Reranking Documents with Antagonistic Terms

Johannes Leveling

FernUniversität in Hagen (University of Hagen)
Intelligent Information and Communication Systems (IICS)
58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de

Abstract. For the participation of the IICS group at the domain-specific task (GIRT) of the CLEF campaign 2006, we employ cooccurrence of search terms and antagonistic terms (antonyms or cohyponyms) in documents to derive values for reranking an initial result set.

A reranking test on GIRT 2004 data showed a significant increase in mean average precision (MAP), i.e. a change from 0.2446 MAP to 0.2986 MAP. Precision for the submitted runs for the domain-specific (DS) task did not change significantly, but the setup for the best experiment included a reranking of result documents (0.3539 MAP). For reranking 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).

1 Introduction

For our participation at the domain-specific information retrieval (IR) task at CLEF 2006, we investigated a method for reranking documents in the initial result set to increase precision. The method determines a set of antagonistic terms, i.e. antonyms or cohyponyms of search terms, and it reduces the score (and subsequently the rank) of documents containing a search term and an antagonistic term. As some terms will frequently occur together with their corresponding antagonistic terms (e.g., day and night, black and white), cooccurrence of search terms and antagonistic terms is considered as well to adapt the calculation of new scores.

The setup for our previous participations at the domain-specific task was used (see [1]) to obtain initial result sets. It includes of a deep linguistic analysis, query expansion with semantically related terms, blind feedback, an entry vocabulary module [2], and several retrieval functions. For the bilingual experiments, a single online translation service, Promt[2], was employed to translate English topic titles and topic descriptions into German.

1 The research described was in part funded by the DFG (Deutsche Forschungsgemeinschaft) in the project IRSAW (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web).

2 Reranking with Information About Antagonistic Terms

There has already been some research on reranking documents to increase precision in IR. Gey et al. [3] 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 [4] 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 performance gain is lower but on the same order as the increase in precision observed in our first reranking test. While this approach is promising, it relies on a controlled vocabulary and therefore will not be portable between domains or even different text corpora.

We introduce the notion of antagonistic terms, meaning terms with a semantics opposed to the semantics of search terms. For a given search term $t$, the set of antagonistic terms contains antonyms of $t$, antonyms of a member in the set of synonyms of $t$, antonyms of hyponyms of $t$, antonyms of hypernyms of $t$, and cohyponyms of $t$. For example, animal and plant are antonyms, vertebrate and plant are antonyms of a hypernym/hyponym, and vertebrate and invertebrate are cohyponyms. Different semantic information resources were combined to create a semantic net holding this background knowledge, including the computer lexicon HaGenLex [5], a mapping of HaGenLex concepts to GermaNet synonym sets [6], the GIRT-Thesaurus (for hyponym relations) and semantic subordination relations semi-automatically extracted from German text corpora.

Based on queries and relevance assessments from the GIRT task in 2005, cooccurrence of query terms and their antagonistic terms in documents assessed as relevant and in other (non-relevant) documents was calculated. The statistics serve to determine to what amount the score of a document in the result set will be adjusted. For example, a document $D$ with a score $S_D$ that contains a search term $A$ as well as its antonym $B$ will have its score decreased. The reranking algorithm, reranking formula, and statistics on cooccurrence of terms are described in more detail in [7].

3 Reranking Results

The reranking algorithm was tested on data and queries from GIRT 2004 (including the corresponding relevance assessments). Our best official run at the DS task in 2004 has 0.2446 MAP. For different values of a factor $c$ in the reranking formula (see [7]), reranking the results always yielded a significantly higher MAP than the official run: 0.2951 ($c=0.005$), 0.2986 ($c=0.008$), 0.2986 ($c=0.01$), 0.2976 ($c=0.05$), and 0.2759 ($c=0.1$).

For the domain-specific task at CLEF 2006, settings for the following parameters were varied: LA: obtain search terms by a linguistic analysis [1], IRM: select