ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval MF/IR 2005

Salvador, Brazil
August 15-19, 2005

Sándor Dominich, Iadh Ounis, Jian-Yun Nie

1 Introduction

The previous five MF/IR workshops showed that the mathematical/formal results achieved in Information Retrieval (IR) could be organized into a coherent theoretical framework, that they brought new knowledge to IR, and that mathematical/formal research in IR can stand as a specialized research area of IR. Therefore the purpose of the MF/IR 2005 was to promote discussion and interaction among those with theoretical and applicative research interests in mathematical/formal aspects of Information Retrieval coming from a large spectrum of different IR fields, and also at being a forum for the presentation of both theoretical and applicative results (e.g., foundational issues; description and/or integration of models; retrieval applications; mathematical/formal techniques, properties and structures in IR; existing and/or new theories and theoretical aspects, interdisciplinary approaches) using formal and mathematical approaches like Sets, Vectors, Similarity Functions, Probability, Algebra, Topology, Metric Spaces, Geometry, Logics, Graph Theory.

2 Papers

The event began with an opening talk given by Jian-Yun Nie on implementing logical retrieval models using language models. The event continued with the presentation of the following papers:

"Parsimonious Translation Models for Information Retrieval" by Seung-Hoon Na, In-Su Kang, Jong-Hyeok Lee.

A modeling approach has a critical problem estimating a query model, which is the probabilistic model that encodes the user's information need. For query expansion in initial retrieval, the translation model had been proposed to involve term co-occurrence statistics. However, the translation model was difficult to apply, because term co-occurrence statistics must be constructed in the offline time. In large collections, constructing such a large matrix of term co-occurrences statistics prohibitively increases time and space complexity. More seriously, reliable retrieval performance cannot be guaranteed because the translation model may comprise noisy non-topical terms in documents. To resolve these problems, this paper investigates an effective method to construct co-occurrence statistics and eliminate noisy terms by employing a parsimonious translation model. The parsimonious translation model is a compact version of a translation model that can reduce the number of terms containing non-zero probabilities by eliminating non-topical terms in documents. Through experimentation on seven different test collections, the authors show that the query model estimated from the parsimonious translation model significantly outperforms not only baseline language modeling but also non-parsimonious models.

"Extracting Template for Knowledge-based Question-Answering Using Conditional Random Fields" by Changki Lee, Ji-Hyun Wang, Hyeon-Jin Kim, Myung-Gil Jang .

In this paper, the authors present an information extraction system that extracts template elements for a question answering (QA) system in the domain of encyclopaedia. They use Conditional Random Fields to extract templates from the texts of an encyclopedia. Using the proposed approach, the authors could achieve a 74.89% precision and a 55.77% F1 in the template extraction. In the question classification, they could archive an 83.6% precision and a 65.4% recall. Finally, in the Knowledge based QA (including template extraction procedure), they could archive an 81.3% precision and a 33.3% recall. The result demonstrated that the approach is feasible and effective for template extraction for QA.

"Searching the Future" by Ricardo Baeza-Yates

In this paper the author presents a new retrieval problem: *future retrieval*. The idea is to use news information to obtain future possible events and then search events related to our current (or future) information needs. In other words, the approach includes time as a formal attribute for a document. The author presents a simple ranking model based on time segments, a prototype for it and some special examples. This work also poses new challenges for natural language processing, information extraction, and answer evaluation.

3 Discussion, conclusion

Due to the a low number of papers, the closing discussion allowed detailed interaction between participants.

The MF/IR 2005 organisers were Sándor Dominich (University of Veszprem, Hungary), and Iadh Ounis (University of Glasgow, Scotland, U.K.), who, on behalf of MF/IR 2005 would like to thank the program committee for their help and time, and all authors and participants for writing and presenting papers as well as attending this event. They also would like to thank Jian-Yun Nie for the keynote talk, and ACM SIGIR for making this event possible.