University of Hagen at CLEF2006: Reranking documents for the domain-specific task

Johannes Leveling
FernUniversität in Hagen (University of Hagen)
Intelligent Information and Communication Systems (IICS)
58084 Hagen, Germany
johannes.leveling@fernuni-hagen.de

Abstract

This paper describes the participation of the IICS group at the domain-specific task (GIRT) of the CLEF campaign 2006. The focus of our retrieval experiments is on trying to increase precision by reranking documents in an initial result set. The reranking method is based on antagonistic terms, i.e. terms with a semantics different from the terms in a query, for example antonyms or cohyponyms of search terms.

We analyzed GIRT data from 2005, i.e. the cooccurrence of search terms and antagonistic terms in documents that were assessed as relevant versus non-relevant documents to derive values for recalculating document scores. Several experiments were performed, using different initial result sets as a starting point. A pre-test with GIRT 2004 data showed a significant increase in mean average precision (a change from 0.2446 mean average precision to 0.2986 MAP). Precision for the official runs for the domain specific task at CLEF 2006 did not change significantly, but the best experiment submitted included a reranking of result documents (0.3539 MAP). In an additional reranking experiment that was run on a result set with an already high MAP (provided by the Berkeley group), a significant decrease in precision was observed (MAP dropped from 0.4343 to 0.3653 after reranking).

There are several explanations for these results: First, a simple and obvious explanation is that improving precision by reranking becomes more difficult the better initial results already are. Second, our calculation of new scores includes a factor with a value that was probably chosen too high. We plan to perform additional experiments with more conservative values for this factor.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods, Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*

General Terms

Experimentation, Performance, Measurement

Keywords

Domain-specific IR,Reranking results

1 Introduction

There are several successful methods for improving performance in information retrieval (IR), such as stemming search terms and document terms to increase recall or expanding a query with related terms to increase precision. For our participation at the domain-specific task in CLEF 2006, a method for reranking documents in the initial result set to increase precision was investigated. The method determines a set of antagonistic terms, i.e. terms that are antonyms or cohyponyms of search terms, and it reduces the score (and subsequently the rank) of documents containing these terms. As some search terms will occur in a text together with their corresponding antagonistic terms frequently (e.g., *"day and night"*, *"man and woman"*, *"black and white"*), the factor of cooccurrence with antagonistic terms is considered as well to adapt the calculation of new scores.

For the retrieval experiments, the setup for our previous participations at the domain-specific task was used. It is described in more detail in [9, 8]. The setup includes of a deep linguistic analysis, query expansion with semantically related terms, blind feedback, an entry vocabulary module (EVM, see [5, 3]), and several retrieval functions implemented in the Cheshire II DBMS: *tf-idf*, Okapi/BM25 [11], and Cori/InQuery [2]. For the bilingual experiments, a single online translation service, Promt[1], was employed to translate English topic titles and topic descriptions into German.

2 Reranking with information about antagonistic terms

2.1 The idea

There has already been some research on reranking documents to increase precision in IR. Gey et al. [4] describe experiments with Boolean filters and negative terms for TREC data. In general, this method does not provide a significant improvement, but an analysis for specific topics shows a potential for performance increase. Our group regards Boolean filters to be too restrictive to help improve precision. Furthermore, the case of a cooccurrence of term and filter term (or antagonistic term) in queries or documents is not addressed.

Kamps [7] describes a method for reranking using a dimensional global analysis of data. The evaluation of this method is based on GIRT (German Indexing and Retrieval Testdatabase) and Amaryllis data. The observed improvement for the GIRT data is lower but on the same order as the increase in precision observed in our pre-test (see Sect. 3). While this approach is promising, it relies on a controlled vocabulary and therefore will not be portable between domains or even different text corpora.

For our experiments in the domain-specific task for CLEF 2006 (GIRT), the general idea is to rerank documents in the result set (1000 documents) by combining information about semantic relations between terms (here: antagonistic relations such as antonymy or cohyponymy) as well as statistic information about the cooccurrence frequency of a term and its antagonistic terms (short: *"a-terms"*). Reranking consists of decreasing a document score whenever a query term and one of its a-terms are found in the document.

2.2 Types of antagonistic relations

We introduce the notion of antagonistic terms, meaning terms with a semantics different from search terms. For a given search term t, the set of antagonistic terms a_t contains terms that are antonyms of t (a_t^{word}), terms that are antonyms of a member in the set of synonyms of t (a_t^{synset}), terms that are antonyms of hyponyms of t (a_t^{hypo}), terms that are antonyms of hypernyms of t (a_t^{hyper}), and cohyponyms of t (a_t^{cohypo}).

Figure 1 shows an excerpt of a semantic net consisting of semantic relations such as synonymy (SYNO), antonymy (ANTO, including subsumed relations for converseness, contrariness, and complement) and subordination (SUB). From this semantic net, it can be inferred that *animal* and *plant* (antonyms), *reptile* and *mammalian* (antonym of synonym), *vertebrate* and *plant* (antonym of hypernym/hyponym), and *vertebrate* and *invertebrate* (cohyponyms) are a-terms.[2] We combine different semantic information resources to cre-

[1] http://www.promt.ru/, http://www.e-promt.com/

[2] Note that in a more complete example, *plant* and *animal* would also be cohyponyms of a more general concept, and *vertebrate* and *invertebrate* might be considered antonyms.

ate the semantic net holding this background knowledge, including the computer lexicon HaGenLex [6], a mapping of HaGenLex concepts to GermaNet synonym sets [1], the GIRT-Thesaurus (for hyponym relations) and semantic subordination relations semi-automatically extracted from German noun compounds in text corpora.

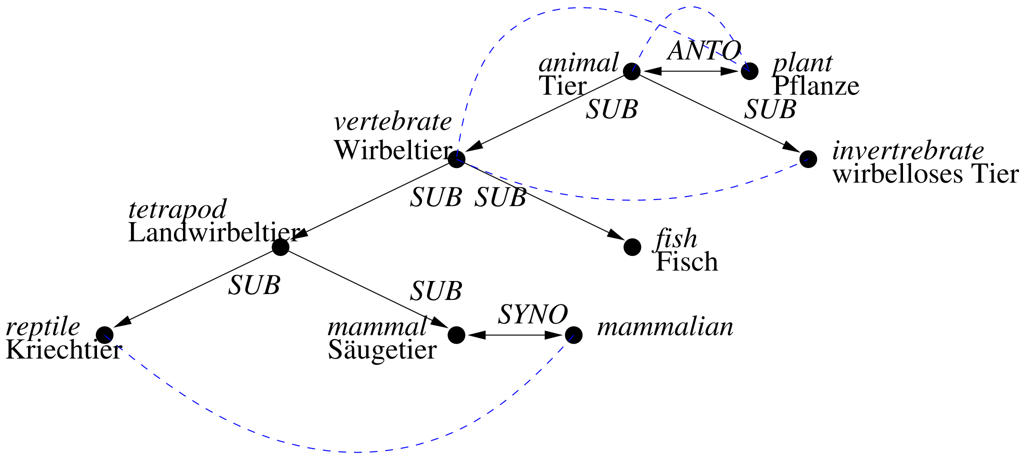

Figure 1: Examples for relations between concepts in a semantic net (*SYNO*:synonymy, *ANTO*:antonymy, *SUB*: subordination/hyponymy) and several inferred antagonistic relations (dashed lines between nodes). Concept names have been translated from German into English.

Using queries and relevance assessments from the GIRT task in 2005, we created a statistics on cooccurrence of query terms and their antagonistic terms in documents assessed as relevant and in other (non-relevant) documents. Table 1 gives an overview over the difference in percentage of term cooccurrence in documents assessed as relevant and other (non-relevant) documents in the GIRT collection. This statistics serves to determine to what amount the score of a document in the result set should be adjusted. For example, a document D with a score S_D that contains a search term A but does contain its cohyponym B will have its score increased.

Table 1: Statistics on differences of percentage of the number of relevant versus the number of non-relevant documents for terms A_i and their antagonistic terms B_j. A total of 6863 cases of antagonistic terms were found for terms in the the GIRT 2005 query topics and documents.

terms			antagonistic type				
D contains			a_t^{word}	a_t^{synset}	a_t^{hypo}	a_t^{hyper}	a_t^{cohypo}
A_i	$\wedge$	B_j	8.0%	4.7%	1.3%	0.4%	0.3%
A_i	$\wedge$	$\neg B_j$	8.0%	18.2%	31.5%	6.0%	19.7%
$\neg A_i$	$\wedge$	B_j	9.0%	2.3%	0.4%	-0.7%	-0.4%
$\neg A_i$	$\wedge$	$\neg B_j$	-26.0%	-25.3%	-33.2%	-5.4%	-19.6%

The data in Table 1 can be briefly interpreted as follows:

- Document scores for documents containing a search term A_i are increased (first and second row), but they are increased less if the document contains an a-term B_j (first row).
- Depending on the type of antagonistic term, document scores are decreased or increased if only A_i's a-term B_j occurs in a document (third row).
- In general, document scores for documents containing neither a term A_i nor its a-term B_j are decreased (last row).

2.3 The reranking formula

Figure 2 shows the algorithm for reranking an initial result set of documents. The following formula for calculating a new score S^{new} from an old score S^{old} for a document D was employed (c is a small constant between 0 and 1):

$$S^{new}_{D_k} = S^{old}_{D_k} + c \cdot \text{freq}(A_i, B_j) \cdot \text{anto\_cooc}(D_k, A_i, B_j) \tag{1}$$

$$\text{anto\_cooc}(D_k, A_i, B_j) = \frac{\text{value for document } D_k \text{ and occurrence of } A_i \text{ and } B_j \text{ /* see Table 1 */}}{1000} \tag{2}$$

$$\text{freq}(A_i, B_j) = \text{number of documents containing term } A_i \text{ and term } B_j \tag{3}$$

```
1. Let $D$ be the initial document set (1000 documents)
2. Let $q$ be the set of query terms $A_i$
3. Let $S_{D_{max}}$ be the highest score of all documents in $D$
4. For each $A_i \in q$:
   • Let $B$ be the set of a-terms $B_j$ for $A_i$
   • For each $B_j \in B$:
     – For each $D_k \in D$:
       * Compute the new score $S^{new}$ of document $D_k$ according to Formula 1 and assign it to $D_k$
5. Normalize all documents scores $S_{D_k}$ so that all values fall into the interval $[0, \cdots, S_{D_{max}}]$
6. Sort $D$ according to new values $S_{D_k}$ and return reranked result set
```

Figure 2: Algorithm for reranking.

3 A pre-test: reranking results from CLEF 2004

In a pre-test of the reranking algorithm, the data consists of queries from GIRT 2004, the corresponding relevance assessments, and the GIRT document corpus. Experiments with different values for the factor c were performed. Table 2 shows the mean average precision (MAP) for our official run from 2004, and MAP for the reranked result set for different values of the factor c. The precision was significantly increased from 0.2446 to 0.2986 MAP for $c = 0.01$.

4 CLEF 2006: reranking results

For the runs submitted for relevance assessment, we employed the experimental setup for the domain-specific task at CLEF in 2005: using query expansion with semantically related terms, and blind feedback for the topic fields title and description. For the bilingual experiments, queries were translated by the Promt online machine translation service. Settings for the following parameters were varied:

- LA: obtain search terms by a linguistic analysis (see [8])

Table 2: MAP for best run for official results at CLEF/GIRT2004 and for monolingual domain-specific pre-tests.

Name/Value of c	MAP
Official CLEF 2004 run	0.2446
c=0.005	0.2951
c=0.008	**0.2986**
c=0.01	**0.2986**
c=0.05	0.2976
c=0.1	0.2759

Table 3: Parameter settings, mean average precision (MAP), and number of relevant and retrieved documents (rel_ret) for monolingual domain-specific experiments. (Additional runs are set in italics.)

Run Identifier	Parameters			Results	
	LA	IRM	RS	rel_ret	MAP
FUHggyydbfl102	Y	tf-idf	N	2444	0.2777
FUHggyydbfl500	Y	BM25	N	2780	0.3205
FUHggyydbfl500R	Y	BM25	Y	2780	0.3179
FUHggyynbfl102	N	if-idf	N	2515	0.2732
FUHggyynbfl500	N	BM25	N	**2920**	0.3525
FUHggyynbfl500R	N	BM25	Y	**2920**	**0.3539**
FUHggyynbfl501	N	Cori	N	1667	0.1374
FUHggyydbfl501	Y	Cori	N	2270	0.2168

- IRM: select retrieval model for Cheshire II (standard *tf-idf*, OKAPI/BM25, Cori/InQuery)
- RS: rerank result set (as described in Sect. 2)

For experiments with reranking, the factor c was set to 0.025. Table 3 and Table 4 show results for official runs and additional runs.

5 A post-test: Reranking results of the Berkeley group

We performed an additional experiment with an even higher MAP of the initial result set, using results of an unofficial run from the Berkeley group.[3] The experiments of the Berkeley group were based on the setup for their participation at the GIRT task in 2005 (see [10]). Reranking applied on the results set

Table 4: Parameter settings and results for bilingual (German-English) domain-specific experiments.

Run Identifier	Parameters			Results	
	LA	IRM	RS	rel_ret	MAP
FUHegpyynl102	Y	tf-idf	N	2134	0.1980
FUHegpyydl102	Y	tf-idf	N	2129	0.1768
FUHegpyydl500	Y	BM25	N	2422	0.2190
FUHegpyydl500R	Y	BM25	Y	2422	0.2180
FUHegpyynl500	N	BM25	N	**2569**	**0.2448**

[3] Thanks to Vivien Petras at UC Berkeley for providing the data.

found by Berkeley, which has an average MAP of 0.4343 for the monolingual German task (3212 rel_ret), significantly lowered performance to 0.3653 MAP.

6 Discussion of results

We performed different sets of experiments with reranking initial result sets for the domain-specific task at CLEF. In a pre-test that was based on the data from CLEF 2004 and results submitted in 2004, reranking increased the MAP from 0.2446 to 0.2976 (+ 21.6%) change). As a single result set is used as input for reranking experiments, recall is not affected.

Results for the official experiments indicate that reranking does not significantly change the MAP. For the monolingual run, MAP dropped from 0.3205 to 0.3179 in one pair of experiments and rose from 0.3525 to 0.3539 in another. For a bilingual pair of comparable experiments, MAP dropped from 0.2190 to 0.2180.

An additional reranking experiment was based on data provided by the Berkeley group with their setup from 2005. The MAP decreased from 0.4343 to 0.3653 when reranking was applied to this data.

There are several explanations as to why precision is affected so differently:

- There may be different intrinsic characteristics for the domain-specific query topics in 2004 and 2006 (the GIRT data did not change), i.e. there may have been fewer antagonistic terms found for the query terms in 2006. We did not have time to test this hypothesis.
- The dampening factor c was not fine-tuned for the retrieval method employed to obtain the initial result set: for the experiments in 2004, we used a database management system with a *tf-idf* IR model, while for GIRT 2006, the OKAPI/BM25 IR model was applied. The corresponding result sets show a different range and distribution of document scores. Thus, the effect of reranking document with the proposed method may depend on the retrieval method employed to obtain the initial results.
- Reranking will obviously become harder the better the initial precision already is. The results from the Berkeley group will be more difficult to improve, as they already have a high precision.
- The dampening factor c should have been initialized with a lower value. Due to time constraints, our group did not have time to repeat reranking experiments with different and more conservative values of c.

7 Conclusion

In this paper, a novel method to rerank documents was presented. It is based on a combination of information about antagonistic relations between terms in queries and documents and their cooccurrence. Different evaluations for this method were presented, showing mixed results.

For a pre-test with CLEF data from 2004, a performance increase in precision was observed. Official results for CLEF 2006 show no major changes, and an additional experiment based on data from the Berkeley group even shows a decrease in precision.

While the pre-test showed that our reranking approach should work in general, the official and additional experiments indicate that it becomes more difficult to increase precision the higher it already is. We plan to complete reranking experiments with different settings and analyze differences in query topics for GIRT 2004–2006.

References

[1] R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. *The CELEX Lexical Database. Release 2 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, 1995.

[2] James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In *Proceedings of the ACM SIGIR 1995*, 1995.

[3] Fredric C. Gey, Michael Buckland, Aitao Chen, and Ray R. Larson. Entry vocabulary – a technology to enhance digital search. In *Proceedings of the First International Conference on Human Language Technology*, March 2001.

[4] Fredric C. Gey, Aitao Chen, Jianzhang He, Liangjie Xu, and Jason Meggs. Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms at TREC-5. In National Institute for Standards and Technology, editor, *Proceedings of TREC-5, the Fifth NIST-DARPA Text REtrieval Conference*, pages 181–190, Washington, DC, 1996.

[5] Fredric C. Gey, Hailing Jiang, Vivien Petras, and Aitao Chen. Cross-language retrieval for the CLEF collections – comparing multiple methods of retrieval. In C. Peters, editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000*, volume 2069 of *Lecture Notes in Computer Science (LNCS)*, pages 116–128. Springer, Berlin, 2001.

[6] Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105, 2003.

[7] Jaap Kamps. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In Sharon McDonald and John Tait, editors, *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR 2004)*, volume 2997 of *Lecture Notes in Computer Science (LNCS)*, pages 283–295. Springer, Heidelberg, 2004.

[8] Johannes Leveling. A baseline for NLP in domain-specific information retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *CLEF 2005 Proceedings*, Lecture Notes in Computer Science (LNCS). Springer, Berlin, 2006. In print.

[9] Johannes Leveling and Sven Hartrumpf. University of Hagen at CLEF 2004: Indexing and translating concepts for the GIRT task. In C. Peters, P. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 271–282. Springer, Berlin, 2005.

[10] Vivien Petras. How one word can make all the difference – using subject metadata for automatic query expansion and reformulation. In Carol Peters, editor, *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop*, Wien, Austria, September 2005. Centromedia.

[11] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In D. Harman, editor, *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. National Institute of Standards and Technology (NIST), Special Publication 500-226, 1994.