The 4th Dutch-Belgium Information Retrieval Workshop

Arjen de Vries
CWI, The Netherlands.
Arjen.de.Vries@cwi.nl

December 8th and 9th, 2003, the Centrum voor Wiskunde en Informatica (CWI) in Amsterdam hosted the 4th Dutch-Belgian Information Retrieval workshop (DIR 2003). The primary aim of the Dutch-Belgium Information Retrieval workshops is to provide a meeting place where researchers from the Netherlands and Belgium (and neighbouring countries) exchange information and present new research developments in the domain of information retrieval and related disciplines. The workshops are student-oriented events, where an important goal is to give PhD students a chance to present their work in an informal setting, to gain experience and prepare for a future performance at larger conferences. Another goal is to share results obtained on benchmark initiatives TREC, INEX and CLEF with those who did not get to attend those workshops. The Programme Committee consisted of Anne Diekema (Syracuse University), Theo Huibers (University of Twente and KPMG), Jaap Kamps (University of Amsterdam), Arjen de Vries (CWI), Erik Tjong Kim Sang (University of Antwerp), and, Maarten de Rijke (University of Amsterdam).

The DIR workshop series started in 2000, as an event collocated with Ruud van der Pol's PhD defense, where Karl Järvelin (one of the committee) has been so nice to give a keynote talk. This excellent idea of using a defense as a source of speakers was then followed in 2001 by Arjen de Vries and Djoerd Hiemstra, when the 2nd DIR (in Twente) was opened by Steve Robertson, who visited for Djoerd's PhD defense. The Belgium part of the conference started in 2002, when Marie-Francine Moens, Djoerd Hiemstra, and Wessel Kraaij organised the 3rd DIR in Leuven, with Karen Spark Jones as (video) keynote. For the first time in the short history of DIR, the extended abstracts were reviewed by a group of well-known, international information retrieval researchers.

The 4th DIR workshop, co-organised by Arjen de Vries, Maarten de Rijke and Jaap Kamps, created another innovation, extending the workshop to two days. We believe this has been a good decision, that allowed for ample discussion. In spite of a longer event, we managed to keep the cost of attendance in student budget (as low as EUR 20, including lunch!), thanks to the Centrum voor Wiskunde en Informatica (CWI) for providing the venue, and generous sponsorship by the IMIX research programme of the Netherlands Organisation for Scientific Research (NWO), the Dutch Research School for Information and Knowledge Systems (SIKS), Cosinus Computing BV, and, the Institute for Logic, Language and Computation (ILLC).

1 Keynotes

The night before DIR started, the Royal Family was extended happily with a new Princess (who will one day be Queen). So, the workshop started with coffee and 'beschuit met muisjes', a Dutch delicacy traditionally handed out when a baby is born.

Thanks to ILLC who flew in the opening speaker, and the (proven successful) strategy of co-locating DIR in time and place with Christof Monz's PhD defense (inviting PhD defense committee members to present their own research at the workshop), we could invite three excellent keynotes: Aleksander Øhrn from FAST

Search & Transfer, Charlie Clarke from University of Waterloo, CA, and Bonnie Webber from University of Edinburgh.

Aleksander Øhrn gave an overview of the many aspects involved in running a search engine business. FAST's view on the retrieval of both structured and unstructured information was very relevant to those with a research interest in the integration of IR and databases. A most impressive range of search strategies has been implemented in the FAST software, combining information retrieval, information extraction, as well as data mining techniques. For the CWI hosting organisation it was nice to hear that a significant part of their software architecture runs on Python, the scripting language developed at CWI. Finally, although AllTheWeb is no longer owned by FAST, Aleksander had been tricked by a very smart audience member to answer a question whether one should use AllTheWeb ('better for specific information needs') or Google ('better for generic ones'). Aleksander recommends PhD students to work on information extraction as enabling technology for information retrieval in an enterprise setting, or distributed systems and fault tolerance to improve software architectures for search.

Charlie Clarke started his keynote by reassuring us that we were not completely insane to bid on SIGIR 2007 in Amsterdam - though he added that it does take some months to recover after the event of organising the conference... He presented preliminary (and secret...) results of a summer workshop sponsored by ARDA NRRC, where a number of groups got together and studied *experimentally* a variety of pseudo-relevance feedback techniques. The idea of getting together for some weeks over the summer seems very productive, and rumours go that some Dutch groups might follow this example - at least interest has been raised! Main lessons from the many experiments were that pseudo-relevance feedback still involves quite some black magic causing big differences across the systems. The most important cause for failing feedback could be identified as the case when systems emphasise the wrong aspect of a query. Another lesson from these experiments is that 'feedback over the WWW' should be routine practise if you want to perform well on TREC-style ad-hoc search tasks.

Bonnie Webber closed the workshop with her keynote, presenting the Question Answering research taking place at University of Edinburgh. She first discussed the freely available 'reading comprehension Q&A' corpus (available through Lisa Ferro of MITRE), which was developed in her research group. Next, she discussed recent research results by her students Dalmas and Leidner. The common denominator in these works is that more use is made of global context: from spatial information in the text, or from the context that relates the answers retrieved in the candidate set. In the near future, she expects to make some progress on the issue of multi-sentence support for identifying the right answers in Q&A tasks.

2 Programme

Twelve high quality papers were submitted to the workshop, each submission reviewed by two of the Programme Committee members. The resulting programme covers many topics in the field of information retrieval, varying from web search to multilingual information retrieval, and from multimedia retrieval to question answering. The papers reflect the different views in the field, presenting purely statistical approaches as well as advanced natural language processing.

The first regular paper was a last-minute improvisation, because the Océ talk was cancelled (due to the flue). Thijs Westerveld was so nice to stand in with a presentation on recent CWI experiments, applying language modelling techniques to content-based video retrieval for the TRECVID search task. Patrick Jeuniaux presented his (NLP oriented) research on improving the accuracy of co-reference resolution. Patrick Watrin followed with a talk on information extraction using lexicon-grammars. Gilad Mishne gave an excellent introduction to the Q&A research at ILLC into Dutch Q&A, one of the tasks at CLEF 2004. He discussed some of the challenges for Dutch Q&A, such as the lack of resources as well as longer sentence spans.

The afternoon opened with a special session on the IMIX research programme, launching their second call – providing an excellent opportunity to extend the Dutch-Belgian IR community with new members. The projects funded in the first call presented their key research questions as well as the first outline of a joint

IMIX demonstrator, a question answering system with multi-modal input and output.

Wessel Kraaij had an early start on Tuesday, presenting a language modelling approach to cross-lingual information retrieval (CLIR). He concluded that transitive generative probabilistic models using dictionaries learnt from the WWW are a viable approach to CLIR. Christof Monz focused on the question whether ad-hoc retrieval for non-English is different from IR on English, concentrating on morphologically rich languages like Dutch. Kees Koster concluded the cross-language session with a study into Spanish-English text categorisation on the ILO collection, a patent database.

Roeland Ordelmans presented the research performed for his (recently concluded) PhD thesis on Dutch SDR, giving an insight in the difficulties of developing speech recognition tools for languages other than English. Floris Wiesman, the organiser of the first DIR workshop, gave an overview of the I^2RP architecture, where information retrieval, question answering, and hypermedia presentation generation are brought together in one system. Kate Byrne then demonstrated the (limitations of) content-based image retrieval techniques in building a retrieval system for the National Monuments Record of Scotland.

Vojkan Mihajlovic bravely faced the challenge to present his research into an XML IR retrieval model based on region algebras, with Charlie Clarke in the audience; succeeding wonderfully! Börkur Sigurbjörnsson explained in the final regular presentation how language modelling approach to information retrieval can be applied to INEX, revealing the rationale behind the impressive ILLC results on INEX.

The (free) online proceedings are available from the DIR 2003 website, `http://lit.science.uva.nl/DIR/`.

3 Future

The next DIR has not been scheduled as yet, but there is a high probability that it will be held early November 2004. The organisation of the series of DIR workshops should be strengthened, to make it less of a chance event whether somebody stands up to actually get his or her act together and organise it. One idea is to seek cooperation with the Werkgemeenschap Informatiewetenschappen (WGI), an association that organises events of interest to the Dutch and Belgian Information Retrieval and Library Science communities. For now, if you are planning a trip to The Netherlands by the end of the year, booking it in the first week of November seems a safe bet if you like to combine your trip with a peek into the kitchen of the Dutch IR community!

Arjen P. de Vries