# Query-Sensitive Similarity Measures for the Calculation of Interdocument Relationships

Anastasios Tombros[1] and C.J. van Rijsbergen

Department of Computing Science, University of Glasgow,
Glasgow G12 8RZ, U.K.

{tombrosa, keith}@dcs.gla.ac.uk

**Abstract.** The application of document clustering to information retrieval has been motivated by the potential effectiveness gains postulated by 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. In this paper we propose that, for any given query, pairs of relevant documents will exhibit an inherent similarity which is dictated by the query itself. Our research describes an attempt to devise means by which this similarity can be detected. We propose the use of query-sensitive similarity measures that bias interdocument relationships towards pairs of documents that jointly possess attributes that are expressed in a query. We experimentally tested query-sensitive measures against conventional ones that do not take the context of the query into account. We calculated interdocument relationships for varying numbers of top-ranked documents for five document collections. Our results show a consistent and significant increase in the number of relevant documents that become nearest neighbours of any given relevant document when query-sensitive measures are used. These results suggest that the effectiveness of a cluster-based IR system has the potential to increase through the use of query-sensitive similarity measures.

## 1. Introduction

C*luster analysis* is a technique that allows the identification of groups, or clusters, of similar objects in multi-dimensional space. It was initially introduced in the field of *Information Retrieval* (IR) as a means of improving the efficiency of serial search (Salton, 1971). Apart from efficiency, effectiveness was also put forward for the use of *hierarchic clustering* in IR (Jardine & Van Rijsbergen, 1971; Croft, 1978). Relevant documents that might have otherwise been ranked low in a traditional *inverted file search* (IFS), will be (through interdocument associations) grouped together with other relevant documents, thus improving the effectiveness of an IR system (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 ones, and therefore tend to appear in the same clusters (Jardine & Van Rijsbergen, 1971). Various tests have been used to quantify the degree to which documents relevant to the same queries (*co-relevant* documents) adhere to the hypothesis (Jardine & Van Rijsbergen, 1971; Voorhees, 1985; El-Hamdouchi & Willett, 1987).

The results of individual tests are interpreted by researchers as being indicative of a specific collection's clustering tendency (Jardine & Van Rijsbergen, 1971; Croft, 1978; Voorhees, 1985; Griffiths *et al*., 1986; El-Hamdouchi-Willett, 1987). If the Cluster Hypothesis holds for a particular document collection, then relevant documents will be well separated from non-relevant ones, and hence a cluster-based search strategy (e.g. Jardine & Van Rijsbergen, 1971; Croft, 1980) will likely be effective.

In this paper we propose that the Cluster Hypothesis should not be viewed as a test for an individual collection's clustering tendency. Instead, we argue that the hypothesis should be valid for every collection, and therefore failure to validate the hypothesis is not caused by properties of the collection under examination but rather by our assumptions about interdocument similarity. We postulate that, for any given query, pairs of relevant documents will exhibit an inherent similarity which is dictated by the query itself. Conventional measures of interdocument relationships (Ellis *et al*., 1993), such as the cosine coefficient for example, can not detect such a similarity, since they do not take into account the specific context (i.e. query) under which the similarity of two objects is judged.

Our research describes an attempt to devise means by which this similarity can be detected. We propose the use of *query-sensitive similarity measures* that bias interdocument relationships towards pairs

[1] Author to whom correspondence should be addressed.

of documents that jointly possess attributes (i.e. terms) that are expressed in a query. In this way we 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 calculating interdocument relationships in IR, purpose can be defined as a per-query adherence to the Cluster Hypothesis.

The aim of this paper is to introduce the notion of query-sensitive similarity, to propose specific formulas for its calculation, and to test its effectiveness against conventional similarity measures. The remainder of the paper first presents some necessary background in section 2. Query-sensitive similarity measures for calculating interdocument relationships are then presented in section 3, followed by a description of the experimental environment under which their effectiveness will be evaluated in section 4. Experimental results are presented and discussed in section 5, and section 6 presents related research. Finally, in section 7 conclusions are drawn, and some points for further research are mentioned.

## 2. Background

There are many measures available for the calculation of interdocument relationships (e.g. Van Rijsbergen, 1979; Jones & Furnas, 1987; Ellis *et al*., 1993), and the choice of a specific measure may influence the outcome of the calculations. Van Rijsbergen, (1979), advised against the use of any measure that is not normalised by the length of the document vectors, something that was experimentally verified by (Willett, 1983). Van Rijsbergen also noted that most of the measures are monotone in respect to each other, and therefore methods that depend only on the rank ordering of the similarity values would give similar results for all such measures. Hubálek, (1982), suggested that each scientific area, after argument and trial, should settle down on those measures most appropriate for its needs. For the field of IR (Ellis *et al*., 1993) have concluded that "the historical attachment to the association coefficients provided by the Dice and cosine formulae is in no need of revision".

Clustering methods, as applied to IR, typically require as input a similarity matrix that contains values of all interdocument associations (Van Rijsbergen, 1979; Willett, 1988). Traditionally, clustering has been applied statically over the whole document collection prior to querying (*static clustering*). Under static clustering, interdocument relationships are also 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.

Equation 1 demonstrates this for the cosine coefficient[2] that is commonly used to measure interdocument relationships (Ellis *et al*., 1993). From equation 1 it follows that the similarity between the two objects depends only on the weights of their constituent terms ($d_{ij}$ and $d_{jk}$). Therefore, for a particular document collection $Sim(D_i, D_j)$ will be the same across all requests.

$$Sim(D_i, D_j) = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} \quad \text{(Eqn.1)}$$

The static notion of similarity has been implicitly[3] challenged by (Hearst & Pedersen, 1996). Hearst and Pedersen viewed the Cluster Hypothesis under the light of *query-specific clustering*, an approach to clustering proposed and tested by (Preece, 1973; Willett, 1985). Query-specific clustering is applied to the search results of an IR system (i.e. the top-*n* ranked documents returned by an IFS).

The re-examination of the Cluster Hypothesis by Hearst and Pedersen postulated that relevant documents tend to appear in the same clusters, but the clusters are created as a function of the documents that are retrieved in response to a query, and therefore have the potential to be more closely tailored to the characteristics of a query than a static clustering (Hearst & Pedersen, 1996).

[2] Our discussion on similarity measures will be based on the cosine coefficient. However, our arguments can easily be extended to other similarity or dissimilarity measures, such as the Dice coefficient or Euclidean distances.

[3] The variation introduced in similarity under query-specific clustering is implicit because it is not explicitly defined by the similarity measure.

A consequence of this is that the similarity between any two documents $D_i$ and $D_j$, assuming that they are both retrieved in the top-*n* documents for different queries, will be different under each query. This difference is attributed 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 ($d_{ij}$ and $d_{jk}$ in equation 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 query-specific clustering, interdocument associations are defined through enumeration of common terms, and a mathematical formulation that quantifies this enumeration (e.g. equation 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 inter-object 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 purposes of clustering relevant objects together.

We view the query as the context under which the similarity of two documents, that are retrieved in response to this query, is judged. The context assigns greater importance to these dimensions (i.e. terms) that are more significant for accomplishing a specific goal. The goal in the context of IR is, for any query, to place relevant documents closer to each other than to non-relevant ones, hence enforcing the validity of the Cluster Hypothesis.

According to this approach, interdocument similarity is dynamic, and changes explicitly depending on the query. Some measure of variability needs then to be introduced in equation 1, so that $Sim(D_i, D_j)$ varies depending on the query that has retrieved the pair of documents. We will call such a class of similarity measures q*uery-sensitive* measures, and we will present them in the following section.

## 3. Query-Sensitive Similarity Measures

Query-sensitive measures can be defined as a function of two components. The first one corresponds to the conventional similarity between two documents $D_i$, $D_j$, and is given by equation 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 we will represent this component by $Sim(D_i, D_j, Q)$. This is the variable component of the similarity. The query-sensitive similarity $Sim(D_i, D_j \mid Q)$ can therefore be defined as:

$$Sim(D_i, D_j \mid Q) = f(Sim(D_i, D_j), Sim(D_i, D_j, Q)) \quad \text{(Eqn.2)}$$

We define $Sim(D_i, D_j, Q)$ by finding all common terms between $D_i$, and $D_j$, and seeing which of the common terms are also terms that appear in the query Q. Similarity between pairs of document 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 3. In this equation $Q=\{q_1, q_2, \ldots, q_l\}$ is the query vector of length *l*, $D_i$ and $D_j$ the two document vectors, and $C = D_i \cap D_j = \{c_1, c_2, \ldots, c_k, \ldots, c_m\}$ is a vector of length *m* containing the common terms of documents $D_i$ and $D_j$.

$$Sim(D_i, D_j, Q) = \frac{\sum_{k=1}^{\max(m,l)} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{m} c_k^2 \cdot \sum_{k=1}^{l} q_k^2}} \quad \text{(Eqn. 3)}$$

We represent the terms of the common vector C by $c_k = (d_{io} + d_{jp}) / 2$, where $d_{io}$, and $d_{jp}$ are the weights of the 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. We investigated other representations of $c_k$ ($\min(d_{io}, d_{jp})$, $\max(d_{io}, d_{jp})$, $(d_{io} * d_{jp})$), and did not find significant differences. We report on this specific form which proved to be the most effective.

Substituting equations 1 and 3 in equation 2, and defining $Sim(D_i, D_j \mid Q)$ as the product of equations 1 and 3, we derive equation 4. We will call this query-sensitive measure *M1*. Measure M1 is a combination of two sources of evidence. The first evidence comes from the conventional static similarity of $D_i$, and $D_j$ (equation 1). The second source comes from the common similarity of the pair of documents

and the query, and it augments the first source. Consequently, for a specific query, pairs of documents that have more terms in common with the query than other pairs will be assigned greater similarity values. This reflects the belief that under the context defined by the query, query terms possess greater salience when determining interdocument relationships.

$$Sim(D_i, D_j \mid Q) = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} \cdot \frac{\sum_{k=1}^{\max(m,l)} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{m} c_k^2 \cdot \sum_{k=1}^{l} q_k^2}} \quad \text{(Eqn.4)}$$

It should be noted that 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 choice reflects the assumption that the presence of query terms is required for a document to be relevant. However, it may cause the similarity between documents that share a large number of terms to equal zero. In the specific case where one of these documents also contains a number of query terms (and hence likely relevant), it may be worth not setting $Sim(D_i, D_j \mid Q)$ to zero as equation 4 does. In this way we may allow for relevant documents that do not contain any of the query terms to be discovered through high association with a relevant document that contains some of the query terms. Further research towards this end is warranted.

As a more extreme form of query term biasing, we also decided to use equation 3 in the experiments reported in this paper. We will call this measure *M2*. Measure M2 only takes into account common terms between the two documents that are also query terms. Like M1, $Sim(D_i, D_j \mid Q)$ will equal zero if none of the common terms is a query term. However, unlike M1, the overall similarity between documents $D_i$, and $D_j$ does not take into account co-occurrence of other terms in the two documents. The effectiveness attained with M2 can be seen as a lower limit of the effectiveness of query-sensitive measures.

Both for M1 and M2, $0 \le Sim(D_i, D_j \mid Q) \le 1$. To preserve the reflexivity of the measures defined by M1 and M2, (i.e. Sim ($D_i$, $D_i$)=1), we define the similarity of a document with itself to be equal to 1. This does not come as a result of either equation 3, or equation 4, but can be introduced by definition. Finally, $Sim(D_i, D_j \mid Q) = Sim(D_j, D_i \mid Q)$ both for M1 and M2 (i.e. query-sensitive similarity is symmetric). These properties are in accordance with those of conventional similarity measures (Van Rijsbergen, 1979).

### *3.1 Limitations*

The assumption that query terms are sufficient indicators of document relevance is made 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, 1970), although it does not automatically result in relevance for users (Barry, 1994).

Measures M1 and M2 do not take into account other dimensions of the concept of relevance. This limitation is not unique to our proposed approach. The majority of IR research to date has focused on the topical aspect of relevance, assuming that query terms provide enough evidence about the user's information need.

A second 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). Both measures defined previously (M1, and M2) regard query terms as the dimensions that acquire significant discriminatory power. If only 2 or 3 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 (e.g. automatic query expansion by lexical-semantic relations (Voorhees, 1994)) could also be applied here. We will further investigate this issue in section 5.2.

## 4. Experimental Details

The experiments reported here aim to investigate the effectiveness of the proposed query-sensitive measures (M1 and M2) in 'forcing' documents that are likely to be co-relevant to be more similar to each other than when using a conventional similarity measure. In other words, we examine the degree to which the Cluster Hypothesis is adhered to. If query-sensitive measures are more effective in placing co-relevant

documents closer to each other than conventional measures, then their application to document clustering can be expected to prove more effective.

Two evaluation tests[4] that measure the degree of separation between relevant and non-relevant documents have been widely applied to IR. Jardine and Van Rijsbergen, (1971), proposed the overlap test, and (Voorhees, 1985) the N-Nearest Neighbour test.

We chose to use the N-Nearest Neighbour test proposed by (Voorhees, 1985) because it fits best with our experimental aims. 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 that neighbourhood. The higher the number of relevant documents, the higher the separation of relevant documents from non-relevant ones. 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) can be obtained when averaging over all of the relevant documents for all the queries in a collection. This single value is calculated and displayed in the results presented in section 5.

The 5NN test 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 with its nearest neighbour (e.g. Griffiths *et al*., 1986; El-Hamdouchi, 1987). Therefore, in addition to the 5NN test we also calculated the percentage of relevant documents whose most similar neighbour is also relevant. We will call this test NN so as to distinguish it from the 5NN test.

### *4.1 Document Collections and Initial Retrieval*

Five document collections were used for the experiments (CACM, CISI, LISA, Medline, and WSJ). WSJ is one of the TREC standard collections (Voorhees & Harman, 1997). The characteristics of the five document collections are given 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[5].

For the WSJ collection, TREC queries 1-50 were used in the experiments. The *Title* section of the queries and a number of manually selected terms from the *Concepts* field were used as query terms. On average 4.4 terms per query were added from the *Concepts* field, yielding an average of 7.6 terms per query for the WSJ collection (table 1). The *Concepts* field usually lists terms and phrases that the creator of the query thinks are related to it (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 terms per query | 13 | 7.6 | 19.4 | 9.9 | 7.6 |
| 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

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 (SMART's *ltc* scheme). The default SMART stoplist and stemming were used in indexing all the collections and queries.

After the initial retrieval, the top-*n* ranked documents were used in order to create the collections that would be investigated. Seven different values of *n* were used: 100, 200, 350, 500, 750, 1000, and full collection (*n* = collection size)[6]. Our motivation for using different values of *n* (as opposed to testing only for the full collection size for example) was twofold. Firstly we were interested in examining how the

[4] El Hamdouchi & Willett, (1987), proposed the density test to measure the clustering tendency of a collection. This test does not quantify the separation between relevant and non-relevant documents, and hence is not considered here.

[5] WSJ documents also tend to be much longer than those of the other four collections.

[6] 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 due to practical limitations.

results would scale for increasing values of *n*, when more non-relevant documents are introduced in the document sets. Secondly, recent research (Tombros *et al*., 2001) has suggested that optimal hierarchical clustering effectiveness occurs for smaller values of *n*.

The same weighting scheme as for the initial retrieval was applied to the document vectors of the sets whose interdocument relationships we were calculating. After initial experimentation with different vector weighting schemes for the cosine coefficient (binary weights, term frequency weights) no significant differences were found - which is in agreement with previous suggestions and findings (Van Rijsbergen, 1979; Willett, 1983; Ellis *et al*., 1993). However, we did not examine the effect of other query weighting schemes on the effectiveness of query-sensitive measures.

## 5. Experimental Results

In tables 2-6 we present the results of the 5NN, and NN tests. Each table comprises eight columns. In the first column the different values of *n* are given for which results were calculated. In the second column we present the average number of relevant documents per query for each value of *n*.

Columns 3-5 contain the results obtained for the 5NN test with the cosine coefficient, M1, and M2 respectively. In columns 4 and 5 we also calculated the percentile difference between the results for M1-Cosine and M2-Cosine respectively. The differences are displayed in brackets. For each of these three columns, the highest value across all values of *n* is displayed in bold.

Columns 6-8 contain the results for the NN test for the cosine, M1, and M2 measures respectively. In each column we present the percentage of relevant documents whose most similar document is also relevant. For each column, the highest percentage is again displayed in bold.

Testing for statistical significance of the results was done using the Wilcoxon signed-ranks test. This test is a powerful statistical tool that makes no assumptions about the distribution of the values that it is comparing (Croft, 1978, pp. 27-29; El-Hamdouchi, 1987, pp. 158-159).

| CACM | Mean Rel. Docs per query | Cosine 5NN | M1 5NN | M2 5NN | Cosine NN (%) | M1 NN (%) | M2 NN (%) |
|---|---|---|---|---|---|---|---|
| **top 100** | 10.46 | **1.621** | 1.924 (18.72%) | 1.663 (2.63%) | **51.65** | 58.46 | 55.51 |
| **top 200** | 11.62 | 1.511 | 1.981 (31.17%) | **1.764** (16.79%) | 45.92 | 58.74 | **61.56** |
| **top 350** | 12.69 | 1.415 | 2.028 (43.27%) | 1.717 (21.29%) | 45.97 | 59.36 | 60.58 |
| **top 500** | 13.21 | 1.393 | 2.039 (46.37%) | 1.688 (21.17%) | 46.35 | 58.48 | 59.65 |
| **top 750** | 13.58 | 1.376 | **2.045** (48.67%) | 1.649 (19.83%) | 44.95 | **60.17** | 57.75 |
| **top 1000** | 13.83 | 1.35 | 2.017 (49.45%) | 1.639 (21.47%) | 43.3 | 59.36 | 55.59 |
| **full** | 15.31 | 1.366 | 1.859 (36.08%) | 1.596 (16.81%) | 43.76 | 54.48 | 50.95 |

**Table 2.** CACM results

| CISI | Mean Rel. Docs per query | Cosine 5NN | M1 5NN | M2 5NN | Cosine NN (%) | M1 NN (%) | M2 NN (%) |
|---|---|---|---|---|---|---|---|
| **top 100** | 16.31 | **1.53** | **1.728** (13.35%) | **1.522** (-0.52%) | **45.44** | **52.11** | 55.79 |
| **top 200** | 24.71 | 1.37 | 1.652 (20.62%) | 1.512 (10.37%) | 39.98 | 49.25 | **56.2** |
| **top 350** | 32.31 | 1.253 | 1.66 (32.51%) | 1.483 (18.39%) | 35.75 | 47.88 | 54.34 |
| **top 500** | 37.06 | 1.203 | 1.625 (35.09%) | 1.411 (17.30%) | 33.87 | 46.53 | 50.85 |
| **top 750** | 42.34 | 1.14 | 1.55 (35.84%) | 1.304 (14.34%) | 32.82 | 45.1 | 44.77 |
| **full** | 49.77 | 1.119 | 1.433 (28.06%) | 1.208 (7.97%) | 32.85 | 41.3 | 37.05 |

**Table 3.** CISI results

| LISA | Mean Rel. Docs per query | Cosine 5NN | M1 5NN | M2 5NN | Cosine NN (%) | M1 NN (%) | M2 NN (%) |
|---|---|---|---|---|---|---|---|
| **top 100** | 7.12 | **0.896** | 1.362 (52.05%) | **1.3** (45.04%) | **30.3** | 46.32 | **49.35** |
| **top 200** | 8.62 | 0.845 | 1.376 (62.84%) | 1.135 (34.35%) | 27.68 | 43.6 | 44.64 |
| **top 350** | 9.17 | 0.784 | **1.449** (84.80%) | 1.17 (49.19%) | 26.27 | 45.89 | 45.89 |
| **top 500** | 9.83 | 0.783 | 1.425 (81.92%) | 1.164 (48.55%) | 27.43 | **47.2** | 45.13 |
| **top 750** | 10.2 | 0.776 | 1.41 (81.68%) | 1.133 (46.02%) | 27.2 | 44.76 | 46.18 |
| **top 1000** | 10.34 | 0.768 | 1.391 (81.18%) | 1.114 (45.08%) | 28.2 | 43.85 | 46.65 |
| **full** | 10.83 | 0.859 | 1.381 (60.73%) | 1.269 (47.64%) | 28.27 | 44.53 | 43.47 |

**Table 4.** LISA results

| MED | Mean Rel. Docs per query | Cosine 5NN | M1 5NN | M2 5NN | Cosine NN (%) | M1 NN (%) | M2 NN (%) |
|---|---|---|---|---|---|---|---|
| **top 100** | 18.97 | **3.143** | **3.569** (13.55%) | 3.361 (6.94%) | **71.88** | **80.49** | 79.44 |
| **top 200** | 20.37 | 3.022 | 3.54 (17.14%) | **3.367** (11.42%) | 67.43 | 76.76 | **80.2** |
| **top 350** | 21.03 | 3.023 | 3.501 (15.81%) | 3.31 (9.49%) | 68.78 | 76.39 | 78.92 |
| **top 500** | 21.13 | 3.003 | 3.475 (15.72%) | 3.305 (10.06%) | 68.35 | 76.06 | 78.58 |
| **top 750** | 21.3 | 3.004 | 3.466 (15.38%) | 3.285 (9.35%) | 68.23 | 76.06 | 78.09 |
| **full** | 23.2 | 3.016 | 3.235 (7.26%) | 3.049 (1.09%) | 68.39 | 72.41 | 69.83 |

**Table 5.** Medline results

| WSJ | Mean Rel. Docs per query | Cosine 5NN | M1 5NN | M2 5NN | Cosine NN (%) | M1 NN (%) | M2 NN (%) |
|---|---|---|---|---|---|---|---|
| **top 100** | 16.63 | **2.122** | 2.357 (11.1%) | **1.872** (-11.78%) | **64.41** | **67.42** | **56.02** |
| **top 200** | 24.02 | 2.051 | 2.446 (19.29%) | 1.827 (-10.92%) | 57.24 | 62.1 | 49.7 |
| **top 350** | 31.88 | 1.909 | **2.468** (29.29%) | 1.832 (-4.03%) | 54.05 | 63.73 | 50 |
| **top 500** | 37 | 1.863 | 2.463 (32.19%) | 1.856 (-0.38%) | 52.65 | 62.9 | 48.64 |
| **top 750** | 43.54 | 1.734 | 2.421 (39.62%) | 1.838 (6%) | 49.19 | 61.82 | 48.18 |
| **top 1000** | 47.75 | 1.711 | 2.416 (41.23%) | 1.799 (5.14%) | 47.6 | 60.43 | 47.73 |

**Table 6.** WSJ results

In section 5.1 we will present results that compare the effectiveness of the three similarity measures used, and in section 5.2 we will report on experiments that study the effect of query length on M1 and M2 for the WSJ collection.

### *5.1 Comparative Effectiveness of the Similarity Measures*

**Query-sensitive measures vs. cosine**

The results obtained for the 5NN test across the five test collections (columns 3-5) show that both the query-sensitive measures, in the vast majority of experimental conditions, are more effective than the cosine coefficient. The only exception to this is noted in the CISI and WSJ collections, where M2 is less effective than the cosine for $n$=100, and for $n$=100, 200, 350, and 500 respectively.

Statistical tests of the results revealed significant improvements of M1 over the cosine (significance level was <0.001 for the majority of cases) for all experimental conditions except for the CISI collection when $n$=100. Measure M2 was significantly more effective than the cosine for the CACM (except for $n$=100), LISA, and Medline (except for $n$=100, 750, full) collections. The levels of significance for M2 were not as low as the ones for M1, though still lower than 0.04 for all significant cases.

If we look at the results across different values of $n$ (across the rows of the tables for column 3), we can see that the cosine coefficient always gives the highest value for $n$=100, and values then follow a decreasing pattern for increasing values of $n$. As values of $n$ increase, so do the numbers of non-relevant documents that are present in the document sets. The cosine coefficient seems to be affected by the non-relevant documents introduced. Recent research has also shown that the decrease of the 5NN values across increasing values of $n$ is, in the majority of cases, statistically significant (Tombros *et al*., 2001).

Measures M1 and M2 (across rows of the tables for columns 4-5) seem to be less affected by the increasing numbers of non-relevant documents introduced. M1 for the CACM, LISA, and WSJ collections shows the highest scores for $n$=750, 350, and 350 respectively. M2 for the CACM and Medline collections displays the highest scores for $n$=200. Statistical tests across different values of $n$ were not performed, as it is not the aim of this paper to examine effectiveness variations for different sets of retrieved documents.

The results for the NN test (columns 6-8) reveal a similar pattern to those for the 5NN test. M1 is significantly more effective than cosine for all experimental conditions (significance levels < 0.02). M2 is significantly more effective for the CACM (except for $n$=100), LISA, and Medline (except for $n$=full) collections (significance levels < 0.03). Therefore, measures M1 and M2 are likely to increase the effectiveness of a clustering system that employs nearest neighbour clusters, such as those proposed by (Griffiths *et al*., 1986). It is worth noting that similar to the 5NN test, for the WSJ collection measure M2 performs worse than the cosine for most values of $n$.

Based on the results of both tests we can conclude that measure M1 is 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.

Results obtained with measure M2, as we discussed in section 3, can be seen as a lower limit for the effectiveness of query-sensitive measures. 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. We view this result as further evidence confirming the applicability of query-sensitive measures to IR.

**Comparative effectiveness of M1 and M2**

The results for the 5NN test in tables 2-6 (columns 4-5) show that M1 achieves higher scores than M2 for all experimental conditions. Statistical testing showed that M1 is significantly more effective than M2 for all experimental conditions (significance levels < 0.001), except for the LISA collection for $n$=100.

The results for the NN test however, show a different pattern. The scores in the tables (columns 7-8) show that in most cases M2 manages to place co-relevant documents as nearest neighbours more often than M1 does. Statistical testing however failed to confirm significance, except for the Medline collection for $n$=200. For the WSJ collection, where as we saw previously M2 performs worse than both M1 and the cosine for the 5NN test, M1 is significantly more effective than M2 for all values of $n$.

For each value of $n$, both measures display large standard deviations in their results for the 5NN test. For example, the deviation for M1 for the WSJ collection ranges from 1.35 ($n$=100) to $n$=1.16 ($n$=1000). This variation suggests that specific properties of the queries may influence the effectiveness of the measures. This is further strengthened by the observation that for specific queries M2 consistently outperforms M1. An analysis of the results on a per query basis may reveal query properties that favour either of the measures. Such an analysis is not within the aims of this paper.

Based on these results it is valid to state that M1 is significantly more effective than M2. In the following section, we compare the effect that query length has on these two measures.

### *5.2 Effect of Query Length on M1 and M2*

In the results for the 5NN test in tables 2-6 (columns 3-5), M2 scored significantly higher 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 CISI and WSJ). This is a consequence of the strong dependence of measure M2 on query terms.

In order to investigate the effect of query length on the effectiveness of M1 and M2, we used an expanded and a shorter version of the 50 TREC queries for the WSJ collection. For the expanded version, terms from the *Title*, *Description*, and *Concepts* fields of each query were used, yielding on average 23.4 terms per query (compared to 7.6 terms initially). For the shorter version of the queries we used only the *Title* field, with an average of 3.2 terms per query.

| | **M1 expanded 5NN** | **M2 expanded 5NN** | **M1 short 5NN** | **M2 short 5NN** |
|---|---|---|---|---|
| **top 100** | 2.457 (4.22%) | 2.372 (26.67%) | **2.32** (-1.57%) | **1.672** (-10.32%) |
| **top 200** | 2.535 (3.63%) | 2.370 (29.69%) | 2.271 (-7.15%) | 1.631 (-10.72%) |
| **top 350** | **2.54** (2.91%) | 2.415 (31.82%) | 2.241 (-9.21%) | 1.536 (-16.18%) |
| **top 500** | 2.54 (3.14%) | **2.425** (30.71%) | 2.195 (-10.89%) | 1.525 (-17.81%) |
| **top 750** | 2.441 (0.83%) | 2.407 (30.93%) | 2.101 (-13.22%) | 1.434 (-21.98%) |
| **top 1000** | 2.437 (0.85%) | 2.399 (40.25%) | 2.064 (-14.6%) | 1.435 (-20.23%) |

**Table 7.** The effect of query length

We repeated the 5NN experiment for both the expanded and shorter versions of the queries on the same sets of documents for each value of *n*. The results are presented in table 7, where highest values for each column are displayed in bold. For columns 2-5 we present in brackets the percentile differences between the reported values and those obtained with the standard queries (table 6, columns 4-5).

The results in table 7 confirm the strong dependence of M2 on query length. M2 with the expanded queries (column 3) is significantly more effective than with the initial queries for all values of *n* (significance levels <0.001). Moreover, it is now significantly more effective than the cosine coefficient for all values of *n* (significance levels <0.03), and is not significantly worse than M1 (either with expanded or initial queries).

Column 5 of table 7 shows a significant drop 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.

Measure M1 on the other hand is 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 table 7 (column 2) are significant. However, when short queries are used (column 4), M1 displays a significant decrease in effectiveness. The decrease is smaller in scale than that reported for M2, but significant (significance levels <0.03) for all values of *n* except for *n*=100. Despite this decrease, M1 using the short queries is still significantly more effective than the cosine (table 6 column 3, significance levels < 0.003).

These results suggest that measure M2 is highly affected by query length, and it would not seem suitable for situations where very short queries are usually input by users unless reliable ways to expand the query could be used. However, as we mentioned in section 5.1, there are specific queries for which M2 outperforms M1, even in the case of short queries. Further research would be needed to investigate whether one can correlate query characteristics with an optimal choice of similarity measure.

Measure M1 is not as much affected by query length as M2, something that would appear useful in an operational environment, like a web search engine for example, where user queries comprise only few terms (Jansesn *et al*., 2000). In our experimental environment, M1 significantly outperformed the cosine when short queries were used. It remains to be seen whether such improvements would occur in operational environments.

## 6. Related Research

The query-sensitive similarity measures we presented in this paper increase the similarity of co-relevant documents on a per-query basis, aiming to increase the probability that such documents will be placed in

the same clusters. A number of approaches that try to 'force' co-relevant documents in the same clusters[7] have been developed in the past under the name of *user-oriented*, or *adaptive clustering* (e.g. Yu *et al*., 1985; Raghavan & Deogun, 1986; Gordon, 1991; Bartell *et al*., 1995). These approaches require user feedback in terms of document relevance as in (Yu *et al*., 1985), or in terms of exhaustive target interdocument similarity values as in (Bartell *et al*., 1995). User supplied information is then used to optimally predict a useful clustering of the documents, by trying to place documents that are likely to be jointly accessed (or jointly assessed as relevant) in response to a set of queries in the same clusters.

This implicitly assumes that there are means of monitoring user activities, collecting usage information, and incorporating this information in the cluster-based 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.

In contrast to adaptive clustering methods, our approach 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 is the query and the document set.

Evidence supporting our view about the salience of specific features for measuring inter-object relationships is provided 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).

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.

## 7. Conclusions & Future Research

In this paper we introduced the notion of query-sensitive similarity measures for the calculation of interdocument relationships. Such measures bias similarity towards pairs of documents that jointly possess terms that are expressed in a query. We presented two query-sensitive measures. The first one takes into account all common terms between a pair of documents, but biases the measure towards those common terms that are also query terms (measure M1). The second one only takes into account common terms that are query terms (measure M2).

Through a series of experiments that assess the degree at which a similarity measure places relevant documents closer to each other than to non-relevant ones, we demonstrated that the query-sensitive measures are significantly more effective than the cosine coefficient. More specifically, measure M1 is always significantly more effective than the cosine, and is not dependent on query length. Measure M2 on the other hand, is sensitive to variations of query length, but despite this it also brought significant improvements over the cosine in a large number of experimental conditions.

Our results demonstrate the applicability of query-sensitive measures to IR. More thorough evaluation of such measures can be performed if one integrates them in a wider application area. An obvious area where query-sensitive measures can be applied is document clustering. We are currently investigating whether the effectiveness improvements introduced by query-sensitive measures in this paper apply to document clustering. We believe that query-sensitive measures have the potential to introduce effectiveness improvements both from a system (intrinsic), and a user (extrinsic) point of view. Further research would be needed to warrant these assumptions.

[7] El-Hamdouchi, (1987), proposed a (static) clustering approach that aimed at forming clusters with a high probability of containing co-relevant documents. El-Hamdouchi's research used a function that ranked documents, or sets of documents, in relation to a query, and did not challenge the use of static interdocument similarity.

In section 3.1 we mentioned two major limitations of the proposed measures. Further work should aim to address such limitations. For example, alternative methods of biasing the similarity measures (e.g. by using user profiles) can be investigated. Furthermore, a more systematic analysis of the dependence of such measures on query length would be appropriate. A specific research issue would be to automatically detect terms in a document set with diagnostic (or classificatory) value (Tversky, 1977), other than the query terms, that could augment the query-sensitive component of the similarity measures.

In conclusion, we view similarity as a dynamic and purpose-sensitive notion. In the context of IR, we demonstrated that query-sensitive measures have the potential to capture the dynamics of similarity for the calculation of interdocument relationships.

## References


Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149-159.

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.

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

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

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.

El-Hamdouchi, A. (1987). Using inter-document relationships in information retrieval. PhD Thesis, 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.

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.

Goodman, N. (1972). Seven strictures on similarity. In Goodman, N. (ed.) *Problems and Projects*, pp. 437-447. Indianapolis and New York: Bobbs-Merrill.

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.

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.

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.

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

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.

Nosofsky, R.M. (1986). Attention, Similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39-57.

Preece, S.E. (1973). Clustering as an output option. *Proceedings of the American Society for Information Science*, 10:189-190.

van Rijsbergen, C.J. (1979). *Information Retrieval*. Butterworths, London, 2nd Edition.

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

Saracevic, T. (1970). The concept of "relevance" in information science: A historical review. In Saracevic, T. (Ed.), *Introduction to Information Science*, 111-151. R.R. Bowker, New York, USA.

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.

Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213-1228.

Tombros, A., Villa, R., van Rijsbergen, C.J. (2001). The effectiveness of query-specific hierarchical clustering in information retrieval. To appear in *Information Processing & Management*.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327-352.

Voorhees, E.M. (1985). The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval. PhD Thesis, Technical Report TR 85-705 of the Department of Computing Science, Cornell University.

Voorhees, E.M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual ACM SIGIR Conference*, pp. 61-69. Dublin, Ireland.

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

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.

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.