Authoritative Re-Ranking in Fusing Authorship-Based Subcollection Search Results

Toine Bogers Antal van den Bosch
ILK / Language and Information Science
Tilburg University, P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands
{A.M.Bogers,Antal.vdnBosch}@uvt.nl

ABSTRACT

We examine the use of authorship information to divide IR test collections into subcollections and we apply techniques from the field of distributed information retrieval to enhance the baseline search results. We base an estimate of an author's expertise on the content of his documents and use this knowledge to construct rankings of the different author subcollections for each query. We go on to demonstrate that these rankings can then be used to re-rank baseline search results and improve performance significantly. We also perform experiments in which we base expertise ratings only on first authors or on all except the final authors and find that these limitations do not further improve our re-ranking method.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Information retrieval, re-ranking, collection fusion, expertise, user modeling

1. INTRODUCTION

Nowadays, many retrieval systems can access a variety of sources and collections to help fulfill a user's information need. Prime examples of such systems are the meta search engines for the Web such as Dogpile [9] and Vivísimo [16], which combine the search results of other search engines and present a merged list of results to the user. Meta search engines do not attempt to perform any actual search themselves. Instead, they respond to user queries by using different search engines to increase their coverage, making use of the fact that not all search engines cover the same parts of the Web. Another type of meta retrieval system that references different sources is an information management assistant that attempts to aid a user in his or her daily writing activities, such as Watson [5] or Syskill & Webert [13]. An important step in any such meta search process is combining the results from the different collections and sources so that they can be presented to the user as if the retrieved documents were all present in one big collection. The entire process, from selecting search engines and submitting the original query to combining and presenting the results, is called *collection fusion.*

In this paper we present a novel method of improving search results within a single collection inspired by collection fusion on a larger scale. The fusion approach we use results in scores for each subcollection that we subsequently use to perform *authoritative re-ranking* of the original search results. For us to be able to apply collection fusion to enhance search in such a way, we need to identify sensible and distinct subcollections within a single collection. The obvious distinction is to equate subcollections with subtopics within the collection, identifying which is a well-researched subfield of information retrieval [15, 18]. Another, more novel way is regarding the sets of papers written by the same author as the subcollections present in the collection. The latter approach utilizes the differences in expertise between authors on certain topics to guide the selection and fusion of the different subcollections. Authors with a lot of documents about a certain topic are more likely to be an expert on that topic and also more likely to have written documents that are relevant for queries about that topic. How we determine this topical expertise is the subject of section 3.

Using author information to identify subcollections requires the parent collection to contain author labels for each document. A typical situation involves, for instance, a research lab where the workgroup is composed of around ten to fifty people and has specific interests. Workgroup members are bound by the common research focus of the workgroup, but each member also has separate interests and may be the group's expert on certain topics. Workgroup collections are also a good testcase because publications of colleagues are often considered to be more trustworthy than random books and articles found in libraries and on the WWW [1]. By adopting a wider perspective and by disregarding institutional or geographical proximity, our method can be ex-

Proceedings of the sixth Dutch-Belgian Information Retrieval workshop (DIR 2006)

tended to scientific communities, e.g. loosely knit groups of people publishing in the same journal or conference proceedings. In general, authoritative fusion can be applied to any collection of documents that represents the research output of a community of sorts, where all the author labels have been preserved. In the remainder of this paper we will refer to such a collection as a *community collection*.

2. BACKGROUND

A major challenge for meta search engines in fulfilling the user's information need is referencing the disparate sources in such a way that it approximates the performance of the hypothetical scenario if all the documents covered by the collections were *all* in a *single* collection [12]. The entire process involves not only selecting the search engines and submitting the original query to these engines but also combining and presenting the results. According to Voorhees, each of these fusion steps has its own peculiar subproblems [17]:

- *Database selection* is concerned with which subcollections to use in responding to an information need. Some collections may charge fees and searching every available collection may be too expensive in terms of resources.
- *Query translation* involves translating the original query to the different formats required by the other search engines used by the meta engine. The utilized search engines may be very different from each other, not only in the retrieval model they use, but also in the type of stemming algorithm used, the use of different stopword lists, or the query processing techniques [6].
- *Document selection* focuses on the question of what kind and how many documents the meta engine should select from the results of every search engine. One problem might be that certain documents may occur in more than one collection but are ranked differently by the search engines. Multiple occurrences of a documents need to be de-duplicated.
- *Results merging* deals with the combining the results into a coherent set to be presented to the user. Not every search engine may return the numerical values used in that specific engine's ranking and some systems might even return results that are not ranked at all.

Different solutions to the collection fusion problem have been proposed over the years. Voorhees et al. [17] propose two different approaches that both use a set of training queries. Their first solution uses relevance feedback information from these training queries to model the distributions of relevant documents over the different collections. They use these distributions to calculate the number of top-ranked documents to be selected from each collection and interleave these ranked result lists. In their second approach they cluster the set of training queries on topic, based on the overlap in relevant documents they retrieve. The new query vector is matched to the cluster centroids and the training weights of the best matching cluster are then retrieved for all collections. These weights are used to determine the number of documents to retrieve from each collection. Callan et al. [6] use a probabilistic approach in the form of an inference network to rank the different collections. They combine these collection-specific weights with the ranking scores assigned to the documents by the retrieval engines of each collection. Documents from collections with high collection weights are favored, but good documents from poor collections can also be ranked higher. Baumgarten [2] also proposes a probabilistic framework for distribution information retrieval, but one that relies less on heuristics and is better motivated theoretically.

In this paper we present a novel method of improving search results where we apply fusion techniques not on disparate collections but on a single collection. We identify different subcollections *within* the parent collection based on the sets of documents written by authors. These documents indirectly represent a subset of the expertise of each author. For each query we derive a ranking of these subcollections based on expertise and use these to re-rank the baseline search results, an approach we call authoritative re-ranking.

This type of *intra-collection* fusion lacks some of the characteristic problems of inter-collection fusion. For instance, searching all the subcollections is not very resource-intensive and since all authors within the parent collection should be considered, the problem of database selection is non-existent. Query translation is also not an issue in our approach since we use one approach for one collection: the same stemming algorithm and stoplist is used for the baseline retrieval and for the ranking of the subcollections. However, our approach does inherit the issues of document selection and results merging; we describe the solutions to these issues within our approach in Section 3.

Constructing rankings of member expertise is a relatively new subfield of information retrieval research. TREC 2005 marked the introduction of the 'Expert Search Task', aimed at solving the problem of identifying employees who are the experts on a certain topic or in a certain situation [14]. Campbell et al. [7] performed similar experiments on a corpus of e-mail messages sent between people in the same company. Neither approach uses these expertise rankings to enhance any kind of information retrieval.

A considerable amount of research has been devoted to improving the search results of information retrieval systems. Among the more successful approaches are query expansion [19] and using cluster analysis [11] or citation analysis for re-ranking purposes [10].

3. AUTHORITATIVE RE-RANKING

As mentioned in the previous sections we try to identify subcollections within a single community collection based on the sets of documents written by an author and the expertise they implicitly represent. We assume that the aggregated content of an author's publications represents his or her expertise. Based on this assumption, we estimate how well a term or phrase points to a certain experts, by calculating the author-term co-occurrence weights in the community collection. We describe a method to create expertise rankings of the members for a query, and use these rankings to re-rank the search results produced by a baseline system. This is similar to producing collection rankings in distributed information retrieval where the collection weights signify the relevance of each collection for a specific query. In our case we combine the original document similarities with subcollection-specific weights: the documents of authors who would be well suited to answer the query will be ranked higher in the final results list.

In addition to this, we also performed some experiments to determine which author rank contributes most to expertise re-ranking. We created special versions of each of our community collections where only the primary authors were included, and versions where the last author was removed from the author listings. Our hypothesis was that, on average, the first author has contributed the most to a paper and the final author the least. This is, in essence, a mild case of database selection by disregarding specific subcollections in the re-ranking process.

We do not use a probabilistic approach, but our approach has much in common with the collection fusion approach of Callan et al. [6]. They too combine the collection-specific weights with the baseline scores assigned to the documents. As in their approach, documents from 'good' collections and good documents from poor collections are favored in the end ranking.

3.1 Baseline approach

Our re-ranking approach was designed to be used on top of a basic vector space model of information retrieval. In our experiments, we used the following formulas for document weights (1) and query weights (2) as proposed by Chisholm et al. [8]:

$$dw_{ij} = \left(\sqrt{f_{ij} - 0.5} + 1\right)\left(\sqrt{\frac{F_i}{n_i}} - 0.9\right) \qquad (1)$$

$$qw_{ij} = (1 + \log(f_{ij}))\left(\log\left(\frac{N}{n_i}\right)\right) \qquad (2)$$

Here, f_{ij} is the frequency of term i in document j, n_i is the number of documents term i appears in, F_i is the frequency of term i throughout the entire collection, and N is the number of documents in the collection. Document-query similarity was calculated by using the cosine measure.

We incorporated some of the tried and tested low-level NLP-techniques in our baseline system, such as stopword filtering and stemming. One-word terms that occurred in the stopword list or in more than a certain percentage of documents were filtered from the documents, and all words were stemmed using the Porter stemming algorithm.

We also experimented with other higher-level techniques such as statistical phrases and using POS tagging and chunking to extract and index syntactic phrases. According to Brants [3], these processing techniques do not always yield improvements and may even result in a decrease in accuracy. Therefore we tested the utility of statistical phrases of different sizes, using syntactic phrases[1], and reweighting based on POS tags. We optimized the use of these techniques for every test collection, as recommended by Brants. We intentionally did not include other techniques such as query expansion in our baseline approach, nor did we distinguish in weighting between the text in the title or the abstract. We intended to measure the effect of our approach as clearly as possible without interference of other possible improvements.

3.2 Test collections

Investigating the merits of authoritative re-ranking required testing our approach on test collections that (a) contain information about the authors of each document, and (b) are a realistic representation of a community, such as a workgroup or a scientific community. We used two well-known test collections, **CACM** and **CISI**, that both represent scientific communities. **CACM** is a reference collection composed of all the 3204 article abstracts published in the Communications of the ACM journal from 1958 to 1979, and **CISI** is made up of 1460 document abstracts selected from a previous collection assembled at ISI [15].

We know of no publicly available IR test collections that represent the body of work published by a workgroup operating in a single institution, which prompted us to create our own: the **ILK** test collection[2]. **ILK** contains 147 document titles and abstracts of publications of current and ex-members of the **ILK** workgroup[3]. The topics of the papers are in the area of machine learning for language engineering and linguistics with subtopics ranging from speech synthesis, morphological analysis, and text analysis & processing to information extraction, text categorization, and information retrieval. We asked the current group members to provide us with queries and the corresponding binary relevance assignments, which resulted in 80 natural language queries.

Table 1: Characteristics of the three main test collections used in the experiments. The total author count ('# total authors') is the sum of the author count over all documents; the total number of unique authors ('# unique authors') is the sum of the author count over all documents with each author counted only once.

	CACM	**CISI**	**ILK**
# documents	3204	1460	147
# queries	52	76	80
# total authors	4392	1971	395
# unique authors	2963	1486	89
avg. # authors per document	1.371	1.350	2.687
avg. # unique authors per doc	0.925	1.018	0.605

Table 1 shows some numeric data characteristics of the three test collections. The four last features listed in the table seem to indicate the type of community collection. **ILK** has a high average number of authors per document but a low average number of unique authors per document, indicating a fairly high degree of cooperation within the community. The distribution of authors in **CACM** is similar to that of **ILK**. This in contrast to, say **CISI**, where these values are lower and higher respectively—it has more cases of solo authorship, and cooperation between the same authors rarely occurs more than once.

[1] We used the Memory-Based Shallow Parser to obtain the POS and chunk tags. See [4] for more information.

[2] Publicly available at http://ilk.uvt.nl/~tbogers/ilk-collection/.

[3] The Induction of Linguistic Knowledge (ILK) workgroup is part of the Department of Language and Information Science of the Faculty of Arts of Tilburg University. It focuses mainly on machine learning for language engineering and linguistics.

Table 2: Author-related characteristics of the six special test collections.

	CACM–first	CISI–first	ILK–first	CACM–m1	CISI–m1	ILK–m1
# total authors	3204	1460	147	3491	1637	278
# unique authors	2155	1112	43	2383	1250	74
avg. # authors per document	1	0.999	0.993	1.090	1.121	1.891
avg. # unique authors per doc	0.673	0.762	0.293	0.744	0.856	0.503

We also performed some experiments to determine which author rank contributes most to expertise re-ranking and created special versions of each collection for this. We created versions where only the primary authors were included (**CACM–first**, **CISI–first**, and **ILK–first**), and versions where the last author was removed from the author listings (**CACM–m1**, **CISI–m1** and **ILK–m1**). This means that, for each community collection, the special versions have the same number of documents and queries. Table 2 lists some characteristics of the six special test collections. The fact that special versions with only the first author have the same number of total authors as documents is not a coincidence. For instance, for **CACM** 3204 documents · 1 author = 3204 total authors.

3.3 Identifying subcollections

Identifying the subcollections in each community collection was a straightforward step. We equate subcollections with the documents written by a member of the community. A document can have multiple authors and can therefore belong to more than one collection—a situation no different from regular distributed information retrieval.

3.4 Determining subcollection weights

Our goal was to determine the expertise of each author to calculate the weights of the different author subcollections: authors with a lot of expertise on a certain query topic were assigned a higher weight. We partitioned the documents into one-vs-all data sets for each author, with each feature vector consisting of the term frequency counts f_{ij} for that document-author combination. In other words, we extracted author-term pairs based on the authorship of a document and the terms appearing in that document, but also the terms appearing in the other documents. We then calculated the co-occurrence weights of each author-term pair for each term (words and phrases) that occurred in the collection. This is similar to Callan's approach, who relies on, among other things, the term occurrence in different collections to calculate collection weights. We examine the co-occurrence of the terms with authors which also involves looking at the occurrence (or lack thereof) in the different author subcollections.

The weights were determined using the following feature selection metrics from text categorization: Information Gain, Chi-Square, and Mutual Information [20]. We also tested using the average TF·IDF value as a measure of term informativeness; collection terms that did not occur in the author's document were assigned a score of zero.

Combining these term weights for each author yielded a matrix of term-author weights which was used to extract the expertise rankings. For each query-author combination an expert score was calculated that signified the expertise of that author on the query topic. Calculating the expert scores was based on the straightforward assumption that if terms characteristic for author X occur in query Q, X is likely to be more of an expert on Q. For each author separately, the informativeness weights were collected for each of the query terms and combined into an expert score. We experimented with taking an unweighted average of the weights and an average weighted by the TF·IDF values of the query terms, so that the differences in the importance of the terms in the query were taken into account. However, there was no appreciable difference between the two, so we chose the intuitively more appealing TF·IDF-weighted average. The end result of this step was ranking of the different subcollections based on the expertise scores[4] for each query. This ranking effectively shows which authors are the biggest experts on the query topic, based on the documents they have authored.

3.5 Document selection & results merging

Document selection and results merging are two issues in collection fusion that are also important for our approach. One issue in document selection is that certain documents may have multiple authors and have different expertise scores. Since our approach works on a single collection and the baseline retrieval also returns a single similarity score for each document-query combination, these documents with multiple expertise scores need to be resolved into a single document score for that query. Merging results involves combining the results into a coherent set to be presented to the user and involves combining the original similarity scores with the expertise weights into a single ranking score. We therefore address both fusion issues simultaneously by re-ranking based on authority.

Our re-ranking is based on the premise that the documents authored by the experts on the current query topic are more likely to be relevant to the query, i.e. more *suitable* to resolve the query. Early experimentation with combining the different expertise scores showed that weighting the scores with the total number of publications of each author gave the best performance. We also investigated abating the influence of high numbers of publications with the square root and the natural logarithm of these counts as weighting factors, which, in general, worked slightly better, but not significantly. After computing this 'suitability' score, which is computed for each query-document combination, it is combined with the original baseline similarity score to form a new score on the basis of which the authoritative re-ranking is performed.

We also performed experiments to determine the optimal way of combining these two scores in order to re-rank the

[4] We will use 'subcollection weights' and 'expertise scores' interchangeably in this paper.

Table 3: Comparison of the re-ranking approaches on R-precision scores. The underlined scores are statistically significant improvements over the baseline.

community collection	re-ranked	baseline	% increase
CACM	<u>0.313</u>	0.233	(+34.3%)
CACM–first	<u>0.302</u>		(+20.2%)
CACM–m1	<u>0.304</u>		(+30.5%)
CISI	<u>0.206</u>	0.203	(+1.5%)
CISI–first	<u>0.206</u>		(+1.5%)
CISI–m1	<u>0.206</u>		(+1.5%)
ILK	0.649	0.647	(+0.3%)
ILK–first	0.650		(+0.5%)
ILK–m1	0.656		(+1.4%)

search results. The most successful combinations involved multiplying the original similarity score with the suitability score (*suit*) and transforming the original similarity score by multiplying it with $1 + suit$ (resulting in a number between 1 and 2). Experiments showed that the optimal re-ranking settings were collection-dependent, so the settings were optimized for each collection, similar to the NLP techniques [3].

4. EVALUATION

We evaluated the performance of our approach using R-precision, the precision at the cut-off rank of the number of relevant documents for a query. R-precision emphasizes the importance of returning more relevant documents earlier. The reliability of the comparisons between our baseline system and the re-ranking approach was determined by performing paired t-tests.

Table 3 shows the results of our experiments. The improvements seem to be very dependent on the community collection used, but improvements were present in each of the nine test collections. Authoritative re-ranking using author-based subcollections produced statistically significant performance improvements on the standard **CACM** test collection and the special versions, ranging from +20.2% to +34.3%. Statistically significant performance improvements were also present in the three versions of the **CISI** test collection, albeit much smaller at +1.5%. Optimal performance on the **ILK** collection yielded very small improvements, but these were not significant. Figures 1–3 show the precision plotted against recall for each cut-off point, both before and after re-ranking, and for each collection. The data points in the lower right half of each graph correspond to the lowest cut-offs. The graphs show that the biggest improvements were made in the top sections of the search results.

A possible reason for these differences in performance might be the topical diversity of the test collections: **CACM** has a much more diverse range of topics than **CISI** and **ILK**, which is likely to make it easier for different areas of expertise to be recognized. Our approach relies on terms that are specific for a certain topic area. This means that our approach has a harder time distinguishing between topics in collections where the different documents are closer together topic-wise.

The experiments with different author selections do not confirm our initial hypothesis: using the expertise of all authors associated with a document yields the best results and using less authors did not increase performance significantly. The difference between the type of community in **CACM** and **CISI** vs. **ILK** might offer an explanation for this, but we have not conducted a more extensive investigation into this matter. These findings suggest that more work is needed to determine the exact influence of author rank.

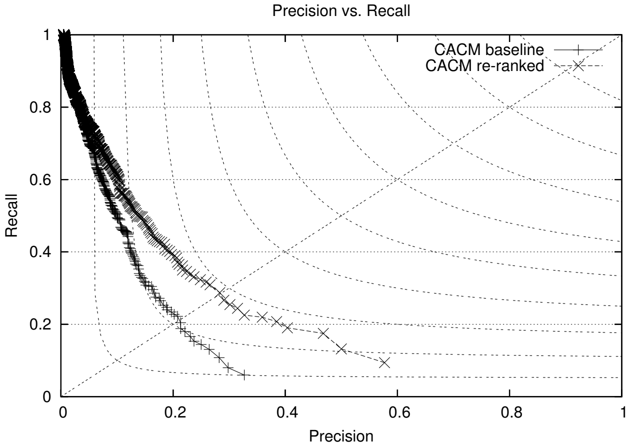

Figure 1: Precision vs. Recall for CACM.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented a novel method of improving search results where we apply fusion techniques on a single collection instead of on disparate collections. We distinguish subcollections based on the sets of documents written by authors and use the content of their documents to produce expertise weights for each query. We use these weights to perform authoritative re-ranking of the baseline search results. Under optimized settings, authoritative re-ranking is able to significantly boost R-precision, especially improving the top search results, with the exact performance increase dependent on the document collection. Therefore, one issue for future research is comparing different ways of constructing expertise rankings such as using clustering, which could also be used to better determine the topical diversity of the three test collections. Another improvement might be the use of citation analysis to improve the expertise scores, similar to the approach taken in [10].

In theory, our approach is equally applicable to the search

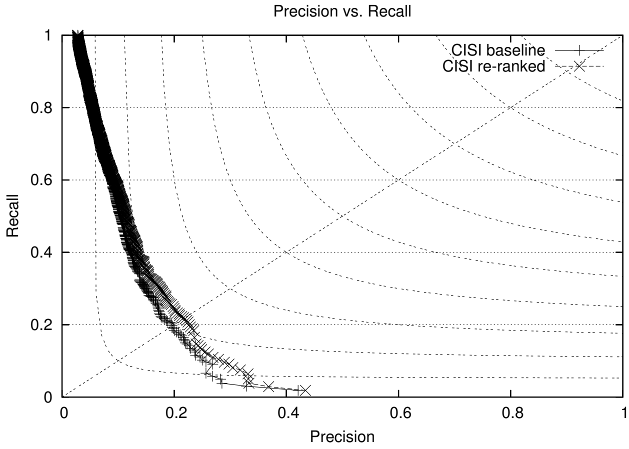

Figure 2: Precision vs. Recall for CISI.

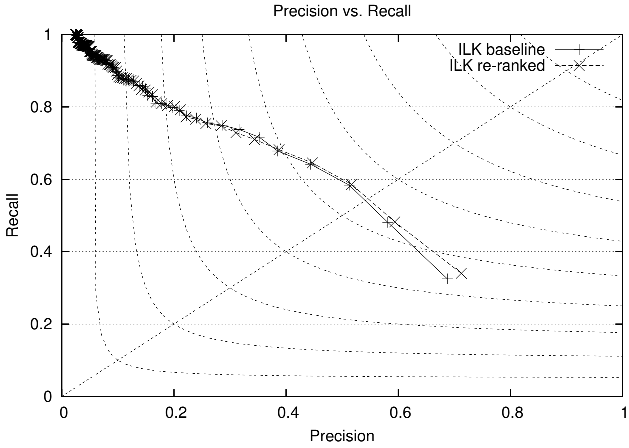

Figure 3: Precision vs. Recall for ILK.

results of, for instance, a probabilistic IR model. However, it would also be interesting to investigate whether using other IR models such as probabilistic retrieval or a language modelling approach indeed show this increase to be universal over the entire range of IR approaches.

Optimal re-ranking performance involves using the expertise of all the authors associated with a document, since considering a smaller number of authors does not increase performance significantly and often decreases it. These findings suggest that more work is needed to determine the exact influence of author rank.

6. ACKNOWLEDGMENTS

This research was funded by the IOP-MMI-program of SenterNovem and the Dutch Ministry of Economic Affairs, as part of the À Propos project. The authors would like to thank Krishna Barat (Google) and Frank Hofstede (Intelli-Gent B.V.) for fruitful discussions and useful comments.

7. REFERENCES

[1] E. Adar, D. Kargar, and L. Stein. Haystack: Per–user Information Environments. In *CIKM '99: Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 413–422, New York, NY, 1999.

[2] C. Baumgarten. A Probabilistic Solution to the Selection and Fusion Problem in Distributed Information Retrieval. In *SIGIR '99: Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 246–253, New York, NY, 1999. ACM Press.

[3] T. Brants. Natural Language Processing in Information Retrieval. In *Proceedings of CLIN 2004*, pages 1–13, Antwerp, Belgium, 2004.

[4] S. Buchholz, J. Veenstra, and W. Daelemans. Cascaded Grammatical Relation Assignment. In *EMNLP-VLC'99, the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

[5] J. Budzik and K. Hammond. Watson: Anticipating and Contextualizing Information Needs. In *62nd Annual Meeting of the American Society for Information Science*, Medford, NJ, 1999.

[6] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, New York, NY, 1995. ACM Press.

[7] C. Campbell, P. Maglio, A. Cozzi, and B. Dom. Expertise Identification using Email Communications. In *Proceedings of CIKM2003*, pages 528–531, New Orleans, LA, 2003.

[8] E. Chisholm and T. Kolga. New Term Weighting Formulas for the Vector Space Method in Information Retrieval. Technical report ORNL/TM-13756, Computer Science and Mathematics Division, Oak Ridge National Laboratory, 1999.

[9] Dogpile. http://www.dogpile.com, 2006. Visited: January 10th, 2006.

[10] C. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of Digital Libraries 98*, pages 89–98, 1998.

[11] K.-S. Lee, Y.-C. Park, and K.-S. Choi. Re-ranking model based on document clusters. *Information Processing & Management*, 37(1):1–14, 2001.

[12] R. M. Losee and L. Church Jr. Information Retrieval with Distributed Databases: Analytic Models of Performance. *IEEE Transactions on Parallel and Distributed Systems*, 14(12):1–10, December 2003.

[13] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill & Webert: Identifying Interesting Web Sites. In *AAAI/IAAI, Vol. 1*, pages 54–61, 1996.

[14] TREC. TREC Enterprise Track, 2005. Visited: October 2005, http://www.ins.cwi.nl/projects/trec-ent/.

[15] C. van Rijsbergen. *Information Retrieval.* Dept. of Computer Science, University of Glasgow, Second edition, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html.

[16] Vivísimo. http://vivisimo.com, 2006. Visited: January 10th, 2006.

[17] E. M. Voorhees, N. Gupta, and B. Johnson-Laird. The Collection Fusion Problem. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 95–104, Gaithersburg, MD, 1995.

[18] W. Wu, H. Xiong, and S. Shekhar. *Clustering and Information Retrieval*. Springer, First edition, 2004.

[19] J. Xu and W. Croft. Query Expansion Using Local and Global Document Analysis. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, New York, NY, 1996. ACM Press.

[20] Z. Zheng and R. Srihari. Optimally Combining Positive and Negative Features for Text Categorization. In *Workshop for Learning from Imbalanced Datasets II, Proceedings of the ICML*, Washington, DC, 2003.