the concentration of useful (i.e. relevant) information into a smaller part of the hierarchy, it would be easier and quicker for users to identify this useful part of the hierarchy.

Another issue to be examined further, is the application of query-sensitive measures to the visualisation of the relationships between documents in a collection. The results in Chapter 7 demonstrated the effectiveness of query-sensitive measures at placing more relevant documents in the neighbourhood of any given relevant document than a conventional static measure. It would therefore be worthwhile to investigate how these measures fit into methods for the visualisation of interdocument relationships, such as for example the ones stemming from Leuski's work (Allan *et al*., 2001; Leuski, 2001).

Query-based clusters can also provide a starting point for a path-based ostensive browsing of the document space (Campbell, 2000). Assuming an initial query creates a set of clusters, the user could gain an overview of the document space by browsing the clusters. At any point the user could select any of the clusters, and the centroid (or any other representation) of the cluster may then be used as a starting point for the path-based browsing. A similar extension would involve investigating the use of query-based clusters for relevance feedback.

### 9.2.4 Other sources of evidence as an indication of the user's information need

In section 7.2.2 I mentioned some limitations of query-sensitive measures. These mainly focused on the use of only query terms as an indication of the user's information need, and also as an indication of a document's relevance. In that section I also mentioned temporal and contextual factors that may affect the user's information need. Consequently, further research to address these issues should be pursued.

A possibility for future work is to investigate semantically enriching the context of the query by employing methods such as Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and the Hyperspace Analog to Language (HAL) (Burgess *et al*., 1998). Such methods generate semantic contexts for terms based on patterns of term co-occurrence in text units. Semantic contexts for query terms could enhance query representations. I believe that investigating the efficacy of such approaches would be worthwhile pursuing.

Some other indications of the user's information need could also stem from sources such as user profiles (Bhatia, 1992), and the user interaction with various aspects of the documents and their contents (e.g. type of documents accessed, time spent on documents, etc.) (Villa & Chalmers, 2001). This type of information may also provide means towards addressing temporal aspects of information needs.

Furthermore, a more systematic analysis of the dependence of such measures on query length would be appropriate. In the experiments carried out in Chapters 7 and 8, I did not employ algorithmic methods of query-expansion. This issue should be further investigated, as the choice of highly discriminating expansion terms may alter the comparative effectiveness of the query-sensitive measures proposed in this work.

### 9.2.5 Efficiency issues

Efficiency issues were deliberately not examined in this thesis. As was explained, this was because I view efficiency issues as more likely to develop in the light of improved effectiveness rather than vice versa. Hierarchic document clustering is a resource-demanding process. If the effectiveness benefits noted in the laboratory experiments in this thesis are to be transferred to operational environments, then efficiency issues should be considered.

One direction for further research is the development of efficient algorithms for the calculation of similarity matrices resulting from query-sensitive measures. Croft (1977) and Willett (1981) have proposed efficient algorithms for the calculation of similarity matrices resulting from static measures. Query-sensitive measures (especially measures M1 and M2) result in sparse matrices

that contain a large number of zero similarities between pairs of documents which have no query terms in common. Investigating the applicability of Croft's and Willett's algorithms for the case of query-sensitive measures, or developing algorithms tailored to query-sensitive measures, is an issue for further development.

A further direction for future research is also related to the finding of Chapter 8, that optimal clusters in query-sensitive hierarchies tend to occur in significantly lower levels of the hierarchy than when using static measures. The implications of this finding for the improvement of time and space efficiency of hierarchic document clustering should be further investigated. For example, the time-efficiency of clustering may improve by terminating the clustering process at an appropriately low level of the generated hierarchy.

# Bibliography and References


Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 18th ACM SIGMOD Conference*, pp. 94-105. Seattle, WA.

Allan, J., Leuski, A., Swan, R., Byrd, D. (2001). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management*, 37(3):435-458.

Allen, R.B., Obry, P., Littman, M. (1993). An interface for navigating clustered document sets returned by queries. In *Proceedings of the ACM Conference on Organizational Computing Systems*, pp. 166-171.

Amba, S., Narasimhamurthi, N., O'Kane, K.C., Turner, P.M. (1996). Automatic linking of thesauri. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 181-186. Zurich, Switzerland.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Anick, P.G. and Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. In *Proceedings of the 20th Annual ACM SIGIR Conference*, pp. 314-323. Philadelphia, PA.

Aslam, J., Pelekhov, K., Rus, D. (1998). Static and dynamic information organization with star clusters. In *Proceedings of the 7th ACM International Conference on Information and Knowledge Management*, pp. 208-217. Washington, DC.

Attar, B. and Fraenkel, A.S. (1977). Local feedback in full-text retrieval systems. *Journal of the ACM*, 24(3):397-417.

Barry, C.L. (1994). User defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149-159.

Barry, C.L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science*, 49(14):1293-1303.

Bartell, B.T., Cottrell, G.W., Belew, R.K. (1995). Representing documents using an explicit model of their similarities. *Journal of the American Society for Information Science*, 46(4):254-271.

Beeferman, D. and Berger, A. (2000). Agglomerative clustering of a search engine log. In *Proceedings of the 6th International Conference on Knowledge Discovery in Data*, pp. 407-416. Boston, MA.

Belew, R.K. (2000). *Finding out about: A cognitive perspective on search engine technology and the WWW.* Cambridge: Cambridge University Press.

Bharat, K., Henzinger, M.R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 104-111. Melbourne, Australia.

Bhatia, S.K. (1992). Selection of search terms based on user profiles. In *Proceedings of the 1992 ACM/SIGAPP Symposium on Applied Computing ( vol. 1)*, pp. 224-233. Kansas City, MO.

Bhatia, S.K. and Deogun, J.S. (1993). Cluster characterization in information retrieval. In *Proceedings of the ACM SIGAPP Symposium on Applied computing: states of the art and practice*, pp. 721-728. Indianapolis, IN.

Blashfield, R.K. (1976). Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchic methods. *Psychological Bulletin*, 83(3):377-388.

Bollmann, P. and Raghavan, V.V. (1993). On the delusiveness of adopting a common space for modeling IR objects: Are queries documents? *Journal of the American Society for Information Science*, 44(10):579-587.

Borlund, P. and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225-250.

Botafogo, R.A. (1993). Cluster analysis for hypertext systems. In *Proceedings of the 16th Annual ACM SIGIR Conference*, pp. 116-125. Pittsburgh, PA.

Broder, A., Glassman, S., Manasse, M., Zweig, G. (1997). Syntactic Clustering of the Web. In *Proceedings of the 6th International World Wide Web Conference*, pp. 391-404. Santa Clara, CA.

Buckley, C., Mitra, M., Walz, J., Cardie, C. (2000). Using clustering and super-concepts within SMART: TREC 6. *Information Processing & Management*, 36(1), 109-131.

Burgess, C., Livesay, K. and Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25(2&3):211-257.

Burgin, R. (1995). The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Society for Information Science*, 46(8):562-572.

Campbell, I. (2000). The ostensive model of developing information needs. Ph.D. Thesis, Department of Computing Science, University of Glasgow.

Can, F. and Ozkarahan, E.A. (1990). Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Transactions on Database Systems*, 15(4): 483-517.

Carey, M., Kriwaczek, F., Rüger, S.M. (2000). A visualization interface for document searching and browsing. In *Proceedings of the ACM CIKM 2000 Workshop on New Paradigms in Information Visualization and Manipulation*. Washington, DC.

Chang, C. and Hsu, C. (1997). Customizable multi-engine search tool with clustering. In *Proceedings of the 6th WWW Conference*. Santa Clara, CA.

Chen, H. and Dumais, S. (2000). Bringing order to the web: Automatically categorizing search results. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 145-152. The Hague, The Netherlands.

Chen, H.M. and Cooper, M.D. (2001). Using clustering techniques to detect usage patterns in a web-based information system. *Journal of the American Society for Information and Technology*, 52(11):888-904.

Cleverdon, C., Mills, J., Keen, M. (1966). *ASLIB Cranfield Research Project: factors determining the performance of indexing systems*. Cranfield Institute of Technology, Cranfield, England.

Cole, A.J. (1969). *Numerical taxonomy: proceedings of the Colloquium in Numerical Taxonomy held in the University of St. Andrews*. London: Academic Press.

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society*, Series A, 134:321-353.

Croft, W.B. (1977). Clustering large files of documents using the single-link method. *Journal of the American Society for Information Science*, 28:341-344.

Croft, W.B. (1978). Organizing and searching large files of document descriptions. Ph.D. Thesis, Churchill College, University of Cambridge.

Croft, W.B. and Harper, D.J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285-295.

Croft, W.B. (1980). A model of cluster searching based on classification. *Information Systems*, 5:189-195.

Croft, W.B., Lucia, T.J., Cringean, J., Willett, P. (1989). Retrieving documents by plausible inference: An experimental study. *Information Processing & Management*, 25(6):599-614.

Crouch, C.J. and Yang, B. (1992). Experiments in automatic statistical thesaurus construction. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pp. 77-88. Copenhagen, Denmark.

Crouch, D.B., Crouch, C.J., Andreas, G. (1989). The use of cluster hierarchies in hypertext information retrieval. In *Proceedings of ACM Hypertext '89*, pp. 225-237. Pittsburgh, PA.

Cunningham, K.M. and Ogilivie, J.C. (1972). Evaluation of hierarchical grouping techniques: a preliminary study. *Computer Journal*, 15(3):209-213.

Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. (1992). Scatter/Gather: A cluster based approach to browsing large document collections. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pp. 126-135. Copenhagen, Denmark.

Defays, D. (1977). An efficient algorithm for a complete link method. *Computer Journal*, 20:93-95.

Deogun, J.S. and Raghavan, V.V. (1986). User-oriented clustering: A framework for learning in information retrieval. In *Proceedings of the 9th Annual ACM SIGIR Conference*, pp. 157-163. Pisa, Italy.

Diday, E. and Simon, J.C. (1976). Clustering Analysis. In Fu, K.S., Keidel, W.D., Wolter, H. (eds.) *Digital Pattern Recognition*. Berlin: Springer-Verlag.

Doyle, L.B. (1964). Some compromises between word grouping and document grouping. In *Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation*, pp. 15-24. U.S. Department of Commerce, national Bureau of Standards, Misc. Publication 269.

Dubes, R. and Jain, A.K. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, 11:235-254.

Dubin, D.S. (1996). Structure in document browsing spaces. Ph.D. Thesis, School of Information Sciences, University of Pittsburgh.

Eguchi, K., Ito, H., Kumamoto, A, Kanata, Y. (2001). Adaptive document clustering using incrementally expanded queries. *Systems and Computers in Japan*, 32(2):64-74.

El-Hamdouchi, A. (1987). Using inter-document relationships in information retrieval. Ph.D. Thesis, Department of Information Studies, University of Sheffield.

El-Hamdouchi, A. and Willett, P. (1987). Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science*, 13:361-365.

El-Hamdouchi, A. and Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220-227.

Ellis, D., Furner-Hines, J., Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, 3(2):128-149.

Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231. Portland, OR.

Everitt, B.S. (1993). *Cluster analysis*. London: E. Arnold, 3rd edition.

Farris, J.S. (1969). On the cophenetic correlation coefficient. *Systematic Zoology*,18:279-285.

Fisher, L. and van Ness, J.W. (1971). Admissible clustering procedures. *Biometrika*, 58:91-104.

Frakes, W.B. and Baeza-Yates, R. (eds.) (1992). *Information Retrieval: Data Structures and Algorithms*. New Jersey: Prentice Hall.

Garey, M.R. and Johnson, D.S. (1979). *Computers and intractability: a guide to the theory of NP-Completeness*. San Francisco: W.H. Freeman.

Garland, K. (1982). An experiment in automatic hierarchical document classification. *Information Processing & Management*, 19(3):113-120.

Goffman, W. (1969). An indirect method of information retrieval. *Information Storage and Retrieval*, 4:363-373.

Good, I.J. (1958). Speculations concerning information retrieval. *Research report PC-78*, IBM Research Centre, Yorktown Heights, New York.

Goodman, N. (1972). Seven strictures on similarity. In Goodman, N. (ed.). *Problems and Projects*, pp. 437-447. Indianapolis and New York: Bobbs-Merrill.

Gordon, A.D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A*, 150(2):119-137.

Gordon, M.D. (1991). User-based clustering by redescribing subject descriptors with a genetic algorithm. *Journal of the American Society for Information Science*, 42(5):311-322.

Gower, J.C. (1974). Maximal predictive classification. *Biometrics*, 30:643-654.

Griffiths, A., Robinson, L.A., Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation*, 40(3):175-205.

Griffiths, A., Luckhurst, H.C., Willett, P. (1986). Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37:3-11.

Guha, S., Rastogi, R., Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*, pp. 73-84. Seattle, WA.

Guillaume, D. and Murtagh, F. (2000). Clustering of XML documents. *Computer Physics Communications*, 127(2-3):215-227.

Harman, D.K. (1992). Relevance feedback revisited. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pp. 1-10. Copenhagen, Denmark.

Harman, D.K. (ed.) (1993). *Proceedings of the First Text Retrieval Conference*. National Institute of Standards and Technology, Gaithersburg, MD.

Harper, D.J., Mechkour, M., Muresan, G. (1999). Document clustering for mediated information access. In *Proceedings of the 21st BCS-IRSG Annual Colloquium on IR Research*, Glasgow, Scotland.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

Hartuv, E., Schmitt, A., Lange, J., Meier-Ewert, S., Lehrach, H., Shamir, R. (1999). An algorithm for clustering cDNAs for gene expression analysis. In *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology*, pp. 188-197. Lyon, France.

Hatzivassiloglou, V., Gravano, L., Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd Annual ACM SIGIR Conference*, pp. 224-231. Athens, Greece.

Hearst, M.A. and Pedersen, J.O. (1996). Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 76-84. Zurich, Switzerland.

Hersh, W. and Over, P. (2001). TREC-9 Interactive Track Report. NIST Special Publication: The Ninth Text Retrieval Conference (TREC 9). (in press).

Hinneburg, A. and Kleim, D.A. (1998). An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pp. 58-65. New York, NY.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews of the Cambridge Philosophical Society*, 57(4):669-689.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the 17th Annual ACM SIGIR Conference*, pp. 101-110. Dublin, Ireland.

Ivie, E.L. (1966). Search procedures based on measures of relatedness between documents. Ph.D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology.

Iwayama, M. (2000). Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. In *Proceedings of the 23rd Annual ACM SIGIR Conference*, pp. 10-16. Athens, Greece.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall..

Jain, A.K., Murty, M.N., Flynn, P.J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323.

Janes, J.W. (1991). Relevance judgements and the incremental presentation of document representations. *Information Processing & Management*, 27(6):629-646.

Jansen, B.J., Spink, A., Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of users on the web. *Information Processing & Management*, 36(2):207-227.

Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, 11(2):177-184.

Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*. New York: Wiley.

Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217-240.

Johnson, A. and Fotouhi, F. (1996). Adaptive clustering of hypermedia documents. *Information Systems*, 21(6):459-473.

Jones, W.P. and Furnas, G.W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420-442.

Karypis, G., Han, E.H., Kumar, V. (1999). CHAMELEON: a hierarchical clustering algorithm using dynamic modelling. *IEEE Computer*, 32(8):68-75.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data: an introduction to cluster analysis*. New York: Wiley.

Keen, E.M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4):491-502.

Kirriemuir, J.W. and Willett, P. (1995). Identification of duplicate and near-duplicate full-text records in database search-outputs using hierarchic cluster analysis. *Program*, 29(3):241-256.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632.

Korpimies, K. and Ukkonen, E. (1998). Term weighting in query-based document clustering. In *Proceedings of the 2nd East European Symposium on Advances in Database Systems, Lecture Notes in Computer Science 1475*, pp. 151-153. Poznan, Poland.

Kuiper, F.K. and Fisher, L.A. (1975). A Monte-Carlo comparison of six clustering procedures. *Biometrics*, 31:777-783.

Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999). Trawling the Web for Emerging Cyber-Communities. In *Proceedings of the 8th WWW Conference*, pp. 403-415. Toronto, Canada.

Kural, Y. (1999). Clustering information retrieval search outputs. Ph.D. Thesis, City University, London.

Kural, Y., Robertson, S.E., Jones, S. (2001). Deciphering cluster representations. *Information Processing & Management*, 37(4):593-601.

Lancaster, F.W. (1968). *Information retrieval systems: characteristics, testing, and evaluation*. New York: Wiley.

Lance, G.N. and Williams, W.T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal*, 9:373-380.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211-240.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 16-22. San Diego, CA.

Leuski, A. and Allan, J. (1998). Evaluating a visual navigation system for a digital library. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL'98)*, pp. 535-554. Heraklion, Greece.

Leuski, A. (2001). Interactive information organization: techniques and evaluation. Ph.D. Thesis, University of Massachusetts, Amherst.

Lewis, D.D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pp. 37-50. Copenhagen, Denmark.

Lewis, D.D. and Sparck Jones, K. (1993). Natural language processing for information retrieval. *Technical Report 307*, University of Cambridge Computer Laboratory.

Ling, R.F. and Killough, G.G. (1976). Probability tables for cluster analysis based on a theory of random graphs. *Journal of the American Statistical Association*, 71(354):293-300.

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159-165.

Maarek, Y.S., Fagin, R., Ben-Shaul, I.Z., Pelleg, D. (2000). Ephemeral document clustering for web applications. *IBM Research Report RJ 10186*, IBM Research, Haifa, Israel.

Macnaughton-Smith, P. (1965). Some statistical and other numerical techniques for classifying individuals, studies in the causes of delinquency and the treatment of offenders. *Home Office Research Unit Report No. 6*, HMSO, London.

Macskassy, S.A., Banerjee, A., Davidson, B.D., Hirsh, H. (1998). Human performance on clustering web pages: a preliminary study. In *Proceedings of The 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pp. 264-268. New York, NY.

Magennis, M. and van Rijsbergen, C.J. (1997). The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th Annual ACM SIGIR Conference*, pp. 324-332. Philadelphia, PA.

Mani, I. and Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35-67.

Mather, L.A. (2000). A linear algebra measure of cluster quality. *Journal of the American Society for Information Science*, 51(7):602-613.

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3-30.

Mechkour, M., Harper, D.J., Muresan, G. (1998). The WebCluster project. Using document clustering for mediating access to the world wide web. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 357-358. Melbourne, Australia.

Milligan, G.W., Soon, S.C., Sokol, L.M. (1983). The effect of cluster size, dimensionality, and the number of cluster on recovery of true cluster structure. *IEEE Transactions on Patter Recognition and Machine Intelligence*, 5(1):40-47.

Milligan, G.W. and Cooper, M.C. (1987). Methodology review: clustering methods. *Applied Psychological Measurement*, 11(4):329-354.

Minker, J., Wilson, G.A., Zimmerman, B.H. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage & Retrieval*, 8:329-348.

Mizzaro, S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science*, 48(9):810-832.

Modha, D.S. and Spangler, W.S. (2000). Clustering hypertext with applications to web searching. In *Proceedings of the ACM Hypertext Conference*, pp. 143-152. San Antonio, TX.

Mukherjea, S. Foley, J.D., Hudson, S.E. (1994). Interactive clustering for navigating in hypermedia systems. In *Proceedings of the 1994 ACM European Conference on Hypermedia Technology*, pp. 136-145. Edinburgh, Scotland.

Mukherjea, S., Hirata, K., Hara, Y. (1998). Using clustering and visualization for refining the results of a WWW image search engine. In *Proceedings of the 1998 Workshop on New Paradigms in Information Visualization and Manipulation*, pp. 29-35. Bethesda, MD.

Mukherjea, S. (2000). Organising topic-specific web information. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia*, pp. 133-141. San Antonio, TX.

Murray, D.M. (1972). Document retrieval based on clustered files. Ph.D. Thesis, Cornell University. Report ISR-20 to National Science Foundation and National Library of Medicine.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *Computer Journal*, 26:354-359.

Murtagh, F. (1984a). Complexities of hierarchic clustering algorithms: state of the art. *Computational Statistics Quarterly*, 1:101-114.

Murtagh, F. (1984b). Structure of hierarchic clusterings: implications for information retrieval and for multivariate data analysis. *Information Processing & Management*, 20(5/6):611-617.

Neto, J.L. and Santos, A.D. (2000). Document clustering and summarization. In *Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 41-56. Blackpool, UK.

Norreault, T., McGill, M., Koll, M.B. (1981). A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In Oddy, R.N., Robertson, S.E., van Rijsbergen, C.J., Williams, P.W. *Information Retrieval Research*. London: Butterworths.

Nosofsky, R.M. (1986). Attention, Similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39-57.

Ottaviani, J.S. (1994). The fractal nature of relevance: A hypothesis. *Journal of the American Society for Information Science*, 45(4):263-272.

Peat, H.J. and Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378-383.

Pirolli, P., Schank, P., Hearst, M., Diehl, C. (1996). Scatter/Gather communicates the topic structure of a very large text collection. In *Proceedings of the ACM Conference on Human Factors and Computer Systems*, pp. 213-220. Vancouver, Canada.

Popescul, A., Flake, G.W., Lawrence, S., Ungar, L.H., Giles, C.L. (2000). Clustering and identifying temporal trends in document databases. In *Proceedings of the IEEE Conference on Advances in Digital Libraries (ADL 2000)*, pp. 173-182. Washington, DC.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program - Automated Library and Information Systems*, 14(3):130-137.

Pratt, W., Hearst, M., Fagan, L.M. (1999). A knowledge-based approach to organizing retrieved documents. In *Proceedings of the 16th National Conference on Artificial Intelligence*, pp. 80-85. Orlando, FL.

Preece, S.E. (1973). Clustering as an output option. *Proceedings of the American Society for Information Science*, 10:189-190.

Radev, D.R., Jing, H., Budzikowska, M. (2000). Summarization of multiple documents: clustering, sentence extraction, and evaluation. In *ANLP/NAACL Workshop on Summarization*. Seattle, WA.

Raghavan, V.V. and Wong, K.M. (1986). A critical analysis of the vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279-287.

Rasmussen, E. (1992). Clustering Algorithms. In Frakes, W.B. and Baeza-Yates, R. (editors) *Information Retrieval: Data Structures and Algorithms*. New Jersey: Prentice Hall.

Rath, G.J., Resnick, A., Savage, T.R. (1961). Comparisons of four types of lexical indicators of content. *American Documentation*, 12(2):126-130.

Reid, J. (2000). A task-oriented non-interactive evaluation methodology for information retrieval systems. *Information retrieval*, 2(1):115-129.

van Rijsbergen, C.J. (1971). An algorithm for information structuring and retrieval. *Computer Journal*, 14:407-412.

van Rijsbergen, C.J. and Sparck Jones, K. (1973). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251-257.

van Rijsbergen, C.J. (1974a). Foundation of evaluation. *Journal of Documentation*, 30(4):365-373.

van Rijsbergen, C.J. (1974b). Further experiments with hierarchic clustering in document retrieval. *Information Storage and Retrieval*, 10:1-14.

van Rijsbergen, C.J. and Croft, W.B. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 Collection. *Information Processing & Management*, 11:171-182.

van Rijsbergen, C.J. (1979). *Information Retrieval*. London: Butterworths, 2nd Edition.

van Rijsbergen, C.J, Harper D.J., Porter, M.F. (1981). The selection of good search terms. *Information Processing & Management*, 17:77-91.

van Rijsbergen, C.J. (1986). A new theoretical framework for information retrieval. In *Proceedings of the 9th Annual ACM SIGIR Conference*, pp. 194-200. Pisa, Italy.

Robertson, S.E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33:294-304.

Robertson, S.E. (1981). The methodology of information retrieval experiment. In Sparck Jones, K. (ed.) *Information Retrieval Experiment*, pp. 9-31. London: Butterworths.

Rocchio, J.J. (1966). Document retrieval systems - Optimization and evaluation. PhD Thesis, *Report ISR-10* to the National Science Foundation, Harvard Computation Laboratory.

Rorvig, M. (1999). Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8):639-651.

Roussinov, D.G. and Chen, H. (2001). Information navigation on the web by clustering and summarizing query results. *Information Processing & Management*, 37(6):789-816.

Ruthven, I., Tombros, A., Jose, J. (2001). A study on the use of summaries and summary-based query expansion for a question-answering task. In *Proceedings of the 23rd Annual BCS*

*European Colloquium on Information Retrieval Research (ECIR 2001)*, pp. 41-53. Darmstadt, Germany.

Salton, G., (ed.) (1971). *The SMART Retrieval System - Experiments in Automatic Document Retrieval.* New Jersey, Englewood Cliffs: Prentice Hall Inc.

Salton, G., Wong, A., Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613-620.

Salton, G. and Wong, A. (1978). Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4):321-346.

Salton, G. and McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513-523.

Saracevic, T. (1969). Comparative effects of titles, abstracts and full texts on relevance judgements (1). *Proceedings of the American Society for Information Science*, 6:293-299.

Saracevic, T. (1975). Relevance: a review of and framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321-343.

Schamber, L., Eisenberg, M.B., Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 26(6):755-776.

Scheibler, D. and Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms: a comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research*, 20:283-304.

Shaw, R.J., and Willett, P. (1993). On the non-random nature of nearest-neighbour document clusters. *Information Processing & Management*, 29(4): 449-452.

Shaw, W.M. Jr. (1990). Subject indexing and citation indexing. Part II: An evaluation and comparison. *Information Processing & Management*, 26: 705-718.

Shaw, W.M. Jr. (1991). Subject and citation indexing. Part II: The optimal cluster-based retrieval performance of composite representations. *Journal of the American Society for Information Science*, 42(9): 676-684.

Shaw, W.M. Jr. (1993). Controlled and uncontrolled subject descriptions in the CF database: A comparison of optimal cluster-based retrieval results. *Information Processing & Management*, 29(6): 751-763.

Shaw, W.M. Jr., Burgin, R., Howell, P. (1997). Performance standards and evaluations in IR test collections: Cluster-Based retrieval models. *Information Processing & Management*, 33(1):1-14.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single link cluster method. *Computer Journal*, 16:30-34.

Siegel, S. and Castellan, N.J. Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Silverstein, C. and Pedersen, J.O. (1997). Almost-constant-time clustering of arbitrary corpus subsets. In *Proceedings of the 20th Annual ACM SIGIR Conference*, pp. 60-66. Philadelphia, PA.

Singhal, A., Buckley, C., Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 21-29. Zurich, Switzerland.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799-813.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W.H. Freeman.

Sokal, R.R. and Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11:33-40.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11-21.

Sparck Jones, K. and van Rijsbergen, C.J. (1976). Information retrieval test collections. *Journal of Documentation*, 32(1):59-75.

Sparck Jones, K. and Willett, P. (eds.) (1997). *Readings in Information retrieval*. San Francisco: Morgan Kaufmann.

Späth, H. (1980). *Cluster Analysis Algorithms For Data Reduction and Classification of Objects*. England: John Ellis Horwood Limited.

Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213-1228.

Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques. In *Proceedings of KDD'2000 International Workshop on TextMining*. Boston, MA.

Tanaka, H., Kumano, T., Uratani, N., Ehara, T. (1999). An efficient document clustering algorithm and its application to a document browser. *Information Processing & Management*, 35(4):541-557.

Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. San Diego: Academic Press.

Tombros, A. and Sanderson, M. (1998). The advantages of query-biased summaries in IR. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 2-10. Melbourne, Australia.

Tombros, A. and Crestani, F. (2000). Users' perception of relevance of spoken documents. *Journal of the American Society for Information Science*, 51(10):929-939.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327-352.

Van Ryzin, J. (ed.) (1977). *Classification and Clustering: Proceedings of the Advanced Seminar Conducted by the Mathematics Research Center, the University of Wisconsin at Madison*. New York, London: Academic Press.

Villa, R. and Chalmers, M. (2001). A framework for implicitly tracking data. In *Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*. Dublin, Ireland.

Voorhees, E.M. (1985a). The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval. Ph.D. Thesis, Technical Report TR 85-705 of the Department of Computing Science, Cornell University.

Voorhees, E.M. (1985b). The cluster hypothesis revisited. In *Proceedings of the 8th Annual ACM SIGIR Conference*, pp. 188-196. Montreal, Canada.

Voorhees, E.M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*, 22(6):465-476.

Voorhees, E.M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual ACM SIGIR Conference*, pp. 61-69. Dublin, Ireland.

Ward, J.H. (1963). Hierarchical grouping to minimize an objective function. *Journal of the American Statistical Association*, 58:236-244.

Watanabe, S. (1969). *Knowing and guessing: a quantitative study of inference and information*. New York: Wiley.

Weiss, R., Velez, B., Sheldon, M. (1996). HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of Hypertext '96*, pp. 180-193. Washington, DC.

Wen, J.R., Nie, J.Y., Zhang, H.J. (2001). Clustering user queries of a search engine. In *Proceedings of the 10th WWW Conference*, pp. 162-168. Hong Kong.

White, R., Ruthven I., Jose, J. (2002). The use of implicit evidence for relevance feedback in web retrieval. In *Proceedings of the 24th Annual BCS European Colloquium on Information Retrieval Research (ECIR 2002)*, pp. 93-109. Glasgow, Scotland.

Wilbur, W.J., and Coffee, L. (1994). The effectiveness of document neighbouring in search enhancement. *Information Processing & Management*, 30(2): 253-266.

Willett, P. (1981). A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing & Management*, 17:53-60.

Willett, P. (1983). Similarity coefficients and weighting functions for automatic document classification: an empirical comparison. *International Classification*, 3:138-142.

Willett, P. (1985). Query specific automatic document classification. *International Forum on Information and Documentation*, 10(2):28-32.

Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577-597.

Williams, W.T. and Clifford, H.T. (1971). On the comparison of two classifications on the same set of elements. *Taxon*, 20:519-522.

Williams, W.T., Clifford, H.T., Lance, G.T. (1971a). Group size dependence: a rationale for choice between numerical classifications. *Computer Journal*, 14:157-162.

Williams, W.T., Lance, G.N., Dale, M.B., Clifford, H.T. (1971b). Controversy concerning the criteria for taxonomic strategies. *Computer Journal*, 14:162-165.

Wishart, D. (1969). An algorithm for hierarchical classification. *Biometrics*, 25:165-170.

Wishart, D. (1998). Classifying single malt whiskies and other business applications of cluster analysis. In *Proceedings of PADD '98 Conference on Knowledge Discovery and Data Mining*. London, England.

Wu, M., Fuller, M., Wilkinson, R. (2001). Using clustering and classification approaches in interactive retrieval. *Information Processing & management*, 37(3):459-484.

Wulfekuhler, M.R. and Punch, W.F. (1997). Finding salient features for personal web page categories. In *Proceedings of the 6th International Conference on the World Wide Web*. Santa Clara, CA.

Xu, J. and Croft, W.B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 4-11. Zurich, Switzerland.

Yang, Y., Pierce, T., Carbonell, J. (1998). A study on retrospective and on-line event detection. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 28-36. Melbourne, Australia.

Yeung, M. and Yeo, B.L (1998). Segmentation of video by clustering and graph analysis. *Computer Vision And Image Understanding*, 71(1):94-109.

Yu, C.T., Wang, Y.T., Chen, C.H. (1985). Adaptive document clustering. In *Proceedings of the 8th Annual ACM SIGIR Conference*, pp. 197-203. Montreal, Canada.

Zamir, O. and Etzioni, O. (1988). Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 46-54. Melbourne, Australia.

Zhang, J. and Rasmussen, E.M. (2001). Developing a new similarity measure from two different perspectives. *Information Processing & Management*, 37(1):279-294.

Zhang, T., Ramakrishnan, R., Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 103-114. Montreal, Canada.

# Appendix A

In this Appendix I present similarity and distance coefficients that are typically used in document clustering.

The formulas in this Appendix calculate the similarity between two documents X=($x_1$, $x_2$, …, $x_n$) and Y=($y_1$, $y_2$, …, $y_n$). For each measure two formulas are given where available: one corresponding to the 2x2 contingency table given below, and one corresponding to the vector representation of the two documents. The formulas have been adapted from (Ellis *et al*., 1993).

| | $y_i$=1 | $y_i$=0 | |
|---|---|---|---|
| $x_i$=1 | a | b | a+b |
| $x_i$=0 | c | d | c+d |
| | a+c | b+d | *n* |

**The 2x2 contingency table**

Dice: $\frac{2a}{2a+b+c}$ $\frac{2\sum_{i=1}^{n}(W_{xi}W_{yi})}{\sum_{i=1}^{n}W_{xi}+\sum_{i=1}^{n}W_{yi}}$

Jaccard: $\frac{a}{a+b+c}$ $\frac{\sum_{i=1}^{n}W_{xi}W_{yi}}{\sum_{i=1}^{n}W_{xi}+\sum_{i=1}^{n}W_{yi}-\sum_{i=1}^{n}(W_{xi}W_{yi})}$

Overlap: $\frac{a}{\min(a+b,a+c)}$ $\frac{\sum_{i=1}^{n}(W_{xi}W_{yi})}{\min(\sum_{i=1}^{n}W_{xi},\sum_{i=1}^{n}W_{yi})}$

Asymmetric $\frac{a}{a+b}$ $\frac{\sum_{i=1}^{n}\min(W_{xi},W_{yi})}{\sum_{i=1}^{n}W_{xi}}$

Cosine: $\frac{a}{\sqrt{(a+b)(a+c)}}$ $\frac{\sum_{i=1}^{n} W_{xi} W_{yi}}{\sqrt{\sum_{k=1}^{n} W_{xk}^2} \sqrt{\sum_{j=1}^{n} W_{jk}^2}}$

Euclidean Distance: $\frac{\sqrt{b+c}}{norm}$ $\sum_{i=1}^{n} \sqrt{(W_{xi} - W_{yi})^2}$

Rogers & Tanimoto: $\frac{a+d}{(a+d)+2(b+c)}$ -

Matching: $\frac{a+d}{a+b+c+d}$ -

# Appendix B

In this Appendix I present results from Chapter 6: “The Effectiveness of Hierarchic Post-Retrieval Clustering”.

| Group Average | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.601 | **0.769** | **0.692** | 0.511 | **0.752** | **0.663** | 0.619 | 0.749 | 0.667 |
| **top200** | 0.606 | 0.787 | 0.701 | 0.514 | 0.778 | 0.685 | 0.604 | 0.741 | 0.663 |
| **top350** | 0.583 | 0.786 | 0.708 | 0.507 | 0.790 | 0.695 | 0.576 | 0.739 | 0.663 |
| **top500** | 0.583 | 0.790 | 0.703 | 0.508 | 0.792 | 0.692 | 0.560 | **0.735** | **0.652** |
| **top750** | 0.576 | 0.785 | 0.710 | 0.488 | 0.785 | 0.699 | 0.562 | 0.745 | 0.657 |
| **top1000** | **0.566** | 0.780 | 0.710 | **0.482** | 0.798 | 0.699 | **0.550** | 0.736 | 0.656 |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | **0.523** | 0.688 | **0.550** | **0.438** | 0.660 | 0.503 | **0.502** | **0.642** | 0.503 |
| **top200** | 0.540 | 0.695 | 0.550 | 0.476 | **0.646** | **0.498** | 0.512 | 0.651 | **0.501** |
| **top350** | 0.543 | 0.681 | 0.550 | 0.469 | 0.647 | 0.503 | 0.52 | 0.667 | 0.501 |
| **top500** | 0.548 | 0.686 | 0.550 | 0.461 | 0.660 | 0.503 | 0.54 | 0.667 | 0.501 |
| **top750** | 0.553 | 0.692 | 0.550 | 0.465 | 0.658 | 0.503 | 0.537 | 0.667 | 0.501 |
| **top1000** | 0.546 | **0.680** | 0.550 | 0.463 | 0.652 | 0.503 | 0.537 | 0.662 | 0.501 |
| **full** | 0.748 | 0.794 | 0.550 | 0.641 | 0.713 | 0.503 | 0.782 | 0.806 | 0.501 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.715 | **0.817** | 0.762 | 0.630 | 0.827 | 0.727 | 0.702 | 0.777 | 0.738 |
| **top200** | 0.697 | 0.827 | 0.750 | 0.609 | 0.820 | 0.729 | 0.658 | **0.741** | 0.699 |
| **top350** | 0.683 | 0.822 | **0.748** | 0.589 | **0.811** | **0.726** | 0.655 | 0.753 | 0.68 |
| **top500** | 0.681 | 0.819 | 0.748 | 0.593 | 0.815 | 0.726 | 0.656 | 0.765 | **0.676** |
| **top750** | **0.667** | 0.823 | 0.748 | **0.567** | 0.818 | 0.726 | **0.649** | 0.776 | 0.676 |
| **full** | 0.842 | 0.840 | 0.748 | 0.790 | 0.873 | 0.726 | 0.798 | 0.824 | 0.676 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.589 | **0.723** | 0.627 | 0.517 | 0.699 | **0.577** | 0.576 | 0.677 | 0.584 |
| **top200** | 0.587 | 0.734 | **0.626** | 0.504 | 0.695 | 0.577 | 0.559 | **0.672** | **0.580** |
| **top350** | 0.566 | 0.744 | 0.626 | 0.493 | **0.693** | 0.577 | 0.553 | 0.698 | 0.580 |
| **top500** | 0.58 | 0.746 | 0.626 | 0.487 | 0.717 | 0.577 | 0.568 | 0.721 | 0.580 |
| **top750** | 0.575 | 0.738 | 0.626 | 0.489 | 0.700 | 0.577 | 0.571 | 0.705 | 0.580 |
| **top1000** | **0.553** | 0.744 | 0.626 | **0.475** | 0.707 | 0.577 | **0.549** | 0.725 | 0.580 |
| **full** | 0.713 | 0.792 | 0.626 | 0.643 | 0.736 | 0.577 | 0.716 | 0.739 | 0.580 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.349 | 0.450 | **0.387** | 0.300 | **0.456** | **0.354** | 0.308 | **0.399** | **0.333** |
| **top200** | 0.326 | 0.455 | 0.387 | 0.281 | 0.468 | 0.354 | 0.294 | 0.413 | 0.333 |
| **top350** | **0.309** | **0.437** | 0.387 | 0.281 | 0.462 | 0.354 | **0.271** | 0.404 | 0.333 |
| **top500** | 0.311 | 0.443 | 0.387 | 0.279 | 0.471 | 0.354 | 0.273 | 0.399 | 0.333 |
| **top750** | 0.311 | 0.446 | 0.387 | **0.276** | 0.462 | 0.354 | 0.272 | 0.400 | 0.333 |
| **full** | 0.744 | 0.494 | 0.387 | 0.682 | 0.596 | 0.354 | 0.711 | 0.403 | 0.333 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.692 | 0.791 | 0.734 | 0.608 | 0.767 | 0.693 | 0.696 | 0.779 | 0.719 |
| **top200** | 0.670 | **0.782** | 0.721 | 0.604 | 0.762 | 0.690 | 0.661 | 0.741 | 0.686 |
| **top350** | 0.671 | 0.784 | 0.716 | 0.603 | **0.760** | **0.689** | 0.650 | 0.742 | 0.666 |
| **top500** | 0.668 | 0.795 | 0.715 | **0.585** | 0.774 | 0.689 | 0.642 | 0.731 | 0.659 |
| **top750** | **0.667** | 0.791 | **0.714** | 0.585 | 0.775 | 0.689 | **0.64** | **0.729** | 0.655 |
| **top1000** | 0.676 | 0.793 | 0.714 | 0.586 | 0.776 | 0.689 | 0.641 | 0.732 | **0.654** |

**Table B1.** Results using the group average method. Highest effectiveness for each column appears in bold.

| Ward | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.626 | **0.760** | **0.692** | 0.537 | **0.763** | **0.663** | 0.637 | 0.742 | 0.667 |
| **top200** | 0.611 | 0.777 | 0.701 | 0.533 | 0.789 | 0.685 | 0.607 | **0.738** | 0.663 |
| **top350** | 0.600 | 0.794 | 0.708 | 0.515 | 0.789 | 0.695 | 0.591 | 0.747 | 0.663 |
| **top500** | 0.607 | 0.794 | 0.703 | 0.527 | 0.805 | 0.692 | 0.588 | 0.752 | **0.652** |
| **top750** | 0.603 | 0.801 | 0.710 | 0.508 | 0.795 | 0.699 | 0.598 | 0.760 | 0.657 |
| **top1000** | **0.592** | 0.796 | 0.710 | **0.502** | 0.804 | 0.699 | **0.577** | 0.749 | 0.656 |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.556 | 0.710 | **0.550** | 0.462 | 0.665 | 0.503 | **0.530** | 0.655 | 0.503 |
| **top200** | 0.564 | **0.674** | 0.550 | 0.488 | 0.656 | **0.498** | 0.546 | **0.643** | **0.501** |
| **top350** | 0.571 | 0.686 | 0.550 | 0.485 | **0.645** | 0.503 | 0.558 | 0.669 | 0.501 |
| **top500** | **0.554** | 0.707 | 0.550 | **0.460** | 0.648 | 0.503 | 0.548 | 0.669 | 0.501 |
| **top750** | 0.560 | 0.693 | 0.550 | 0.476 | 0.663 | 0.503 | 0.546 | 0.685 | 0.501 |
| **top1000** | 0.572 | 0.692 | 0.550 | 0.479 | 0.665 | 0.503 | 0.570 | 0.685 | 0.501 |
| **full** | 0.742 | 0.786 | 0.550 | 0.641 | 0.695 | 0.503 | 0.773 | 0.760 | 0.501 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.727 | 0.824 | 0.762 | 0.645 | 0.825 | 0.727 | 0.711 | 0.766 | 0.738 |
| **top200** | 0.701 | **0.809** | 0.750 | 0.621 | 0.816 | 0.729 | 0.663 | **0.745** | 0.699 |
| **top350** | 0.695 | 0.823 | **0.748** | **0.596** | **0.811** | **0.726** | **0.651** | 0.752 | 0.680 |
| **top500** | 0.694 | 0.830 | 0.748 | 0.597 | 0.814 | 0.726 | 0.655 | 0.762 | **0.676** |
| **top750** | **0.688** | 0.835 | 0.748 | 0.601 | 0.827 | 0.726 | 0.659 | 0.763 | 0.676 |
| **full** | 0.844 | 0.869 | 0.748 | 0.785 | 0.877 | 0.726 | 0.796 | 0.817 | 0.676 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.598 | 0.740 | 0.627 | 0.520 | 0.709 | **0.577** | 0.585 | 0.700 | 0.584 |
| **top200** | 0.604 | 0.740 | **0.626** | 0.513 | 0.705 | 0.577 | 0.581 | **0.688** | **0.580** |
| **top350** | 0.582 | 0.731 | 0.626 | 0.506 | 0.715 | 0.577 | 0.574 | 0.708 | 0.580 |
| **top500** | **0.568** | **0.728** | 0.626 | **0.490** | **0.699** | 0.577 | 0.559 | 0.695 | 0.580 |
| **top750** | 0.568 | 0.744 | 0.626 | 0.491 | 0.723 | 0.577 | **0.555** | 0.703 | 0.580 |
| **top1000** | 0.574 | 0.745 | 0.626 | 0.500 | 0.702 | 0.577 | 0.568 | 0.716 | 0.580 |
| **full** | 0.715 | 0.797 | 0.626 | 0.643 | 0.738 | 0.577 | 0.726 | 0.780 | 0.580 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.439 | 0.525 | **0.387** | 0.352 | 0.480 | **0.354** | 0.394 | 0.448 | **0.333** |
| **top200** | 0.391 | 0.462 | 0.387 | 0.330 | 0.484 | 0.354 | 0.367 | **0.422** | 0.333 |
| **top350** | 0.376 | **0.453** | 0.387 | 0.322 | **0.445** | 0.354 | 0.360 | 0.425 | 0.333 |
| **top500** | **0.373** | 0.454 | 0.387 | **0.314** | 0.445 | 0.354 | **0.351** | 0.424 | 0.333 |
| **top750** | 0.375 | 0.453 | 0.387 | 0.319 | 0.448 | 0.354 | 0.359 | 0.428 | 0.333 |
| **full** | 0.765 | 0.531 | 0.387 | 0.681 | 0.615 | 0.354 | 0.753 | 0.431 | 0.333 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** | **MK1** | **MK1-k** | **MK3** |
| **top100** | 0.701 | 0.795 | 0.734 | 0.629 | 0.774 | 0.693 | 0.705 | 0.775 | 0.719 |
| **top200** | 0.689 | **0.777** | 0.721 | 0.614 | 0.770 | 0.690 | 0.676 | 0.734 | 0.686 |
| **top350** | 0.685 | 0.778 | 0.716 | 0.616 | **0.769** | **0.689** | 0.661 | 0.726 | 0.666 |
| **top500** | **0.679** | 0.781 | 0.715 | 0.614 | 0.775 | 0.689 | 0.656 | **0.722** | 0.659 |
| **top750** | 0.685 | 0.784 | **0.714** | 0.608 | 0.774 | 0.689 | **0.652** | 0.730 | 0.655 |
| **top1000** | 0.681 | 0.780 | 0.714 | **0.606** | 0.775 | 0.689 | 0.656 | 0.736 | **0.654** |

**Table B2.** Results using Ward's method. Highest effectiveness for each column appears in bold.

| *Complete Link* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.622 | **0.761** | **0.692** | 0.534 | 0.741 | **0.663** | 0.630 | 0.751 | 0.667 |
| **top200** | 0.614 | 0.792 | 0.701 | 0.527 | **0.701** | 0.685 | 0.603 | 0.747 | 0.663 |
| **top350** | 0.618 | 0.785 | 0.708 | 0.536 | 0.702 | 0.695 | 0.603 | 0.743 | 0.663 |
| **top500** | 0.620 | 0.795 | 0.703 | 0.536 | 0.697 | 0.692 | 0.598 | 0.746 | **0.652** |
| **top750** | 0.610 | 0.791 | 0.710 | 0.521 | 0.679 | 0.699 | 0.589 | **0.735** | 0.657 |
| **top1000** | **0.599** | 0.795 | 0.710 | **0.505** | 0.687 | 0.699 | **0.580** | 0.741 | 0.656 |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | **0.560** | 0.710 | **0.550** | **0.461** | 0.650 | 0.503 | **0.542** | **0.665** | 0.503 |
| **top200** | 0.588 | **0.696** | 0.550 | 0.490 | 0.662 | **0.498** | 0.577 | 0.665 | **0.501** |
| **top350** | 0.596 | 0.697 | 0.550 | 0.487 | 0.659 | 0.503 | 0.601 | 0.682 | 0.501 |
| **top500** | 0.573 | 0.700 | 0.550 | 0.466 | **0.632** | 0.503 | 0.594 | 0.686 | 0.501 |
| **top750** | 0.601 | 0.704 | 0.550 | 0.495 | 0.654 | 0.503 | 0.624 | 0.700 | 0.501 |
| **top1000** | 0.604 | 0.713 | 0.550 | 0.492 | 0.655 | 0.503 | 0.628 | 0.706 | 0.501 |
| **full** | 0.743 | 0.774 | 0.550 | 0.640 | 0.699 | 0.503 | 0.761 | 0.745 | 0.501 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.723 | 0.819 | 0.762 | 0.640 | **0.808** | 0.727 | 0.714 | 0.771 | 0.738 |
| **top200** | 0.708 | 0.822 | 0.750 | 0.623 | 0.820 | 0.729 | 0.667 | 0.738 | 0.699 |
| **top350** | **0.705** | **0.808** | **0.748** | 0.619 | 0.834 | **0.726** | **0.656** | **0.736** | 0.680 |
| **top500** | 0.716 | 0.839 | 0.748 | 0.612 | 0.826 | 0.726 | 0.671 | 0.754 | **0.676** |
| **top750** | 0.718 | 0.838 | 0.748 | **0.609** | 0.844 | 0.726 | 0.691 | 0.784 | 0.676 |
| **full** | 0.841 | 0.890 | 0.748 | 0.786 | 0.874 | 0.726 | 0.796 | 0.843 | 0.676 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.616 | 0.733 | 0.627 | 0.532 | **0.697** | **0.577** | 0.605 | **0.686** | 0.584 |
| **top200** | 0.589 | **0.726** | **0.626** | 0.493 | 0.706 | 0.577 | 0.600 | 0.699 | **0.580** |
| **top350** | 0.589 | 0.749 | 0.626 | 0.501 | 0.697 | 0.577 | 0.596 | 0.723 | 0.580 |
| **top500** | 0.588 | 0.755 | 0.626 | **0.482** | 0.716 | 0.577 | 0.604 | 0.750 | 0.580 |
| **top750** | **0.577** | 0.753 | 0.626 | 0.491 | 0.706 | 0.577 | **0.582** | 0.735 | 0.580 |
| **top1000** | 0.584 | 0.758 | 0.626 | 0.489 | 0.700 | 0.577 | 0.606 | 0.769 | 0.580 |
| **full** | 0.699 | 0.773 | 0.626 | 0.630 | 0.718 | 0.577 | 0.715 | 0.796 | 0.580 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.428 | 0.505 | **0.387** | 0.345 | 0.514 | **0.354** | **0.395** | **0.413** | **0.333** |
| **top200** | 0.416 | 0.485 | 0.387 | **0.331** | 0.489 | 0.354 | 0.405 | 0.440 | 0.333 |
| **top350** | **0.411** | **0.481** | 0.387 | 0.331 | **0.464** | 0.354 | 0.413 | 0.442 | 0.333 |
| **top500** | 0.411 | 0.499 | 0.387 | 0.331 | 0.467 | 0.354 | 0.401 | 0.443 | 0.333 |
| **top750** | 0.413 | 0.490 | 0.387 | 0.335 | 0.465 | 0.354 | 0.399 | 0.433 | 0.333 |
| **full** | 0.786 | 0.623 | 0.387 | 0.681 | 0.610 | 0.354 | 0.783 | 0.526 | 0.333 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.704 | 0.788 | 0.734 | 0.619 | 0.778 | 0.693 | 0.709 | 0.769 | 0.719 |
| **top200** | 0.696 | 0.783 | 0.721 | 0.620 | 0.770 | 0.690 | 0.686 | 0.733 | 0.686 |
| **top350** | **0.690** | **0.782** | 0.716 | 0.614 | **0.769** | **0.689** | **0.670** | 0.739 | 0.666 |
| **top500** | 0.699 | 0.791 | 0.715 | 0.616 | 0.777 | 0.689 | 0.675 | **0.732** | 0.659 |
| **top750** | 0.703 | 0.791 | 0.714 | **0.609** | 0.773 | 0.689 | 0.677 | 0.732 | 0.655 |
| **top1000** | 0.713 | 0.804 | **0.714** | 0.621 | 0.774 | 0.689 | 0.694 | 0.729 | **0.654** |

**Table B3.** Results using the complete link method. Highest effectiveness for each column appears in bold.

| Single Link | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.656 | **0.778** | **0.692** | 0.581 | **0.784** | **0.663** | 0.658 | 0.739 | 0.667 |
| **top200** | 0.647 | 0.784 | 0.701 | 0.566 | 0.796 | 0.685 | 0.640 | **0.735** | 0.663 |
| **top350** | **0.635** | 0.799 | 0.708 | 0.546 | 0.804 | 0.695 | **0.630** | 0.759 | 0.663 |
| **top500** | 0.653 | 0.815 | 0.703 | 0.567 | 0.813 | 0.692 | 0.649 | 0.771 | **0.652** |
| **top750** | 0.647 | 0.826 | 0.710 | 0.549 | 0.810 | 0.699 | 0.646 | 0.777 | 0.657 |
| **top1000** | 0.647 | 0.822 | 0.710 | **0.540** | 0.827 | 0.699 | 0.652 | 0.783 | 0.656 |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | **0.565** | 0.710 | **0.550** | **0.480** | **0.656** | 0.503 | **0.543** | **0.659** | 0.503 |
| **top200** | 0.585 | **0.709** | 0.550 | 0.510 | 0.662 | **0.498** | 0.571 | 0.693 | **0.501** |
| **top350** | 0.604 | 0.718 | 0.550 | 0.514 | 0.666 | 0.503 | 0.600 | 0.688 | 0.501 |
| **top500** | 0.597 | 0.714 | 0.550 | 0.496 | 0.667 | 0.503 | 0.611 | 0.708 | 0.501 |
| **top750** | 0.608 | 0.718 | 0.550 | 0.514 | 0.670 | 0.503 | 0.626 | 0.708 | 0.501 |
| **top1000** | 0.612 | 0.711 | 0.550 | 0.522 | 0.678 | 0.503 | 0.627 | 0.706 | 0.501 |
| **full** | 0.760 | 0.795 | 0.550 | 0.660 | 0.725 | 0.503 | 0.790 | 0.816 | 0.501 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.749 | **0.814** | 0.762 | 0.677 | **0.818** | 0.727 | 0.733 | 0.772 | 0.738 |
| **top200** | **0.719** | 0.820 | 0.750 | 0.657 | 0.830 | 0.729 | 0.669 | **0.745** | 0.699 |
| **top350** | 0.723 | 0.827 | **0.748** | 0.661 | 0.825 | 0.726 | **0.666** | 0.748 | 0.680 |
| **top500** | 0.728 | 0.835 | 0.748 | 0.661 | 0.833 | **0.726** | 0.677 | 0.755 | **0.676** |
| **top750** | 0.735 | 0.837 | 0.748 | **0.659** | 0.832 | 0.726 | 0.685 | 0.766 | 0.676 |
| **full** | 0.876 | 0.898 | 0.748 | 0.822 | 0.884 | 0.726 | 0.825 | 0.842 | 0.676 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | **0.664** | 0.766 | 0.627 | **0.587** | 0.753 | **0.577** | **0.647** | **0.709** | 0.584 |
| **top200** | 0.674 | 0.790 | **0.626** | 0.590 | 0.758 | 0.577 | 0.666 | 0.764 | **0.580** |
| **top350** | 0.686 | 0.780 | 0.626 | 0.598 | 0.756 | 0.577 | 0.684 | 0.766 | 0.580 |
| **top500** | 0.684 | **0.758** | 0.626 | 0.591 | 0.739 | 0.577 | 0.697 | 0.760 | 0.580 |
| **top750** | 0.706 | 0.789 | 0.626 | 0.603 | 0.741 | 0.577 | 0.732 | 0.777 | 0.580 |
| **top1000** | 0.696 | 0.774 | 0.626 | 0.595 | **0.729** | 0.577 | 0.722 | 0.765 | 0.580 |
| **full** | 0.746 | 0.804 | 0.626 | 0.670 | 0.734 | 0.577 | 0.766 | 0.814 | 0.580 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | **0.395** | **0.476** | **0.387** | 0.323 | 0.460 | **0.354** | **0.376** | **0.420** | **0.333** |
| **top200** | 0.410 | 0.498 | 0.387 | 0.317 | 0.455 | 0.354 | 0.417 | 0.454 | 0.333 |
| **top350** | 0.397 | 0.482 | 0.387 | **0.305** | 0.447 | 0.354 | 0.414 | 0.454 | 0.333 |
| **top500** | 0.401 | 0.487 | 0.387 | 0.309 | 0.452 | 0.354 | 0.416 | 0.456 | 0.333 |
| **top750** | 0.408 | 0.493 | 0.387 | 0.312 | **0.451** | 0.354 | 0.417 | 0.451 | 0.333 |
| **full** | 0.791 | 0.572 | 0.387 | 0.704 | 0.646 | 0.354 | 0.776 | 0.484 | 0.333 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** | **MK1** | **MK1-K** | **MK3** |
| **top100** | 0.733 | 0.796 | 0.734 | 0.654 | 0.777 | 0.693 | 0.724 | 0.772 | 0.719 |
| **top200** | 0.733 | 0.795 | 0.721 | **0.653** | 0.776 | 0.690 | 0.709 | **0.740** | 0.686 |
| **top350** | 0.725 | **0.794** | 0.716 | 0.656 | **0.766** | **0.689** | 0.701 | 0.747 | 0.666 |
| **top500** | **0.724** | 0.796 | 0.715 | 0.654 | 0.775 | 0.689 | **0.693** | 0.744 | 0.659 |
| **top750** | 0.736 | 0.803 | **0.714** | 0.655 | 0.796 | 0.689 | 0.704 | 0.746 | 0.655 |
| **top1000** | 0.736 | 0.807 | 0.714 | 0.653 | 0.807 | 0.689 | 0.715 | 0.756 | **0.654** |

**Table B4.** Results using the single link method. Highest effectiveness for each column appears in bold.

| AP | | | | WSJ | | | |
|---|---|---|---|---|---|---|---|
| **n** | **β=1** | **β=2** | **β=0.5** | **n** | **β=1** | **β=2** | **β=0.5** |
| 100 | 0.633 | 0.628 | 0.550 | 100 | 0.715 | 0.712 | 0.645 |
| 200 | **0.626** | 0.613 | **0.543** | 200 | 0.701 | 0.679 | 0.640 |
| 350 | 0.631 | 0.611 | 0.552 | 350 | 0.696 | 0.659 | 0.638 |
| 500 | 0.629 | 0.605 | 0.552 | 500 | 0.694 | 0.651 | 0.636 |
| 750 | 0.629 | 0.605 | 0.548 | 750 | **0.693** | 0.647 | **0.633** |
| 1000 | 0.629 | **0.604** | 0.548 | 1000 | 0.693 | **0.646** | 0.633 |

**Table B5.** Results for the MK4 measure using AP and WSJ. Highest effectiveness for each column appears in bold.

| CACM | | | | LISA | | | |
|---|---|---|---|---|---|---|---|
| **n** | **β=1** | **β=2** | **β=0.5** | **n** | **β=1** | **β=2** | **β=0.5** |
| 100 | 0.537 | 0.500 | 0.448 | 100 | 0.575 | 0.570 | 0.438 |
| 200 | 0.540 | 0.497 | 0.448 | 200 | 0.559 | 0.549 | 0.420 |
| 350 | **0.535** | **0.492** | **0.444** | 350 | 0.541 | **0.529** | **0.400** |
| 500 | 0.535 | 0.492 | 0.444 | 500 | **0.540** | 0.529 | 0.400 |
| 750 | 0.535 | 0.492 | 0.444 | 750 | 0.540 | 0.529 | 0.400 |
| 1000 | 0.535 | 0.492 | 0.444 | 1000 | 0.540 | 0.529 | 0.400 |
| full | 0.535 | 0.492 | 0.444 | full | 0.540 | 0.529 | 0.400 |

**Table B6.** Results for the MK$ measure using CACM and LISA. Highest effectiveness for each column appears in bold.

| CISI | | | | MED | | | |
|---|---|---|---|---|---|---|---|
| **n** | **β=1** | **β=2** | **β=0.5** | **n** | **β=1** | **β=2** | **β=0.5** |
| 100 | 0.726 | 0.717 | 0.651 | top100 | **0.381** | **0.331** | **0.327** |
| 200 | 0.710 | 0.674 | 0.651 | top200 | 0.381 | 0.331 | 0.327 |
| 350 | **0.704** | 0.653 | 0.639 | top350 | 0.381 | 0.331 | 0.327 |
| 500 | 0.704 | 0.649 | 0.639 | top500 | 0.381 | 0.331 | 0.327 |
| 750 | 0.704 | **0.648** | **0.638** | top750 | 0.381 | 0.331 | 0.327 |
| full | 0.704 | 0.648 | 0.638 | full | 0.381 | 0.331 | 0.327 |

**Table B7.** Results for the MK4 measure using CISI and MED. Highest effectiveness for each column appears in bold.

| AP | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
|---|---|---|---|---|---|---|
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.601 | 0.704 | 0.511 | 0.731 | 0.619 | 0.744 |
| **top200** | 0.606 | 0.727 | 0.514 | 0.76 | 0.604 | 0.763 |
| **top350** | 0.583 | 0.739 | 0.507 | 0.771 | 0.576 | 0.769 |
| **top500** | 0.583 | 0.755 | 0.508 | 0.788 | 0.560 | 0.793 |
| **top750** | 0.576 | 0.778 | 0.488 | 0.804 | 0.562 | 0.807 |
| **top1000** | 0.566 | 0.803 | 0.482 | 0.818 | 0.550 | 0.826 |
| **CACM** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.523 | 0.671 | 0.438 | 0.619 | 0.502 | 0.608 |
| **top200** | 0.54 | 0.718 | 0.476 | 0.651 | 0.512 | 0.672 |
| **top350** | 0.543 | 0.739 | 0.469 | 0.666 | 0.52 | 0.712 |
| **top500** | 0.548 | 0.749 | 0.461 | 0.674 | 0.54 | 0.736 |
| **top750** | 0.553 | 0.759 | 0.465 | 0.682 | 0.537 | 0.757 |
| **top1000** | 0.546 | 0.765 | 0.463 | 0.690 | 0.537 | 0.767 |
| **full** | 0.748 | 0.828 | 0.641 | 0.757 | 0.782 | 0.831 |
| **CISI** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.715 | 0.755 | 0.63 | 0.727 | 0.702 | 0.722 |
| **top200** | 0.697 | 0.753 | 0.609 | 0.732 | 0.658 | 0.685 |
| **top350** | 0.683 | 0.775 | 0.589 | 0.748 | 0.655 | 0.691 |
| **top500** | 0.681 | 0.795 | 0.593 | 0.761 | 0.656 | 0.712 |
| **top750** | 0.667 | 0.814 | 0.567 | 0.772 | 0.649 | 0.741 |
| **full** | 0.842 | 0.869 | 0.79 | 0.821 | 0.798 | 0.815 |
| **LISA** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.589 | 0.690 | 0.517 | 0.633 | 0.576 | 0.650 |
| **top200** | 0.587 | 0.694 | 0.504 | 0.631 | 0.559 | 0.664 |
| **top350** | 0.566 | 0.686 | 0.493 | 0.625 | 0.553 | 0.667 |
| **top500** | 0.58 | 0.696 | 0.487 | 0.636 | 0.568 | 0.681 |
| **top750** | 0.575 | 0.701 | 0.489 | 0.642 | 0.571 | 0.695 |
| **top1000** | 0.553 | 0.707 | 0.475 | 0.646 | 0.549 | 0.704 |
| **full** | 0.713 | 0.758 | 0.643 | 0.711 | 0.716 | 0.751 |
| **MED** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.349 | 0.652 | 0.3 | 0.609 | 0.308 | 0.506 |
| **top200** | 0.326 | 0.732 | 0.281 | 0.661 | 0.294 | 0.619 |
| **top350** | 0.309 | 0.768 | 0.281 | 0.687 | 0.271 | 0.688 |
| **top500** | 0.311 | 0.777 | 0.279 | 0.692 | 0.273 | 0.711 |
| **top750** | 0.311 | 0.781 | 0.276 | 0.701 | 0.272 | 0.719 |
| **full** | 0.744 | 0.816 | 0.682 | 0.748 | 0.711 | 0.767 |
| **WSJ** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.692 | 0.776 | 0.608 | 0.735 | 0.696 | 0.748 |
| **top200** | 0.67 | 0.790 | 0.604 | 0.758 | 0.661 | 0.744 |
| **top350** | 0.671 | 0.802 | 0.603 | 0.770 | 0.650 | 0.748 |
| **top500** | 0.668 | 0.818 | 0.585 | 0.783 | 0.642 | 0.759 |
| **top750** | 0.667 | 0.837 | 0.585 | 0.799 | 0.640 | 0.780 |
| **top1000** | 0.676 | 0.849 | 0.586 | 0.808 | 0.641 | 0.797 |

**Table B8.** Random and actual effectiveness values using the group average method

| AP | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
|---|---|---|---|---|---|---|
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.626 | 0.704 | 0.537 | 0.731 | 0.637 | 0.745 |
| **top200** | 0.611 | 0.727 | 0.533 | 0.761 | 0.607 | 0.763 |
| **top350** | 0.600 | 0.738 | 0.515 | 0.770 | 0.591 | 0.770 |
| **top500** | 0.607 | 0.756 | 0.527 | 0.786 | 0.588 | 0.794 |
| **top750** | 0.603 | 0.776 | 0.508 | 0.805 | 0.598 | 0.807 |
| **top1000** | 0.592 | 0.803 | 0.502 | 0.819 | 0.577 | 0.825 |
| **CACM** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.556 | 0.671 | 0.462 | 0.618 | 0.530 | 0.608 |
| **top200** | 0.564 | 0.717 | 0.488 | 0.650 | 0.546 | 0.672 |
| **top350** | 0.571 | 0.738 | 0.485 | 0.665 | 0.558 | 0.711 |
| **top500** | 0.554 | 0.749 | 0.460 | 0.673 | 0.548 | 0.735 |
| **top750** | 0.560 | 0.759 | 0.476 | 0.682 | 0.546 | 0.756 |
| **top1000** | 0.572 | 0.765 | 0.479 | 0.690 | 0.570 | 0.767 |
| **full** | 0.742 | 0.828 | 0.641 | 0.758 | 0.773 | 0.832 |
| **CISI** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.727 | 0.756 | 0.645 | 0.728 | 0.711 | 0.722 |
| **top200** | 0.701 | 0.753 | 0.621 | 0.731 | 0.663 | 0.685 |
| **top350** | 0.695 | 0.775 | 0.596 | 0.748 | 0.651 | 0.692 |
| **top500** | 0.694 | 0.795 | 0.597 | 0.762 | 0.655 | 0.711 |
| **top750** | 0.688 | 0.814 | 0.601 | 0.772 | 0.659 | 0.740 |
| **full** | 0.844 | 0.870 | 0.785 | 0.821 | 0.796 | 0.815 |
| **LISA** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.598 | 0.691 | 0.520 | 0.633 | 0.585 | 0.652 |
| **top200** | 0.604 | 0.692 | 0.513 | 0.629 | 0.581 | 0.663 |
| **top350** | 0.582 | 0.686 | 0.506 | 0.626 | 0.574 | 0.667 |
| **top500** | 0.569 | 0.695 | 0.490 | 0.635 | 0.559 | 0.681 |
| **top750** | 0.568 | 0.702 | 0.491 | 0.642 | 0.555 | 0.695 |
| **top1000** | 0.574 | 0.707 | 0.500 | 0.646 | 0.568 | 0.705 |
| **full** | 0.715 | 0.756 | 0.643 | 0.712 | 0.726 | 0.751 |
| **MED** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.439 | 0.652 | 0.352 | 0.606 | 0.394 | 0.507 |
| **top200** | 0.391 | 0.733 | 0.330 | 0.662 | 0.367 | 0.619 |
| **top350** | 0.376 | 0.768 | 0.322 | 0.685 | 0.360 | 0.688 |
| **top500** | 0.373 | 0.776 | 0.314 | 0.692 | 0.351 | 0.710 |
| **top750** | 0.375 | 0.781 | 0.319 | 0.700 | 0.359 | 0.719 |
| **full** | 0.765 | 0.814 | 0.681 | 0.747 | 0.753 | 0.766 |
| **WSJ** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.701 | 0.775 | 0.629 | 0.734 | 0.705 | 0.748 |
| **top200** | 0.689 | 0.790 | 0.614 | 0.758 | 0.676 | 0.744 |
| **top350** | 0.685 | 0.803 | 0.616 | 0.770 | 0.661 | 0.748 |
| **top500** | 0.679 | 0.816 | 0.614 | 0.783 | 0.656 | 0.758 |
| **top750** | 0.685 | 0.837 | 0.608 | 0.799 | 0.652 | 0.780 |
| **top1000** | 0.681 | 0.849 | 0.606 | 0.808 | 0.656 | 0.797 |

**Table B9.** Random and actual effectiveness values using Ward's method

| AP | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
|---|---|---|---|---|---|---|
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.622 | 0.705 | 0.534 | 0.731 | 0.630 | 0.744 |
| **top200** | 0.614 | 0.729 | 0.527 | 0.761 | 0.603 | 0.764 |
| **top350** | 0.618 | 0.738 | 0.536 | 0.772 | 0.603 | 0.770 |
| **top500** | 0.620 | 0.758 | 0.536 | 0.787 | 0.598 | 0.795 |
| **top750** | 0.610 | 0.777 | 0.521 | 0.804 | 0.589 | 0.806 |
| **top1000** | 0.599 | 0.804 | 0.505 | 0.820 | 0.580 | 0.825 |
| **CACM** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.560 | 0.671 | 0.461 | 0.620 | 0.542 | 0.607 |
| **top200** | 0.588 | 0.715 | 0.490 | 0.650 | 0.577 | 0.671 |
| **top350** | 0.596 | 0.738 | 0.487 | 0.665 | 0.601 | 0.712 |
| **top500** | 0.573 | 0.749 | 0.466 | 0.674 | 0.594 | 0.736 |
| **top750** | 0.601 | 0.759 | 0.495 | 0.682 | 0.624 | 0.756 |
| **top1000** | 0.604 | 0.765 | 0.492 | 0.690 | 0.628 | 0.767 |
| **full** | 0.743 | 0.828 | 0.640 | 0.759 | 0.761 | 0.833 |
| **CISI** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.723 | 0.756 | 0.640 | 0.729 | 0.714 | 0.722 |
| **top200** | 0.708 | 0.753 | 0.623 | 0.731 | 0.667 | 0.685 |
| **top350** | 0.705 | 0.775 | 0.619 | 0.749 | 0.656 | 0.692 |
| **top500** | 0.716 | 0.794 | 0.612 | 0.762 | 0.671 | 0.712 |
| **top750** | 0.718 | 0.815 | 0.609 | 0.772 | 0.691 | 0.741 |
| **full** | 0.841 | 0.870 | 0.786 | 0.822 | 0.796 | 0.814 |
| **LISA** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.616 | 0.691 | 0.532 | 0.633 | 0.605 | 0.651 |
| **top200** | 0.589 | 0.691 | 0.493 | 0.630 | 0.600 | 0.659 |
| **top350** | 0.589 | 0.686 | 0.501 | 0.625 | 0.596 | 0.666 |
| **top500** | 0.588 | 0.695 | 0.482 | 0.635 | 0.604 | 0.680 |
| **top750** | 0.577 | 0.702 | 0.491 | 0.642 | 0.582 | 0.696 |
| **top1000** | 0.584 | 0.707 | 0.489 | 0.646 | 0.606 | 0.705 |
| **full** | 0.699 | 0.755 | 0.630 | 0.713 | 0.715 | 0.751 |
| **MED** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.428 | 0.651 | 0.345 | 0.612 | 0.395 | 0.506 |
| **top200** | 0.416 | 0.734 | 0.331 | 0.662 | 0.405 | 0.620 |
| **top350** | 0.411 | 0.768 | 0.331 | 0.685 | 0.413 | 0.690 |
| **top500** | 0.410 | 0.776 | 0.330 | 0.692 | 0.401 | 0.711 |
| **top750** | 0.413 | 0.782 | 0.335 | 0.699 | 0.399 | 0.719 |
| **full** | 0.786 | 0.814 | 0.681 | 0.746 | 0.783 | 0.765 |
| **WSJ** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.704 | 0.776 | 0.619 | 0.735 | 0.709 | 0.749 |
| **top200** | 0.696 | 0.790 | 0.620 | 0.758 | 0.686 | 0.744 |
| **top350** | 0.690 | 0.803 | 0.614 | 0.770 | 0.670 | 0.748 |
| **top500** | 0.699 | 0.817 | 0.616 | 0.782 | 0.675 | 0.758 |
| **top750** | 0.703 | 0.836 | 0.609 | 0.799 | 0.677 | 0.779 |
| **top1000** | 0.713 | 0.849 | 0.621 | 0.806 | 0.694 | 0.798 |

**Table B10.** Random and actual effectiveness using the complete link method.

| AP | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
|---|---|---|---|---|---|---|
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.656 | 0.707 | 0.581 | 0.740 | 0.658 | 0.744 |
| **top200** | 0.647 | 0.732 | 0.566 | 0.774 | 0.640 | 0.766 |
| **top350** | 0.635 | 0.741 | 0.546 | 0.786 | 0.630 | 0.773 |
| **top500** | 0.653 | 0.759 | 0.567 | 0.802 | 0.649 | 0.799 |
| **top750** | 0.647 | 0.783 | 0.549 | 0.815 | 0.646 | 0.808 |
| **top1000** | 0.647 | 0.808 | 0.540 | 0.833 | 0.652 | 0.826 |
| **CACM** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.565 | 0.698 | 0.480 | 0.655 | 0.543 | 0.625 |
| **top200** | 0.585 | 0.743 | 0.510 | 0.681 | 0.571 | 0.691 |
| **top350** | 0.604 | 0.767 | 0.514 | 0.696 | 0.600 | 0.735 |
| **top500** | 0.597 | 0.778 | 0.496 | 0.704 | 0.611 | 0.762 |
| **top750** | 0.608 | 0.787 | 0.514 | 0.709 | 0.626 | 0.785 |
| **top1000** | 0.612 | 0.793 | 0.522 | 0.716 | 0.627 | 0.797 |
| **full** | 0.760 | 0.837 | 0.660 | 0.776 | 0.790 | 0.849 |
| **CISI** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.749 | 0.762 | 0.677 | 0.740 | 0.733 | 0.723 |
| **top200** | 0.719 | 0.764 | 0.657 | 0.751 | 0.669 | 0.692 |
| **top350** | 0.723 | 0.789 | 0.661 | 0.769 | 0.666 | 0.700 |
| **top500** | 0.728 | 0.809 | 0.66 | 0.780 | 0.677 | 0.720 |
| **top750** | 0.735 | 0.832 | 0.659 | 0.795 | 0.685 | 0.753 |
| **full** | 0.876 | 0.884 | 0.821 | 0.843 | 0.825 | 0.831 |
| **LISA** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.664 | 0.725 | 0.587 | 0.672 | 0.647 | 0.676 |
| **top200** | 0.674 | 0.735 | 0.590 | 0.675 | 0.666 | 0.698 |
| **top350** | 0.686 | 0.736 | 0.598 | 0.676 | 0.684 | 0.715 |
| **top500** | 0.684 | 0.737 | 0.591 | 0.676 | 0.697 | 0.725 |
| **top750** | 0.706 | 0.747 | 0.603 | 0.684 | 0.732 | 0.745 |
| **top1000** | 0.696 | 0.755 | 0.595 | 0.692 | 0.722 | 0.759 |
| **full** | 0.746 | 0.782 | 0.670 | 0.739 | 0.766 | 0.794 |
| **MED** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.395 | 0.650 | 0.323 | 0.633 | 0.376 | 0.495 |
| **top200** | 0.410 | 0.742 | 0.317 | 0.690 | 0.417 | 0.611 |
| **top350** | 0.397 | 0.780 | 0.305 | 0.709 | 0.414 | 0.685 |
| **top500** | 0.401 | 0.790 | 0.309 | 0.714 | 0.416 | 0.711 |
| **top750** | 0.408 | 0.795 | 0.312 | 0.720 | 0.417 | 0.720 |
| **full** | 0.791 | 0.827 | 0.704 | 0.769 | 0.776 | 0.788 |
| **WSJ** | $\beta$=1 | | $\beta$=0.5 | | $\beta$=2 | |
| | **MK1** | **Random** | **MK1** | **Random** | **MK1** | **Random** |
| **top100** | 0.733 | 0.780 | 0.654 | 0.746 | 0.724 | 0.748 |
| **top200** | 0.733 | 0.795 | 0.653 | 0.769 | 0.709 | 0.745 |
| **top350** | 0.725 | 0.810 | 0.656 | 0.785 | 0.701 | 0.750 |
| **top500** | 0.724 | 0.823 | 0.654 | 0.797 | 0.693 | 0.761 |
| **top750** | 0.736 | 0.844 | 0.655 | 0.813 | 0.704 | 0.784 |
| **top1000** | 0.736 | 0.855 | 0.650 | 0.819 | 0.715 | 0.801 |

**Table B11.** Random and actual effectiveness using the single link method.

# Appendix C

In this Appendix I present results from Chapter 7: “Query Sensitive Similarity Measures”.

## I. The effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$

| n | 4:1 | 1:1 | 1:2 | 1:4 | 1:5 | 1:7 | 1:9 | M2 |
|---|---|---|---|---|---|---|---|---|
| 100 | **2.440** | **2.561** | **2.633** | **2.652** | **2.640** | **2.574** | **2.566** | **2.079** |
| 200 | 2.218 | 2.321 | 2.354 | 2.404 | 2.377 | 2.354 | 2.316 | 1.834 |
| 350 | 2.155 | 2.244 | 2.339 | 2.359 | 2.351 | 2.313 | 2.298 | 1.671 |
| 500 | 2.143 | 2.263 | 2.353 | 2.387 | 2.398 | 2.393 | 2.319 | 1.663 |
| 750 | 2.167 | 2.293 | 2.372 | 2.431 | 2.412 | 2.390 | 2.333 | 1.605 |
| 1000 | 2.070 | 2.209 | 2.290 | 2.337 | 2.334 | 2.270 | 2.232 | 1.517 |

**Table C1.** Effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$ for AP

| n | 4:1 | 1:1 | 1:2 | 1:4 | 1:5 | 1:7 | 1:9 | M2 |
|---|---|---|---|---|---|---|---|---|
| 100 | **1.706** | **1.840** | 1.890 | 1.923 | 1.934 | 1.911 | 1.899 | 1.754 |
| 200 | 1.578 | 1.739 | 1.922 | 1.995 | 2.036 | 2.040 | 2.045 | **1.902** |
| 350 | 1.531 | 1.766 | 1.987 | **2.058** | **2.070** | **2.073** | **2.074** | 1.875 |
| 500 | 1.540 | 1.757 | **2.007** | 2.049 | 2.037 | 2.051 | 2.040 | 1.850 |
| 750 | 1.520 | 1.768 | 1.987 | 2.022 | 2.019 | 2.006 | 2.001 | 1.761 |
| 1000 | 1.506 | 1.772 | 1.971 | 1.998 | 2.003 | 1.987 | 1.966 | 1.731 |
| full | 1.443 | 1.534 | 1.687 | 1.850 | 1.868 | 1.873 | 1.78 | 1.655 |

**Table C2.** Effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$ for CACM

| n | 4:1 | 1:1 | 1:2 | 1:4 | 1:5 | 1:7 | 1:9 | M2 |
|---|---|---|---|---|---|---|---|---|
| 100 | **1.578** | **1.698** | **1.722** | **1.746** | **1.744** | 1.761 | **1.757** | 1.703 |
| 200 | 1.456 | 1.574 | 1.631 | 1.724 | **1.744** | **1.789** | 1.719 | **1.733** |
| 350 | 1.334 | 1.494 | 1.626 | 1.674 | 1.703 | 1.692 | 1.697 | 1.555 |
| 500 | 1.284 | 1.476 | 1.593 | 1.651 | 1.669 | 1.652 | 1.636 | 1.436 |
| 750 | 1.227 | 1.442 | 1.538 | 1.599 | 1.593 | 1.575 | 1.555 | 1.357 |
| full | 1.224 | 1.315 | 1.321 | 1.330 | 1.338 | 1.442 | 1.391 | 1.328 |

**Table C3.** Effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$ for CISI

| n | 4:1 | 1:1 | 1:2 | 1:4 | 1:5 | 1:7 | 1:9 | M2 |
|---|---|---|---|---|---|---|---|---|
| 100 | **0.990** | 1.206 | 1.352 | 1.383 | 1.384 | 1.402 | 1.392 | **1.395** |
| 200 | 0.972 | 1.195 | 1.311 | 1.327 | 1.372 | 1.390 | 1.391 | 1.269 |
| 350 | 0.930 | 1.199 | 1.335 | **1.418** | 1.430 | **1.429** | **1.420** | 1.315 |
| 500 | 0.938 | **1.237** | **1.374** | 1.415 | **1.446** | 1.423 | 1.403 | 1.317 |
| 750 | 0.940 | 1.208 | 1.358 | 1.395 | 1.405 | 1.421 | 1.413 | 1.287 |
| 1000 | 0.910 | 1.204 | 1.341 | 1.384 | 1.388 | 1.393 | 1.385 | 1.303 |
| full | 0.946 | 1.177 | 1.303 | 1.332 | 1.346 | 1.388 | 1.341 | 1.289 |

**Table C4.** Effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$ for LISA

| n | 4:1 | 1:1 | 1:2 | 1:4 | 1:5 | 1:7 | 1:9 | M2 |
|---|---|---|---|---|---|---|---|---|
| 100 | **3.255** | **3.463** | **3.550** | **3.566** | **3.564** | **3.576** | **3.537** | 3.361 |
| 200 | 3.198 | 3.405 | 3.507 | 3.525 | 3.525 | 3.532 | 3.528 | **3.367** |
| 350 | 3.190 | 3.352 | 3.470 | 3.482 | 3.478 | 3.476 | 3.461 | 3.310 |
| 500 | 3.187 | 3.340 | 3.456 | 3.442 | 3.450 | 3.436 | 3.424 | 3.305 |
| 750 | 3.190 | 3.346 | 3.452 | 3.429 | 3.440 | 3.431 | 3.421 | 3.285 |
| full | 3.111 | 3.116 | 3.201 | 3.204 | 3.210 | 3.216 | 3.213 | 3.124 |

**Table C5.** Effectiveness of M3 as a function of the ratio $\vartheta_1$:$\vartheta_2$ for Medline

| n | 4:1 | 1:1 | 1:2 | 1:4 | 1:5 | 1:7 | 1:9 | M2 |
|---|---|---|---|---|---|---|---|---|
| 100 | **2.177** | **2.314** | 2.344 | 2.354 | 2.348 | 2.317 | 2.249 | **1.872** |
| 200 | 2.123 | 2.306 | **2.431** | **2.443** | **2.390** | 2.328 | **2.280** | 1.827 |
| 350 | 1.989 | 2.226 | 2.391 | 2.389 | 2.370 | 2.318 | 2.255 | 1.832 |
| 500 | 1.958 | 2.190 | 2.351 | 2.377 | 2.355 | **2.329** | 2.252 | 1.856 |
| 750 | 1.840 | 2.087 | 2.262 | 2.300 | 2.299 | 2.257 | 2.204 | 1.838 |
| 1000 | 1.790 | 2.046 | 2.218 | 2.269 | 2.267 | 2.222 | 2.162 | 1.799 |

**Table C6.** Effectiveness of M3 as a function of the ratio $\vartheta_1$:$\vartheta_2$ for WSJ

## II. Results for the 1NN test

| AP | | | | | CISI | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **n** | **Cosine** | **M1** | **M2** | **M3** | **n** | **Cosine** | **M1** | **M2** | **M3** |
| 100 | **68.98%** | **71.53%** | **50.18%** | **71.17%** | 100 | **45.44%** | **52.11%** | 55.79% | **55.79%** |
| 200 | 66.29% | 71.72% | 45.33% | 70.33% | 200 | 39.98% | 49.25% | **56.20%** | 55.39% |
| 350 | 64.16% | 69.40% | 44.14% | 68.16% | 350 | 35.75% | 47.88% | 54.34% | 52.92% |
| 500 | 64.06% | 70.07% | 43.78% | 68.43% | 500 | 33.87% | 46.53% | 50.85% | 51.08% |
| 750 | 62.39% | 67.86% | 42.88% | 65.47% | 750 | 32.82% | 45.10% | 44.77% | 48.21% |
| 1000 | 62.12% | 66.97% | 42.15% | 64.25% | full | 32.85% | 41.30% | 37.05% | 42.79% |

**Table C7.** Results for the 1NN test using AP and CISI

| CACM | | | | | LISA | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **n** | **Cosine** | **M1** | **M2** | **M3** | **n** | **Cosine** | **M1** | **M2** | **M3** |
| 100 | **51.94%** | 58.78% | 55.82% | 60.26% | 100 | **30.30%** | 46.32% | **49.35%** | 47.62% |
| 200 | 45.92% | 58.74% | 59.90% | 63.56% | 200 | 27.68% | 43.60% | 44.64% | 48.79% |
| 350 | 45.97% | 59.36% | **60.58%** | **65.60%** | 350 | 26.27% | 45.89% | 45.89% | **49.05%** |
| 500 | 46.35% | 58.48% | 59.65% | 65.20% | 500 | 27.43% | **47.20%** | 45.13% | 47.20% |
| 750 | 44.95% | **60.17%** | 57.75% | 64.44% | 750 | 27.20% | 44.76% | 46.18% | 48.44% |
| 1000 | 43.30% | 59.36% | 55.59% | 62.15% | 1000 | 28.21% | 43.85% | 46.65% | 47.77% |
| full | 43.76% | 54.48% | 50.95% | 56.22% | full | 28.27% | 44.53% | 43.47% | 46.19% |

**Table C8.** Results for the 1NN test using CACM and LISA

| Medline | | | | | WSJ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **n** | **Cosine** | **M1** | **M2** | **M3** | **n** | **Cosine** | **M1** | **M2** | **M3** |
| 100 | **71.88%** | **80.49%** | 79.44% | **80.84%** | 100 | **64.41%** | **67.42%** | **56.02%** | **65.16%** |
| 200 | 67.43% | 76.76% | **80.20%** | 79.87% | 200 | 57.24% | 62.10% | 49.70% | 61.67% |
| 350 | 68.78% | 76.39% | 78.92% | 78.76% | 350 | 54.05% | 63.73% | 50.00% | 60.72% |
| 500 | 68.35% | 76.06% | 78.58% | 78.58% | 500 | 52.65% | 62.90% | 48.64% | 58.83% |
| 750 | 68.23% | 76.06% | 78.09% | 78.56% | 750 | 49.19% | 61.82% | 48.18% | 57.32% |
| full | 68.39% | 72.41% | 69.83% | 73.62% | 1000 | 47.60% | 60.43% | 47.73% | 55.76% |

**Table C9.** Results for the 1NN test using Medline and WSJ

# Appendix D

In this Appendix I present results from Chapter 8: “Hierarchic Document Clustering Using Query-Sensitive Similarity Measures”.

| Group Average | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.605 | 0.645 | 0.601 | 0.527 | 0.573 | 0.520 | 0.613 | 0.643 | 0.618 |
| **top200** | 0.583 | 0.620 | 0.589 | 0.509 | 0.567 | 0.511 | 0.576 | 0.603 | 0.580 |
| **top350** | 0.560 | 0.611 | 0.564 | 0.494 | **0.557** | 0.497 | 0.545 | 0.580 | 0.551 |
| **top500** | 0.552 | 0.611 | 0.560 | 0.477 | 0.562 | 0.486 | 0.522 | 0.575 | 0.536 |
| **top750** | **0.530** | 0.610 | 0.543 | 0.452 | 0.564 | 0.477 | **0.505** | 0.569 | 0.518 |
| **top1000** | 0.531 | **0.606** | **0.532** | **0.448** | 0.561 | **0.467** | 0.513 | **0.565** | **0.512** |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.506 | 0.545 | 0.521 | 0.435 | 0.468 | 0.438 | 0.472 | 0.509 | 0.490 |
| **top200** | 0.500 | 0.505 | 0.503 | 0.426 | 0.417 | 0.418 | 0.476 | 0.480 | 0.484 |
| **top350** | 0.482 | 0.495 | 0.488 | 0.412 | 0.423 | 0.408 | 0.450 | 0.480 | 0.475 |
| **top500** | 0.478 | **0.493** | **0.484** | 0.404 | **0.412** | 0.417 | 0.440 | **0.478** | 0.468 |
| **top750** | **0.474** | 0.505 | 0.488 | **0.400** | 0.427 | **0.407** | **0.442** | 0.483 | 0.470 |
| **top1000** | 0.484 | 0.498 | 0.488 | 0.405 | 0.417 | 0.412 | 0.445 | 0.479 | **0.462** |
| **full** | 0.747 | 0.749 | 0.747 | 0.639 | 0.642 | 0.641 | 0.787 | 0.788 | 0.787 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.712 | 0.721 | 0.711 | 0.635 | 0.667 | 0.642 | 0.702 | 0.710 | 0.708 |
| **top200** | 0.667 | 0.707 | 0.673 | 0.592 | 0.648 | 0.604 | 0.649 | 0.671 | 0.649 |
| **top350** | 0.649 | **0.694** | 0.662 | 0.578 | **0.641** | 0.598 | 0.614 | 0.642 | 0.623 |
| **top500** | 0.650 | 0.697 | **0.655** | 0.570 | 0.648 | 0.593 | 0.615 | **0.639** | **0.615** |
| **top750** | **0.648** | 0.699 | 0.655 | **0.561** | 0.643 | **0.577** | **0.609** | 0.642 | 0.615 |
| **full** | 0.844 | 0.846 | 0.845 | 0.787 | 0.787 | 0.787 | 0.797 | 0.796 | 0.797 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.538 | 0.605 | 0.570 | 0.463 | 0.524 | 0.492 | 0.524 | 0.584 | 0.550 |
| **top200** | 0.510 | 0.566 | 0.517 | 0.423 | 0.478 | 0.438 | 0.503 | 0.550 | 0.507 |
| **top350** | **0.493** | **0.529** | **0.500** | **0.411** | **0.451** | **0.432** | **0.490** | **0.496** | 0.477 |
| **top500** | 0.506 | 0.539 | 0.533 | 0.425 | 0.465 | 0.458 | 0.497 | 0.503 | 0.507 |
| **top750** | 0.513 | 0.553 | 0.526 | 0.446 | 0.475 | 0.449 | 0.490 | 0.523 | 0.500 |
| **top1000** | 0.518 | 0.545 | 0.516 | 0.450 | 0.466 | 0.444 | 0.496 | 0.520 | **0.490** |
| **full** | 0.716 | 0.715 | 0.717 | 0.641 | 0.644 | 0.642 | 0.713 | 0.716 | 0.714 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.320 | 0.381 | 0.358 | 0.264 | 0.324 | 0.294 | 0.296 | 0.335 | 0.319 |
| **top200** | 0.313 | 0.353 | 0.333 | 0.264 | 0.300 | 0.287 | 0.277 | **0.319** | 0.300 |
| **top350** | **0.305** | **0.345** | 0.321 | **0.254** | **0.288** | **0.274** | **0.269** | 0.321 | 0.294 |
| **top500** | 0.306 | 0.345 | **0.316** | 0.258 | 0.291 | 0.275 | 0.270 | 0.322 | **0.291** |
| **top750** | 0.306 | 0.346 | 0.320 | 0.259 | 0.292 | 0.278 | 0.271 | 0.324 | 0.295 |
| **full** | 0.750 | 0.752 | 0.748 | 0.682 | 0.687 | 0.684 | 0.734 | 0.740 | 0.736 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.686 | 0.694 | 0.690 | 0.585 | 0.609 | 0.586 | 0.679 | 0.682 | 0.682 |
| **top200** | 0.655 | 0.662 | 0.658 | 0.560 | 0.594 | 0.558 | 0.620 | 0.633 | 0.629 |
| **top350** | 0.637 | 0.648 | 0.648 | 0.541 | 0.591 | 0.546 | 0.582 | 0.590 | 0.583 |
| **top500** | **0.629** | 0.635 | 0.637 | **0.534** | 0.581 | **0.535** | 0.568 | 0.575 | 0.574 |
| **top750** | 0.631 | **0.633** | 0.644 | 0.537 | **0.569** | 0.537 | 0.563 | 0.573 | 0.567 |
| **top1000** | 0.631 | 0.640 | **0.636** | 0.541 | 0.572 | 0.536 | **0.559** | **0.569** | **0.557** |

**Table D1.** Results using the group average method. Highest effectiveness for each column appears in bold.

| Ward | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.631 | 0.653 | 0.630 | 0.552 | 0.583 | 0.551 | 0.638 | 0.655 | 0.634 |
| **top200** | 0.612 | 0.634 | 0.598 | 0.527 | 0.563 | 0.520 | 0.613 | 0.619 | 0.597 |
| **top350** | 0.603 | 0.623 | 0.590 | 0.524 | 0.564 | 0.510 | 0.588 | 0.606 | 0.571 |
| **top500** | 0.603 | 0.620 | 0.589 | 0.506 | 0.561 | 0.499 | 0.587 | 0.595 | 0.570 |
| **top750** | 0.586 | **0.619** | 0.573 | 0.490 | 0.558 | 0.502 | 0.573 | 0.594 | 0.555 |
| **top1000** | **0.580** | 0.619 | **0.568** | **0.481** | **0.550** | **0.492** | **0.565** | **0.591** | **0.547** |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.535 | 0.522 | 0.524 | 0.460 | 0.438 | 0.432 | 0.520 | 0.503 | 0.503 |
| **top200** | 0.540 | 0.522 | 0.530 | 0.447 | 0.424 | 0.435 | 0.529 | 0.501 | 0.515 |
| **top350** | 0.518 | 0.517 | 0.521 | 0.421 | 0.426 | 0.430 | 0.520 | 0.510 | 0.513 |
| **top500** | 0.514 | 0.517 | 0.514 | 0.417 | 0.422 | 0.419 | 0.511 | 0.517 | 0.504 |
| **top750** | 0.510 | **0.506** | 0.510 | 0.417 | **0.418** | **0.416** | 0.504 | 0.504 | 0.498 |
| **top1000** | **0.505** | 0.506 | **0.506** | **0.416** | 0.418 | 0.416 | **0.495** | **0.503** | **0.497** |
| **full** | 0.741 | 0.745 | 0.743 | 0.638 | 0.639 | 0.640 | 0.773 | 0.772 | 0.773 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.714 | 0.723 | 0.713 | 0.639 | 0.664 | 0.645 | 0.706 | 0.710 | 0.708 |
| **top200** | 0.681 | 0.708 | 0.684 | 0.606 | **0.640** | 0.602 | 0.653 | 0.671 | 0.654 |
| **top350** | **0.673** | 0.711 | 0.676 | **0.596** | 0.653 | 0.594 | **0.636** | 0.660 | **0.632** |
| **top500** | 0.673 | **0.703** | 0.679 | 0.602 | 0.649 | 0.600 | 0.638 | **0.652** | 0.639 |
| **top750** | 0.682 | 0.709 | **0.673** | 0.607 | 0.653 | **0.592** | 0.643 | 0.664 | 0.642 |
| **full** | 0.844 | 0.848 | 0.844 | 0.789 | 0.789 | 0.788 | 0.801 | 0.803 | 0.799 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.577 | 0.609 | 0.575 | 0.488 | 0.517 | 0.490 | 0.570 | 0.602 | 0.570 |
| **top200** | 0.535 | 0.561 | 0.532 | 0.442 | 0.471 | 0.450 | 0.531 | 0.548 | 0.520 |
| **top350** | 0.519 | 0.550 | 0.522 | 0.435 | 0.464 | 0.442 | 0.522 | 0.536 | 0.514 |
| **top500** | **0.504** | **0.537** | 0.526 | **0.412** | **0.457** | 0.439 | **0.502** | **0.531** | **0.511** |
| **top750** | 0.538 | 0.553 | **0.520** | 0.451 | 0.473 | **0.434** | 0.537 | 0.538 | 0.514 |
| **top1000** | 0.525 | 0.553 | 0.533 | 0.449 | 0.467 | 0.439 | 0.534 | 0.537 | 0.525 |
| **full** | 0.713 | 0.715 | 0.712 | 0.640 | 0.645 | 0.644 | 0.729 | 0.731 | 0.727 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.421 | 0.420 | 0.403 | 0.334 | 0.355 | 0.335 | 0.404 | 0.378 | 0.372 |
| **top200** | **0.381** | 0.378 | 0.368 | 0.298 | 0.335 | 0.320 | 0.391 | **0.344** | 0.342 |
| **top350** | 0.386 | **0.371** | **0.359** | 0.312 | 0.326 | **0.311** | **0.387** | 0.344 | **0.328** |
| **top500** | 0.385 | 0.371 | 0.361 | **0.309** | **0.325** | 0.317 | 0.389 | 0.348 | 0.333 |
| **top750** | 0.386 | 0.372 | 0.368 | 0.309 | 0.325 | 0.320 | 0.390 | 0.345 | 0.350 |
| **full** | 0.769 | 0.767 | 0.767 | 0.679 | 0.683 | 0.679 | 0.753 | 0.758 | 0.752 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.695 | 0.690 | 0.681 | 0.609 | 0.618 | 0.594 | 0.702 | 0.696 | 0.692 |
| **top200** | 0.669 | 0.661 | 0.656 | 0.590 | 0.597 | 0.578 | 0.656 | 0.649 | 0.645 |
| **top350** | 0.648 | 0.643 | 0.635 | 0.567 | 0.585 | 0.559 | 0.621 | 0.612 | 0.620 |
| **top500** | 0.649 | 0.641 | **0.613** | 0.580 | 0.579 | 0.550 | 0.620 | 0.600 | 0.596 |
| **top750** | **0.637** | 0.633 | 0.623 | 0.565 | **0.564** | **0.544** | 0.613 | 0.598 | 0.594 |
| **top1000** | 0.645 | **0.628** | 0.620 | **0.563** | 0.564 | 0.546 | **0.604** | **0.593** | **0.588** |

**Table D2.** Results using Ward's method. Highest effectiveness for each column appears in bold.

| *Complete Link* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.632 | 0.650 | 0.630 | 0.539 | 0.577 | 0.537 | 0.634 | 0.652 | 0.629 |
| **top200** | 0.618 | 0.627 | 0.614 | 0.521 | 0.562 | 0.532 | 0.608 | 0.619 | 0.602 |
| **top350** | 0.601 | **0.614** | 0.597 | 0.508 | 0.557 | 0.512 | 0.588 | 0.604 | 0.582 |
| **top500** | 0.594 | 0.621 | 0.601 | 0.505 | 0.562 | 0.517 | 0.578 | 0.602 | 0.583 |
| **top750** | 0.580 | 0.618 | **0.580** | **0.480** | 0.562 | 0.495 | 0.569 | 0.595 | 0.569 |
| **top1000** | **0.577** | 0.620 | 0.580 | 0.491 | **0.552** | **0.486** | **0.568** | **0.594** | **0.574** |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.536 | 0.530 | 0.526 | 0.456 | 0.446 | 0.435 | 0.503 | 0.502 | 0.515 |
| **top200** | 0.537 | 0.510 | 0.521 | 0.449 | 0.414 | 0.428 | 0.516 | 0.497 | 0.516 |
| **top350** | 0.516 | 0.522 | 0.516 | 0.427 | 0.434 | 0.424 | **0.500** | 0.505 | 0.506 |
| **top500** | **0.509** | 0.513 | 0.516 | 0.417 | 0.422 | 0.427 | 0.509 | 0.500 | 0.502 |
| **top750** | 0.520 | 0.500 | **0.506** | 0.423 | 0.418 | **0.412** | 0.508 | **0.486** | **0.485** |
| **top1000** | 0.513 | **0.494** | 0.509 | **0.414** | **0.406** | 0.414 | 0.506 | 0.490 | 0.496 |
| **full** | 0.743 | 0.745 | 0.743 | 0.636 | 0.642 | 0.639 | 0.759 | 0.760 | 0.758 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.717 | 0.722 | 0.715 | 0.639 | 0.660 | 0.644 | 0.704 | 0.708 | 0.706 |
| **top200** | 0.681 | 0.702 | 0.683 | 0.600 | 0.637 | 0.602 | 0.655 | 0.663 | 0.652 |
| **top350** | 0.672 | 0.709 | 0.685 | 0.591 | 0.649 | 0.609 | 0.632 | 0.654 | 0.642 |
| **top500** | **0.668** | **0.700** | **0.672** | **0.580** | **0.639** | 0.603 | **0.626** | 0.647 | **0.624** |
| **top750** | 0.671 | 0.703 | 0.681 | 0.597 | 0.648 | **0.596** | 0.630 | **0.643** | 0.632 |
| **full** | 0.844 | 0.845 | 0.842 | 0.788 | 0.792 | 0.786 | 0.799 | 0.803 | 0.797 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.575 | 0.610 | 0.581 | 0.496 | 0.522 | 0.497 | 0.558 | 0.595 | 0.568 |
| **top200** | 0.534 | 0.550 | 0.543 | 0.445 | **0.461** | 0.445 | 0.535 | 0.550 | 0.536 |
| **top350** | 0.522 | **0.545** | 0.516 | 0.443 | 0.463 | **0.432** | 0.516 | **0.523** | 0.505 |
| **top500** | 0.506 | 0.546 | **0.515** | **0.424** | 0.462 | 0.432 | 0.506 | 0.538 | **0.489** |
| **top750** | **0.502** | 0.547 | 0.535 | 0.429 | 0.463 | 0.446 | **0.501** | 0.523 | 0.522 |
| **top1000** | 0.513 | 0.549 | 0.526 | 0.437 | 0.467 | 0.437 | 0.506 | 0.528 | 0.515 |
| **full** | 0.702 | 0.704 | 0.700 | 0.629 | 0.633 | 0.627 | 0.719 | 0.718 | 0.715 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.415 | 0.420 | 0.416 | 0.327 | 0.352 | 0.342 | 0.370 | 0.378 | 0.375 |
| **top200** | 0.392 | 0.392 | **0.381** | 0.317 | 0.319 | **0.310** | 0.367 | 0.367 | 0.371 |
| **top350** | 0.388 | 0.374 | 0.385 | **0.315** | **0.313** | 0.311 | **0.362** | 0.354 | **0.361** |
| **top500** | **0.385** | **0.373** | 0.383 | 0.315 | 0.316 | 0.315 | 0.363 | **0.353** | 0.368 |
| **top750** | 0.386 | 0.373 | 0.389 | 0.318 | 0.317 | 0.315 | 0.363 | 0.353 | 0.375 |
| **full** | 0.789 | 0.793 | 0.787 | 0.688 | 0.688 | 0.683 | 0.778 | 0.779 | 0.780 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.679 | 0.696 | 0.689 | 0.599 | 0.616 | 0.599 | 0.687 | 0.693 | 0.698 |
| **top200** | 0.660 | 0.660 | 0.657 | 0.591 | 0.597 | 0.576 | 0.647 | 0.645 | 0.646 |
| **top350** | **0.636** | 0.642 | 0.642 | 0.563 | 0.580 | 0.566 | 0.616 | 0.617 | 0.618 |
| **top500** | 0.637 | 0.639 | 0.643 | **0.557** | 0.575 | 0.568 | **0.611** | 0.605 | 0.622 |
| **top750** | 0.642 | 0.630 | 0.656 | 0.564 | **0.558** | 0.573 | 0.612 | **0.599** | 0.628 |
| **top1000** | 0.638 | **0.626** | **0.637** | 0.559 | 0.559 | **0.547** | 0.614 | 0.599 | **0.614** |

**Table D3.** Results using the complete link method. Highest effectiveness for each column appears in bold.

| Single Link | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **AP** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.632 | 0.661 | 0.643 | 0.558 | 0.608 | 0.566 | 0.638 | 0.650 | 0.647 |
| **top200** | 0.622 | 0.663 | 0.618 | 0.553 | 0.631 | 0.547 | 0.613 | 0.630 | 0.600 |
| **top350** | 0.606 | 0.661 | 0.614 | 0.551 | 0.635 | 0.544 | 0.590 | 0.612 | 0.581 |
| **top500** | 0.598 | 0.662 | 0.606 | 0.535 | 0.638 | 0.538 | 0.576 | **0.605** | 0.572 |
| **top750** | 0.598 | 0.662 | 0.588 | 0.531 | 0.637 | 0.524 | 0.568 | 0.609 | 0.558 |
| **top1000** | **0.589** | **0.656** | **0.573** | **0.521** | **0.629** | **0.502** | **0.560** | 0.609 | **0.546** |
| **CACM** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.513 | 0.587 | 0.542 | 0.442 | 0.519 | 0.469 | 0.482 | 0.549 | 0.512 |
| **top200** | 0.493 | 0.568 | 0.522 | 0.423 | 0.502 | 0.443 | 0.469 | 0.534 | 0.499 |
| **top350** | 0.487 | 0.536 | 0.503 | 0.416 | 0.484 | 0.434 | 0.452 | 0.500 | 0.469 |
| **top500** | 0.498 | 0.533 | 0.499 | 0.422 | 0.485 | **0.433** | 0.461 | 0.492 | **0.463** |
| **top750** | 0.485 | **0.526** | 0.499 | **0.411** | 0.464 | 0.435 | 0.450 | **0.491** | 0.465 |
| **top1000** | **0.479** | 0.527 | **0.498** | 0.412 | **0.462** | 0.433 | **0.446** | 0.492 | 0.469 |
| **full** | 0.759 | 0.764 | 0.763 | 0.664 | 0.663 | 0.663 | 0.787 | 0.788 | 0.788 |
| **CISI** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.750 | 0.746 | 0.726 | 0.707 | 0.714 | 0.691 | 0.725 | 0.719 | 0.704 |
| **top200** | 0.723 | 0.752 | 0.727 | 0.676 | 0.718 | 0.678 | 0.676 | 0.691 | 0.679 |
| **top350** | 0.714 | 0.744 | 0.720 | 0.665 | **0.701** | 0.672 | 0.663 | 0.676 | 0.659 |
| **top500** | 0.717 | 0.741 | 0.719 | 0.676 | 0.703 | 0.669 | 0.656 | 0.668 | 0.652 |
| **top750** | **0.702** | **0.738** | **0.713** | **0.651** | 0.709 | **0.661** | **0.646** | **0.662** | **0.646** |
| **full** | 0.879 | 0.882 | 0.878 | 0.823 | 0.824 | 0.823 | 0.824 | 0.824 | 0.823 |
| **LISA** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.597 | 0.653 | 0.602 | 0.522 | 0.580 | 0.539 | 0.560 | 0.622 | 0.578 |
| **top200** | 0.578 | 0.653 | **0.586** | 0.497 | **0.567** | **0.496** | 0.568 | 0.631 | **0.577** |
| **top350** | **0.561** | 0.653 | 0.628 | **0.473** | 0.571 | 0.531 | **0.550** | 0.625 | 0.612 |
| **top500** | 0.613 | 0.654 | 0.607 | 0.516 | 0.583 | 0.524 | 0.607 | 0.620 | 0.581 |
| **top750** | 0.601 | 0.665 | 0.613 | 0.509 | 0.594 | 0.530 | 0.585 | 0.620 | 0.593 |
| **top1000** | 0.596 | **0.642** | 0.608 | 0.515 | 0.572 | 0.529 | 0.575 | **0.609** | 0.593 |
| **full** | 0.744 | 0.746 | 0.741 | 0.665 | 0.672 | 0.669 | 0.763 | 0.766 | 0.764 |
| **MED** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.330 | 0.400 | 0.366 | 0.268 | 0.368 | 0.330 | 0.302 | 0.341 | 0.319 |
| **top200** | 0.316 | 0.352 | 0.328 | **0.267** | 0.318 | 0.292 | 0.274 | 0.307 | 0.284 |
| **top350** | **0.314** | **0.334** | **0.323** | 0.269 | 0.305 | **0.287** | **0.272** | **0.296** | **0.281** |
| **top500** | 0.315 | 0.335 | 0.325 | 0.275 | **0.304** | 0.289 | 0.272 | 0.301 | 0.285 |
| **top750** | 0.316 | 0.336 | 0.325 | 0.271 | 0.304 | 0.288 | 0.274 | 0.302 | 0.286 |
| **full** | 0.790 | 0.795 | 0.792 | 0.700 | 0.702 | 0.700 | 0.773 | 0.778 | 0.773 |
| **WSJ** | **$\beta$=1** | | | **$\beta$=0.5** | | | **$\beta$=2** | | |
| | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** | **M1** | **M2** | **M3** |
| **top100** | 0.686 | 0.694 | 0.690 | 0.611 | 0.656 | 0.619 | 0.684 | 0.683 | 0.688 |
| **top200** | 0.655 | 0.662 | 0.658 | 0.595 | 0.635 | 0.591 | 0.635 | 0.632 | 0.636 |
| **top350** | 0.637 | 0.648 | 0.648 | 0.584 | 0.627 | 0.593 | 0.605 | 0.599 | 0.609 |
| **top500** | 0.629 | 0.635 | 0.637 | 0.576 | 0.618 | 0.586 | 0.588 | 0.578 | 0.595 |
| **top750** | 0.631 | 0.633 | 0.644 | 0.584 | 0.618 | 0.595 | 0.580 | 0.566 | 0.589 |
| **top1000** | 0.631 | 0.640 | 0.636 | 0.582 | 0.622 | 0.580 | 0.577 | 0.570 | 0.584 |

**Table D4.** Results using the single link method. Highest effectiveness for each column appears in bold.