Exploiting Structure for Information Retrieval

Jaap Kamps and **Maarten Marx** and **Christof Monz** and **Maarten de Rijke**
Language & Inference Technology Group
University of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam
E-mail: {kamps,marx,christof,mdr}@science.uva.nl
http://www.illc.uva.nl/LIT/

Abstract

Structured elements are pervasive in digital libraries, product catalogs, scientific data collections and on the Internet. One of our research aims is to investigate the ways in which the additional structure of a collection can be brought to bear on retrieval effectiveness. This paper reports on our experiments on the use of manually assigned keywords in domain specific collections; on the use of URL and link structure on the Internet; and on the use of XML-structure in annotated scientific collections.

1 Introduction

In 2002, the LIT Group at the University of Amsterdam participated in a number of tasks that contain different types of structural information:

- the usage of (manually assigned) keywords in the scientific collections GIRT and Amaryllis used at CLEF (CLEF, 2002);
- the mark-up, URL and link structure in the .GOV collection used at TREC's Web Track (Web Track, 2002); and
- the XML-structure in the IEEE Computer Society collection used at INEX (INEX, 2002).

These three evaluation exercises are loosely related in that they all go beyond the traditional plain-text collection. In all cases, there is some additional structure available that may help to improve the effectiveness of information retrieval, be it that the type of structure differs greatly between tasks.

The outline of this paper is as follows. First, we'll briefly discuss our experimental set up. Then, in three separate sections, we give a brief overview of our experiences during each of the evaluation campaigns. Finally, we discuss our results and draw some tentative conclusions.For further information on our experiments, we refer to (Monz et al., 2002, 2003; Marx et al., 2002).

2 Experimental Set-up

All experiments were carried out with the FlexIR system developed at the University of Amsterdam (Monz and de Rijke, 2002), using the Lnu.ltc weighting scheme.

CLEF Scientific Collections. The domain-specific collections at CLEF are the GIRT collection of German social science literature, and the Amaryllis collection containing French scientific literature. We built free-text only indexes for the GIRT and Amaryllis collections. For both French and German we used a lexical-based stemmer (Schmid, 1994). For German we applied a compound splitter. All morphological runs use blind feedback. Additionally, we built keyword-only indexes of the manually assigned keywords. The keywords were indexed as given, indexing the keywords or keyword-phrases as a single token. Blind feedback was switched off for keyword runs. For GIRT's English to German bilingual runs, we used the Ding dictionary (Ding, 2002). For Amaryllis' English to French bilingual task, we used the on-line Systran translator (Systran, 2002).

TREC Web Track. This year's collection, aptly named .GOV, is based on a crawl of the .gov Internet domain in early 2002. We built a free-text index of the collection using the Porter stemmer (Porter, 1980). Additionally, we built three different anchor-text only indexes, assigning the anchor texts to the linked documents. We made runs on the text and anchors indexes, using Lnu.ltc and a weighting scheme based on minimal matching spans (Monz et al., 2003). None of our runs used blind feedback.

INEX. The collection for INEX consists of IEEE Computer Society journals and proceedings. We built three free-text indexes: using plain words; using the Porter stemmer (Porter, 1980); and using an ngram approach. We preserved the XML-structure in the inverted index by indexing each tag as a single

token. We made initial retrieval runs, using Lnu.ltc weighting and blind feedback. For the content-and-structure queries, we used an XML-parser to extract the required XML-elements from the initially retrieved set of documents.

3 Exploiting Keyword Structure

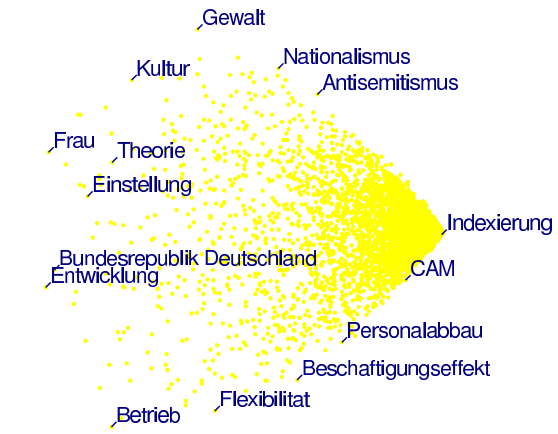
Many domain-specific collections, such as the scientific collections of GIRT and Amaryllis, contain meta-information such as keywords. Special dictionaries or thesauri for the meta-information are not always available. Our strategy for CLEF 2002 was to compute the similarity of keywords based on their occurrence in the collection, and explore whether the resulting keyword space can be used to improve retrieval effectiveness.

The GIRT collection contains $76,128$ documents from German social science literature published between 1978 and 1996 (Kluck and Gey, 2001). The documents are also classified by keywords assigned by human indexers. The average number of keywords in a document is 9.91. A total of $6,745$ different keywords are used in the collection. The Amaryllis collection contains $148,688$ documents in French from various scientific fields. The average number of manually assigned keywords in a document is 10.75. A total of $125,360$ different keywords are used in the collection. We decided to focus on the $10,274$ keywords that occur ≥ 25 times in the collection. We determined the number of occurrences of keywords and pairs of keywords, and used these to define a distance metric (Gower and Legendre, 1986). We reduced the matrix to 10 dimensions using metric multi-dimensional scaling techniques (Cox and Cox, 1994). For all calculations, we used the best approximation of the keyword distance matrix on 10 dimensions (the plots in Figure 1 show the 2 principal dimensions).

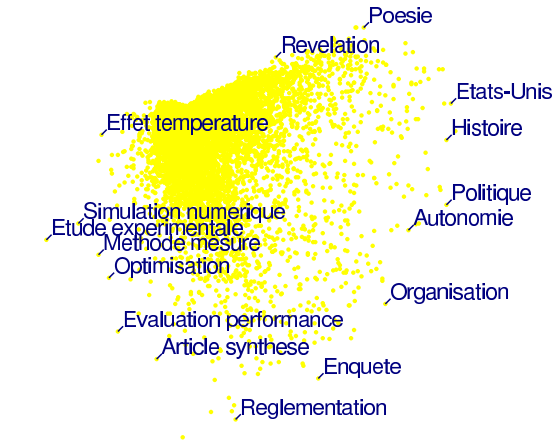
We experimented with the use of the resulting keyword spaces for keyword recovery and document reranking. The resulting keyword space has a 10-dimensional vector for each of the keywords. Vectors for documents and topics are based on the initially retrieved documents from a morphological base run (not using the keywords). For each of these documents, we collect the keywords, and determine a document vector by taking the mean of the keyword vectors. Next, we determine a topic vector by taking the weighted mean of the document vectors for the top 10 documents. We can recover keywords for a topic by selecting, from the keywords used in the top 10 documents, the ten closest to the topic vector. In the monolingual Amaryllis task, the topic authors have assigned keywords for the topics in the narrative field. Table 2 compares these manually assigned keywords to our recovered keywords.

Table 1: Manual Keyword Space.

GIRT 6745 keywords



Amaryllis 10274 keywords



We used the keyword space for recovering keywords, and for document reranking. The recovered keywords are used in keyword-only runs. We created combined runs of the morphological base runs and the keyword-only runs. For document reranking, we simply reranked the documents retrieved in the base run by the distance between the document and topic vectors. The morphological base runs and rerank runs are also used in combined runs. The runs were combined in the following manner. Following Lee (1995), the scores are normalized using $RSV_i' = \frac{RSV_i - min_i}{max_i - min_i}$. We as-

Table 2: Provided versus Recovered Keywords.
Amaryllis topic 001
⟨FR-title⟩ *Impact sur l'environnement des moteurs diesel* ⟨FR-desc⟩ *Pollution de l'air par des gaz d'échappement des moteurs diesel et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2, CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution* ⟨EN-title⟩ The impact of diesel engine on environment ⟨EN-desc⟩ Air pollution by the exhaust of gas from diesel engines and methods of controlling air pollution. Pollutant emissions (NOX, SO2, CO, CO2, unburned product, ...) and air pollution control
Provided keywords
Concentration et toxicité des polluants *Mécanisme de formation des polluants* *Réduction de la pollution* *Choix du carburant* *Réglage de la combustion* *Traitement des gaz d'échappement* *Législation et réglementation*
Recovered keywords
Moteur diesel *Qualité air* *Azote oxyde* *Norme ISO* *Produit pétrolier* *Lutte antipollution air* *Véhicule à moteur* *Gas oil* *Consommation carburant* *Carburant*

signed new weights to the documents using the summation function used by Vogt and Cottrell (1998): $RSV_{new} = \lambda \cdot RSV_1 + (1-\lambda) \cdot RSV_2$. All combination of the morphological base run with a rerank run use interpolation factor 0.6, all combination with a keyword-only run use factor 0.7. These factors were obtained from pre-submission experiments on the GIRT collection.

Results. Table 3 lists our non-interpolated average precision scores for the morphological base runs, and for the best combined runs. The figure in brackets indicates the improvement over the best underlying run. Results for the GIRT monolingual morphological run are disappointing (German monolingual 0.4476, GIRT01 0.3083, GIRT00 0.3145). The GIRT bilingual runs score even worse; the morphological base run has only 1.4 relevant documents in the top 10. This explains the decrease in performance for the run combined with a keyword-only run.

Table 3: NI average precision scores.	
GIRT monolingual	
Morphological	0.1639
Morph/Keyword	0.1687 ($+2.9\%$)
Morph/Rerank	0.1906 ($+16.3\%$)
GIRT bilingual	
Morphological	0.0666
Morph/Keyword	0.0620 (-6.9%)
Morph/Rerank	0.0704 ($+5.7\%$)
Amaryllis monolingual	
Morphological	0.2681
Morph/Keyword provided	0.3401 ($+26.7\%$)
Morph/Keyword recovered	0.2923 ($+9.0\%$)
Morph/Rerank	0.2796 ($+4.3\%$)
Amaryllis bilingual	
Morphological	0.2325
Morph/Keyword	0.2660 ($+14.4\%$)
Morph/Rerank	0.2537 ($+9.1\%$)

For the monolingual Amaryllis task, the provided keywords score remarkably well (keyword-only run 0.2684), the recovered keywords score 0.1120. In combination, both improve the morphological base run: the combination with recovered keywords scores 0.2923 ($+9.0\%$); and with provided keywords scores 0.3401 ($+26.7\%$). The chosen combination factors were generally close to the optimal values for recovered keywords and rerank runs. They proved far from optimal for the provided keywords in monolingual Amaryllis; with 0.4 the combination scores 0.4175 ($+55.6\%$).

4 Exploiting Link Structure

TREC's Web Track featured two tasks: named-page finding and topic distillation. For the text index, we indexed all of the documents' textual contents, decoding special html-characters into plain ASCII, and replacing diacritics with the unmarked characters. The resulting plain-text index covers 1.25 million documents. Arguably, pages that do not receive links from other sites will rarely be key resources. This motivated experiments with anchor-text only runs on three different indexes:

1. Only extracting complete link descriptions in the collection, which includes all links between pages on different sites. All unique anchor-texts are assigned to the document to which the link points. We remove repeated occurrences

of the same anchor-text. The resulting index covers only 15% of the collection.

2. Here we try to recover as many links as possible, including links within a site. We again remove repeated occurrences of the same anchor-texts. The resulting index covers 54% of the collection.

3. We use the same procedure as for the second anchors index, but now retain repeated occurrences, similar to (Craswell et al., 2001).

For the named-page finding task, we experimented with plain text runs, anchor-text runs, and their combinations. The text and anchor-only runs were combined in the following manner: We only consider the first ten results of both runs. The scores are normalized, and we assign new weights to the documents using the summation function used by Fox and Shaw (1994): $RSV_{new} = RSV_1 + RSV_2$.

We performed extensive experiments with link and URL structure for topic distillation. For topic distillation, only the best documents in the collection will be regarded as relevant. We experimented with the following approach for exploiting the URL information: Since there will rarely be more than one key resource per site, we cluster pages by their base URL, and return the page with the lowest URL depth. Specifically, we assign the top 100 documents to the first 10 different base URLs. Next, we return the page with the lowest URL depth or slash-count per cluster.

We also experimented with the use of the link structure of the documents. There are two established ways of exploiting link structure: page-rank (Brin and Page, 1998) uses the global link structure; Hyperlink Induced Topic Search (HITS) (Kleinberg, 1999) uses the local link structure surrounding an initially retrieved set of documents. We implemented an approach that combines both global and local link structure by comparing how much of the links of a page are present in the local set of initially retrieved documents.

We carried out pre-submission experiments using Kleinberg (1999)'s HITS for the topic distillation task. Table 4 shows the top 10 authorities over the top 100, top 200, and top 500 initially retrieved documents for the test topic '*obesity in the U.S.*'. HITS is successful at isolating key resources, but shows considerable topic drift toward generally good 'authorities.' A loosely-related authority can easily infiltrate due to the correlation between authorities and pages with many inlinks (Kleinberg, 1999; Amento et al., 2000). Our link-based method tries to avoid such topic drift. A general good authority may have many links in the local set, but the proportion of inlinks that is in the local set of documents will remain low. The top 10 results are also shown in Table 4 as 'Realized Indegree Top 100/200/500.' Informal evaluation shows that our combined approach is more robust than HITS: when considering the top 500 initially retrieved documents HITS authorities are unrelated to the topics, whereas the 'realized indegree' method remains on topic.

Results. The official run results are shown in Table 5: column 'MRR' lists the mean reciprocal rank of the first correct answer (the official measure); column 'Top 10' lists the number of topics with at least one correct named page in the top 10; and column 'Unknown' lists the number of topics for which no named page was found in the top 50. The combined

Table 5: Official named page finding run results.

Run	MRR	Top 10	Unknown
Text-only	0.4254	82 (54.7%)	46 (30.7%)
Anchors	0.3279	69 (46.0%)	70 (46.7%)
Combined	**0.4317**	99 (66.0%)	35 (23.3%)

text and anchor run performed the best with a MRR of 0.4317. The anchor-text only run, only indexing half the documents, scores 77.08% of the text only run. The combination of both runs improves the MRR by 1.48% over the text only run, the number of topics in the top 10 is improved by 20.73% over the text only run.

For the topic distillation task, we made runs on the text-only and anchors-only collections. Furthermore, we experimented with approaches to exploiting the URL information and link structure of the documents. The results of our official runs are

Table 6: Official topic distillation run results.

Run	Prec. at 10, 20, and 30		
1. Text-only	**0.1755**	0.1245	0.1020
2. Realized indegree 1	0.0673	0.0582	0.0463
3. Anchors	0.1000	0.0714	0.0558
4. Realized indegree 3	0.0633	0.0469	0.0381
5. base URL clusters 3	0.0653	0.0786	0.0660

shown in Table 6. The official measure is precision at 10, at which the text-only run scores best with 0.1755. The anchor-text only run scores 56.98% of the text only run. A text-only run using Lnu.ltc weighting, not submitted, scored better than the official run with a precision at 10 of 0.2102. The run

Table 4: Test topic "obesity in the U.S."

HITS Top 100	Realized Indegree Top 100
www.nih.gov/icd/od/foia/	www.niddk.nih.gov/health/nutrit/pubs/unders.htm
www.nlm.nih.gov/	www.nlm.nih.gov/medlineplus/obesity.html
www.nlm.nih.gov/medlineplus/obesity.html	hin.nhlbi.nih.gov/bmi_palm.htm
www.nlm.nih.gov/accessibility.html	www.ahcpr.gov/research/may00/0500RA6.htm
www.nlm.nih.gov/contacts/	www.nhlbi.nih.gov/guidelines/obesity/bmi_tbl.htm
www.nlm.nih.gov/disclaimer.html	www.nlm.nih.gov/medlineplus/diabetes.html
www.nichd.nih.gov/	www.fitness.gov/Reading_Room/reading_room.html
www.nlm.nih.gov/medlineplus/diabetes.html	www.cdc.gov/nccdphp/dnpa/dnpalink.htm
www.nlm.nih.gov/medlineplus/highbloodpressure.h	response.restoration.noaa.gov/photos/dispers/
www.nlm.nih.gov/medlineplus/sleepdisorders.html	www.fda.gov/bbs/topics/NEWS/NEW00575.html
HITS Top 200	**Realized Indegree Top 200**
www.nih.gov/icd/od/foia/	www.nhlbi.nih.gov/health/public/heart/obesity/
www.nlm.nih.gov/	www.nlm.nih.gov/medlineplus/obesity.html
www.nlm.nih.gov/medlineplus/obesity.html	hin.nhlbi.nih.gov/bmi_palm.htm
www.nichd.nih.gov/	www.niddk.nih.gov/health/nutrit/pubs/unders.htm
www.nlm.nih.gov/disclaimer.html	www.ftc.gov/bcp/conline/pubs/health/setgoals.htm
www.nlm.nih.gov/accessibility.html	www.cdc.gov/nccdphp/dnpa/
www.nlm.nih.gov/contacts/	www.cdc.gov/health/obesity.htm
www.nlm.nih.gov/medlineplus/diabetes.html	whi.nih.gov/health/prof/heart/
www.nlm.nih.gov/medlineplus/highbloodpressure.h	www.ahcpr.gov/research/may00/0500RA6.htm
www.nlm.nih.gov/medlineplus/respiratorydiseases	www.ftc.gov/bcp/conline/pubs/health/setgoals.pdf
HITS Top 500	**Realized Indegree Top 500**
www.disability.gov/	www.nhlbi.nih.gov/health/public/heart/obesity/
www.nhlbi.nih.gov/health/public/heart/obesity/	whi.nih.gov/health/public/heart/
www.business.gov/	www.niddk.nih.gov/health/nutrit/win.htm
www.seniors.gov/	hin.nhlbi.nih.gov/bmi_palm.htm
www.tradenet.gov/	www.niddk.nih.gov/health/nutrit/pubs/binge.htm
www.workers.gov/	www.nlm.nih.gov/medlineplus/obesity.html
www.students.gov/	www.niddk.nih.gov/health/nutrit/pubs/unders.htm
www.seniors.gov/	www.niddk.nih.gov/health/diabetes/pubs/afam/
www.npr.gov/	www.cdc.gov/health/obesity.htm
www.cio.gov/	www.healthfinder.gov/news/

using the base-URL clusters fails to improve the anchor text base run, although it improves precision at 20 and 30. The runs based on link information all perform worse than the underlying base runs.

Table 7: Anchors only run results.

Run	Index	MRR	Prec. at 10
Named page	Anchors 1	0.1391	
Named page	Anchors 2	**0.3279**	
Named page	Anchors 3	0.3098	
Distillation	Anchors 1		0.0673
Distillation	Anchors 2		**0.1000**
Distillation	Anchors 3		0.0837

The post-submission experiments shown in Table 7 show the performance of anchor-text only runs using the three anchor-text indexes as described above. The second anchor-text index, which was used for our official runs, shows the best performance. This index contains unique occurrences of links between and within sites.

5 Exploiting XML Structure

The INEX collection, 21 IEEE Computer Society journals from 1995–2002, consists of $12,135$ documents with extensive XML-markup. The INEX initiative for the evaluation for XML retrieval featured two types of topics: traditional content-only topics, and content-and-structure topics. Our aims at INEX were to set up a baseline system on which we plan to build in future editions of this task. Some of our more ambitious plans failed to be realized due to the inconvenience of crashing XML-parsers, or the inability to produce the required Xpath-location. Our baseline system uses a two-stage strategy. In the first stage, we use the content words in the query to retrieve an initial set of documents. In the second stage, we subject this set of potentially relevant documents to greater scrutiny. In particular, for the content-and-structure queries, we used an XML-parser to extract the required XML-elements from the initially retrieved set of documents.

Our official runs experiment with the effectiveness of different types of morphological normalization for structured corpora. Morphological normalization proved successful for plain text collections (Monz and de Rijke, 2002; Monz et al., 2002). The XML retrieval tasks departs from the strict boolean query matching used in traditional database theory,

allowing for various gradations of relevance. In particular, related words like morphological variants should share some of their relevance. In order to study the precise effect of morphological normalization, we created plain-word, stemmed, and ngrammed indexes that preserve the XML-structure of the original documents. This allows for both the content-only and content-and-structure topics to be evaluated against all three indexes. Informal evaluation shows that morphological normalization helps to retrieve relevant documents missed out by the plain text run. At the time of writing, relevance assessment for INEX is still in progress.

6 Discussion and Conclusions

The three sets of experiments described in this paper, are loosely connected in that they all go beyond the traditional plain-text collection. In all our experiments some additional structure is brought to bear on the information retrieval task, be it that the type of structure differs greatly between tasks. Still, it is worth to discuss some points of agreement between the experiments. There are similarities in the used techniques, the HITS approach uses the same multi-dimensional scaling techniques we applied to the keyword space. MDS techniques give the best approximation of a high-dimensional space in a small number of dimensions. HITS authorities and hubs are based on the principal dimension only, whereas we focus on approximations on ten dimensions.

Both the `.GOV` collection used at Web Track and the IEEE Computer Society collection used at INEX have extensive mark-up in HTML and XML, respectively. Assigning different weights of importance to words occurring in specific tags (such as bold-faced words of headings) can be effective for improving retrieval effectiveness (Cutler et al., 1997). We did not apply this technique yet, since estimating the relative weights of different tags requires a set of test topics. These were not yet available, because both the `.GOV` and IEEE Computer Society collections are used for the first time in 2002. Using the sets of test topics of this year's evaluation, we plan to look at this in more detail.

Most of the journals in the IEEE Computer Society collection have keywords assigned to the documents. Thus, the same techniques we used on the GIRT and Amaryllis collection, i.e., keyword recovery and reranking documents, can be directly applied to the XML retrieval task. Since it was unclear whether these techniques were relevant for the particular type of relevance judgments used at INEX, we did not implement this for our official runs. Again, we plan to address this in future research when the evaluation sets for XML retrieval come available.

It is not the case that using some additional structure will always help to improve the retrieval effectiveness over a highly sophisticated plain-text base run. Our experiments with link-based methods for Web Track's topic distillation task show a decrease in precision at 10. This is in line with earlier attempts at exploiting link structure in the ad hoc task (Hawking and Craswell, 2002). A possible explanation could be the topics used for the distillation task. These are more specific than the very general topics used in Kleinberg (1999), such as '*java*,' '*censorship*,' '*search engines*,' and '*Gates*.' Also, after stopping, the test topic '*obesity in the U.S.*' results in the one-word query '*obesity*.' For such general queries, relevant documents will dominate the top 10, top 100, or even top 200 of initially retrieved documents. Under this assumption, link-based approaches, which ignore the content of documents and solely consider the link topology, can be effective. If non-relevant documents dominate the initially retrieved set of documents, one cannot expect link-based methods to deliver. For the named page finding task, a genuine needle-in-a-haystack task, we experimented with text-only and anchor-text only runs, and their combinations. Here, the combined text/anchor-text run slightly improves the mean reciprocal rank, but significantly improves the number of topics with the named page in the top 10.

The experience on CLEF's scientific collections is that recovered keywords and reranking runs score worse than the morphological base runs. The lower performance of the keyword-only runs is no surprise considering the lack of information contained in the documents' textual parts. The lower performance of the reranking runs is probably due to the unsophisticated reranking strategy that, for example, does not take keyword frequency into account. Having said that, the combined runs with keywords and reranking show a significant improvement of retrieval effectiveness. It is interesting to note that for the GIRT task the combined reranking runs outperform the combined keyword runs, whereas for the Amaryllis task, the combined keyword runs outperform the combined reranking runs. This may be due to the difference in the numbers of keywords used to characterize the documents, which is much

more fine-grained in the case of Amaryllis. The fact that the combined runs significantly improve over the best underlying base runs gives us some confidence in the effectiveness of our approach. It shows that extracting the meaning of keywords from their usage in the collection itself can be a viable alternative for manually constructed, domain-dependent dictionaries and thesauri. Additionally, the keyword space can be useful for providing visualizations of keywords, documents, and topics (Hearst, 1999).

Acknowledgments

The paper benefitted greatly from the comments of the anonymous reviewers. We want to thank Willem van Hage and Vera Hollink for their technical support. Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO), grant # 400-20-036. Maarten Marx received support from NWO grant 612.000.106. Christof Monz was supported by the Physical Sciences Council with financial support from NWO, project 612-13-001. Maarten de Rijke was supported by grants from NWO, under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, and 612.000.207.

References

Amento, B., L. Terveen, and W. Hill (2000). Does 'authority' mean quality? predicting expert quality ratings of web documents. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong, and P. Ingwersen (Eds.), *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 296–303. ACM Press, New York NY, USA.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pp. 107–117. Elsevier Science, New York.

CLEF (2002). Cross language evaluation forum. `http://www.clef-campaign.org/`.

Cox, T. F. and M. A. A. Cox (1994). *Multidimensional Scaling*. Chapman & Hall, London UK.

Craswell, N., D. Hawking, and S. Robertson (2001). Effective site finding using link anchor information. In D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel (Eds.), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 250–257. ACM Press, New York NY, USA.

Cutler, M., Y. Shih, and W. Meng (1997). Using the structure of html documents to improve retrieval. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*.

Ding (2002). Ding: A dictionary lookup program. `http://dict.tu-chemnitz.de/`.

Fox, E. A. and J. A. Shaw (1994). Combination of multiple searches. In D. K. Harman (Ed.), *The Second Text Retrieval Conference (TREC-2)*, pp. 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215.

Gower, J. C. and P. Legendre (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification 3*, 5–48.

Hawking, D. and N. Craswell (2002). Overview of the TREC-2001 web track. In E. M. Voorhees and D. K. Harman (Eds.), *The Tenth Text Retrieval Conference (TREC 2001)*, pp. 25–31. National Institute for Standards and Technology. NIST Special Publication 500-250.

Hearst, M. A. (1999). User interfaces and visualization. In *Modern Information Retrieval*, Chapter 10, pp. 257–324. ACM Press, New York and Addison Wesley Longman, Harlow.

INEX (2002). Initiative for the evaluation of XML retrieval. `http://qmir.dcs.qmw.ac.uk/INEX/`.

Kleinberg, J. M. (1999). Authoritative structures in a hyperlinked environment. *Journal of the ACM 46*, 604–632.

Kluck, M. and F. C. Gey (2001). The domain-specific task of CLEF - specific evaluation strategies in cross-language information retrieval. In C. Peters (Ed.), *Cross-Language Information Retrieval and Evaluation, CLEF 2000*, Volume 2069 of *Lecture Notes in Computer Science*, pp. 48–56. Springer.

Lee, J. H. (1995). Combining multiple evidence from different properties of weighting schemes. In E. A. Fox, P. Ingwersen, and R. Fidel (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 180–188. ACM Press, New York NY, USA.

Marx, M., J. Kamps, and M. de Rijke (2002). The University of Amsterdam at INEX-2002. In N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas (Eds.), *INEX: Initiative for the Evaluation of XML retrieval*.

Monz, C. and M. de Rijke (2002). Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck (Eds.), *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, Volume 2406 of *Lecture Notes in Computer Science*, pp. 262–277. Springer.

Monz, C., J. Kamps, and M. de Rijke (2002). The University of Amsterdam at CLEF-2002. In C. Peters (Ed.), *Results of the CLEF 2002 Cross-Language System Evaluation Campaign*, pp. 73–84.

Monz, C., J. Kamps, and M. de Rijke (2003). The University of Amsterdam at TREC 2002. In E. M. Voorhees and D. K. Harman (Eds.), *The Eleventh Text Retrieval Conference (TREC 2002)*. National Institute for Standards and Technology.

Porter, M. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Systran (2002). Systran Online Translator. `http://www.systransoft.com/`.

Vogt, C. C. and G. W. Cottrell (1998). Predicting the performance of linearly combined ir systems. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 190–196. ACM Press, New York NY, USA.

Web Track (2002). Web track at TREC. `http://www.ted.cmis.csiro.au/TRECWeb/`.