Report on the SIGIR 2007 Workshop on Focused Retrieval

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Shlomo Geva
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

Jaap Kamps
Archives and Information Studies
University of Amsterdam,
Amsterdam, The Netherlands
kamps@science.uva.nl

Abstract

On the 27th July 2007 the SIGIR 2007 Workshop on Focused Retrieval was held as part of SIGIR in Amsterdam, the Netherlands. Nine papers were presented in three sessions and in a fourth session there was a panel discussion. This 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 tries to remove the onus on the end-user, by providing more direct access to relevant information. That is, focused retrieval is addressing *information* retrieval. Focused retrieval is becoming increasingly important. Question Answering has been examined by TREC, CLEF, and NTCIR for many years, and is arguably the ultimate goal of semantic web research for interrogative information needs. Passage retrieval has an even longer history and is 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 sections from academic documents, the application to text book searching is obvious and such commercial systems already exist.

Although on initial inspection these focused retrieval paradigms appear quite different, they share much in common. As with traditional document-centric information 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.

These issues are unique to focused retrieval, and to date have not been examined as general problems. For example, the metrics used for passage retrieval at TREC and those for XML-IR at INEX were developed independently even though they arguably measure the same thing: an XML element is a passage.

This workshop focused on theory and methodology of focused retrieval, independent of the evaluation forums specifics. This report outlines the events of the workshop on a session by session basis and concludes with the major outcomes.

2 Sessions

The workshop was divided into four sessions, three in which a total of 9 papers were presented, and a fourth which was a panel discussion. Speakers were allotted 30 minutes to present their paper and lead a discussion on it. There were 32 registered participants with a mixed background including regular participants of TREC, and INEX, as well several with a QA background. In some cases discussion went over the time limits. We present the papers here grouped into their original topic-centered sessions.

2.1 Session 1: Passages and Elements

Trotman *et al*. [8] discuss an assessment experiment held during a session of INEX 2006 (the first at-INEX experiment). They generated shallow pools of about 100 documents per topic for assessment in an hour and a half. During the experiment 41 assessors judged between 2 and 154 documents with a mean of 87. When relative performance of runs submitted to the INEX 2006 relevance-in-context task was measured using both the shallow pools and the official pools (of about 500 documents) a Spearman's correlation of 0.97 was seen. When the same was done with just the top 10 runs no such correlation was seen. The conclusion is that shallow pools give a good idea of performance but are not sufficient to accurately measure performance.

They merged their new assessments with those of the INEX double-assessment experiment resulting in between 3 and 5 assessors for a total of 15 topics. They then examined agreement levels and plotted the intersection and union. Their conclusion is that with as few as 8 assessors there would be no consensus on where within a document the relevant content can be found, but it takes 19 assessors before there is no common relevant document.

An analysis of a questionnaire that accompanied the experiment suggests that factors other than content are used to determine relevance, and that about 40% of assessors changed their mind while assessing. The typical relevant passage was one element or smaller and about half the assessors had a preferred result size.

Kamps & Koolen [5] examine search result granularity for focused retrieval. There is ongoing debate over whether XML elements or passages of text are best. Specifically they examine three questions: How long are relevant passages? How well do passages and XML element correlate? Do passage boundaries match XML element boundaries? They extensively analyzed the assessments from 114 topics at INEX 2006 in which assessors identified relevant content within Wikipedia documents using an electronic yellow highlighter.

On average there were 1.6 relevant passages per document. The mean passage length was 1090 characters but the median was 297 characters. There is no clear topic-based influence in passage length and the article length did not appear to influence passage length. When they correlated passages with XML-elements they found that passages typically both started on and ended on (or very close to) an element boundary, but that half the time that was not a single element.

They conclude that relevant passages are typically short (mean a paragraph and median a sentence). Half the passages were a near perfect fit to a single element. Passages start boundaries are typically element boundaries, passage end boundaries are typically close to or on an element boundary. Their result suggests that the performance of passage retrieval systems may be enhanced by using document structures.

Itakura & Clarke [3] believe that passages (or elements derived from passages) are better results than XML elements. They show that elements are passages, but that passages are not elements. To go from passages to elements either additional content must be added (TG+), or some content must be lost (TG-).

They propose ranking every possible passage in a document and converting into elements for comparison to other element retrieval systems. Of those passages, some are XML elements making a direct comparison of element and passage retrieval possible. Using the INEX IEEE collection and 39 INEX 2005 topics they compared two systems using the nxCG metric. They found that element retrieval out performed passage retrieval. Examining the reason why they show that the extension of passages to elements tends to result in an increase in recall but a decrease in precision. Overall this has a negative effect in performance.

In conclusion, the best elements to return are those that best fit the results. Since the results tend to be passages it may be necessary to return multiple consecutive elements to cover a passage.

2.2 Session 2: QA and Snippet Retrieval

Jijkoun & De Rijke [4] discuss Question Answering at TREC & CLEF and providing a brief history of the discipline. QA started with short factoid questions with questions like "when was Mozart born?" but has now progressed to topic driven QA where, given a topic like "John William King convicted of murder", a question like "who was the victim of the murder?" might be seen. At CLEF mono-lingual and cross lingual QA is seen. The experiments are real-time and systems are expected to supply support snippets.

They suggest it is time to re-examine the first-principles of QA. Their aim is to define a user-centric QA task for which evaluation makes sense. This includes examining the exactness of an answer; knowing the Simpson family come from Springfield is a long running joke in The Simpsons (Springfield exists in half the US states). Prior knowledge of the user is needed to know how exact an answer should be. With respect to the answer size, traditionally a very small number (between 1 and 3) had been used. They observe that it is necessary to have background on the user and to understand the topic before such a limit can be give. In answer to the question "list the airports in London", the purpose of the travel (recreational, domestic, European, or international) affects the answer so the list length is dependent of user purpose. The proportions of questions that are procedural, descriptive, explanation, and factoid has been determined from web logs. It is important to match these proportions in evaluation forums. Of course it is also important to match the document collection with the nature of the question. Asking opinion questions of the Wikipedia is inappropriate – a blog corpus is more suited to this kind of question.

There is confusion in the task definition and Jijkoun & De Rijke identify two different tasks: a user-driven task, and an NLP evaluation task. They go on to propose scenarios for the user-driven task and ask that as a community embrace the user and evaluate QA on a user basis.

Takechi *et al*. [7] examine question-type identification from multi-sentence queries in Japanese. This might be used by a dialog system to separate questions from dialog and to determine the best ways to answer them. 2,234 queries were extracted from articles in 21 categories of *Oshiete! goo*. The average query length was 5.7 sentences with a standard deviation of 3.9. Sentences had a mean length

of 73.9 bytes. These questions were then tagged with 10 query types including how-to, yes-no, location, description and so on.

Initially they segment article sentences. Then they group the sentences into chunks and label each with a type. Finally they extract the question chunks. This method identifies those questions that take more than one sentence to ask. Conditional random fields (CRFs) were used to learn how to label sentences of 954 questions from their collection. Features included simple word features, uni-grams and bi-grams. 2-fold cross-validation was used. Accuracy was low suggesting that simple word features cannot be used to accurately classify sentences.

Their failure analysis shows that errors often occur in the boundaries of adjacent questions, and when chunks contained more than one question. In further work they intend to separate segmentation from question type identification. For reasons not well understood the first mention of *Oshiete! goo* was met with enthusiasm by the Australasian participants.

Zotos *et al*. [9] examine snippet selection for the web. They argue that snippets are not presently semantically generated and lack coherence. For them, better snippets would be semantic and cohesive, and they show that this is the case for users too.

The algorithms they use first disambiguates the query meaning by looking for conflicts in WordNet and asking the user to resolve ambiguity. Candidate snippets are then selected based on semantic correlation with the query. Snippets are selected based on their individual coherence and collective expressiveness. Finally they are presented to the user in the search results.

Evaluation with 15 users using two systems (the proposed system and Alicante) showed a 3.5% improvement in the precision of the top-10 results. An accompanying survey suggests that users prefer query-relevant snippets to those of the alternate system.

2.3 Session 3: Evaluating Focused Retrieval Tasks

Ali *et al*. [1] introduce the Structural Relevance (SR) metric for XML retrieval. Specifically they address the question of introducing a measure (that works in the presence of arbitrary overlap) without the need to make any *a priori* assumptions in order to create an ideal recall base. Their metric measures the expectation of encountering a relevant element in the results lists. An element might be encountered in isolation or as part of a larger element and the metric captures this. Structural relevance can be substituted for number-relevant in traditional metrics such as precision and recall.

Using the INEX 2006 Wikipedia collection and 3 thorough runs, a comparison of SR-Precision to nxXCG is given (for the top-10 results). The comparison shows how overlap is naturally included in the metric and the application of the metric to this task.

Pehcevski & Thom [6] formally introduce the new MAgP metric used for the evaluation of the relevance-in-context task at INEX. This task is of particular interest because it is believed to be the most user-grounded task in XML-IR. The search engine must first identify and rank documents relevant to a user's query and then within those documents it must identify relevant XML-elements (or text passages). The use case is relevant-passage highlighting.

Any metric for measuring the performance of relevance-in-context results must reflect the performance of the document rank order, and also how well the highlighted document content matches known relevant passages. Additionally, the metric should be easy to interpret and be an extension to already well understood metrics. The proposed metric computes the precision and recall within a single document and takes the harmonic mean (F). From this the generalized precision (gP) is computed as the sum of F over the relative rank. Average generalized precision is defined as the mean of gP at natural recall points, and the mean over a number of topics is MAgP. Fidelity tests on simulated runs suggest that this measure correctly captures both document rank order and within-document highlighting.

This metric might be used to measure both the performance of element-retrieval systems and passage-retrieval systems. As elements are passages (but not *vice versa*, see [3]) and at INEX assessors are identifying relevant passages, a comparison of the two technologies is possible for the first time.

Huang *et al*. [2] discuss the new link-the-wiki task running at INEX 2007. The purpose of this task is to examine automatic identification of the anchor text of a hyper-text link in a Wikipedia document and the identification of the destination of that link (as a best-entry-point into another document). In 2007 they propose a reduced version of this task – the identification of links between documents.

First a collection of documents is extracted from the Wikipedia. These documents are stripped of all links, and the remainder of the collection is stripped of links to these documents. These orphaned documents form the topics of the task. Participants will submit runs based on these topics and, in 2007, they will be automatically evaluated based on how well a run matches the links extracted from the orphans. In future years pooling and an evaluation tool will be used.

Performance measures for both accuracy and throughput remain unaddressed problems. Although there are metrics for measuring accuracy of best entry points, there are none for measuring the performance of source-to-destination linking. Throughput is also important because potentially hundreds of queries per document might be required. Of particular interest to the authors is the trade-off between cost and precision.

3 Session 4: Panel Discussion

The final session of the workshop was a panel discussion on “What is (not) focused retrieval?” The workshop's call for participation hinted at similarities between retrieval tasks that return selected parts of documents but without giving a very precise definition of focused retrieval. The final slot at the workshop was a perfect time to discuss what focused retrieval is, why it is useful, how to evaluate it, and what the main challenges are. The panel was chaired by Jaap Kamps (Amsterdam), and panelists were: Charlie Clarke (Waterloo); Valentin Jijkoun (Amsterdam); Mounia Lalmas (Queen Mary); Jamie Thom (RMIT); and Arjen de Vries (CWI).

The panel was organized around four questions. The first question was *why*: Are “focused retrieval” methods actually useful to improve information access? For what tasks / domains / collections would it be most helpful? This resulted in a lively discussion, both with the audience as well as among panelists. Although the motivation varied, there was broad consensus on the usefulness of focused retrieval techniques.

The second question was *what*: What are the essential characteristics of "focused retrieval", and are there important differences between the particular focused retrieval tasks? Focused retrieval was almost equivocally defined as sub-document retrieval, providing results tailored to the request at hand. At the same time, there was interesting discussion on differences between focused retrieval as a preprocessing step feeding into a QA-pipeline, and focused retrieval systems operated directly by end-users.

The third question was *evaluation*: Should we try to define a suitable generic task, or rather embrace some of the specific task contexts? On the one hand, there was little support for trying to come up with a single abstract task, but to bring task-specific elements into the evaluation. On the other hand, as Voorhees reminded us, bringing task-specific features into the evaluation may not be gratuitous, as it may affect the reusability of the resulting test collection.

The fourth question was *research challenges*. A wide range of interesting research areas was mentioned – ranging from technical issues, interface design and user studies, to evaluation issues – witnessing the research potential of focused retrieval. The lively discussion was brought to an abrupt end when the local organizer of SIGIR, Maarten de Rijke, entered and declared SIGIR over.

4 Major Outcomes

The outcome of a workshop comes not only from the presented papers, but also from the discussions before, during, and after the workshop. From all of these there were several major outcomes.

Pehcevski & Thom formally introduced the MAgP measure and demonstrated its fidelity. This metric is a step towards the unification of passage retrieval and element retrieval. Itakura & Clarke demonstrated that a passage retrieval system could be used to select elements and demonstrated that elements are passages.

Huang *et al*. formally defined the link-the-wiki task that is likely to require a combination of data-minding, natural language processing, and element retrieval techniques. A simplified version of this task is being run for the first time at INEX 2007, with the expectation of a full task in 2008 and beyond.

Jijkoun & De Rijke proposed a complete re-examination of QA from a user perspective. This could form the basis of an entirely new methodology for question answering.

Perhaps the most important outcome was that there turned out to be more common ground between specific focused retrieval tasks (QA, element retrieval, and passage retrieval) than most of the workshop's participants anticipated. We hope and expect that this will lead to further discussion and collaboration in the future.

5 Acknowledgements

We would like to thank ACM and SIGIR for hosting this workshop. We would also like to thank the program committee, the paper authors and the participants for a great workshop. Some workshop paper authors contributed to this paper prior to submission. The University of Otago is hosting the workshop proceedings which are online (http://www.cs.otago.ac.nz/sigirfocus/).

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).

6 References

[1] Ali, S., Consens, M., & Lalmas, M. (2007). Structural relevance in XML retrieval evaluation. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 1-8.

[2] Huang, W. C., Trotman, A., & Geva, S. (2007). Collaborative knowledge management: Evaluation of automated link discovery in the Wikipedia. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 9-16.

[3] Itakura, K., & Clarke, C. (2007). From passages into elements in XML retrieval. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 17-22.

[4] Jijkoun, V., & De Rijke, M. (2007). The task first, please. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 23-27.

[5] Kamps, J., & Koolen, M. (2007). On the relation between relevant passages and XML document structure. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 28-32.

[6] Pehcevski, J., & Thom, J. A. (2007). Evaluating focused retrieval tasks. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 33-40.

[7] Takechi, M., Tokunaga, T., & Matsumoto, Y. (2007). Chunking-based question type identification for multi-sentence queries. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 41-48.

[8] Trotman, A., Pharo, N., & Jenkinson, D. (2007). Can we at least agree on something? In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 49-56.

[9] Zotos, N., Tzekou, P., Tsatsaronis, G., Kozanidis, L., Stamou, S., & Varlamis, I. (2007). To click or not to click? The role of contextualized and user-centric web snippets. In *Proceedings of the SIGIR 2007 Workshop on Focused Retrieval*, 57-64.