The sixth Dutch-Belgian Information Retrieval workshop (DIR 2006)

Wessel Kraaij, TNO ICT
Franciska de Jong, TNO ICT and University of Twente
The Netherlands

1 Introduction

The 6th issue of the Dutch-Belgian Information Retrieval workshop took place at March 13 and 14, 2006 and was hosted by TNO Information and Communication Technology in Delft, The Netherlands. The primary aim of the DIR workshops is to provide an international meeting place where researchers from the domain of information retrieval and related disciplines, can exchange information and present new research developments. This year, there was a special focus on contributions focusing on domain-specific retrieval tasks.

The workshop is organised under the auspices of the Dutch Working Community on Information Sciences (WGI). TNO ICT is the organising institute and additional support comes from the Human Media Interaction group (HMI University of Twente), the Dutch Research School for Information and Knowledge Systems (SIKS), Netherlands Bioinformatics Centre (NBIC), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO Exacte Wetenschappen), and the Taalunie-programme STEVIN.

2 Presentations

The first keynote by ChengXiang Zhai (UIUC) focused on domain specific IR, in particular bioinformatics. His presentation started out with a tutorial on some of the problems in bioinformatics and demonstrated the relevance of IR techniques. He concluded his presentation with several challenges for IR inspired by bioinformatics.

Leif Azzopardi unfortunately could not make it to the workshop, but submitted his presentation about query intention acquisition, evaluated using data from the enterprise track. Several LM based retrieval models were tested, some of which were less robust than others.

Vojkan Mihaljovic (University of Twente) presented work on vague element selection in the context of the INEX. The extensions to the TIJAH framework

resulted in a higher recall for vague element name selection and a higher precision for rewriting techniques.

Kees van der Meer (Delft University) discussed some solutions to practical problems related to construct an XML based data container including metadata for long-term storage of hydrology datasets with inherent authenticity guarantee. The presentation raised an interesting discussion on the scope and usage scenarios of Dublin Core versus metadata element schemes such as OAIS.

Willem van Hage announced the OAEI an initiatve aiming at a controlled evaluation of the alignment of ontologies. The workshop will be run for the third time this year `http://oaei.ontologymatching.org/2006/`.

Anni Jarvelin presented research from the University of Tampere on Cross-language Information Retrieval between closely related languages. Experiments showed that for related language pairs, translation dictionaries might not be required, since n-gram term matching can yield quite acceptable performance. Skipgrams turned out to be the most effective.

Fotis Lazarinis submitted work about information extraction techiques for Greek texts, but did not attend the workshop.

Gijs Geleijnse presented a pattern based approach to information extraction. A system has been designed and evaluated to query a web search engine and fill templates in an ontology, e.g. attributes of a movie (title, director, actors). Also an algorithm to determine the most famous people was presented. Most of them had German nationality, which gave plenty of room for discussion on the quality of the algorithm.

The second workshop day started with a keynote lecture of Maarten de Rijke entitled "facing restrictions in QA". Maarten first discussed the XIRAF architecture, which extends XQuery with stand-off XML. Xiraf enables structured queries that leverage multiple annotation levels. StandOff axes. As a second example of restrictions, he presented specialized presentation methods such as WikiTimeLine, where time information helps to structure different facts `www.wikitimeline.net`. Finally Maarten presented an idea to use the highly structured wikipedia biography pages as a skeleton for generating biographical summaries, by means of sentence voting.

Toine Bogers discussed expert search.in workgroups. His method computes the informativeness of a term for an author. Subsequently a baseline cosine system with an author based authority score. The method had significant improvement on CACM, but not on two other small collections.

Claudia Hauff discussed scale free networks to describe citation networks. Her Hypothesis was: importance indicator and reference graph help users. A user study showed a positive tendency, although not unanimous.

Vera Hollink presented an experiment comparing different methods for search result presentation. Three methods were compared: list vs, hierarchical structure vs. Information Gain. Conclusion is that IG achieves the best results because it balances "using information" vs. "collecting information". Label quality is very important, since poor labels will induce backtracking search behaviour.

Jaap Kamps (University of Amsterdam) presented work on Wikipedia. He first advertised its use as a research corpus: "It's large, free and FUN!" , general, highly structured and densely linked. Secondly he presented an experiment where a baseline (page based) search system was compares with a focused system returning section level snippets. The evaluation did not provide a clear answers whether users like focused access than the baseline system. But time to solve task was decreased.

3 Conclusions

The DIR 2006 workshop was a success, with some 40 attendees including many PhD students and several information professionals and even representatives form industry. The full proceedings of the workshop including the presentation slides can be found on: `http://hmi.ewi.utwente.nl/dir2006/program.php`. In 2007, DIR will be hosted by the Katholieke Universiteit Leuven in Belgium.