Similarity Measures for Short Queries

Ross Wilkinson Justin Zobel Ron Sacks-Davis
Department of Computer Science, RMIT
GPO Box 2476V, Melbourne 3001, Australia
{ross,jz,rsd}@cs.rmit.edu.au

October 1995

Abstract

Ad-hoc queries are usually short, of perhaps two to ten terms. However, in previous rounds of TREC we have concentrated on obtaining optimal performance for the long TREC topics. In this paper we investigate the behaviour of similarity measures on short queries, and show experimentally that two successful measures—which give similar, good performance on long TREC topics—do not work well for short queries. We explore methods for achieving greater effectiveness for short queries, and conclude that a successful approach is to combine these similarity measures with other evidence. We also briefly describe our experiments with the Spanish data.

1 Introduction

Users of interactive text retrieval systems often pose short queries, typically of a few words only. Our experience with users of real systems who pose such queries is that they are unsatisfied with the behaviour of standard similarity measures such as cosine, reporting that the answers often have little apparent correspondence to the query.

These problems are despite consistent positive results for such similarity measures on test databases, including the first three rounds of TREC experiments. One feature of the TREC experiments has, however, been the length of the topics, which are intended to be careful descriptions of information needs rather than ad-hoc queries. Problems that are specific to short queries did not emerge; in particular, with short queries there is a tendency for documents containing several occurrences of the rarest query term to be ranked first, regardless of whether the other query terms occur. It is not at all clear, therefore, that a similarity measure that is successful for TREC topics 1 to 200 will work well for short queries.

As part of other work, Zobel and Moffat have been using automatic (or "hill-climbing") techniques to identify effective similarity measures [6]. These experiments identified—for the full TREC topics—effective formulations of cosine, in broad terms confirming current wisdom about which formulations work well, but also confirming that the "Okapi" measure (or 2-Poisson method) [3, 4] is slightly more effective.

In this paper we explore similarity measures designed to address the problems of short queries. We show that when the cosine and Okapi measures are applied to short forms of the TREC topics, very different behaviour emerged, with the cosine measure in particular performing far worse. We propose that these measures be combined with other evidence, and that in particular it is advantageous to highly rank documents that contain all the query terms, even if they are not particularly highly ranked by the similarity measure. These conjectures are confirmed by our results in the current round of TREC experiments.

Hill-climbing is briefly summarised in Section 2. Approaches to similarity for short queries are discussed in Sections 3 and 4, which also gives the results of TREC experiments for this year. Initial experiments with the Spanish data are reported in Section 5. We conclude in Section 6.

2 Hill climbing

There has been continued refinement of similarity measures such as cosine, with several changes over recent years that result in better performance: use of logs in within-document frequency, modification to term weighting in queries, and so on. These changes not only impact on effectiveness, but have implications for efficient implementation.

Zobel and Moffat have, as part of separate work, been using automatic methods to search for new similarity measures [6], with the primary aim of identifying effective techniques that can be evaluated efficiently. Given the number of different proposed formulations for similarity measures, and of formulations of their components such as term weight, it is quite possible that some more effective formulation is as yet undiscovered; and equally possible that there is some formulation that is not only effective, but easier to evaluate. Standard cosine requires, for example, a table of document weights based on collection-wide information, and a cosine that did not require this table could be evaluated more efficiently.

The technique used by Zobel and Moffat was to break similarity measures into several components: combining function, relative term frequency, global term weight, in-document and in-query frequency, and document and query length. Several formulations were made available for each component, giving 311,040 combinations, of which about 70,000 are mathematically distinct. These formulations were incorporated into the MG text retrieval system [1]. Using WSJ and the relevance judgements for topics 51 to 200 (and using AP in separate experiments to confirm the results), effective formulations were found by the hill-climbing method of search for the best variant of each combination of components. Interestingly—and vindicating the value of this approach—the most successful formulations were not precisely those expected. In particular, the standard formulation of document length was no more effective than using approximated lengths such as the square root of the number of distinct terms in the document.

The hill-climbing approach identified many variants of the cosine measure that were about equally effective. We chose one of these for use in our experiments, as shown in the next section. Slightly more effective was a variant on the Okapi measure, also shown in the next section.

3 Short queries versus long queries

Users tend to issue short queries when examining text databases online. Queries may be extended with relevance feedback, or queries by example may be developed during a query session. Yet the first query can be crucial, perhaps the most important in a retrieval session. However in most test collections, and notably in TREC, the supplied queries have been long, and experimental systems have been optimised with regard to these queries.

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
CDM	0.464	0.218	0.143	0.096	0.069	0.047	0.029	0.008	0.008	0.002	0.000	0.078
WCM	0.625	0.404	0.309	0.241	0.193	0.145	0.112	0.046	0.046	0.021	0.004	0.177
COS	0.803	0.607	0.483	0.395	0.327	0.262	0.210	0.092	0.092	0.040	0.006	0.286
OKA	0.770	0.537	0.417	0.331	0.271	0.218	0.166	0.074	0.074	0.033	0.006	0.244

Table 1: Precision of similarity measures for full TREC topics (WSJ)

In feedback from users of the Structured Information Manager SIM, a commercial SGML-conformant document management system developed at CITRI [5], we have found that people find it more acceptable to see documents that have many matching terms, compared to one term matching many times. This is true even if the document is irrelevant. [1] That is, they would find some kind of *coordination match* as an important component of any similarity measure.

We considered four measures of similarity in these investigations:

Coordination match:

$$\mathrm{CDM}(q,d) = |\{t : t \in q \wedge t \in d\}|$$

Weighted coordination match:

$$\mathrm{WCM}(q,d) = \sum_{t \in q \wedge t \in d} \log(N/f_t)$$

Cosine measure:

$$\mathrm{COS}(q,d) = \frac{\sum_{t \in q \wedge t \in d}(w_{q,t} \cdot w_{d,t})}{\sqrt{\left(\sum_{t \in q}(w_{q,t})^2 \cdot \sum_{t \in d}(w_{d,t})^2\right)}}$$

A variant of the Okapi measure:

$$\mathrm{OKA}(q,d) = \sum_{t \in q \wedge t \in d} \log\left(\frac{N - f_t}{f_t}\right)\left(\frac{f_{d,t}}{f_{d,t} + \sqrt{f_d}/av(\sqrt{f_d})}\right)$$

where $w_{q,t} = log(N/f_t + 1)$ and $w_{d,t} = log(f_{d,t} + 1)$, $f_{x,t}$ is the frequency of t in x, f_d is the number of terms in document d, N is the number of documents in the collection, and f_t is the number of documents containing t.

We used topics 51 to 200 against the Wall Street Journal sub-collection of the TREC collection to guide our investigations. We first give the results for each of these measures using standard stopping and stemming of terms. No special processing of the data to find index terms was used in any of these experiments. The top 1000 ranked documents were used for analysis. Table 1 gives effectiveness for each of the above similarity measures on these topics. We can see the clear advantage of the two more sophisticated measures over the co-ordination match measures. This is to be expected of long queries.

[1]Thus there is a human-factors question about whether retrieval effectiveness is the appropriate property to optimise.

Query set	Terms
TREC topics	77
Titles	4
Hand	6

Table 2: Query lengths

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
Titles												
CDM	0.400	0.222	0.182	0.139	0.120	0.097	0.079	0.034	0.034	0.023	0.009	0.102
WCM	0.420	0.259	0.217	0.182	0.159	0.131	0.111	0.074	0.074	0.054	0.016	0.133
COS	0.531	0.379	0.300	0.255	0.211	0.164	0.119	0.050	0.050	0.026	0.004	0.175
OKA	0.706	0.514	0.414	0.331	0.261	0.192	0.148	0.055	0.055	0.031	0.009	0.230
Hand												
CDM	0.502	0.330	0.262	0.214	0.180	0.144	0.099	0.039	0.039	0.022	0.009	0.147
WCM	0.525	0.371	0.322	0.274	0.233	0.182	0.153	0.090	0.090	0.059	0.020	0.191
COS	0.652	0.471	0.395	0.333	0.270	0.215	0.164	0.067	0.067	0.032	0.006	0.226
OKA	0.776	0.598	0.512	0.434	0.371	0.310	0.252	0.138	0.138	0.076	0.018	0.316

Table 3: Precision of similarity measures for short queries 51-200 (WSJ)

To investigate behaviour on short queries, a source of short queries is required. One possibility is to use the titles of the TREC topics. Since it was not clear that titles field were intended to stand alone as queries, we also created with our own short query statement for each topic, which was typically a list of the most important keywords. The average number of terms in each form of the query sets is given in Table 2. The results of using these short query expressions are given in Table 3.

We see in these experiments that coordination match approaches do not work well in comparison to more sophisticated methods. WCM did however identify a larger pool of relevant documents. It is also interesting to note how much better the Okapi measure did than the cosine measure for short queries. It seems clear that there is potential for taking some account of coordination match. We look below at a number of ways of combining evidence.

We found another telling reason for looking at some form of combination: each of the methods are producing significantly different collections of high-ranking documents. When the topics 201–250 where evaluated against the TREC collection and the top 1,000 documents retrieved for each of CDM, COS, OKA, we examined how many documents were ranked highly by pairs of methods. If the measures were identical, we would expect the overlap of the sets of answers of the different methods to be 50,000 documents. If they were completely incompatible, we would get an overlap of 0 documents.

We found that the overlap between CDM and COS was 9,106, that between CDM and OKA was 21,101, and remarkably, that between COS and OKA was only 10,561. Intuitively we would thus expect that combining evidence would help by identifying those documents found by all methods.

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
COS	0.652	0.471	0.395	0.333	0.270	0.215	0.164	0.067	0.067	0.032	0.006	0.226
(CDM>COS)	0.724	0.512	0.423	0.345	0.283	0.213	0.148	0.066	0.066	0.032	0.014	0.235
(COS>CDM)	0.726	0.525	0.443	0.367	0.307	0.241	0.191	0.088	0.088	0.045	0.013	0.257
(WCM>COS)	0.701	0.487	0.409	0.337	0.281	0.215	0.176	0.101	0.101	0.063	0.021	0.242
(COS>WCM)	0.711	0.510	0.439	0.367	0.309	0.250	0.211	0.103	0.103	0.049	0.010	0.260
(CDM+COS)	0.727	0.526	0.444	0.367	0.307	0.241	0.191	0.088	0.088	0.045	0.013	0.258
(WCM+COS)	0.733	0.526	0.454	0.383	0.318	0.259	0.216	0.103	0.103	0.050	0.010	0.270
r(WCM+COS)	0.727	0.533	0.463	0.404	0.337	0.279	0.229	0.121	0.121	0.070	0.017	0.285

Table 4: Precision of similarity measures for combined evidence

4 Combined approaches

How might we combine the evidence of the different similarity measures? If we accept the user's view that the document with the most matches is best, we simply order by coordination match, and then use the cosine measure to order documents with the same number of terms in common. In subsequent results in this section we give results for topics 51–200 with the hand–generated short queries against the Wall Street Journal Collection. We see in Table 4 that this approach (CDM>COS) gives a small improvement—of 4% over using the cosine measure alone.

The (CDM>COS) approach was applied only to the top 1,000 documents ranked by CDM. It is possible to use the corresponding approach of reordering the top 1,000 documents ranked by COS according to coordination match. This experiment, (COS>CDM), gave a further improvement of 9.7% simply due to the better pool of documents that were selected. Note that, should all documents in the database be ranked, then the results would be identical. However this method of combination provides a 13.7% improvement compared to COS for the set of short queries.

The weighted coordination match WCM gave better results than the unweighted coordination match, so we could combine in the same way and look for further improvements. It was possible, however, that improvements would not be available as this weighting is already provided by the cosine measure. Nevertheless (WCM>COS), the WCM equivalent of (CDM>COS), and (COS>WCM), the WCM equivalent of (COS>CDM), gave small improvements as shown in Table 4 .

Instead of using this two-tiered approach to ranking documents, we could regard the coordination match and the cosine measure simply as two different pieces of evidence that documents are relevant. In this case, we simply add the evidence in some way. Since the cosine measure takes values between 0 and 1, and the coordination match gives integer values, the measures had to be normalised, or the result of addition would simply be (CDM>COS) or (WCM>COS). One possibility is to divide the coordination match figures by the maximum possible value–that is, the number of terms in the query. Another possibility is to normalise each piece of evidence by dividing by the score of the highest-ranked document. In both cases, there remains the question of what relative weights to apply to the two pools of evidence. Linear regression is the obvious way of providing relative weights. (CDM+COS) gives the result of normalising the coordination match, and then adding to the cosine score, and (WCM+COS) gives the result with the weighted coordination match. r(WCM+COS) gives the result of normalising both the

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
CDM	0.270	0.094	0.058	0.033	0.020	0.011	0.007	0.001	0.000	0.000	0.000	0.028
WCM	0.381	0.224	0.176	0.155	0.121	0.099	0.064	0.046	0.031	0.016	0.005	0.100
COS	0.612	0.362	0.268	0.220	0.175	0.119	0.073	0.033	0.011	0.003	0.001	0.147
OKA	0.595	0.384	0.320	0.261	0.213	0.171	0.119	0.070	0.031	0.010	0.003	0.178
(WCM+COS)	0.668	0.422	0.336	0.278	0.234	0.160	0.100	0.065	0.038	0.015	0.003	0.187
(WCM+OKA)	0.570	0.328	0.281	0.236	0.185	0.151	0.106	0.059	0.040	0.016	0.005	0.156
(OKA+COS)	0.648	0.447	0.356	0.287	0.238	0.179	0.116	0.066	0.029	0.009	0.002	0.196

Table 5: Precision of similarity measures for on TREC queries 202–250

cosine and the weighted coordination match, and then regressing against the relevance results. While this last experiment is not in practice a valid way of combining evidence, since it can only be applied a postiori, it does give an indication of the maximal results achievable using this approach.

The next stage of experimentation involved using the Okapi measure instead of the cosine measure in combination with coordination measures. All of these experiments failed to give any improvement over using the Okapi measure on its own. This result was quite surprising. However, the cosine measure and Okapi have very different behaviour on small queries that helps explain the behaviour.

Suppose that we have two documents, A and B. Document A has a rare term that matches the query 4 times. Document B has two more common terms that match once each. Assuming that A and B are of average size, $N = 10^7$, $f_{t1} = 10$, and $f_{t2} = 100$, we find that

$$OKA(Q, A) \simeq log(10^6) \times 0.8 \simeq 11$$

$$OKA(Q, B) \simeq 2 \times log(10^5) \times 0.5 \simeq 11.5$$

By contrast,

$$COS(Q, A) \simeq log(10^6) \times log(5)/D \simeq 22.2/D$$

$$COS(Q, B) \simeq 2 \times log(10^5) \times log(2)/D \simeq 12.8/D$$

Thus the Okapi measure is much more like the coordination match than the cosine is. It thus became appropriate to regard the Okapi measure as a form of co-ordination match.

Topics 201–250 are much shorter than the previous TREC topics, at about 9 terms each. This size is still larger than we believe that people usually ask, and we did not expect that pure coordination match approaches were likely to be helpful. Thus we decided to use the Okapi measure, which had proved robust on smaller queries as our baseline. This was our CITRI1 run. We combined the Okapi measure and the cosine measure for our second run, to allow us to see whether the fact that these two measures produce very different result sets could be combined with good effect. The result of adding normalised Okapi scores to cosine scores are given in (OKA+COS) in Table 5, and constitute our CITRI2 run. We see that there is a significant gain in combining these two measures.

However, we are really interested in what happens with very short queries. Thus the topics 202–250 were again hand edited into queries with just 3–4 terms. The experiments were run against the Wall Street Journal database. The results are given in Table 6. It is remarkable how badly the cosine measure does on this data.

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
CDM	0.374	0.242	0.178	0.132	0.102	0.087	0.074	0.020	0.020	0.008	0.006	0.096
WCM	0.386	0.244	0.182	0.143	0.132	0.101	0.094	0.032	0.032	0.016	0.011	0.107
COS	0.226	0.177	0.127	0.089	0.068	0.055	0.044	0.013	0.013	0.008	0.004	0.066
OKA	0.497	0.337	0.242	0.196	0.146	0.116	0.094	0.033	0.033	0.019	0.013	0.139
(WCM+COS)	0.342	0.272	0.210	0.158	0.134	0.106	0.090	0.031	0.031	0.015	0.009	0.114
(WCM+OKA)	0.503	0.332	0.238	0.194	0.132	0.100	0.093	0.041	0.041	0.018	0.013	0.135
(OKA+COS)	0.361	0.273	0.223	0.158	0.124	0.103	0.084	0.025	0.025	0.016	0.012	0.116

Table 6: Precision of similarity measures for short queries 202-250 (WSJ)

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
S	0.851	0.657	0.542	0.435	0.374	0.314	0.246	0.095	0.095	0.022	0.000	0.315
SS	0.864	0.630	0.547	0.437	0.366	0.303	0.249	0.085	0.085	0.038	0.000	0.315
SSP	0.818	0.625	0.527	0.443	0.379	0.319	0.258	0.099	0.099	0.030	0.000	0.318
(S+SSP)	0.866	0.677	0.571	0.461	0.400	0.340	0.272	0.117	0.117	0.027	0.000	0.334

Table 7: Spanish Experiments (Queries 1–25)

5 Spanish experiments

In this, our first attempt at the Spanish track, we were principally interested in how well the indexing approaches that are used in English work in Spanish. Thus we developed a simple stop-list and a stemmer. However we were also interested exploring the notion of Church [2], that standard stemming approaches are too strong as normalisation processes. We felt that by looking at results where the data is both stemmed and not stemmed we might be able to gain, continuing our theme of combining evidence.

A stop-list of 316 words was created, and we used a stemmer that recursively removed 30 word endings. The suffixes were principally the regular verb suffixes. We then implemented an adjacent pair finder, with the constraint that the combined length of the stems was less than 10 characters. (This was to reduce the number of index terms to something less than a million terms.)

We were interested to see how we might combine the results from experiments where we stopped (S), stopped and stemmed (SS), and stopped, stemmed and paired (SSP). The results on queries 1–25 are given in Table 7. We looked at several ways of combining: S with SS, SS with SSP, S with SSP, and finally S with both SS and SSP. The best of these results was S with SSP, or (S+SSP). Thus our submitted runs are a run with just stopped data (CITRI-SP1), and then a run with the stopped results combined with stopped, stemmed and paired (CITRI-SP2). The results on queries 26–50 are given in Table 8.

Given the success of the Okapi measure, we give the same results using this measure in Table 9. We note that these experiments seem to indicate that a similar methodology to English works quite well for Spanish text.

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
S	0.595	0.337	0.263	0.206	0.165	0.121	0.081	0.014	0.014	0.000	0.000	0.143
SS	0.674	0.383	0.282	0.211	0.169	0.132	0.083	0.017	0.017	0.000	0.000	0.157
SSP	0.649	0.407	0.286	0.205	0.165	0.121	0.080	0.014	0.014	0.001	0.000	0.155
(S+SSP)	0.642	0.377	0.295	0.222	0.178	0.132	0.089	0.019	0.019	0.001	0.000	0.158

Table 8: Spanish Experiments (Queries 25–50)

Test	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Av.
S	0.725	0.398	0.302	0.238	0.201	0.161	0.116	0.041	0.041	0.016	0.000	0.178
SS	0.834	0.456	0.331	0.274	0.212	0.178	0.141	0.057	0.057	0.020	0.000	0.206
SSP	0.740	0.435	0.319	0.263	0.205	0.167	0.135	0.053	0.053	0.016	0.001	0.195
(S+SSP)	0.698	0.455	0.323	0.266	0.225	0.188	0.137	0.045	0.045	0.017	0.001	0.202

Table 9: Spanish Experiments (Queries 25–50) with Okapi Measure

6 Conclusions

We have investigated similarity measures for short queries. The preliminary results of our investigation are that some form of coordination match is important to information retrieval with such queries, but that coordination match should not be the primary factor determining the order of answers. Our results also confirm those of the City University group last year, that the Okapi method gives an excellent similarity measure.

It is important to note that we have seen very different performance from different measures as the average query length is varied. The coordination measures appear to introduce new useful documents. Different measures are finding very different result sets, so there is reason to believe that good improvement is possible.

Acknowledgements

This work was supported by the Australian Research Council. Descriptions of the systems used for this work are found at `http://www.kbs.citri.edu.au/sim/sim.html` and `http://www.kbs.citri.edu.au/mg`.

References

[1] T.C. Bell, A. Moffat, I.H. Witten, and J. Zobel. The MG retrieval system: Compressing for space and speed. *Communications of the ACM*, 38(4):41–42, April 1995.

[2] K. Church. One term or two? In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 310–318, Seattle, Washington, July 1995.

[3] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson method for probabilistic weighted retrieval. In W.B. Croft and C.J. van Rijsbergen, editors,

Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval, pages 232–241, Dublin, Ireland, July 1994.

[4] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D.K. Harman, editor, *Proc. Text Retrieval Conference (TREC)*, pages 109–126, Gaithersburg, Maryland, 1994. NIST Special Publication 500-225.

[5] R. Sacks-Davis, T. Arnold-Moore, and A. Kent. A standards-based approach to combining information retrieval and database functionality. International Journal of Information Technology, Vol. 1, No. 1, pages 1–15, 1995.

[6] J. Zobel and A. Moffat. Similarity measures explored. Technical Report TR-95–3, Collaborative Information Technology Research Institute, RMIT and The University of Melbourne, 1995.