Principled Query Processing

Jussi Karlgren, Magnus Sahlgren, Rickard Cöster
Swedish Institute of Computer Science, Stockholm
{jussi, mange, rick@sics.se

Abstract

This year, the SICS team decided to concentrate on query processing and on the internal topical structure of the query: we have identified this as one of the major bottlenecks for cross-lingual access systems. Previous years, the SICS team has investigated, among other issues, how to translate compounds. Compound translation is non-trivial due to dependencies between compound elements and has been treated in various ways in the treatment of compounding languages such as Swedish. We decided this year to investigate the topical dependencies between query terms, under the hypothesis that the complexity of translating compounds is a special case of the more general case of understanding the respective topicality of query terms.

The question under investigation is how much each query term contributes in terms of topicality in the documents of the collection under consideration. If a query term happens to be non-topical or noise, it should be discarded or given a low weight when ranking retrieved documents; if a query term shows high topicality its weight should be boosted. Our base system is used with two different enhancements to test the hypothesis that boosting topically active terms is beneficial for retreival results.

Both schemes are based on the analysis of the distributional character of query terms: one using similarity of occurrence context between query terms; the other using the likelihood of individual terms to appear topically in text. These are two different avenues of analysis and will most likely provide different results if pursued further than these initial experiments.

The results of the boosting schemes delivered uncontroversially improved results. These results will provide impetus for the further study of translation of complex terms — the question which first prompted this set of experiments in the first place.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; I.2.7 Natural Language Processing

General Terms

Experimentation, Performance, Theory

Keywords

Query processing, Distributional data, Term topicality

1 Query Terms and Their Internal Relations

This year, the SICS team decided to concentrate on query processing and on the internal topical structure of the query: we have identified this as one of the major bottlenecks for cross-lingual access systems. Previous years, the SICS team has investigated, among other issues, how to translate compounds [2]. Compound translation is non-trivial due to dependencies between compound elements and has been treated in various ways in the treatment of compounding languages such as Swedish [1, 3, e.g.]. We decided this year to investigate the topical dependencies between query terms, under the hypothesis that the complexity of translating compounds is a special case of the more general case of understanding the respective topicality of query terms.

The question under investigation is how much each query term contributes in terms of topicality in the documents of the collection under consideration. If a query term happens to be non-topical or noise, it should be discarded or given a low weight when ranking retrieved documents; if a query term shows high topicality its weight should be boosted. Our base system is used with two different enhancements to test the hypothesis that boosting topically active terms is beneficial for retreival results.

2 Base System

The French target collection and the French topics were lemmatized and normalized using the commercially available FDG tools from Connexor Oy, described in several publications [8, e.g.].

The text retrieval engine used for our experiments is based on a standard retrieval system being developed at SICS. The system is described in more detail in our CLEF paper from 2002 [7]. The French target collection was indexed by the system and the translated French queries were used to retrieve texts from the French collection without manual intervention.

In retrieval, query terms were weighted by standard tf.idf metrics, and, in contrast with earlier years, documents with several of the query terms were given a boost higher than documents with the equivalent number of occurrences but distributed over fewer different query terms.

3 Distributional Statistics using Vector Space Models

In this experiment, we use distributional information to weight words selected from the query description field. The idea is to select words with similar distributional properties, since they are assumed to indicate similar topics. As an example, consider query number 251, where supposedly the term *"médecine"* is a good descriptor. We would then want to boost the weight of query words that are topically similar to *"médecine"* but that occur in other documents (it would be no point in selecting words that occur in exactly the same documents, since we retrieve those documents anyway by using the term *"médecine"*). Considering the example query, we would supposedly like to include words such as *"homéopathie", "chiropractie", "acupuncture", and "thérapie"*. Our hypothesis is that we can use second order co-occurrence information to find such query words.

Our approach is based on Random Key Indexing[4, 5], a technique for the efficient and tractable analysis of cooccurrence statistics. Random Key Indexing incrementally collects distributional data for terms in the text collection under consideration and can be used to build a vector space based on those data. In this experiment we use Random Key Indexing[1] to collect second order co-occurrences to accumulate a word space in which words with similar distributional properties are located close to each other. We compute distributional similarity between words using the cosine of the angles between "context vectors" that represent their distributional profiles. The cosine values are then used to weight the words in the query description field.

Note that we use *second order* co-occurrences, since we want to find words with similar distributional statistics that *do not* occur in the same documents, but that occur in the same *type*

[1] Parameters settings for the Random Key Indexing process: 1000-dimensional vectors; 1% non-zero elements in the index vectors; 2+2-sized distance weighted context window.

of contexts. Using first order co-ocurrences (as is done in, e.g., Latent Semantic Analysis) would merely find words that occur in similar documents, which is not beneficial for the adhoc Information Retrieval task, since those documents are found by the system by default.

4 Probabilistic Models: Katz' γ

Using an analysis of query term distribution in the target collection, Katz' γ is calculated for each term in the query. This can be understood as the estimated probability for the term to appear at least twice in any given text and is calculated by as the relative frequency of texts with at least two occurrences of the term under consideration to texts with only one occurrence of it. The intuition underlying Katz' γ is that singleton occurrences may be happenstance noise whereas repeated occurrences of a term are likely to be topical [6]; the intuition behind our use of the measure is that terms that often are likely to be topical are likely to be of more interest as regards query relevance than terms that often occur non-topically.

5 Three Submissions and their Results

The first submission (V) used the baseline system without modification. The second submission (B) boosted query terms according to their location in the vector space as provided by random key indexing by multiplying the standard tf.idf score with the cosine between it and the closest neighbor of the other query terms. The third submission (K) boosted terms that are likely to be topical by multiplying the standard tf.idf score with its γ.

	Average precision	Precision at 20	Above median	At median	Below median
V	0.3135	0.420	14	9	27
B	0.3174	0.421	15	11	24
K	0.3271	0.427	21	1	27

The results were reasonably good with half of the fifty queries on or above median. The two boosting schemes proved to deliver improved results.

6 Discussion of First Impressions

The results of the boosting schemes delivered uncontroversially improved results. One scheme examined the individual character of the terms; the other the relation between query terms. These are two different avenues of analysis and will most likely provide different (and even better) results if pursued further. These results will also provide impetus for the further study of translation of complex terms — the question which first prompted this set of experiments in the first place.

References

[1] Per Ahlgren. *The Effects of Indexing Strategy-Query Term Combination on Retrieval Effectiveness in a Swedish Full Text Database.* PhD thesis, Department of Library and Information Science, University College of Borås, Borås, Sweden, 2004.

[2] Rickard Cöster, Magnus Sahlgren, and Jussi Karlgren. Selective compound splitting of swedish queries for boolean combinations of truncated terms. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003).* Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2004.

[3] Turid Hedlund. *Dictionary-Based Cross-Language Information Retrieval: Principles, System Design and Evaluation.* PhD thesis, Department of Information Science, University of Tampere, Tampere, Finland, 2003.

[4] P. Kanerva, J. Kristofersson, and A. Holst. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036. Erlbaum, 2000.

[5] J. Karlgren and M. Sahlgren. From words to understanding. In Y. Uesaka, P. Kanerva, and H. Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications, 2001.

[6] Slava Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2:15–60, 1996.

[7] Magnus Sahlgren, Jussi Karlgren, Rickard Cöster, and Timo Järvinen. Automatic query expansion using random indexing. In *Proceedings of CLEF 2002*, 2002.

[8] Pasi Tapanainen and Timo Jrvinen. A non-projective dependency parser. pages 64–71, 1997.