Report on the SIGIR 2008 Workshop on Focused Retrieval

Jaap Kamps[1] **Shlomo Geva**[2] **Andrew Trotman**[3]

[1]University of Amsterdam, Amsterdam, The Netherlands, *kamps@uva.nl*
[2]Queensland University of Technology, Brisbane, Australia, *s.geva@qut.edu.au*
[3]University of Otago, Dunedin, New Zealand, *andrew@cs.otago.ac.nz*

Abstract

On July 24, 2008 the SIGIR Workshop on Focused Retrieval was held as part of SIGIR in Singapore. Nine paper were presented in three sessions and in a fourth session—joint with the SIGIR 2008 Workshop on Aggregate Search—there was a panel discussion. The report outlines the events of the workshop and summarizes the major outcomes.

1 Introduction

Standard document retrieval finds atomic documents, and leaves it to the end-user to then locate the relevant information inside the document. Focused retrieval removes this latter task from the end-user by providing more direct access to relevant information. That is, focused retrieval addresses *information* retrieval proper, and not simply *document* retrieval. Focused retrieval is becoming increasingly important in all areas of information retrieval. Question Answering has been examined by TREC, CLEF, NTCIR, and TAC for many years, and is arguably the ultimate goal of semantic web research for interrogative information needs. Passage retrieval has an even longer history including INEX and the genomics track at TREC, but is also important when searching long documents of any kind. Element retrieval (XML-IR) has been examined by INEX where it has been used to extract relevant fragments from academic documents, and from the online encyclopedia Wikipedia.

Although on initial inspection these focused retrieval paradigms appear quite different, they share much in common. As with traditional document-centric information retrieval, with focused retrieval the user need is loose, linguistic variations are frequent, and answers are a ranked list of relevant results. Furthermore, in focused retrieval the size of the unit of information retrieved is variable and results within a single document may naturally overlap. Focused Retrieval is an exciting field appealing to researchers from three previously distinct disciplines.

This Workshop was a continuation of the successful Workshop at SIGIR 2007, and addresses theory and methodology of focused retrieval, independent of the evaluation forums specifics. In particular, it provided an opportunity for IR researchers who have been working in different areas of focused retrieval to collaborate and to share ideas. A total of nine papers were selected by the program committee from fourteen submissions (a 64% acceptance rate). The papers covered a range of focused retrieval topics:

- Question answering
- Passage (or sentence/snippet) retrieval
- Link detection
- XML retrieval

We will discuss these papers in four separate sections corresponding the areas of focused retrieval. Then, we summarize the panel discussion, and end by summarizing the major outcomes of the workshop.

2 Question answering

Bilotti et al. [1] introduce a theoretical framework for focused retrieval based on "annotation graphs," rich text representations consisting of information elements, and relations between pairs of elements. These annotation graphs can support focused retrieval tasks, such as Question Answering (QA), in the following way. Structured documents, or documents with generated annotations, can be represented as annotation graphs. Also search requests can be represented as annotation graphs, allowing for expressing arbitrary constraints on types or relations, and for matching these constraints at query-time against the documents.

The framework is foremost a theoretical model, furthering understanding of the issues involved the matching problem of richly annotated documents and queries by allowing a high level of abstraction. For all practical purposes, the annotation graphs are mapped onto the XML element retrieval functionality in the Indri search engine. The main challenge is how to integrate all different aspects of the matching between annotated graphs into a single ranking. Assuming a non-Boolean system, how do we deal with partial matches? How do we combine evidence from multiple partial matches? How do we aggregate over multiple matching structures? Etc. The paper elegantly highlights the deep relations between QA and XML-IR. This also emphasizes both the importance of, and difficulty of, the ranking problem in QA, having to balance partial matches on textual content and various structural types and relations.

Lin and Liu [5] investigate multiple-focus questions in question answering. This is a new question type of the form *What are the populations of the countries in the world?* that can be decomposed two different sub-questions. Namely, a list question *What are the countries in the world?* resulting in a list of answers $country_i$, with an embedded factoid question, for each of the countries in the list, *What is the population of* $country_i$*?* The original question has both countries and populations as foci, a natural form for presenting the answer would be as a table of countries and populations. These questions form a special case of scenario questions.

Inspection of a sample of questions from three Chinese QA sites revealed that more than 8 percent were multiple-focus questions. These real questions exhibit a range of different relations between the foci, here called coordinate, entity-attribute, entity-relationship, and thematic relations, that can be instrumental in query decomposition. The paper identifies an interesting query type beyond the well-known factoid and list questions. Answering multiple-focus questions suggests the use of richly annotated corpora, and also reminds us of the classic QA work on natural language front-ends to databases.

3 Sentence and snippet retrieval

Losada [6] investigates query expansion methods for sentence retrieval (at the workshop, the work was presented by Ronald T. Fernández). Sentence retrieval is important for various other tasks such as question answering, information extraction, and query-biased summarization or snippet generation. The author assumes a two-tier system that first retrieves a document, and secondly retrieves sentences from high ranking documents. Hence, query expansion could be done both during the initial document retrieval (i.e., before sentence retrieval), or after sentence retrieval, or both. Two methods of query expansion are used: standard Rocchio pseudo-relevance feedback, and local context analysis (see their paper for details).

A comprehensive set of experiments is performed on TREC 2002–2004 Novelty track data, leading to a range of conclusions. The most noteworthy finding is that standard blind feedback on the document does perform as good as local context analysis on initially retrieved sentences. Given that sentence retrieval is computationally expensive, and feedback on sentences requires the more expensive local context analysis, this can be a substantial gain in efficiency.

Ratkiewicz and Menczer [7] explore the idea that the document structure of Web pages provides valuable cues about the information it contains. Specifically, by considering the Document Object Model (DOM) of (X)HTML pages and hyperlinks between different pages, a massive graph, called DomGraph, can be constructed. That is, in comparison with a traditional link graph between web pages, each individual page now consists of a tree with hundreds of nodes.

The DomGraph was used in a snippet generation experiment with, as baseline, a standard text-based snippet generation method. The DomGraph-based method, in contrast, selects the highest ranking node in the DOM trees of retrieved documents. Both methods were evaluated in a Web-based user study, asking participants which of the two snippets was preferred. As it turned out, the DomGraph-based snippets were preferred in almost 2/3 of the cases, suggesting that the DOM structure provides additional information that can be exploited for IR.

4 Link detection

Huang et al. [4] present an overview of link discovery in Wikipedia, reporting the setup of the INEX 2007 Link the Wiki track, its results, and the plans for the track at INEX 2008. The scenario is a user creating a new Wikipedia page, and needing link that page to existing Wikipedia pages (outlinks from the new page), and at the same time where appropriate insert links in existing pages linking to the page (inlinks to the new page). The test collection is build from 90 “orphaned” Wikipedia pages (pages with all links to and from these pages removed), using the removed links as ground truth for evaluation. Effective techniques used either page titles or existing link anchors, and relatively straightforward matching gave very good performance.

At INEX 2008, there will be experiments with links to arbitrary entry points in the documents. In addition, it will be possible to provide multiple links per anchor. Both of these features are not supported by current Web browsers. In addition to the “removed” links, there will be assessments of the link candidates. These assessments will also shed further light on the nature of link discovery.

Zhang and Kamps [9] also investigate link discovery and look at repeated links in the same documents. This aspect was ignored in the INEX track (focusing instead on the sets of pages where an orphaned page links to or from) but is relatively frequent in the Wikipedia. In the 90 topics, over 1/3 of all the outlinks are repeated occurrences in the same document.

Based on the Wikipedia's Manual of Style, three indicators of link repetition are investigated. First, document length, assuming that it is more likely to repeat a link in longer documents. Second, anchor candidate distance, assuming that it is more likely to repeat a link at greater distance from the first link occurrence. Third, number of repeated candidate links, assuming that links having greater numbers of anchor candidates in the same document indicate important links for the document's topic, and hence are more likely to be repeated. Experiments show that document length is an indicator of link repetition, but fails to single out which of the links is actually repeated. The other two factors, anchor candidate distance and number of link candidates, can be used to predict when links are repeated.

5 XML Element retrieval

Doucet and Lehtonen [2] discuss the importance of phrases in general and at INEX, provide various analysis, and provide suggestions on how to elicit more phrases in INEX topics. Early INEX collections had abundant phrases occurring in up to 70% of the topics, but in recent years explicitly marked up phrases were infrequent only occurring in less than 10% of the topics. This limits the impact of phrase-based methods, and makes their investigation unattractive. This is witnessed by a diminishing interest in phrases by INEX researchers over the years.

According to an analysis by the authors, in 75% to 90% of the topics "implicit" phrases—sequences of adjacent words that form a multi-word lexical unit are present—are present. Hence, the INEX collections could be a valuable resource for evaluating phrase-base retrieval techniques. Special instructions could be provided to the topic authors, so that a much larger fraction of phrases is explicitly marked up. Alternatively, an alternative query set with implicit phrases marked up explicitly could be constructed by an analysis similar to the one of the paper, so that both query sets can be compared.

Focused retrieval methods may require more expressive queries in order to narrow down the exact bits of relevant text. Phrases, whether from the query or from phrase detection, are a proven technique to create more expressive search requests.

Hiemstra et al. [3] introduce a notion of "soundness" for ranking algorithms for structured queries, such as NEXI or XQuery Full-text, based on two conditions: 1) ranking should reflect both the content and structure constraints, and 2) ranking should be the same for equivalent queries in terms of the answer set of a standard XPath or XQuery rendition of the query.

The soundness of 200 ranking approaches is analysed, by testing a full matrix of 5 score combination methods (add, multiply, max, min, mean); 5 score aggregation methods (sum, product, max, min, mean); 4 retrieval models (language model, non-smoothed language model, normalized log-likelihood ratio, BM25); and 2 ranking semantics (ranked list or Boolean set). As it turns out, almost all ranking approaches fail to satisfy the soundness conditions! Only 4 approaches satisfy the matching semantics (Boolean set), all of them using the language model without smoothing—known to be not a very effective approach to ranked retrieval. Only 13 approaches satisfy the ranking semantics, among which several using the standard language model.

The paper addresses the semantics of structured IR query head-on. In experiments, researchers have often found that a strict interpretation of the query was ineffective and resorted to various “unsound” ways of processing the queries. However, an unclear relation between the query and how it is executed, creates potential problems for users having to formulate effective queries in the first place.

Wu et al. [8] demonstrate the applicability of probabilistic Datalog for book-page retrieval. Probabilistic Datalog is a high-level language, allowing the expression of retrieval strategies as intuitive programs. This gives the flexibility to experiment with various refinements of retrieval strategies without the need to recode the retrieval system, or re-index the collection.

The approach is used to develop a particular model for book-page retrieval that takes the back of the book-index into account. The book index contains (best) entry pages for a range of topics. In the INEX 2007 Book Search track collection, individual book pages and book indexes with individual entries and page references are all marked up. The book index is used by two approaches for “tf-boosting,” which resembles anchor text propagation in Web IR. Here, the term frequency of a word on a page is incremented if it occurs in the book index. In the naive approach, the book index words are added to the destination pages, leading to a mild increase of term frequency. In the voter approach, the book frequency of the term is equally divided over the destination pages, leading to a more radical boost. These book search experiments demonstrate the ease and flexibility of probabilistic Datalog to model and prototype different retrieval strategies.

6 Panel discussion: Zooming in, zooming out

The workshop continued with a panel discussion, jointly organized with the neighboring SIGIR Workshop on Aggregated Search organized by Mounia Lalmas (University of Glasgow) and Vanessa Murdock (Yahoo! Research Barcelona). The panel’s topic was “Beyond document retrieval: zooming in, zooming out.” It is remarkable that, after 50 years of Information Retrieval, the general solution is in fact Document Retrieval, which is about returning whole documents to users. This is in sharp contrast with the neighboring field of databases where results of any granularity, and infinite aggregate results, can be retrieved. Document Retrieval makes an implicit assumption that whole documents are the most appropriate unit of retrieval, but does this assumption hold in all contexts? Would there be value in direct access to relevant information in documents (zooming in)? Or should we provide an overview of relevance in different documents (zooming out)? Or should we do both, that is providing an overview of relevance within documents (zooming in and out)?

The panelists were: Bruce Croft, University of Massachusetts, Amherst; Djoerd Hiemstra, University of Twente; Peter Ingwersen, Royal School of Library and Information Science; Jaap Kamps, University of Amsterdam (Chair); Ray Larson, University of California, Berkeley; and Cecile Paris, CSIRO, Sydney. The panel addressed a range of issues on focused/aggregated search, covering:

Systems Can’t we build systems that zoom in or out? What would be the crucial components of such systems?

Users Are users not willing to work with such systems? What would be the main contexts of use? What would be the barriers of use?

User Interfaces Can't we effectively communicate more complex retrieval results? What sort of user interface would cater for this? How important is interaction?

Research Challenges What are the main research challenges for "zooming in and out"?

Although its difficult to summarize the lively discussion that started, we will note here a few of the main points.

Peter Ingwersen reminded us of complex systems developed in the 1980s, and that systems require complex interfaces and the design must be highly task dependent. For example, the temporal dimension may be very important. The main reason for the abundance of straightforward document retrieval systems is that its easy to build them.

Bruce Croft took a more critical position, being unhappy about the introduction of new terminology where there is no clear new research problem. In real-world scenarios, where there is just a short query and little context, its very difficult to guess the intended result granularity (exact answer? passage? document? combination/overview of different resources?). It is also unclear how to evaluate aggregated search.

Cecile Paris stressed that we need systems that zoom in or out for more complex tasks, and that the exact system depends on the task—there is no one-size-fits-all solution. It is an open research issue how to combine information in a coherent way, a single ranked list is not enough for solving complex tasks. Evaluation should go beyond topical relevance, and could focus on users being able to solve problems effectively.

Ray Larson stated that all the components for such a system are available, but we don't know yet the exact way to put them together. The main issue is how to retrieve results in a meaningful context: when retrieving a part of a scientific article, some contextual information about the article (title/author/abstract) is necessary. This presents a range of challenges, from evaluation to interface design. This may lead to the creation of a portal-like interface for every query.

Djoerd Hiemstra took a system-building perspective, and discussed how some emerging systems allow abstract specification of the retrieval strategy as a high-level program, instead of it being hard-coded. This will allow for enormous flexibility even changing the total system behavior based on user preferences or an analysis of the query.

The discussion touched upon a range of other issues, including a separation between academia and industry for certain types of research (such as Web search), and the writing of funding proposals bringing all the new terminology discussed at the Workshops together.

7 Major outcomes

Some of the results of the contributed papers will have a lasting impact. However, the major outcome of Workshop was to foster ties between researchers working on different tasks, and and to discover a range of hitherto unknown research links. These include the following. The intimate relations between QA and XML-IR, as witnessed by the discussion following [1], and to which the INEX 2008 Question Answering track will contribute. The relations between W3C standards XPath and XQuery FT and XML-IR, as discussed in [3], aiding to our understanding of the relations between the fields of IR and DB in this area. The links between passage retrieval, sentence retrieval, and snippet generation, as highlighted by [6, 7], which are crucial for building effective user interfaces for focused retrieval systems. The close relation between link detection and XML-IR, as discussed by [4]. As well as some more links discussed in the sections above.

The collaboration between the two workshops at the panel discussion has resulted in a special issue on "Focused Retrieval and Result Aggregation" in the Information Retrieval journal published by Springer (with Mounia Lalmas and Andrew Trotman as guest editors). Submissions are due on May 1, 2009.

Acknowledgments

We would like to thank ACM and SIGIR for hosting this Workshop. We would also like to thank the program committee, consisting of James Allan, Charles Clarke, Norbert Fuhr, Gabriella Kazai, Alistair Knott, Mounia Lalmas, Maarten de Rijke, Liu Wenyin, and the three program chairs. Final thanks are due to the paper authors, the panelists, and the participants for a great and lively workshop. Some of the authors contributed to this report prior to submission. The University of Otago is hosting the workshop proceedings which are on-line at `http://www.cs.otago.ac.nz/sigirfocus2008/`.

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

References

[1] M. Bilotti, L. Zhao, J. Callan, and E. Nyberg. Focused retrieval over richly-annotated collections. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 1–8, 2008.

[2] A. Doucet and M. Lehtonen. Let's phrase it: INEX topics need keyphrases. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 9–14, 2008.

[3] D. Hiemstra, S. Klinger, H. Rode, J. Flokstra, and P. Apers. Sound ranking algorithms for XML search. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 15–21, 2008.

[4] D. Huang, A. Trotman, and S. Geva. Experiments and evaluation of link discovery in the wikipedia. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 22–29, 2008.

[5] C.-J. Lin and R.-R. Liu. An analysis of multi-focus questions. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 30–36, 2008.

[6] D. Losada. A study of statistical query expansion strategies for sentence retrieval. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 37–44, 2008.

[7] J. Ratkiewicz and F. Menczer. Text snippets from the domgraph. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 45–50, 2008.

[8] H. Wu, G. Kazai, and T. Roelleke. Modelling anchor text retrieval in book search based on back-of-book index. In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 51–58, 2008.

[9] J. Zhang and J. Kamps. Link density in XML documents: What about repeated links? In *Proceedings of the SIGIR 2008 Workshop on Focused Retrieval*, pages 59–66, 2008.