Query-sensitive similarity measures for information retrieval

Anastasios Tombros 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 an axiomatic view of the hypothesis, by suggesting that documents relevant to the same query (co-relevant documents) display an inherent similarity to each other which is dictated by the query itself. Because of this inherent similarity, the cluster hypothesis should be valid for any document collection. 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 three query-sensitive measures against conventional ones that do not take the context of the query into account, and we also examined the comparative effectiveness of the three query-sensitive measures. We calculated interdocument relationships for varying numbers of top-ranked documents for six 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 a space that is typically assumed to be multi-dimensional. 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 conceptually lies in the heart of the clustering process. If relevant documents are indeed more similar to each other than to non-relevant ones, then the effectiveness of document clustering should indeed be high, as the likelihood of placing documents relevant to the same requests (*co-relevant* documents) in the same clusters will also be high.

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

In this paper we propose an alternative view of the cluster hypothesis. According to this view, the hypothesis should not be seen as a test for an individual collection's clustering tendency. Instead, we argue that the hypothesis should be valid for every collection, and should therefore be seen as an axiom of cluster-based retrieval.

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

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

Conventional measures of interdocument relationships, such as the cosine coefficient for example, can not detect the inherent similarity between co-relevant documents, 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* (QSSM) that bias interdocument relationships towards pairs 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. Specific formulas for the calculation of QSSM are then presented in section 3, followed by a description of the experimental environment under which their effectiveness is evaluated in section 4. Experimental results are presented and discussed in section 5, and

section 6 presents some 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; Rorvig, 1999), 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[1] which 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.

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

The static notion of similarity has been implicitly (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).

[1] Our discussion on similarity measures is based on the cosine coefficient. However, our arguments can easily be extended to other similarity or dissimilarity measures.

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).

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.

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

We view the query as the context under which the similarity of two documents, that are retrieved in response to this query, is judged in IR applications. 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 and 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 measure. The query-sensitive similarity $Sim(D_i, D_j \mid Q)$ can therefore be defined as:

$$Sim(D_i, D_j \mid Q) = f(Sim(D_i, D_j), Sim(D_i, D_j, Q)) \qquad (2)$$

The similarity given by the variable component $Sim(D_i, D_j, Q)$ can be defined by finding all common terms between documents D_i and D_j, and seeing which of these common terms are also terms that appear in the query Q. The similarity between pairs of documents that have a large number of common terms that are query terms should then be accordingly augmented. This idea can be defined in terms of the cosine coefficient in Equation 7.3. In this equation $Q=\{q_1, q_2, \ldots, q_n\}$ is the query vector, D_i and D_j are the two document vectors, and $C = D_i \cap D_j = \{c_1, c_2, \ldots, c_k, \ldots, c_n\}$ is a vector which contains the common terms of documents D_i and D_j. The terms of the common vector C can be represented by $c_k = (d_{io} + d_{jp}) / 2$, where d_{io}, and d_{jp} are the weights of each of the common terms in D_i and D_j respectively. Vector C then contains the set of common terms of the two documents, and each term of C is weighted by the average of the weights of the common terms. Other representations of c_k were also investigated ($min(d_{io}, d_{jp})$, $max(d_{io}, d_{jp})$, $(d_{io} \cdot d_{jp})$), but no significant differences were found. We report this specific form which proved to be consistently the most effective.

$$Sim(D_i, D_j, Q) = \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \qquad (3)$$

Having established ways to define the two components of equation 2, what remains is to define the function that combines these two sources of evidence. One way to do so is by using a linear combination of the two sources: $Sim(D_i, D_j \mid Q) = \vartheta_1 Sim(D_i, D_j) + \vartheta_2 Sim(D_i, D_j, Q)$, where $\vartheta_1 + \vartheta_2 = 1$. By substituting equations 1 and 3 in the above, we derive equation 4 which gives the query-sensitive similarity between D_i and D_j. We will call this measure *M3*. It should be noted that a linear combination of sources of evidence is commonly used in IR applications (Wen *et al*., 2001).

Intuitively, if one bases the calculation of interdocument similarities on measure M3, then for a specific query, pairs of documents that have more terms in common with the query than other pairs will be assigned higher similarity values (assuming that they have the same number of non-query terms in common). This reflects the idea that under the context defined by the query, query terms possess greater salience when

determining interdocument relationships. The relative importance of each of the components of equation 4 can be determined by assigning appropriate values to parameters ϑ_1 and ϑ_2.

$$Sim(D_i, D_j \mid Q) = \vartheta_1 \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} + \vartheta_2 \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \qquad (4)$$

More specifically, the first parameter (ϑ_1) determines the importance assigned to the conventional, static similarity of the documents under comparison, while the other parameter determines the importance assigned to the varying component of Equation 7.4. If ϑ_2 is set equal to zero, then the similarity given by equation 4 is simply the cosine coefficient between the two documents D_i and D_j adjusted by the parameter ϑ_1. The same effect can be achieved when none of the common terms between the two documents is a query term; in this case equation 3 will give a similarity value of zero.

On the other hand, if parameter ϑ_1 is set equal to zero, then the query-sensitive similarity between the two documents becomes equivalent to the one given by equation 3. In this case, the effect of the static similarity is ignored, and the resulting formula can be seen as the most extreme form of query-biasing. We will call this measure *M2*. Measure M2 only takes into account common terms between the two documents that are also query terms. Unlike the measure defined by equation 4, M2 will equal zero if none of the common terms between the documents is a query term. Also unlike Equation 7.4, the overall similarity between D_i and D_j does not take into account the co-occurrence of other terms (apart from query terms) in the two documents. The effectiveness attained with M2 can be seen as a lower limit of the effectiveness of query-sensitive measures.

A note that should be made regarding the value of these two parameters is that their absolute value is of no practical significance. Instead, it is the ratio of one parameter over the other that is of importance. The reason for this, is that it is not the absolute value of interdocument similarities that affects the clustering process, but rather the relative ranking of these similarities (Van Rijsbergen, 1979). The constraint set earlier ($\vartheta_1+\vartheta_2=1$) reflects this. In section 5.1 we investigate the selection of appropriate values for the two parameters.

$$Sim(D_i, D_j \mid Q) = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \cdot \sum_{k=1}^{n} d_{jk}^2}} \cdot \frac{\sum_{k=1}^{n} c_k \cdot q_k}{\sqrt{\sum_{k=1}^{n} c_k^2 \cdot \sum_{k=1}^{n} q_k^2}} \qquad (5)$$

One more QSSM will be defined in this section, its definition being highly similar to the one of M3. This third measure differs in the way that it combines the two sources of evidence given by equations 1 and 3. Instead of a linear combination of the two components (equation 4), the new measure is defined as the product of the two sources of information. This is presented in equation 5; we will call this measure *M1*. The rationale behind measure M1 is exactly the same as for M3, i.e. for a specific query, pairs of documents that have more terms in common with the query than other pairs will be assigned higher similarity values.

However, there is one significant difference between the two measures. When using M1, if none of the common terms between the two documents is a query term (i.e. $Sim(D_i, D_j, Q) = 0$), then the overall similarity $Sim(D_i, D_j \mid Q)$ will equal zero. This is in contrast to when using M3, where $Sim(D_i, D_j \mid Q)$ will be equal to the conventional similarity of the two documents (adjusted by the parameter ϑ_1). The aim of query-sensitive measures is to increase, on a per-query basis, the similarity of documents that are likely to be co-relevant. Measure M1 attempts to do so in a rather "greedy" way, by setting the similarity of pairs of documents that do not possess any query terms in common to zero. This choice for M1 reflects the assumption that the presence of query terms is required for a document to be relevant.

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

Table 1. Query term statistics for the six test collections

This is verified by the behaviour of the test collections used in this experimental environment. Table 1 presents in the first row the percentage of relevant documents which contain at least one query term for each of the six collections[2] used in this work, and in the second row the average number of query terms contained in a relevant document. The figures in the first row of this table all exceed 91%, an exceptionally high value that verifies the highly topical and algorithmic nature of relevance that is employed in standard IR evaluation (Schamber *et al*., 1990). The implication of this for the query-sensitive measures presented here, and especially for M1, is that the likelihood for pairs of co-relevant documents to contain at least one query term in common is high.

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

[2] Details of the test collections are given in section 4.1. Calculations are based on stemmed forms of terms.

this decision is that if two documents contain non-overlapping sets of query terms, this may be an indication that the documents are discussing these terms under different topics.

For measures M1 and M2, $0 \leq Sim(D_i, D_j \mid Q) \leq 1$. For measure M3 this property can be retained by appropriate selection of parameters ϑ_1 and ϑ_2 (e.g. by constraining the parameters so that $\vartheta_1+\vartheta_2=1$). To preserve the reflexivity of the measures defined by M1, M2 and M3 (i.e. $Sim(D_i, D_i)=1$), the similarity of a document with itself is defined to be equal to 1. This does not follow as a result of either equations 3, 4, or 5, but can be introduced by definition. Finally, for all three measures $Sim(D_i, D_j \mid Q) = Sim(D_j, D_i \mid Q)$ (i.e. query-sensitive similarity is symmetric). These properties are in accordance with those of conventional similarity measures (Van Rijsbergen, 1979).

3.1 Limitations

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

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

Campbell (2000) suggested that there is a temporal aspect to the notion of relevance, and this temporal aspect should be incorporated in the retrieval model. In the same way, one can argue that the similarity between two objects may change over time due to new evidence presented, or due to the contextual effect of other objects (Tversky, 1977).

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

These limitations are not unique to the approach proposed in this work. The majority of IR research to date has focused on the topical aspect of relevance, taking the view that query terms offer the only evidence about the user's information need. As far as this work is concerned, the choice not to consider factors such as the

ones mentioned previously was taken on the basis that in a non-interactive laboratory-based environment it is difficult to model such factors.

A further limitation relates to the problem of short queries, the type usually encountered in web search engines, averaging about 2-3 terms per query (Jansen *et al*., 2000). The three measures defined previously, regard query terms as the dimensions that acquire significant discriminatory power. If only 2 or 3 such terms are supplied by the user, it is doubtful whether these measures (especially M2) will have enough information to effectively bias similarity. This is a well-known research problem in IR, and methods that have been used to tackle it previously (Voorhees, 1994; Xu & Croft, 1996) could also be applied here. The effect of query length on the effectiveness of these measures is investigated in section 5.4.

4 Experimental Details

The experiments reported here mainly aim to investigate the effectiveness of the proposed query-sensitive measures (M1, M2 and M3) in 'forcing' documents that are likely to be co-relevant to be more similar to each other than when using a conventional similarity measure (i.e. the cosine coefficient). 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 which 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

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

Statistics for the six document collections are presented in Table 2. It should be noted that the four smallest collections (CACM, CISI, LISA, and Medline) are homogeneous, treating one major subject area (e.g. Library and Information Science, Biomedicine, etc.), and such topical homogeneity may effect the experimental results. The AP and WSJ collections, on the other hand, cover in their documents a wide variety of topics, providing two collections with different characteristics.

For these two collections, TREC topics (i.e. queries) 1-50 were randomly chosen and used in the experiments. 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 2). The *Concepts* field usually lists terms and phrases that the creator of the query thinks are related to it (Harman, 1993).

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

Table 2. 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 (Salton & Buckley, 1988). 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 document sets that would be investigated. Seven values of *n* were used: 100, 200, 350, 500, 750, 1000, and full collection (*n* = collection size)[3]. 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 results would scale for increasing values of *n*, when more non-relevant documents are introduced in the document sets. Secondly,

[3] The value of 1000 was not used in CISI and Medline collections because their sizes are 1460 and 1033 documents respectively. The full AP and WSJ collections were also not considered.

recent research (Tombros *et al*., 2002) 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 this section we report and analyse results that are obtained regarding the effectiveness of the query-sensitive measures. The presentation of the results consists of four parts. First, in section 5.1 we examine how the effectiveness of measure M3 varies as a function of the two parameters (ϑ_1 and ϑ_2). Subsequently, in section 5.2 we investigate the comparative effectiveness of the QSSM and the cosine coefficient, in section 5.3 we study the comparative effectiveness of the three query-sensitive measures, and in section 5.4 we consider the effect of the query length on M1, M2 and M3.

5.1 Selecting parameters for M3

In this section, the selection of appropriate values for the parameters ϑ_1 and ϑ_2 of M3 (section 3, equation 4) is examined. As it was explained in that section, it is not the absolute values of these parameters that is of interest, but rather, their ratio. By varying the ratio of these parameters, one can investigate the effect of assigning different importance to the two components of Equation 7.4. More specifically, ϑ_1 determines how much importance is associated to the static similarity of the two documents, whereas ϑ_2 how much importance is associated to the common similarity of the two documents and the query. It should also be reminded that for ϑ_1=1 and ϑ_2=0 M3 becomes equivalent to the cosine coefficient (equation 1), and also that for ϑ_1=0 and ϑ_2=1 M3 becomes equivalent to M2 (equation 3).

Intuitively, one would expect the results of the 5NN test to resemble those attained by the cosine coefficient when the values of the parameter ϑ_1 are much higher than those of ϑ_2. Then, by decreasing the difference in the values of the two parameters (and hence their ratio), the results should start to differ to those obtained by the cosine. This is evident in Table 3, where the results of the 5NN test are presented for four different ratios of the two parameters (9:1, 4:1, 2:1, 1:1) when using LISA (highest value for each ratio is in bold). The results presented in Table 3 are representative of the results obtained using the other five test collections.

Results for all six collections for varying ratios of the two parameters are presented in the Appendix, Tables A1-A6. For comparison, the results for this collection using the cosine coefficient are reported in Table 6.

n	*9:1*	*4:1*	*2:1*	*1:1*
100	**0.946**	**0.99**	1.055	1.206
200	0.926	0.972	1.027	1.195
350	0.821	0.93	1.029	1.199
500	0.849	0.938	1.037	**1.237**
750	0.859	0.94	1.041	1.208
1000	0.83	0.91	**1.076**	1.204
full	0.913	0.946	1.028	1.177

Table 3. Results of the 5NN test for the LISA collection by varying the ϑ_1: ϑ_2 ratio in favour of ϑ_1

By observing the results, it becomes evident that significant improvements are introduced by reducing the importance of the component of equation 4 that corresponds to the static similarity between the two documents and the query terms (i.e. by increasing the importance of ϑ_2). If one compares, for instance, the results obtained when the static component of equation 4 is nine times more important than the variable component (i.e. ϑ_1:ϑ_2=9:1) to the results obtained when both components are weighted equally, the differences range between 27.5 and 46% in favour of the latter ratio. The differences in the majority of cases are statistically significant, especially as the relative importance assigned to ϑ_1 is reduced.

n	*1:2*	*1:4*	*1:7*	*1:9*	*1:10*	*M2*
100	1.352	1.383	1.402	1.392	1.404	**1.395**
200	1.311	1.327	1.390	1.391	1.389	1.269
350	1.335	1.418	**1.429**	**1.42**	**1.428**	1.315
500	**1.374**	**1.415**	1.423	1.403	1.406	1.317
750	1.358	1.395	1.421	1.413	1.392	1.287
1000	1.341	1.384	1.393	1.385	1.380	1.303
full	1.303	1.332	1.376	1.354	1.341	1.269

Table 4. Results of the 5NN test for the LISA collection by varying the ϑ_1: ϑ_2 ratio in favour of ϑ_2

Having established that the results obtained by the 5NN test significantly increase when the ratio of the two parameters increases in favour of ϑ_2, what remains to be established is whether there is a specific ratio for each collection that displays the highest effectiveness. In Table 4, the results of the 5NN test are presented when using the LISA collection, and when the ratio of the two parameters is varied in favour of ϑ_2. The last column of this table contains the results obtained when using ϑ_1=0 and ϑ_2=1; as we mentioned earlier this corresponds to the M2 measure (equation 3). The results of this table demonstrate that, in general, the

effectiveness of M3 tends to increase as the weight assigned to ϑ_2 increases. When M3 becomes equivalent to M2 (last column of the table), there seems to be a rather significant drop in the effectiveness of the measure. For the specific case of the LISA collection, the peak in effectiveness seems to occur between the ratios of 1:7 and 1:10. However, the differences in effectiveness at this region are not statistically significant.

It should also be noted that the behaviour when using LISA with increasing importance assigned to ϑ_2 is not typical of the two larger collections (AP and WSJ). In general, when using LISA, as the data in Table 4 demonstrate, when the importance attributed to the common similarity between the documents and the query increases it does not seem to significantly impair the effectiveness of the measure, at least not until M3 becomes equivalent to M2 (i.e. the difference in effectiveness when using ratios 1:7, 1:9, 1:10 are small). This is especially evident for small values of *n* (i.e. 100, 200, 350).

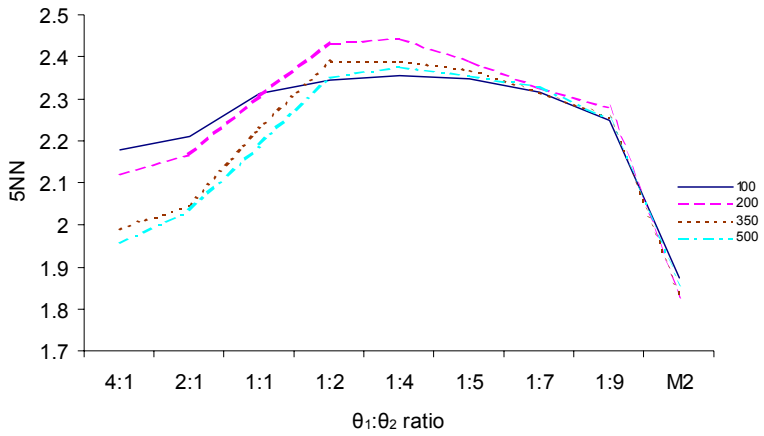

Figure 1. The effectiveness of M3 as a function of ϑ_1 and ϑ_2 for the WSJ collection

Figure 1 demonstrates the variation in the effectiveness of M3 for varying ratios of the two parameters for *n*=100, 200, 350 and 500 when using the WSJ collection. The pattern of the results for the WSJ collection is for the effectiveness of M3 to peak when the ratio between the two parameters is in the region of 1:4. The results display a consistent decrease past this point as the weight assigned to ϑ_2 increases (i.e. ratios 1:7, 1:9 yield lower results).

A reason for the rather different behaviour of the two databases can be given in terms of their characteristics. Documents of the LISA collection are rather short, with 39.7 terms on average per document. The length of the queries for this collection is large (almost 20 terms per query on average, Table 1), almost half the average document size. Moreover, as it was mentioned in section 3, relevant documents in this collection contain on average 4.5 query terms (Table 1). Taking these characteristics into account, it can be appreciated why query influence in this database is strong: the combination of short documents, long queries and relatively large

number of query terms per relevant document increases the likelihood of pairs of co-relevant documents to be assigned high similarity by M3.

The WSJ collection on the other hand, is characterised by long documents (377 terms on average per document), and shorter queries than LISA (7.6 terms on average). In this collection, as the weight assigned to the static similarity is decreased and calculations are increasingly biased towards common query terms between documents, the effectiveness of the measure seems to be obscured by the length of the documents and the relatively few query terms (especially comparatively to document length). In addition, documents of the WSJ are more topically diverse than those of the smaller collections, and therefore query terms can be used under a varying number of contexts in such documents. M3, in such an environment, is more likely to reach a higher effectiveness when the importance assigned to common query terms and common "content" terms is more balanced (but still in favour of the former) than in more topically homogeneous collections. The other TREC collection (AP) displays a similar behaviour (Appendix, Table A1).

As far as the other three collections are concerned (CACM, CISI and Medline), the effectiveness of M3 seems to peak when the ratio of the two parameters is set to around 1:7 (Appendix, Tables A2, A3, A5). This behaviour is similar to the one noted for LISA. These four collections are topically homogeneous, treating mainly a single subject area (e.g. library and information science for LISA).

As a conclusion regarding the selection of parameters for M3, the data obtained support the view that this is heavily dependent on the characteristics of the test collection under investigation. What was noted for all six collections was that the effectiveness of M3 for the 5NN test increases as the relative importance of ϑ_2 over ϑ_1 increases, and it reaches its peak when the ratio between the two parameters is considerably in favour of ϑ_2. The effectiveness of the measure then tends to drop past this point, and when ϑ_1 becomes equal to zero M3 generally displays its lowest effectiveness.

It should also be emphasised that as the ratio of the two parameters increases in favour of ϑ_2, the differences in the effectiveness of M3 are generally not statistically significant. For the two TREC collections there are significant differences as the ratios of the values move past the peak point (1:4), i.e. the differences between the ratios of 1:4 and 1:7, 1:9 are significant in favour of the former.

5.2 Comparative effectiveness of the query-sensitive measures and the cosine coefficient

In Tables 5-7 the results of the 5NN test for each of the six test collections and each of the three QSSM are presented. Each table comprises five columns[4]. In the first column the different values of *n* are given for

[4] Each table contains results for two collections, so each table contains ten columns. Data corresponding to each collection are treated as a separate table.

which results are calculated. Columns 2-5 contain the results obtained for the 5NN test with the cosine coefficient, and measures M1, M2 and M3 respectively. In columns 3-5 the percentage difference between the results for M1-cosine, M2-cosine and M3-cosine, respectively, are also calculated. The differences are displayed in brackets. For each of the four columns (2-5), the highest value for the 5NN test across all values of *n* is displayed in bold.

AP	*Cosine 5NN*	*M1 5NN*	*M2 5NN*	*M3 5NN*	*WSJ*	*Cosine 5NN*	*M1 5NN*	*M2 5NN*	*M3 5NN*
top100	**2.447**	**2.619** (7.02%)	**2.079** (-15.06%)	**2.652** (8.35%)	top100	**2.122**	2.357 (11.1%)	**1.872** (-11.74%)	2.354 (10.95%)
top200	2.184	2.406 (10.18%)	1.834 (-16.02%)	2.404 (10.07%)	top200	2.051	2.446 (19.29%)	1.827 (-10.88%)	**2.443** (19.15%)
top350	2.111	2.39 (13.22%)	1.671 (-20.84%)	2.349 (11.26%)	top350	1.909	**2.468** (29.29%)	1.832 (-4.01%)	2.389 (25.15%)
top500	2.085	2.442 (17.1%)	1.663 (-20.25%)	2.387 (14.49%)	top500	1.863	2.463 (32.19%)	1.856 (-0.39%)	2.377 (27.61%)
top750	2.11	2.457 (16.41%)	1.605 (-23.93%)	2.431 (15.18%)	top750	1.734	2.421 (39.62%)	1.838 (6.01%)	2.3 (32.63%)
top1000	2.01	2.37 (17.95%)	1.517 (-24.52%)	2.337 (16.28%)	top1000	1.711	2.416 (41.23%)	1.799 (5.17%)	2.269 (32.6%)

Table 5. AP and WSJ results

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	*Cosine 5NN*	*M1 5NN*	*M2 5NN*	*M3 5NN*	*LISA*	*Cosine 5NN*	*M1 5NN*	*M2 5NN*	*M3 5NN*
top100	**1.621**	1.924 (18.72%)	1.754 (8.24%)	1.911 (17.93%)	top100	**0.896**	1.362 (52.05%)	**1.395** (55.74%)	1.402 (56.5%)
top200	1.511	1.981 (31.17%)	**1.902** (25.89%)	2.04 (35.03%)	top200	0.845	1.376 (62.84%)	1.269 (50.13%)	1.39 (64.53%)
top350	1.415	2.028 (43.27%)	1.875 (32.45%)	**2.073** (46.45%)	top350	0.784	**1.449** (84.8%)	1.315 (67.75%)	**1.429** (82.21%)
top500	1.393	2.039 (46.37%)	1.85 (32.87%)	2.051 (47.24%)	top500	0.783	1.425 (81.92%)	1.317 (68.17%)	1.423 (81.64%)
top750	1.376	**2.045** (48.67%)	1.761 (28.04%)	2.006 (45.82%)	top750	0.776	1.41 (81.68%)	1.287 (65.81%)	1.421 (83.09%)
top1000	1.35	2.017 (49.45%)	1.731 (28.25%)	1.987 (47.26%)	top1000	0.768	1.391 (81.18%)	1.303 (69.71%)	1.393 (81.49%)
full	1.366	1.859 (36.08%)	1.655 (21.2%)	1.873 (37.11%)	full	0.859	1.381 (60.73%)	1.289 (49.97%)	1.388 (61.5%)

Table 6. CACM and LISA results

As far as measure M3 is concerned, the values presented here are the ones resulting from a single setting of the ratio of the two parameters ϑ_1 and ϑ_2 for each test collection. The ratio selected is the one that displayed the highest effectiveness for each collection across values of *n* based on the results reported in the previous section. For the four smaller collections (CACM, CISI, LISA and Medline) the ratio selected is that of 1:7,

whereas the ratio selected for the two TREC collections is 1:4. In cases where there is not a clear best ratio for all values of *n*, the ratio that displays the best average rank among all ratios is selected.

An alternative procedure for reporting results for M3 would have been to select, for each value of *n*, that ratio that gives the highest effectiveness. This strategy would have resulted in the best possible values for M3. However, it was deemed as more realistic to select values from a single ratio for all values of *n*, rather than to do so selectively from the best ratio for each value of *n*. Moreover, it was mentioned in the previous section that the differences in the effectiveness of M3 for the ratios that give the highest values are not significantly different.

The results obtained for the 5NN test across all test collections show that QSSM, in the majority of experimental conditions, are more effective than the cosine coefficient at placing co-relevant documents in the same "neighbourhood". The only exception to this is noted when using the M2 measure in the two TREC collections (AP and WSJ), where M2 is less effective than the cosine for all values of *n* when using the AP collection, and for $n \leq 500$ when using the WSJ (Table 5).

CISI	*Cosine 5NN*	*M1 5NN*	*M2 5NN*	*M3 5NN*	*MED*	*Cosine 5NN*	*M1 5NN*	*M2 5NN*	*M3 5NN*
top100	**1.53**	**1.728** (12.96%)	1.703 (11.34%)	1.761 (15.13%)	top100	**3.143**	**3.569** (13.57%)	3.361 (6.94%)	**3.576** (13.79%)
top200	1.37	1.652 (20.62%)	**1.733** (26.49%)	**1.789** (30.61%)	top200	3.022	3.54 (17.13%)	**3.367** (11.4%)	3.532 (16.86%)
top350	1.253	1.66 (32.51%)	1.555 (24.13%)	1.692 (35.09%)	top350	3.023	3.501 (15.8%)	3.31 (9.5%)	3.476 (14.98%)
top500	1.203	1.625 (35.09%)	1.436 (19.38%)	1.652 (37.36%)	top500	3.003	3.475 (15.71%)	3.305 (10.06%)	3.436 (14.2%)
top750	1.14	1.55 (35.84%)	1.357 (19.01%)	1.575 (38.12%)	top750	3.004	3.466 (15.4%)	3.285 (9.37%)	3.431 (14.23%)
full	1.119	1.433 (28.06%)	1.328 (18.69%)	1.442 (28.87%)	full	3.016	3.235 (7.26%)	3.124 (3.57%)	3.216 (6.63%)

Table 7. CISI and Medline results

Statistical tests of the results reveal significant improvements of M1 and M3 over the cosine (significance level <0.001 for the majority of cases) for all experimental conditions except for the CISI collection when n=100. Measure M2 is significantly more effective than the cosine for the CACM (except for n=100), LISA (all values of n), and Medline (except for n=100, 750, full) collections. It is also significantly more effective than the cosine when using the WSJ collection for n=750, 1000. Significance levels for M2 are not as low as the ones for M1 and M3, but they are still lower than 0.04 for all significant cases.

The gains in effectiveness introduced by using QSSM are in most cases "material", i.e. over 10%, which confirms the significance of the results (Keen, 1992). The largest differences occur when using the LISA collection, where all three query-sensitive measures are over 50% more effective than the cosine in all experimental conditions. Even M2, which relies only on common terms between documents that are query terms, introduces improvements of that magnitude. This behaviour for LISA can be explained on the basis of

its characteristics: on average, queries contain as much as half the number of terms that documents do, and also relevant documents for this collection are strongly characterised by the presence of query terms. CACM, that possesses similar properties, also displays high effectiveness gains for all three QSSM.

Regarding the two TREC collections, it is perhaps not surprising that the use of M2 does not introduce effectiveness gains. The documents of the two TREC collections are large (370 and 377 terms on average per document for AP and WSJ respectively), and the queries relatively short (7.6 terms per query). Moreover, as mentioned previously, these two collections are topically diverse, and therefore terms that appear in queries are likely to be used in documents under many different contexts, not necessarily under the ones dictated by the query. M2 does not use any further contextual information (i.e. the rest of the content overlap between documents), and hence the topical diversity of these collections may mislead the similarity calculations. In such a setting it would seem unlikely that the use of only common query terms between documents can improve the effectiveness of the cosine coefficient.

As far as the AP collection is concerned, this is verified: the use of M2 is always significantly lower than that of the cosine. However, when using the WSJ collection, for n=750 and 1000, M2 is significantly more effective than the cosine coefficient. This is result is surprising, given that for large numbers of top-ranked documents one would expect the confounding effect of non-relevant documents that contain query terms to be stronger on the effectiveness of M2. As this result is not confirmed when using the other TREC collection, it should be seen with caution since it is more likely to be attributed to particular characteristics of the WSJ documents rather than to the actual effectiveness of M2.

If we look at the results across different values of n (across the rows of Tables 5-7 for column 2), 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*., 2002).

Measures M1, M2 and M3 (across rows of Tables 5-7 for columns 3-5) seem to be less affected by the increasing numbers of non-relevant documents introduced. 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.

In section 4 we mentioned that the 5NN test does not provide any information on the number of immediate co-relevant nearest neighbours. To provide information at this level of detail, a variation of the 5NN test (the NN test) is performed. The results for this test are presented in the Appendix, Tables A7-A9. In these tables the percentage of documents whose nearest neighbour is also relevant is displayed when using the cosine

coefficient, and when using each of the three QSSM. For measure M3 the same best ratio $\vartheta_1:\vartheta_2$ for each collection is used as for the calculations in Tables 5-7.

These results reveal a similar pattern to those for the 5NN test. M1 and M3 are 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). 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. Therefore, all three QSSM (especially M1 and M3) are likely to increase the effectiveness of a clustering system that employs nearest neighbour clusters, such as those proposed by (Griffiths *et al*., 1986).

The results of both the 5NN and NN tests suggest that measures M1 and M3 are significantly more effective than the cosine at placing co-relevant documents closer to each other. In this way, the likelihood of a more effective clustering of the document space is increased. Augmenting term co-occurrence similarity with query-term co-occurrence information in a pair of documents, is shown to be an effective way of detecting the similarity of co-relevant documents.

The results obtained with measure M2, as we discussed in section 3, can be seen as a lower limit for the effectiveness of QSSM. However, despite the extreme form of query biasing that M2 employs, it manages to introduce significant improvements over the cosine in a large number of cases. This result can be seen as providing further evidence for the applicability of query-sensitive measures to IR.

5.3 Comparative effectiveness of M1, M2 and M3

The results of the 5NN test in Tables 5-7 (columns 3-5) show that measures M1 and M3 achieve higher scores than M2 for the majority of experimental conditions. The only two exceptions are noted when using CISI for n=200, and when using LISA for n=100; in both cases M2 is more effective than M1 (though not significantly more effective). Statistical testing showed that M1 and M3 are significantly more effective than M2 for all values of $n\neq$200 when using CACM, for n>200 when using CISI, and only for n=750 and full when using LISA. For the two TREC collections and Medline, all differences are significant.

The results regarding the comparatively lower effectiveness of M2 are not surprising, given that this measure uses less information than the other two measures. Especially when using the topically diverse TREC collections, the lower effectiveness of M2 compared to M1 and M3 is attributed to its reliance only on common query terms between documents. M2 ignores other common terms between documents that may define the context under which query terms are used within documents.

The other issue to be examined here is the comparative effectiveness of M1 and M3. The results in Tables 5-7 reveal that the effectiveness of these measures is comparable in most experimental conditions. When using CACM, CISI, LISA or Medline, the differences between the two measures are generally negligible, and never

statistically significant. Moreover, none of the two measures consistently outperforms the other in these collections so as to offer an indication of superior effectiveness. For example, when using CACM M3 is more effective than M1 in 4 out of 7 possible values of n; there is no pattern to relate smaller values of n with superior effectiveness of one measure over the other. The only consistent behaviour noted is when using CISI where M3 is always more effective than M1, and when using Medline where M1 is always more effective than M3.

The only indication of superior performance comes when using the two TREC collections. When using AP, M1 is more effective than M3 for all but one (n=100) values of n, and when using WSJ it is more effective than M3 for all values of n. Significant differences occur for n=750 when using AP, and for n>200 when using WSJ.

To appreciate why any significant differences in performance occur between these two measures, one has to look at the way they use information from the query to augment interdocument similarity values. Both measures use information from the content overlap and from the query-term overlap between documents. Consequently, when query terms are common between documents, both measures will augment the content similarity value between those documents by a factor that is incorporated differently for each measure (product for M1, linear combination for M3).

More important than the way similarity values are augmented, is the behaviour of the two measures when no common terms between the two documents are query terms (i.e. when equation 3 outputs zero): M1 sets the similarity of the two documents to zero, whereas M3 sets it equal to a value corresponding to the static similarity between the two documents, adjusted by the parameter ϑ_1.

Let us consider the case of a relevant document D_i that contains a few query terms. According to the static component of the similarity, this document will be similar to other documents with which it shares a large number of content terms (not necessarily including query terms). M1 and M3 will re-order this initial similarity ranking in such a way so as to promote documents that share a large number of content terms and query terms with document D_i. The re-ordering generated by M1 will remove documents with no query-term overlap with D_i from the top of the list in a rather crude way, by setting their similarities to D_i to zero. The reordering generated by M3 will promote documents with query-term overlap with D_i, but may not promote such documents sufficiently to "force" them to obtain a similarity to D_i higher than documents with no query-term overlap (but significant content term overlap) may have. This is also more likely to occur for TREC documents because of their length: it is more likely to have documents with a strong (non-query term) content overlap than it is for documents of shorter lengths, as those of the other four collections.

Based on the results presented in this section, it is valid to state that M1 and M3 are both more effective than M2 at placing co-relevant documents at close proximity to each other. This is especially evident when using document collections with short queries, since M2 relies only on the information supplied by the query terms.

5.4 Effect of query length on the query-sensitive measures

In the results for the 5NN test in Tables 5-7 (columns 3-5), M2 was more effective than the cosine for the CACM, LISA and Medline collections, where the average query length is relatively large (on average, 13 terms for CACM, 19.4 for LISA, and 10 for Medline, compared to 7.6 for AP, CISI and WSJ). This is a consequence of the strong dependence of M2 on query terms.

In order to investigate the effect of query length on the effectiveness of all three measures, an expanded and a shorter version of the 50 TREC topics for the AP and WSJ collections were used. For the expanded version, terms from the *Title*, *Description*, and *Concepts* fields of each topic were used (see section 4.1), yielding on average 23.4 terms per query (compared to 7.6 terms initially). For the shorter version of the queries only the *Title* field was used, with an average of 3.2 terms per query.

The expansion terms for the TREC topics are not generated algorithmically, and this can perhaps be seen as a point of criticism. For example, a query expansion algorithm might have selected terms that are better discriminators than the ones selected manually, by analysing distribution patterns over an entire document corpus, or locally over a set of retrieved documents (Xu & Croft, 1996). However, it is felt that the experimental procedure followed in this section is sufficient to demonstrate the behaviour of the query-sensitive measures when variations in query length occur, as any research relating to query-expansion issues is not pursued here.

n	*M1 expanded*	*M2 expanded*	*M3 expanded*	*M1 short*	*M2 short*	*M3 short*
100	**2.67** (1.95%)	**2.364** (13.75%)	**2.687** (1.34%)	**2.459** (-6.45%)	**1.616** (-22.25%)	**2.541** (-4.18%%)
200	2.39 (-0.67%)	2.128 (16.05%)	2.402 (-0.05%)	2.095 (-12.92%)	1.313 (-28.39%)	2.254 (-6.23%%)
350	2.408 (0.75%)	2.1 (25.65%)	2.384 (1.5%)	2.081 (-12.92%)	1.21 (-27.59%)	2.216 (-5.66%)
500	2.422 (-0.83%)	2.124 (27.71%)	2.401 (0.57%)	2.08 (-14.81%)	1.199 (-27.88%)	2.228 (-6.68%)
750	2.494 (1.52%)	2.191 (36.5%)	2.464 (1.38%)	2.137 (-13%)	1.192 (-25.72%)	2.237 (-7.95%)
1000	2.428 (2.41%)	2.129 (40.39%)	2.387 (2.14%)	2.05 (-13.51%)	1.127 (-25.7%)	2.167 (-7.25%)

Table 8. The effect of query length for AP: results of the 5NN test

The 5NN test was repeated for both the expanded and shorter versions of the queries, on the same sets of documents as for the original queries[5], for each value of *n*. For measure M3 the best ratio (1:4) of parameters ϑ_1 and ϑ_2 was used for both collections so as to allow these results to be compared to the results reported in Table 5. Other ratios were tried in order to examine whether query length would change the most effective

[5] This choice was made so as to be able to compare the results between the modified and the original queries.

ratio for these collections, but there were no significant deviations from the pattern of the results presented in Tables A1 and A6. The results using the modified queries for the AP and WSJ collections are presented in Tables 8 and 9 respectively, where the highest values for each column are displayed in bold. For columns 2-7 the percentage differences between the reported values and those obtained with the standard queries (Table 5, columns 3-5) are displayed in brackets.

The results in Tables 8 and 9 confirm the strong dependence of M2 on query length. M2 with the expanded queries (column 3) is significantly more effective than using the initial queries for all values of *n* (significance levels <0.001). Moreover, when using WSJ, M2 is significantly more effective than the cosine coefficient for all values of *n* (significance levels <0.03), and is not significantly worse than M1 or M3 (either with expanded or initial queries). It is also more effective that M3 for *n*=750 and 1000, but not significantly so. When using the AP collection, M2 exceeds the cosine for some values of *n* (500, 750 and 1000) but not significantly, and it is also not significantly worse than the cosine for the other values of *n*. In contrast to when using WSJ, M2 with the expanded queries is still significantly worse than both M1 and M3 for all values of *n*.

n	*M1 expanded*	*M2 expanded*	*M3 expanded*	*M1 short*	*M2 short*	*M3 short*
top 100	2.457 (4.22%)	2.372 (26.67%)	2.414 (2.54%)	**2.32** (-1.59%)	**1.672** (-12%)	**2.295** (-2.52%)
top 200	2.535 (3.63%)	2.37 (29.69%)	**2.474** (1.25%)	2.271 (-7.7%)	1.631 (-12.04%)	2.236 (-8.49%)
top 350	**2.54** (2.91%)	2.415 (31.82%)	2.448 (2.46%)	2.241 (-10.14%)	1.536 (-19.31%)	2.159 (-9.52%)
top 500	2.54 (3.14%)	**2.425** (30.71%)	2.44 (2.65%)	2.195 (-12.22%)	1.525 (-21.67%)	2.173 (-8.59%)
top 750	2.441 (0.83%)	2.407 (30.93%)	2.344 (1.9%)	2.101 (-15.24%)	1.434 (-28.17%)	2.05 (-10.88%)
top 1000	2.437 (0.85%)	2.399 (33.35%)	2.325 (2.47%)	2.064 (-17.09%)	1.435 (-25.36%)	2.022 (-10.88%)

Table 9. The effect of query length for WSJ: results of the 5NN test

The behaviour of M2 for expanded queries can be explained on the basis of the role that the added query terms play for this measure. Because M2 relies only on common query terms between documents, it lacks the contextual information provided by other common terms between documents. The addition of terms to the query provides more information to M2 to effectively assess the likelihood of two documents to be jointly relevant to the same query.

Column 6 of Tables 8 and 9 shows a significant decrease in effectiveness for M2 when average query length is decreased to 3.2 terms. The decrease in effectiveness is sizeable if one considers that the difference in query length between the initial and the short queries is on average just 4.4. terms.

Measures M1 and M3, on the other hand, are less affected by the increase in query length from 7.6 terms per query (initial queries) to 23.4 (expanded). None of the differences in effectiveness reported in Tables 8 and 9

(columns 2 and 4) between the expanded and the initial form of the queries are significant. In some cases when using AP, there is even a minor decrease in the effectiveness of the measure when expanded queries are used (n=200 and 500 for M1 and n=200 for M3).

When short queries are used (columns 5 and 7 of Tables 8 and 9), both measures (M1 and M3) display a significant decrease in effectiveness. The decrease is smaller in scale than that reported for M2, but significant (significance levels <0.03) for both collections and all values of n, except for n=100. Despite this decrease, M1 and M3 using the short queries are still significantly more effective than the cosine when using the WSJ collection (Table 5, column 2, significance levels <0.003). When using AP, M1 is more effective than the cosine for n=100, 750 and 1000, and more effective than M3 for all values of n. However, no significant differences between these measures and the cosine are noted.

The results presented here suggest that M2 is highly affected by query length, and it would therefore not seem suitable to be applied to environments where very short queries are usually input by users, unless effective ways to expand the query could be used. Assessing the likelihood of two documents to be jointly relevant to a query based on the amount of information provided by approximately 3 terms on average is not likely to be effective.

The other two measures do not seem as much affected by variations in query length. This is due to that they combine contextual information (the whole set of terms between documents) with increased weight assigned to query information. In this way, M1 and M3 are more likely to cope well when query length is decreased: the contextual information may be a good indicator of whether the few query terms are used in the same topic between documents. It is for this same reason that the effectiveness of the two measures does not significantly benefit from the addition of terms to the query.

This behaviour of measures M1 and M3 might appear useful in an operational environment, like a web search engine for example, where user queries comprise only few terms (Jansen *et al*., 2000). In the specific experimental environment used, M1 and M3 outperformed the cosine coefficient in a large number of cases when short queries were used. It remains to be seen whether such improvements would occur in operational environments.

It should also be mentioned that the results reported in this section regarding the effect of query length, may have been affected by the way that the expanded forms of the queries were obtained. If the expanded terms were chosen in a different way, then a different picture regarding the effectiveness of the measures for varying query lengths might have been obtained. If, for example, expansion terms were obtained algorithmically, then the effectiveness of M2 compared to M1 and M3 may improve. Query terms added algorithmically may be better at discriminating between relevant and non-relevant documents than the ones used here.

The way that the expanded forms of queries were obtained may also be a contributing factor for the different effect of query length when using the two TREC collections. Tables 8 and 9 show that the results of the 5NN test in each of the two collections are differently affected by variations in query length. The discriminating power of the query terms in each query form examined is likely to be different for each collection, and therefore likely to have a different effect on the effectiveness of the query-sensitive measures.

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 have been developed in the past under the name of *user-oriented*, or *adaptive clustering* (e.g. Yu *et al*., 1985; Deogun & Raghavan, 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 (QSSM) for the calculation of interdocument relationships. Query-sensitive measures bias similarity towards pairs of documents that jointly possess terms that are expressed in a query. This is based on the view that similarity is a dynamic and purpose-sensitive notion, and that QSSM have the potential to capture the dynamics of similarity for the calculation of interdocument relationships.

We presented three such measures. Two of them take into account all common terms between a pair of documents, but bias the similarity measure towards those common terms that are also query terms (measures M1 and M3). Each of these two measures uses a different function to combine static and variable similarity (M1 uses a product of the two sources, where M3 uses a linear combination). The third measure only takes into account common terms between documents that are query terms (measure M2).

Through a series of experiments that assess the degree at which similarity measures place relevant documents at close proximity to each other, we demonstrated that QSSM are significantly more effective than the cosine coefficient. More specifically, measures M1 and M3 are always significantly more effective than the cosine, and are not strongly 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.

The main conclusion from this research is that the use of query-sensitive measures for the calculation of interdocument relationships is highly effective. Regarding the motivation behind the introduction of QSSM to IR, the per-query adherence to the cluster hypothesis, the results presented in this paper demonstrate that, compared to static measures, query-sensitive measures achieve a significantly higher adherence to the hypothesis. A perfect per-query adherence is not achieved, and it would seem unlikely that considering only topical aspects of relevance would achieve this.

A more thorough evaluation of QSSM can be performed if one integrates them in a wider application area. An obvious area where QSSM 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.

In section 3.1 we mentioned some 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.

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.

8 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.

Campbell, I. (2000). The ostensive model of developing information needs. Ph.D. Thesis, Department of Computing Science, University of Glasgow.

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

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. Ph.D. 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.

Harman, D.K. (ed.) (1993). *Proceedings of the First Text Retrieval Conference*. National Institute of Standards and Technology, Gaithersburg, MD.

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.

Keen, E.M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4):491-502.

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.

Rorvig, M. (1999). Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8):639-651.

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

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513-523.

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. (2002). The effectiveness of query-specific hierarchical clustering in information retrieval. *Information Processing & Management*, 38(4):559-582.

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. Ph.D. 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.

Wen, J.R., Nie, J.Y., Zhang, H.J. (2001). Clustering user queries of a search engine. In *Proceedings of the 10th WWW Conference*, pp. 162-168. Hong Kong.

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.

Xu, J. and Croft, W.B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 4-11. Zurich, Switzerland.

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.

Appendix

n	4:1	1:1	1:2	1:4	1:5	1:7	1:9	M2
100	**2.440**	**2.561**	**2.633**	**2.652**	**2.640**	**2.574**	**2.566**	**2.079**
200	2.218	2.321	2.354	2.404	2.377	2.354	2.316	1.834
350	2.155	2.244	2.339	2.359	2.351	2.313	2.298	1.671
500	2.143	2.263	2.353	2.387	2.398	2.393	2.319	1.663
750	2.167	2.293	2.372	2.431	2.412	2.390	2.333	1.605
1000	2.070	2.209	2.290	2.337	2.334	2.270	2.232	1.517

Table A1. Effectiveness of M3 as a function of the ratio ϑ_1:ϑ_2 for AP

n	4:1	1:1	1:2	1:4	1:5	1:7	1:9	M2
100	**1.706**	**1.840**	1.890	1.923	1.934	1.911	1.899	1.754
200	1.578	1.739	1.922	1.995	2.036	2.040	2.045	**1.902**
350	1.531	1.766	1.987	**2.058**	**2.070**	**2.073**	**2.074**	1.875
500	1.540	1.757	**2.007**	2.049	2.037	2.051	2.040	1.850
750	1.520	1.768	1.987	2.022	2.019	2.006	2.001	1.761
1000	1.506	1.772	1.971	1.998	2.003	1.987	1.966	1.731
full	1.443	1.534	1.687	1.850	1.868	1.873	1.78	1.655

Table A2. Effectiveness of M3 as a function of the ratio ϑ_1:ϑ_2 for CACM

n	4:1	1:1	1:2	1:4	1:5	1:7	1:9	M2
100	**1.578**	**1.698**	**1.722**	**1.746**	**1.744**	1.761	**1.757**	1.703
200	1.456	1.574	1.631	1.724	**1.744**	**1.789**	1.719	**1.733**
350	1.334	1.494	1.626	1.674	1.703	1.692	1.697	1.555
500	1.284	1.476	1.593	1.651	1.669	1.652	1.636	1.436
750	1.227	1.442	1.538	1.599	1.593	1.575	1.555	1.357
full	1.224	1.315	1.321	1.330	1.338	1.442	1.391	1.328

Table A3. Effectiveness of M3 as a function of the ratio ϑ_1:ϑ_2 for CISI

n	4:1	1:1	1:2	1:4	1:5	1:7	1:9	M2
100	**0.990**	1.206	1.352	1.383	1.384	1.402	1.392	**1.395**
200	0.972	1.195	1.311	1.327	1.372	1.390	1.391	1.269
350	0.930	1.199	1.335	**1.418**	1.430	**1.429**	**1.420**	1.315
500	0.938	**1.237**	**1.374**	1.415	**1.446**	1.423	1.403	1.317
750	0.940	1.208	1.358	1.395	1.405	1.421	1.413	1.287
1000	0.910	1.204	1.341	1.384	1.388	1.393	1.385	1.303
full	0.946	1.177	1.303	1.332	1.346	1.388	1.341	1.289

Table A4. Effectiveness of M3 as a function of the ratio ϑ_1:ϑ_2 for LISA

n	4:1	1:1	1:2	1:4	1:5	1:7	1:9	M2
100	**3.255**	**3.463**	**3.550**	**3.566**	**3.564**	**3.576**	**3.537**	3.361
200	3.198	3.405	3.507	3.525	3.525	3.532	3.528	**3.367**
350	3.190	3.352	3.470	3.482	3.478	3.476	3.461	3.310
500	3.187	3.340	3.456	3.442	3.450	3.436	3.424	3.305
750	3.190	3.346	3.452	3.429	3.440	3.431	3.421	3.285
full	3.111	3.116	3.201	3.204	3.210	3.216	3.213	3.124

Table A5. Effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$ for Medline

n	4:1	1:1	1:2	1:4	1:5	1:7	1:9	M2
100	**2.177**	**2.314**	2.344	2.354	2.348	2.317	2.249	**1.872**
200	2.123	2.306	**2.431**	**2.443**	**2.390**	2.328	**2.280**	1.827
350	1.989	2.226	2.391	2.389	2.370	2.318	2.255	1.832
500	1.958	2.190	2.351	2.377	2.355	**2.329**	2.252	1.856
750	1.840	2.087	2.262	2.300	2.299	2.257	2.204	1.838
1000	1.790	2.046	2.218	2.269	2.267	2.222	2.162	1.799

Table A6. Effectiveness of M3 as a function of the ratio $\vartheta_1:\vartheta_2$ for WSJ

AP					CISI				
n	Cosine	M1	M2	M3	n	Cosine	M1	M2	M3
100	**68.98%**	**71.53%**	**50.18%**	**71.17%**	100	**45.44%**	**52.11%**	55.79%	**55.79%**
200	66.29%	71.72%	45.33%	70.33%	200	39.98%	49.25%	**56.20%**	55.39%
350	64.16%	69.40%	44.14%	68.16%	350	35.75%	47.88%	54.34%	52.92%
500	64.06%	70.07%	43.78%	68.43%	500	33.87%	46.53%	50.85%	51.08%
750	62.39%	67.86%	42.88%	65.47%	750	32.82%	45.10%	44.77%	48.21%
1000	62.12%	66.97%	42.15%	64.25%	full	32.85%	41.30%	37.05%	42.79%

Table A7. Results for the 1NN test using AP and CISI

CACM					LISA				
n	Cosine	M1	M2	M3	n	Cosine	M1	M2	M3
100	**51.94%**	58.78%	55.82%	60.26%	100	**30.30%**	46.32%	**49.35%**	47.62%
200	45.92%	58.74%	59.90%	63.56%	200	27.68%	43.60%	44.64%	48.79%
350	45.97%	59.36%	**60.58%**	**65.60%**	350	26.27%	45.89%	45.89%	**49.05%**
500	46.35%	58.48%	59.65%	65.20%	500	27.43%	**47.20%**	45.13%	47.20%
750	44.95%	**60.17%**	57.75%	64.44%	750	27.20%	44.76%	46.18%	48.44%
1000	43.30%	59.36%	55.59%	62.15%	1000	28.21%	43.85%	46.65%	47.77%
full	43.76%	54.48%	50.95%	56.22%	full	28.27%	44.53%	43.47%	46.19%

Table A8. Results for the 1NN test using CACM and LISA

Medline					WSJ				
n	**Cosine**	**M1**	**M2**	**M3**	**n**	**Cosine**	**M1**	**M2**	**M3**
100	**71.88%**	**80.49%**	79.44%	**80.84%**	100	**64.41%**	**67.42%**	**56.02%**	**65.16%**
200	67.43%	76.76%	**80.20%**	79.87%	200	57.24%	62.10%	49.70%	61.67%
350	68.78%	76.39%	78.92%	78.76%	350	54.05%	63.73%	50.00%	60.72%
500	68.35%	76.06%	78.58%	78.58%	500	52.65%	62.90%	48.64%	58.83%
750	68.23%	76.06%	78.09%	78.56%	750	49.19%	61.82%	48.18%	57.32%
full	68.39%	72.41%	69.83%	73.62%	1000	47.60%	60.43%	47.73%	55.76%

Table A9. Results for the 1NN test using Medline and WSJ