The Effectiveness of Query-Specific Hierarchic Clustering in Information Retrieval

Anastasios Tombros[*], Robert Villa, C.J. van Rijsbergen

Department of Computing Science
University of Glasgow, G12 8RZ, Scotland

Abstract

Hierarchic document clustering has been widely applied to Information Retrieval (IR) on the grounds of its potential improved effectiveness over inverted file search. However, previous research has been inconclusive as to whether clustering does bring improvements. In this paper we take the view that if hierarchic clustering is applied to search results (query-specific clustering), then it has the potential to increase the retrieval effectiveness compared both to that of static clustering and of conventional inverted file search. We conducted a number of experiments using five document collections and four hierarchic clustering methods. Our results show that the effectiveness of query-specific clustering is indeed higher, and suggest that there is scope for its application to IR.

Keywords: Information Retrieval; Query-Specific Hierarchic Clustering; Effectiveness Evaluation

1. Introduction

C*luster analysis* is a technique that allows the identification of groups, or clusters, of similar objects in multi-dimensional space (Cormack, 1971). It was initially introduced in the field of *Information Retrieval* (IR) as a means of improving the efficiency of serial search (Salton, 1968, p.258). Apart from efficiency, effectiveness was also put forward for the use of *hierarchic clustering* in IR (Jardine & Van Rijsbergen, 1971; Croft, 1978). 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 the documents in a collection (Croft, 1980).

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 the non-relevant ones. Relevant documents that might have otherwise been ranked low in an *inverted file search* (IFS), will be (through inter-document associations) grouped together with other

[*] Corresponding author. Tel: +44-141-3304971; fax: +44-141-3304913. E-mail addresses: tombrosa@dcs.gla.ac.uk (A. Tombros), villar@dcs.gla.ac.uk (R. Villa), keith@dcs.gla.ac.uk (C.J. van Rijsbergen).

relevant documents, thus improving the effectiveness of an IR system. The actual effectiveness of hierarchic clustering can be gauged by *cluster-based searches* that retrieve the cluster that best matches the query (Croft, 1978).

Traditionally, clustering has been applied statically over the whole document collection prior to querying. Many researchers have examined the retrieval effectiveness of hierarchic clustering methods, and have compared it to that of conventional IFS (e.g. Croft, 1980; Grifftihs *et al*., 1984; Voorhees, 1985). On the other hand, the behaviour and effectiveness of clustering methods when applied to the search results of an IR system (i.e. dynamic, or *query-specific* clustering) have not been extensively investigated. Only a few researchers have looked into this issue (Preece, 1973; Willett, 1985; Hearst & Pedersen, 1996), leaving scope for further research and experimentation.

This paper makes a case for the use of hierarchic query-specific clustering in IR on the grounds of improved retrieval effectiveness. The case is intuitively appealing, since a query-specific document hierarchy has the potential to capture the essence of the Cluster Hypothesis: the hierarchy will be adjusted to a specific query, increasing the likelihood of placing together in clusters documents that are relevant to the query (Hearst & Pedersen, 1996). It is the purpose of this paper to establish the validity of this case experimentally, by systematically investigating the effectiveness of query-specific hierarchic clustering.

This paper is structured as follows: section 2 provides some background on document clustering, and reports on related work in the area of query-specific clustering. Section 3 will provide details of the environment and procedure under which the experiments were conducted, while Section 4 will present the results. The results will be analysed and discussed in Section 5, and finally, Section 6 will outline the conclusions and present areas for future work.

2. Related Work

Two different types of clustering methods have been applied to IR: *partitioning* and *hierarchic* (Van Rijsbergen, 1979; Willett, 1988; Rasmussen, 1992). Partitioning methods were first used because of their low computational requirements (typically in the order of O(N) to O(NlogN) for the clustering of N documents) (Salton, 1968).

These methods suffer on a theoretical basis, as they generally require a number of arbitrarily determined experimental parameters (Willett, 1988). A typical limitation of such methods is the requirement to determine *a priori* the number of clusters that will be generated, something that makes them inappropriate for query-specific clustering, when the number of topics actually present in a document set can not be estimated *a priori*. Effectiveness of searches based on document partitions has proven to be significantly inferior to that based on searches of the unclustered file (Salton, 1971). Recent applications of partitioning algorithms to IR (Cutting *et al*., 1992; Silverstein & Pedersen, 1997; Zamir &

Etzioni, 1998) have also focused on efficiency issues for on-the-fly clustering rather than on effectiveness issues.

Hierarchic methods[1] on the other hand, result in tree-like classifications in which small clusters of documents that are found to be strongly similar to each other are nested within larger clusters that contain less similar documents (Willett, 1988). Two main methods, and many variants of them, for matching a query against a document hierarchy have been proposed (Jardine & Van Rijsbergen, 1971; Van Rijsbergen, 1974; Van Rijsbergen & Croft, 1975; Croft, 1980; Voorhees, 1985): a *top-down* search, and a *bottom-up* search. In both types of search, a single cluster that satisfies a retrieval criterion is retrieved. Jardine and Van Rijsbergen, (1971), called this type of retrieval *cluster based retrieval*. Comparative studies of the two search strategies suggest that, in general, a bottom-up method is more effective than a top-down one (Van Rijsbergen & Croft, 1975; Croft, 1980; El Hamdouchi & Willett, 1989).

Hierarchic clustering was applied to IR based on its potential to increase the effectiveness of IR systems, a potential supported by the Cluster Hypothesis. A lot of experimental work has therefore been carried out in order to examine the comparative effectiveness of cluster-based searches and IFS. Jardine and Van Rijsbergen were the first ones to introduce the notion of an *optimal cluster search* (i.e. the cluster in the hierarchy that if retrieved would give the maximum possible value for a retrieval criterion out of all the clusters in the hierarchy). Such a search has the advantage that it isolates cluster effectiveness from the bias introduced by any specific search method, but it provides a value that will only be reached if a search strategy infallibly selects the best cluster for each request.

In experiments where optimal cluster effectiveness was compared against IFS effectiveness (Jardine & Van Rijsbergen, 1971; Croft, 1978) it was demonstrated that cluster searches have the potential to outperform IFS. However, in these experiments the effectiveness of the initial retrieval is likely to have been low due to the simple query matching and term weighting schemes employed.

However, when actual cluster-based search strategies were used in the comparison results have been inconclusive. Particular types of top-down or bottom-up searches have been shown to perform better than IFS (e.g. Croft, 1978; Croft, 1980; Griffiths *et al.*, 1986; Voorhees, 1985) especially for precision-oriented searches, whereas other experimental work has suggested that non-clustered searches are in general more effective (e.g. El Hamdouchi & Willett, 1989).

In all the work reported in the previous paragraphs, clustering has been applied to entire document collections in a static manner (i.e. once, before querying). Preece, (1973) was one of the first researchers to suggest the use of clustering after an IFS has retrieved a limited set of documents, in order to allow a more convenient analysis of the search results.

[1] For a comprehensive review of such methods the interested reader can refer to (Willett, 1988).

Willett, (1985), to the best of the authors' knowledge, is the only researcher to have examined the effectiveness of query-specific hierarchic clustering in IR. He noted three novel features of such an approach: the need to generate a new hierarchy for each query, the fact that there is no need to employ any updating strategies for the hierarchy, and the substantial efficiency gains of such an approach since only a relatively small subset of the collection needs be clustered.

Willett then experimentally tested whether the effectiveness of such an approach would be comparable to that of static clustering. In order to determine the set of documents to be clustered for each query he used a *coordination level search*; levels of 0, 1, 2, and 3 were used[2]. Willett's results showed that the effectiveness of the dynamic method was not substantially inferior to that of static clustering.

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. Acknowledging the fact, Willett (1985, p.30) reports 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.

Hearst and Pedersen, (1996), viewed the Cluster Hypothesis under the light of query-specific clustering. They performed experiments in which they clustered the top-*n* documents (100, 250, 500, 1000) returned from a similarity search, in order to examine whether the Scatter/Gather system (Cutting *et al*., 1992) succeeds in placing relevant documents in the same partitions. Scatter/Gather employs a partitioning clustering method that inherits from this class of algorithms the problems mentioned in earlier parts of this section (e.g. documents are arbitrarily clustered into 5 partitions).

By observing the distribution of the percentage of relevant documents in each of the five partitions, Hearst and Pedersen conjectured that the Cluster Hypothesis holds for the Scatter/Gather system, since the best partition always contains at least 50% of the relevant documents retrieved.

Hearst and Pedersen did not compare the effectiveness of the query-specific partitions to that of a static partitioning. No information is also provided on the varying degree at which the Cluster Hypothesis holds for different values of *n*. Moreover, no statistics of the partitions generated are given (e.g. size), and the effect that partition size may have on the number of relevant documents in the best partition was not investigated. For example, for large values of *n* one expects the mean size of each partition to increase (since there are always 5 partitions), increasing at the same time the probability of having more relevant documents placed in the same partition.

[2] A level of 0 corresponds to the entire collection, a level of 1 corresponds to documents that have at least 1 term in common with the query, etc.

Hearst and Pedersen also compared a ranking of documents in a best cluster to an equivalent cut-off in the original top-ranked documents. Their results showed that if for every query the best cluster was selected, then the effectiveness of the clustering would be higher than IFS effectiveness. As we shall discuss in Section 3.3, this type of comparison is not perfectly fair on IFS, since it compares a theoretical maximum effectiveness (attained only if a user always selects the best partition) against a value that is not optimal. In section 3.3 a fairer comparison will be suggested.

A number of other researchers have performed clustering on document sets returned by a similarity search, arbitrarily choosing a particular value for *n* without further justification (e.g. Allen *et al*., 1993; Kirriemuir & Willett, 1995; Leouski & Allan, 1998; Zamir & Etzioni, 1998). Kirriemuir and Willett, (1995), for example, were interested in examining the effect of hierarchical clustering methods and similarity coefficients on the effectiveness 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 clustering and IFS. Leouski and Allan, (1998), examined two and three dimensional visualisations of the top-50 retrieved documents returned from an IFS. They did not cluster the retrieved document sets, instead they investigated how the visualisations 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.

The research reported in this paper is an attempt to investigate the effectiveness of query-specific hierarchic clustering in the context of IR. We believe that query-specific clustering has the potential to increase the effectiveness of hierarchic clustering by taking into account the relationships of documents that have a higher probability of being relevant to a specific query. Through systematic experimentation we aim to investigate the validity of this statement.

3. Experimental Details

The two main goals of our work are to examine the variation of the optimal cluster effectiveness achieved from the application of hierarchic clustering to progressively larger numbers of top-ranked documents returned from an IFS, and to compare this effectiveness with the effectiveness of an IFS. Five document collections, four hierarchic agglomerative methods, and seven different numbers of top-ranked documents to be clustered are employed in the experiments. A by-product of the experimental procedure will inevitably be a comparison of the effectiveness of the four clustering methods. This is in addition to the main focus of the paper, since the comparative effectiveness of these four agglomerative methods has been extensively studied in the past (e.g. Griffiths *et al.*, 1984; Voorhees, 1985; El-Hamdouchi & Willett, 1989; Burgin, 1995), albeit under different experimental settings (i.e. static clustering).

In the following sections we provide details on the conditions under which the experiments were carried out. Section 3.1 presents the 5 document collections used and their properties, as well as the details of the IR system used for the initial retrieval. Section 3.2 will discuss the choice of the four clustering methods used, and finally section 3.3 will provide information about the conditions under which the optimal cluster evaluation was conducted.

3.1 Document Collections and Initial Retrieval

Five document collections were used in the experiments. Four of them (CACM, CISI, LISA, and Medline) have been used before for experimentation with hierarchic clustering methods (Griffiths *et al.*, 1986; Voorhees, 1985; El Hamdouchi & Willett, 1989; Burgin, 1995), and the fifth is one of the TREC standard collections (Voorhees & Harman, 1997).

	CACM	**CISI**	**LISA**	**MED**	**WSJ**
Number of docs.	3204	1460	6004	1033	74520
Mean terms per doc.	22.5	43.9	39.7	51.6	377
Number of queries	52	35	35	30	50
Mean relevant docs per query	15.3	49.8	10.8	23.2	71.4
Total relevant docs.	796	1742	379	696	3572

Table 1 . Collection statistics

Statistics for the five document collections are presented in Table 1. It should be noted that the first four collections are homogeneous, treating one major subject area (e.g. Library and Information Science, Biomedicine, etc.), and such topical homogeneity might distort the experimental results. The WSJ collection, although specialising on financial issues, covers in its documents a wide variety of topics, providing a collection with different characteristics. For the WSJ collection, TREC topics 1-50 were randomly chosen and used in the experiments.

	CACM	**CISI**	**LISA**	**MED**	**WSJ**
11 pt. Avg.	0.3778	0.1945	0.3115	0.5699	0.2546
3 pt. Avg.	0.365	0.1678	0.2991	0.5836	0.2259

Table 2. Initial retrieval evaluation

The SMART IR system (Salton, 1971) was used in order to perform the initial retrieval. Initial retrieval for all collections was performed using a *tf-idf* weighting scheme for document and query terms that involves cosine normalisation[3] (SMART's *ltc* scheme). The default SMART stoplist and stemming

[3] WSJ documents 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. However, we feel that the effectiveness of the initial retrieval for WSJ is sufficient for our experiments.

were used in indexing all the collections and queries. Table 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 five collections.

After the initial retrieval, the top-*n* ranked documents were used to create the collections that were clustered. Seven different values of *n* were used: 100, 200, 350, 500, 750, 1000, and full collection (*n* = collection size)[4]. The same weighting scheme as for the initial retrieval (*ltc*) was applied to the document vectors of the collections that were to be clustered. After initial experimentation with different vector weighting schemes (binary, term frequency weights) for clustering, no significant differences were found - which is in agreement with previous findings (Norreault *et al*., 1981; Willett, 1983).

3.2 The Clustering Algorithms

Four hierarchical methods were employed in the experiments: single link, complete link, group average, and Ward's method. The main reason behind the choice of these four methods is the fact that they have been extensively used and examined in the context of IR (e.g. Van Rijsbergen & Croft, 1975; Griffiths *et al*., 1984; Voorhees, 1985; El Hamdouchi & Willett, 1989). The implementation used was based on the algorithms given in (Späth, 1980). Apart from Ward's method that requires a specific form of distance measure that minimises the within group variance (Ward, 1963), the association measure used for the other three methods was the cosine coefficient. Experiments with the normalised Euclidean distance, and the Dice coefficient, did not produce significantly different results - again in agreement with previous suggestions and findings (Van Rijsbergen, 1979; Norreault *et al*., 1981; Willett, 1983; Ellis *et al*., 1993).

	Group Average	**Ward**	**Complete Link**	**Single Link**
top100	12.6	8.8	8	28.3
top200	16.7	10.4	9.4	54.1
top350	21	11.9	10.6	91.2
top500	24.3	12.7	11.6	129.7
top750	28.6	13.6	13	196.7
top1000	31.8	14.5	14.5	263.4

Table 3. Average cluster sizes for the four methods using the WSJ collection

In Table 3 we present average cluster sizes for the WSJ collection, for 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 *n* increases, is single link. The other three methods produce hierarchies that are unaffected by the increase in the number of documents clustered. This behaviour is typical of the four methods used (Murtagh, 1984), and is consistent across the five document collections.

[4] The value of 1000 was not used in CISI and Medline collections because their sizes are 1460 and 1033 documents respectively. The full WSJ collection (74520 documents) was not clustered for practical reasons.

Hierarchic agglomerative algorithms usually have a time complexity of $O(n^2)$, something that makes them an inefficient solution for the clustering of large data sets. A dynamic, query-specific clustering method should have efficiency as a high priority (Zamir & Etzioni, 1998). However, efficiency issues are not tackled in this paper for two reasons. The first reason is that for query-specific 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. The second, and main, reason is based on the authors' belief that effectiveness is of primary importance, whereas efficiency is a factor that is heavily dependent on technological advances. One may also view the improved effectiveness that can be achieved as a motivation for the development of more efficient algorithms and/or hardware that would exploit hierarchic clustering. The present research is therefore focused solely on issues of effectiveness.

3.3 Optimal Cluster Evaluation

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 (Van Rijsbergen, 1979). Cluster-based retrieval strategies, on the other hand, perform a ranking of clusters instead of individual documents in response to each request (Jardine & Van Rijsbergen, 1971). The generation of precision-recall graphs is thus not possible in such systems, and in order to derive an evaluation function for clustering systems the E effectiveness function was proposed by (Jardine & Van Rijsbergen, 1971).

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 for precision and recall (over the set of documents of a specific cluster), and β is a parameter that reflects the relative importance attached to precision and recall. Three values of this parameter are usually used: 1, 0.5 and 2, the first value attributing equal importance to precision and recall, the second deeming precision twice as important as recall, and the third treating recall twice as important as precision. The E effectiveness measure and these three values of the parameter β are used in the experiments reported in this paper.

The optimal cluster of a hierarchy for any given query is the cluster that yields the least E value for that query. Therefore, the optimal cluster effectiveness represents the maximum effectiveness that is attainable by a cluster-based search strategy that selects a single cluster in response to each query. Jardine and Van Rijsbergen named this measure MK1. It is used to measure optimal cluster effectiveness in our experiments.

In order to compare optimal cluster effectiveness with IFS effectiveness two measures are employed in our experiments. The first one stems 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 IFS. We will call this measure MK1-k.

Intuitively, this measure captures the degree at which IFS effectiveness matches cluster effectiveness, for the number of documents for which cluster effectiveness is optimal. MK1-k yields different values for each one of the four clustering algorithms employed.

The comparison based on MK1 and MK1-k measures can be thought of as being unfair on IFS, since 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, the second measure used in our experiments for comparison between optimal cluster effectiveness and IFS effectiveness is based on the above, i.e. the optimal IFS effectiveness. Jardine and Van Rijsbergen, (1971), call this measure MK3. It represents a measurement of the maximum effectiveness that is attainable using an IFS strategy, its value being calculated irrespective of a particular clustering method.

We believe that by using two distinct measures to gauge IFS effectiveness (MK1-k, MK3), and by comparing them to cluster-based effectiveness (MK1), we can obtain an accurate and fair picture of effectiveness variation across experimental conditions. It should be noted that in order for a cluster-based search strategy to achieve an effectiveness equal to MK1, it would have to search the hierarchy in such a way that would always retrieve the optimal cluster. Similarly, for an IFS to achieve an effectiveness equal to MK3 it would have to identify for each request the optimal threshold at which to retrieve documents (Jardine & Van Rijsbergen, 1971).

Optimal cluster evaluation has been widely employed in the past (Jardine & Van Rijsbergen, 1971; Croft, 1978; Griffiths *et al.*, 1984; Shaw, 1991; Burgin, 1995). 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).

Our motivation for using optimal evaluation measures was based on this advantage. Optimal measures eliminate any bias that may be introduced from sources *external* to the hierarchy. External sources include the choice of a particular cluster-based search strategy that matches queries to clusters, and the ability of a user during a browsing 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 include the type of search, e.g. bottom-up, top-down, narrow, wide, (Jardine & Van Rijsbergen, 1971; Croft 1980; Voorhees, 1985; El-Hamdouchi & Willett, 1989), the type and length of cluster representative against which queries are matched (Croft, 1978; Voorhees, 1985), the entry point in the hierarchy in the case of a bottom-up search (El-Hamdouchi & Willett, 1989), etc.

In the case of a user browsing a clustered document collection (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; Leouski & Allan 1998), the way that cluster contents are summarised and displayed (Hearst & Pedersen, 1996; Radev *et al*., 2000; Kural *et al*., 2001), etc.

By eliminating such external parameters from our experimental design, we can infer that the variation in effectiveness across our experimental conditions is attributed to the different conditions themselves (*internal* parameters, i.e. different numbers of top-ranked documents), and not to any form of bias that may have been introduced by any of the external parameters.

Finally, in order to test the statistical significance of results (E values) across different experimental conditions, we used the Wilcoxon signed-ranks test. This test has been suggested as appropriate by (Croft, 1978; pp. 27-29) based on its statistical strength and on the reasonable assumptions it makes about the distribution of E values.

4. Experimental Results

Based on the settings detailed in the previous section a number of experiments were conducted in order to investigate the effectiveness of query-specific clustering, the results of which we present in this section.

4.1 Clustering tendency of the collections

The experiments reported here aim to examine the degree at which the clustering tendency of the five document collections is affected by the different values of *n* used in the clustering of the top-*n* ranked documents. In 1971 Jardine and Van Rijsbergen postulated the Cluster Hypothesis, which we presented in the introductory section of this paper. The more a document collection is characterised by this Hypothesis, the higher the effectiveness of a cluster search strategy is expected to be for that collection.

Three different methods have been proposed in order to test whether or not the Cluster Hypothesis holds for a specific collection: (Jardine & Van Rijsbergen, 1971) proposed the overlap test, (Voorhees, 1985) the Nearest Neighbour (NN) test, and (El Hamdouchi & Willett, 1987) the density test[5].

We chose to use the method proposed by Voorhees, because we believe that it fits better with our experimental goal, which is to test whether query-specific clustering results at a better structured collection with a better clustering behaviour than static clustering. The NN test consists of finding the N nearest neighbours (i.e. most similar documents) for each relevant document for a specific request, and of counting the number of relevant documents in that neighbourhood. The higher the number of relevant documents, the higher the probability that the Cluster Hypothesis holds for the collection.

The results for the NN test are displayed in Table 4. For each of the five collections, and for each value of *n*, we display a single value that corresponds to the number of relevant documents contained in the NN set (we used a value of 5 for the test, the same that Voorhees used for her experiments) when

[5] The interested reader can refer to the original papers for details of the tests.

averaged over all of the relevant documents for all the queries in a collection. The highest value in each column is displayed in bold.

These results suggest that the clustering tendency of each collection tends to decrease for increasing values of *n*. This leads us to expect a higher effectiveness for those collections derived with lower values of *n*. Statistical analysis of the results using the Wilcoxon signed-rank test showed significant differences for the WSJ collection (across all combinations of *n*), CACM (across all combinations between *n*=100 and the rest, and *n*=200 and the rest), Medline (between *n*=100 and the rest), and CISI (across all combinations of *n*). No significance was found for LISA.

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

Table 4. Results of the NN test. Highest values in bold

One explanation for these results is that by reducing the number of top-ranked documents one eliminates larger numbers of non-relevant than relevant documents. 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. This behaviour 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 that we presented. However, as we shall see in the next section, these results do not translate into significantly improved effectiveness for smaller values of *n*.

4.2 Optimal cluster evaluation results

Optimal cluster evaluation results based on the MK1 measure for cluster effectiveness, and MK1-k and MK3 measures for IFS, are presented here. These results allow us to examine the behaviour of optimal cluster effectiveness when the number of top-ranked documents changes, and to examine whether optimal cluster effectiveness is higher than IFS effectiveness. Table 5 presents the results (E values) for the group average method for all five document collections, and for all three values of β. The results for the other three clustering methods are presented in the Appendix (Tables A1, A2, A3).

		β= 1			**β= 0.5**			**β= 2**		
CACM	**Mean rel. docs per query**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**
top-100	10.46	**0.523**	0.688	**0.55**	**0.438**	0.66	0.503	**0.502**	**0.642**	0.503
top-200	11.62	0.54	0.695	0.55	0.476	**0.646**	**0.498**	0.512	0.651	**0.501**
top-350	12.69	0.543	0.681	0.55	0.469	0.647	0.503	0.52	0.667	0.501
top-500	13.21	0.548	0.686	0.55	0.461	0.66	0.503	0.54	0.667	0.501
top-750	13.58	0.553	0.692	0.55	0.465	0.658	0.503	0.537	0.667	0.501
top-1000	13.83	0.546	**0.68**	0.55	0.463	0.652	0.503	0.537	0.662	0.501
full	15.31	0.748	0.794	0.55	0.641	0.713	0.503	0.782	0.806	0.501
CISI	**Mean rel. docs per query**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**
top-100	16.31	0.715	**0.817**	0.762	0.63	0.827	0.727	0.702	0.777	0.738
top-200	24.71	0.697	0.827	0.75	0.609	0.82	0.729	0.658	**0.741**	0.699
top-350	32.31	0.683	0.822	**0.748**	0.589	**0.811**	**0.726**	0.655	0.753	0.68
top-500	37.06	0.681	0.819	0.748	0.593	0.815	0.726	0.656	0.765	**0.676**
top-750	42.34	**0.667**	0.823	0.748	**0.567**	0.818	0.726	**0.649**	0.776	0.676
full	49.77	0.842	0.84	0.748	0.79	0.873	0.726	0.798	0.824	0.676
LISA	**Mean rel. docs per query**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**
top-100	7.12	0.589	**0.723**	0.627	0.517	0.699	**0.577**	0.576	0.677	0.584
top-200	8.62	0.587	0.734	**0.626**	0.504	0.695	0.577	0.559	**0.672**	**0.58**
top-350	9.17	0.566	0.744	0.626	0.493	**0.693**	0.577	0.553	0.698	0.58
top-500	9.83	0.58	0.746	0.626	0.487	0.717	0.577	0.568	0.721	0.58
top-750	10.2	0.575	0.738	0.626	0.489	0.7	0.577	0.571	0.705	0.58
top-1000	10.34	**0.553**	0.744	0.626	**0.475**	0.707	0.577	**0.549**	0.725	0.58
full	10.83	0.713	0.792	0.626	0.643	0.736	0.577	0.716	0.739	0.58
MED	**Mean rel. docs per query**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**
top-100	18.97	0.349	0.45	**0.387**	0.3	**0.456**	**0.354**	0.308	**0.399**	**0.333**
top-200	20.37	0.326	0.455	0.387	0.281	0.468	0.354	0.294	0.413	0.333
top-350	21.03	**0.309**	**0.437**	0.387	0.281	0.462	0.354	**0.271**	0.404	0.333
top-500	21.13	0.311	0.443	0.387	0.279	0.471	0.354	0.273	0.399	0.333
top-750	21.3	0.311	0.446	0.387	**0.276**	0.462	0.354	0.272	0.4	0.333
full	23.2	0.744	0.494	0.387	0.682	0.596	0.354	0.711	0.403	0.333
WSJ	**Mean rel. docs per query**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**	**MK1**	**MK1-k**	**MK3**
top-100	16.63	0.692	0.791	0.734	0.608	0.767	0.693	0.696	0.779	0.719
top-200	24.02	0.67	**0.782**	0.721	0.604	0.762	0.69	0.661	0.741	0.686
top-350	31.88	0.671	0.784	0.716	0.603	**0.76**	**0.689**	0.65	0.742	0.666
top-500	37	0.668	0.795	0.715	**0.585**	0.774	0.689	0.642	0.731	0.659
top-750	43.54	**0.667**	0.791	**0.714**	0.585	0.775	0.689	**0.64**	**0.729**	0.655
top-1000	47.75	0.676	0.793	0.714	0.586	0.776	0.689	0.641	0.732	**0.654**

Table 5. Evaluation results for the group average method. Highest effectiveness (the lowest E value) for each column appears in bold

The values in Table 5 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 we saw in Table 3, average cluster size does not always increase in proportion to the number of top-ranked 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. Consequently, for comparisons across different values of *n* (across rows of MK1 values) it is necessary to consider only the definition of recall over all relevant documents; one can view this as an attempt to normalise the results.

There seems to be a small degradation of effectiveness for decreasing values of *n* for all three values of β. However, static clustering effectiveness (i.e. *n* = full) is significantly lower than that obtained at any query-specific level (i.e. any other value of *n*). These results will be further analysed in Section 5.1.

As far as the comparison between cluster effectiveness and IFS effectiveness is concerned (column MK1 vs. MK1-k and MK3), there are some differences depending on the definition of recall used. If recall is defined over relevant and retrieved documents, then results are more in favour of cluster effectiveness than those obtained with the conventional definition of recall. Since there is no distortion of the results we will evolve our discussion around the definition of recall (i.e. over all relevant documents) that corresponds to the values shown in Table 5. From these results one can notice that query-specific cluster effectiveness is higher than IFS effectiveness for all values of *n* (100, 200, 350, 500, 750, 1000). However, this is not the case for static clustering (*n* = full). The results reported here will be analysed and discussed in section 5.2.

As we shall point out in section 5.3, the group average method proved to be the best of the four clustering methods. However, the pattern of the results is similar for the other three methods: optimal cluster effectiveness does not significantly increase as the value of *n* increases, query-specific effectiveness is significantly better than static clustering effectiveness, and optimal query-specific cluster effectiveness is better than IFS (with the exception of the single link method that rarely outperforms IFS).

4.3 Optimal cluster characteristics

This section aims to give details about some characteristics of optimal clusters. Table 6, columns 3-6, present 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). The optimal clusters in Table 6 have been generated by the group average method for β=0.5, and β=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 β as for the optimal clusters) returned by an IFS (MK3 measure). The optimal clusters and IFS sets in Table 6 correspond to the E values presented in Table 5.

We chose to present the data for these two collections (LISA & WSJ) 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 five collections we tested, whereas LISA the least.

LISA	**Mean rel. per query**	**MK1 β=0.5**		**MK1 β=2**		**MK3 β=0.5**		**MK3 β=2**	
		Avg. size	**Avg. rel.**	**Avg. size**	**Avg. rel.**	**Avg. size**	**Avg. rel.**	**Avg. size**	**Avg. rel.**
top100	7.1	4.7	2.9	23.5	5.3	10	3.1	34.8	5.9
top200	8.6	4	2.6	30.7	5.8	15.2	3.1	42.8	6.3
top350	9.2	4.3	2.7	27.4	5.2	21.3	3.1	52	6.6
top500	9.8	3.5	2.4	37.7	5.4	21.3	3.1	52	6.6
top750	10.2	3.9	2.7	38.3	5.3	21.3	3.1	52	6.6
top1000	10.3	4.1	2.8	20.3	4.7	21.3	3.1	52	6.6
full	10.8	2.2	1.2	17.7	1.9	21.3	3.1	52	6.6
WSJ	**Mean rel. per query**	**MK1 β=0.5**		**MK1 β=2**		**MK3 β=0.5**		**MK3 β=2**	
		Avg. size	**Avg. rel.**	**Avg. size**	**Avg. rel.**	**Avg. size**	**Avg. rel.**	**Avg. size**	**Avg. rel.**
top100	16.6	25.1	12.3	55.8	16.1	36.5	12	70.2	16
top200	24	32.5	15.1	98.3	22.5	48.6	14.5	125.5	22.6
top350	31.9	48.2	19.1	129.3	28.1	60.7	16.2	188.4	29
top500	37	37.2	15.6	165.1	32.1	63.8	16.4	233.3	32.7
top750	43.5	25	13.5	224.5	36.3	69	16.9	269.7	35.7
top1000	47.7	22.6	12.5	245.4	37.9	73.2	17.2	285.4	36.8

Table 6. Average size and average number of relevant documents for optimal clusters using the group average method (MK1), and for optimal IFS (MK3), for β = 0.5 & β = 2 for LISA and WSJ collections

By definition, an optimal cluster (or an optimal set returned by an IFS) is the one that best combines P and R values. For a collection with a small number of relevant documents per query (such as LISA) we expect 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), we expect the size of optimal clusters to be large. This is confirmed by the data presented in Table 6. 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, is that optimal cluster and optimal IFS size depend on the value of the parameter β that we investigate. Precision oriented searches (β=0.5) lead to smaller sizes, both for clusters and IFS sets, than recall oriented (β=2) searches do.

Burgin, (1995), listed a number of factors imposed by experimental test collections that may affect the level of performance of cluster-based systems (e.g. number of relevant documents per query, number of terms per document). Our results support this view, and indicate that experimental results should be examined in the context of the specific experimental environment that generated them.

4.4 Bottom-level optimal clusters

In the results presented in the previous sections, we have not imposed any constraints on the characteristics of optimal clusters. Croft (1978, 1980), Griffiths *et al.* (1986), and El Hamdouchi and Willett (1989) suggested that if a bottom-up search considers only the *bottom-level clusters* of a document hierarchy (bottom level cluster being the cluster through which a document first joins the hierarchy), its effectiveness exceeds all other types of cluster searches. Table 7 presents some statistics about the size of the bottom-level clusters of the full LISA hierarchy (6003 clusters in total) that was generated by the four clustering methods. 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. Results for the other 4 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	2946	901	104	33	3	2
Ward	3561	2.4	3287	274	0	0	0	0
Complete link	3722	2.6	3248	463	11	0	0	0
Single link	4915	2076.2	1522	525	109	58	19	2682

Table 7. Bottom level cluster size statistics for LISA hierarchies

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

Table 8 presents the percentage of optimal clusters that are bottom-level for the WSJ and LISA collections. We present the results for 2 values of β (0.5, 2). For WSJ 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 P and R (given the different values of β). For these three methods bottom-level clusters have a consistently small size (Table 7), so for a collection with a large number of relevant documents per query, such as WSJ, they wouldn't 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	
	β=2	β=0.5	β=2	β=0.5	β=2	β=0.5	β=2	β=0.5
top100	32.3	67.7	33.3	80	29	71	55.2	89.7
top200	25.8	77.4	29	80.6	38.7	90.3	71	93.5
top350	33.3	76.7	35.5	71	41.9	67.7	83.9	96.8
top500	36.7	76.7	29	67.7	48.4	83.9	77.4	100
top750	67.7	67.7	29	74.2	38.7	74.2	77.4	93.5
top1000	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**	
	β=2	β=0.5	β=2	β=0.5	β=2	β=0.5	β=2	β=0.5
top100	20.8	35.4	6.3	31.3	8.3	20.8	62.5	72.9
top200	6.3	25	0	22.9	2.1	20.8	64.6	62.5
top350	12.5	25	2.1	25	4.2	25	60.4	58.3
top500	10.4	25	2.1	20.8	8.3	25	66.7	54.2
top750	14.6	35.4	4.2	22.9	10.4	31.3	58.3	56.3
top1000	12.5	39.6	6.3	20.8	6.3	29.2	54.2	54.2

Table 8. Percentage of optimal bottom-level clusters

The most notable result from Table 8 is that for the full (static) LISA hierarchy all optimal clusters are bottom-level (the same happens for the other 3 collections). From our 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 is either in the region of 2-6 documents or in the region of a few hundred documents (when a document joins the hierarchy at a low level of similarity). Therefore, optimality for static clustering is reached only in extreme cases where too many or too few documents are contained in the cluster. This happens because the quality of the clustering is not good enough to allow a different behaviour[6].

We therefore believe that previous research that suggested the benefits of bottom-level clusters for retrieval is restricted to the case of static clustering. Query-specific clustering, on the other hand, tends to reach optimality in much more practical settings. Based on these findings we will not report any evaluation for query-specific clustering that constrains optimality to bottom-level clusters.

5. Discussion of the Results

In this section we analyse the experimental results, and discuss their implications.

[6] For example, average sizes of optimal clusters for the CACM collection (β=1) range from 10 to 27 documents for all four methods and values of *n*, whereas static optimal clusters have an average size of 3.6, 7.2, 3.6, and 3.2 documents for group average, complete link, single link and Ward methods respectively.

5.1 Results for different numbers of top-ranked documents

The results that were presented in Table 5 allow us to reach two conclusions: First that there is no significant degradation of effectiveness for decreasing values of *n*, and secondly that static clustering is significantly inferior to any level of query-specific clustering.

The effectiveness for different values of *n* (across rows for the MK1 column at Table 5) appears to increase as *n* increases, but the gains in effectiveness did not prove to be statistically significant. In fact, the only statistically significant cases were with the group-average method and when MK1 at *n*=100 is compared against MK1 at other values of *n*. No statistical significance existed for any other combinations of values of *n* (except for *n*=500 & 750 in the CISI collection). Table 9 gives a summary of the cases in which significance was achieved for β=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.

n	CACM	CISI	MED	WSJ
100 - 200	0.049	-	-	<0.0001
100 - 350	-	0.0328	0.0012	<0.0001
100 - 500	0.05	-	0.0021	0.0002
100 - 750	0.0364	0.0089	0.0027	0.0003
100 - 1000	-	N/A	N/A	0.0011

Table 9. Significance levels for comparisons across values of *n*.
Results are for the group average method, β=1

The results suggest that, with the exception of the smallest value of *n* for the group-average method, there is no significant increase in effectiveness for larger numbers of top-ranked documents. Consequently, 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 was 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 query-specific clustering, for all clustering methods, all collections and all values of β. In fact statistical tests gave significance at level < 0.0001 for all combinations.

We believe that this is due to the very nature of query-specific clustering. It customises the document space to the request, increasing the chance of relevant documents being placed in the same clusters (Hearst & Pedersen, 1996). We also believe that the choice of static clustering in previous research in hierarchic clustering has been a major reason for a series of results that have seen it outperformed by IFS (e.g. El Hamdouchi & Willett, 1989), an issue that we discuss in the next section.

5.2 Cluster vs. inverted file search effectiveness

If MK1-k is used to gauge IFS effectiveness, then all four clustering methods significantly outperform conventional search, for all values of β, and for all values of *n* (including static clustering, although not always significantly). As we mentioned in section 3.3 this comparison is not fair on IFS, however it shouldn't be dismissed since it offers a comparison demonstrating that IFS does not do as well as optimal clusters under the conditions that 'define' cluster optimality. In the rest of this section we will focus on comparisons with the MK3 measure.

n	**Average**	**Ward**	**Complete**
100	0.0002	0.006	0.012
200	<0.0001	0.0035	0.042
350	0.006	0.0062	0.0142
500	0.0013	0.0049	-
750	0.0028	0.022	-
1000	0.011	0.076	-

Table 10. Significance levels for β=1 using the WSJ collection

In Table 10 we present significance levels at which query-specific clustering outperforms IFS for the WSJ collection and for β=1. The single link method did not significantly outperform IFS for β=1. The results for precision-oriented searches (β=0.5) using WSJ show significance for all values of *n* for the other three methods (at levels < 0.001), and for *n*=100 and 200 for the single link method. For recall-oriented searches (β=2) significance is rarely achieved, with the exception of the group-average method in Medline (*n*=200, 350, 500, 750), CISI (*n*=100), and WSJ (*n*=100, 200). Results for the other collections are not presented here but the patterns follow the ones displayed in Table 10. We found the size and characteristics of the WSJ collection more suitable for presentation in this section.

From these results it follows that in our experimental environment optimal query-specific cluster effectiveness significantly outperforms optimal IFS effectiveness for most combinations of clustering methods, values of *n*, and values of β, 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).

The results shown in Table 10 suggest that significance levels do not decrease for increasing values of *n*. On the contrary, the best significance is often achieved for *n*=100 or *n*=200. This result further strengthens our suggestions from the previous section, and suggests that it may be advantageous to use smaller numbers of top-ranked documents as an input to a hierarchic clustering system (i.e. somewhere in the order of 200-350 documents since section 5.1 suggested significant effectiveness losses for *n*=100).

Moreover, from our results it follows that static clustering effectiveness is lower than IFS effectiveness for all experimental conditions, and for the vast majority of cases it is significantly lower. We view this result as a further supporting argument for the application of query-specific clustering to IR.

Based on these results, we can conclude 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 query-specific clusters can prove useful in an operational environment by, for example, triggering a relevance feedback process (Buckley *et al*., 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 we mentioned in section 3.3 (e.g. type of search strategy, cluster visualisation, cluster summarisation method, etc.)[7]. These parameters are external to the document hierarchies, and form separate research issues in their own right.

We believe that our study, as well as that of other researchers that investigate effectiveness issues, can motivate research into areas that are related to these external parameters. Two such areas, that have long been neglected in IR, are cluster-based search strategies and cluster representatives. It has also been acknowledged (Kural *et al*., 2001) that users have difficulty in recognising 'good' clusters based on conventional representations of cluster contents. More research is warranted in this area to investigate more effective cluster representations (e.g. Radev *et al*., 2000).

5.3 Comparison of the four clustering methods

As we mentioned at Section 2, it is not the main aim of this paper to compare the effectiveness of the four clustering methods. However, our results over all experimental conditions indicated that the group average method was the most effective in terms of optimal cluster evaluation. Complete link and Ward methods were close to each other with often negligible differences in effectiveness, while single link displayed the poorest effectiveness of the four. If one looks at static clustering effectiveness, then complete link slightly outperforms the group average method for the majority of experimental conditions. This behaviour for static clustering seems to be in agreement with previous findings (e.g. Willett, 1988).

It is not in our intention to draw any conclusions from these informal remarks, but our data does provide a pool of evidence through which, should one wish, an extensive comparison of the four methods in the context of query-specific clustering can be made.

[7] In fact, the same can be said on whether a user using an IFS system will be able to benefit from its optimal MK3 effectiveness in an operational environment. The way that document contents will be presented to the user as relevance clues, for example, will highly determine the actual effectiveness (Tombros & Sanderson, 1998).

6. Future Work & Conclusions

The research reported in this paper demonstrated that, in the specific experimental environment used, gains in retrieval effectiveness occur from the application of query-specific clustering to IR. We investigated two main issues: firstly whether there is significant variation in the effectiveness of clustering when different numbers of top-ranked documents are clustered and when the full collection is clustered; and secondly whether query-specific clustering effectiveness is higher than that of a conventional IFS.

Our results suggest that there is not a statistically significant variation in query-specific cluster effectiveness for different values of top-ranked documents (with the exception perhaps of the 100 case), and more importantly, that query-specific clustering significantly outperforms static clustering for all experimental conditions. Furthermore, query-specific clustering does outperform IFS, and for most of the experimental conditions this is statistically significant, especially for precision-oriented searches. Static clustering fails to do so.

The main implication of our results is that they provide evidence for the application of hierarchic query-specific clustering to IR based on improved effectiveness, something that had not been formally established in the past. More thorough investigation with other test collections (e.g. larger and more diverse document collections should be investigated), and with end-user cluster-based systems is needed to validate our findings in a wider context.

As the research reported here covers the optimal effectiveness of clusters, an extension of this work would be to exploit this effectiveness in an interactive environment. The results in this paper suggest that optimal clusters contain high proportions of useful documents. It should be the purpose of future research to find methods that will 'guide' users towards the optimal clusters in an interactive search setting. Methods that provide relevance clues to users are likely to prove effective in this direction. Such methods include automatic summarisation (Radev *et al*., 2000), and more specifically query-biased automatic summarisation (Tombros & Sanderson, 1998). It would also be worth comparing the effectiveness of query-specific hierarchic clustering in an interactive search task to that of partitioning methods, such as the Scatter/Gather system (Cutting *et al*., 1992). Alternative methods for customising document hierarchies to user queries are also worth investigating. One such alternative would be to take the query context into account when generating the hierarchies through the similarity measure used.

To conclude, our results show that the application of query-specific hierarchic clustering to IR can introduce a number of improvements, compared to both static clustering effectiveness, and inverted file search effectiveness.

Acknowledgements

The authors would like to thank Ian Ruthven and Dr. Iain Campbell for their comments on earlier drafts of this paper, and the anonymous reviewers for their helpful suggestions. Anastasios Tombros is funded by a University of Glasgow Postgraduate Scholarship. Robert Villa is funded by an EPSRC studentship.

References

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. Milpitas, 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.

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.

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society*, Series A, 134, 321-353.

Croft, W.B. (1978). Organizing and searching large files of document descriptions. Ph.D. Thesis, Churchill College, University of Cambridge.

Croft, W.B. (1980). A model of cluster searching based on classification. *Information Systems*, 5, 189-195.

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.

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.

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.

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.

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.

Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7, 217-240.

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.

Kural, Y., Robertson, S.E., Jones, S. (2001). Deciphering cluster representations. *Information Processing & Management*, 37(4), 593-601.

Leouski, A. and Allan, J. (1998). Evaluating a visual navigation system for a digital library. In *Proceedings of the Second European Conference on Research and Technology for Digital Libraries*, pp. 535-554. Heraklion, Greece.

Murtagh, F. (1984). Structure of hierarchic clusterings: implications for information retrieval and for multivariate data analysis. *Information Processing & Management*, 20(5/6), 611-617.

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.

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.

Rasmussen, E. (1992). Clustering Algorithms. In Frakes, W.B. and Baeza-Yates, R. (editors) *Information Retrieval: Data Structures and Algorithms*. New Jersey: Prentice Hall.

van Rijsbergen, C.J. (1974). 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.

Salton, G. (1968). *Automatic Information Organization and Retrieval*. New York: McGraw-Hill.

Salton, G. (1971). *The SMART Retrieval System - Experiments in Automatic Document Retrieval.* Englewood Cliffs, New Jersey: Prentice Hall Inc.

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.

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.

Späth, H. (1980). *Cluster Analysis Algorithms For Data Reduction and Classification of Objects*. Chichester: John Ellis Horwood Limited.

Voorhees, E.M. (1985). 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. and Harman, D.K. (1997). *Proceedings of the Fifth Text Retrieval Conference*. National Institute of Standards and Technology Special Publication 500-238.

Tombros, A. and Sanderson, M. (1998). The advantages of query-biased summaries in information retrieval. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 2-10. Melbourne, Australia.

Ward, J.H. (1963). Hierarchical grouping to minimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

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.

Zamir, O. and Etzioni, O. (1998). Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 46-54. Melbourne, Australia.

Appendix

Ward									
CACM	***β*=1**			***β*=0.5**			***β*=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	***β*=1**			***β*=0.5**			***β*=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	***β*=1**			***β*=0.5**			***β*=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	***β*=1**			***β*=0.5**			***β*=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	***β*=1**			***β*=0.5**			***β*=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 A1. Evaluation results for Ward's method. Highest effectiveness (the lowest value of E) for each column appears in bold.

Complete Link									
CACM	**β=1**			**β=0.5**			**β=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	**β=1**			**β=0.5**			**β=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	**β=1**			**β=0.5**			**β=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	**β=1**			**β=0.5**			**β=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	**β=1**			**β=0.5**			**β=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 A2. Evaluation results for the complete link method. Highest effectiveness (the lowest value of E) for each column appears in bold.

Single Link									
CACM	β=1			β=0.5			β=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	β=1			β=0.5			β=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	β=1			β=0.5			β=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	β=1			β=0.5			β=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	β=1			β=0.5			β=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 A3. Evaluation results for the single link method. Highest effectiveness (the lowest value of E) for each column appears in bold.