Master Modélisation Informatique des Connaissances et du Raisonnement (MICR)

Cours Recherche et Extraction d'information

Liste des articles à présenter, pour la partie Recherche d'Information

- Chris Buckley and Ellen M. Voorhees
  Evaluating evaluation measure stability
  Proceedings of the 23rd Annual International ACM SIGIR conference on Research and Development in Information Retrieval, p. 33-40, 2000
  http://citeseer.ist.psu.edu/buckley00evaluating.html

  Abstract: This paper presents a novel way of examining the accuracy of the evaluation measures commonly used in information retrieval experiments. It validates several of the rules-of-thumb experimenters use, such as the number of queries needed for a good experiment is at least 25 and 50 is better, while challenging other beliefs, such as the common evaluation measures are equally reliable. As an example, we show that Precision at 30 documents has about twice the average error rate as Average Precision has. These results can help information retrieval researchers design experiments that provide a desired level of confidence in their results. In particular, we suggest researchers usingWeb measures such as Precision at 10 documents will need to use many more than 50 queries or will have to require two methods to have a very large difference in evaluation scores before concluding that the two methods are actually different.

- John Ho Lee
  Analyses of multiple evidence combination
  Proceedings of the 20th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, p. 267-276, 1997
  http://citeseer.nj.nec.com/152190.html

  Abstract: It has been known that different representations of a query retrieve different sets of documents. Recent work suggests that significant improvements in retrieval performance can be achieved by combining multiple representations of an information need. However, little effort has been made to understand the reason why combining multiple sources of evidence improves retrieval effectiveness. In this paper we analyze why improvements can be achieved with evidence combination, and investigate how evidence should be combined. We describe a rationale for multiple evidence combination, and propose a combining method whose properties coincide with the rationale. We also investigate the effect of using rank instead of similarity on retrieval effectiveness.

- Rada Mihalcea and Dan Moldovan
  Semantic Indexing Using WordNet Senses
  Proceedings of ACL Workshop on IR & NLP, 2000
  http://citeseer.ist.psu.edu/417656.html

  Abstract: We describe in this paper a boolean Information l~.etrieval system that adds word semantics to the classic word based indexing. Two of the main tasks of our system, namely the indexing and retrieval components, are using a combined word-based and sense-based approach. The key to our system is a methodology for building semantic representations of open text, at word and collocation level. This new technique, called semantic indexing, shows improved effectiveness over the classic word based indexing techniques.

- Lisa Ballesteros and W. Bruce Croft
  Resolving Ambiguity for Cross-Language Retrieval
  Research and Development in Information Retrieval p. 64-71, 1998
  http://citeseer.ist.psu.edu/ballesteros98resolving.html

  Abstract: One of the main hurdles to improved CLIR effectiveness is resolving ambiguity associated with translation. Availability of resources is also a problem. First we present a technique based on co-occurrence statistics from unlinked corpora which can be used to reduce the ambiguity associated with phrasal and term translation. We then combine this method with other techniques for reducing ambiguity and achieve more than 90\% monolingual effectiveness. Finally, we compare the co-occurrence method with parallel corpus and machine translation techniques and show that good retrieval effectiveness can be achieved without complex resources.

- Hua Cheng, Yan Qu, Jesse Montgomery, David A. Evans
  Exploring Semantic Constraints for Document Retrieval
  Proceedings of the Workshop on *How Can Computational Linguistics Improve Information Retrieval?* ACL/COLING 2006
  http://acl.ldc.upenn.edu/W/W06/W06-0800.pdf

  Abstract:In this paper, we explore the use of structured content as semantic constraints for enhancing the performance of traditional term-based document retrieval in special domains. First, we describe a method for automatic extraction of semantic content in the form of attribute-value (AV) pairs from natural language texts based on domain models constructed from a semistructured web resource. Then, we explore the effect of combining a state-of-the-art term-based IR system and a simple constraint-based search system that uses the extracted AV pairs. Our evaluation results have shown that such combination produces some improvement in IR performance over the term-based IR system on our test collection.

- Carmen Alvarez, Philippe Langlais and Jian-Yun Nie
  Word Pairs in Language Modeling for Information Retrieval
  RIAO 2004 Conference Proceedings, p. 686--705, 2004
  http://citeseer.ist.psu.edu/alvarez04word.html

  Abstract: Previous language modeling approaches to information retrieval have focused primarily on single terms. The use of bigram models has been studied, but the restriction on word order and adjacency may not be justified for information retrieval. We propose a new language modeling approach to information retrieval that incorporates lexical affinities, or pairs of words that occur near each other, without a constraint on word order. The use of compound terms in the vector space model has been shown to outperform the vector model with only single terms (Nie & Dufort, 2002). We explore the use of compound terms in a language modeling approach, and compare our results with the vector space model, and unigram and bigram language model approaches.

- C. Zhai and J. Lafferty
  Two-stage language models for information retrieval
  Proceedings of the 25rd Annual International ACM SIGIR conference on Research and Development in Information Retrieval, 2002
  http://citeseer.ist.psu.edu/717851.html

  Abstract: The optimal settings of retrieval parameters often depend on both the document collection and the query, and are usually found through empirical tuning. In this paper, we propose a family of two-stage language models for information retrieval that explicitly captures the different influences of the query and document collection on the optimal settings of retrieval parameters. As a special case, we present a two-stage smoothing method that allows us to estimate the smoothing parameters completely automatically. In the first stage, the document language model is smoothed using a Dirichlet prior with the collection language model as the reference model. In the second stage, the smoothed document language model is further interpolated with a query background language model. We propose a leave-one-out method for estimating the Dirichlet parameter of the first stage, and the use of document mixture models for estimating the interpolation parameter of the second stage. Evaluation on five different databases and four types of queries indicates that the twostage smoothing method with the proposed parameter estimation methods consistently gives retrieval performance that is close to or better than the best results achieved using a single smoothing method and exhaustive parameter search on the test data.

- Stefan Klink and Armin Hust and Markus Junker and Andreas Dengel
  Improving Document Retrieval by Automatic Query Expansion Using Collaborative Learning of Term-Based Concepts
  Proceedings of DAS 2002, 5th International Workshop on Document Analysis Systems, LNCS, p. 376—387, 2002
  http://citeseer.ist.psu.edu/klink02improving.html

  Abstract: Query expansion methods have been studied for a long time with debatable success in many instances. In this paper, a new approach is presented based on using term concepts learned by other queries. Two important issues with query expansion are addressed: the selection and the weighing of additional search terms. In contrast to other methods, the regarded query is expanded by adding those terms which are most similar to the concept of individual query terms, rather than selecting terms that are similar to the complete query or that are directly similar to the query terms. Experiments have shown that this kind of query expansion results in notable improvements of the retrieval effectiveness if measured the recall/precision in comparison to the standard vector space model and to the pseudo relevance feedback. This approach can be used to improve the retrieval of documents in Digital Libraries, in Document Management Systems, in the WWW etc.

- Marti A. Hearst and Jan O. Pedersen
  Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results
  Proceedings of 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, 1996
  http://citeseer.ist.psu.edu/hearst96reexamining.html

  Abstract: We present Scatter/Gather, a cluster-based document browsing method, as an alternative to ranked titles for the organization and viewing of retrieval results. We systematically evaluate Scatter/Gather in this context and find significant improvements over similarity search ranking alone. This result provides evidence validating the cluster hypothesis which states that relevant documents tend to be more similar to each other than to non-relevant documents. We describe a system employing Scatter/Gather and demonstrate that users are able to use this system close to its full potential.

- Boldareva, L. and de Vries, A. P. and Hiemstra, D.
  Monitoring User-System Performance in Interactive Retrieval Tasks
  RIAO 2004 Conference Proceedings, p. 474—486, 2004
  http://www.riao.org/sites/RIAO-2004/Proceedings-2004/papers/0520.pdf

  Abstract: Monitoring user-system performance in interactive search is a challenging task. Traditional measures of retrieval evaluation, based on recall and precision, are not of any use in real time, for they require a priori knowledge of relevant documents. This paper shows how a Shannon entropy-based measure of user-system performance naturally falls in the framework of (interactive) probabilistic information retrieval. The value of entropy of the distribution of probability of relevance associated with the documents in the collection can be used to monitor search progress in live testing, to allow for example the system to select an optimal combination of search strategies. User profiling and tuning parameters of retrieval systems are other important applications.