WORKSHOP/CONFERENCE/KEYNOTE REPORT

Workshop on Novel Methodologies for Evaluation in Information Retrieval – Held at ECIR 2008, Glasgow, UK, 30th March, 2008

Mark Sanderson
Department of Information Studies,
University of Sheffield,
Regent Court, 211 Portobello St,
Sheffield, S1 4DP, UK
m.sanderson@shef.ac.uk,
http://dis.shef.ac.uk/mark/

Abstract

Information retrieval is an empirical science; the field cannot move forward unless there are means of evaluating the innovations devised by researchers. However the methodologies conceived in the early years of IR and used in the campaigns of today are starting to show their age and new research is emerging to understand how to overcome the twin challenges of scale and diversity. With such challenges in mind it was decided to hold the first Workshop on Novel Methodologies for Evaluation in Information Retrieval. The workshop was composed of two invited talks as well as long and short papers covering a range of important evaluation methods and tools. The workshop was chaired by Mark Sanderson; with co-organization from Julio Gonzalo, Nicola Ferro and Martin Braschler.

1 Invited talks

The invited talks were from Tetsuya Sakai (NewsWatch) and Martin Braschler (Zurich University of Applied Science). In both talks, the speakers described approaches to evaluation that did not involve the traditional use of test collections. Tetsuya spoke on his experience evaluating search engines working at NewsWatch. The extensive use of query logs was a key part of his talk. Sakai showed the way in which use of such logs allows examination of more complex search behaviors beyond the initial search covered by test collections. In the same vein, Martin Braschler detailed a study of the search facilities on a large number of enterprise web sites. Like Sakai, Braschler choose to look beyond traditional approaches of evaluation by not just examining precision and recall, but other factors such as speed of response and coverage of the search engine of structured data sources held by the enterprise.

2 Refereed papers

Eleven short and long papers were presented at the workshop. The papers are grouped under common themes.

2.1 Beyond generic information seeking and binary relevance

Arguments against the use of binary relevance judgments in test collections are as old as test collections themselves; papers suggesting other forms of evaluation tasks date back to the 1960s. The workshop had a set of papers describing further innovation in this area.

2.1.1 Towards the Evaluation of Literature Based Discovery

Beresi, Baillie and Ruthven described a pilot study examining how best to evaluate the success of literature based discovery (LBD). They examined the relevance criteria used by people engaging in LBD and showed that unlike the binary relevance judgments common to most test collections, the criteria were broader: encompassing notions of the depth, scope and specificity of documents. There was a specific criteria specified by people to find a generic overview document that introduces a new subject to the searcher.

2.1.2 Changing the subject — one way of measuring trust in information

Jussi Karlgren then described preliminary work on how one determines a user's trust in the information they seek. Karlgren described experiments where users were asked explicitly to describe their trust in documents on particular topics. He reported that with this methodology he was able to determine such levels, but questioned how easy it would be to extend this methodology to very large data sets and groups of users.

2.1.3 To separate or not to separate: reflections about current GIR practice

Cardoso and Santos examined how to measure the effectiveness of a geographical search engine as typified in the GeoCLEF evaluations (Gey et al. 2006). They asked which was better to do when evaluating: use a straightforward notion of relevance that is essentially a catch all for both the thematic and spatial qualities of a geographic search engine? Or instead, as the current orthodoxy in this area of IR suggests, study the thematic and spatial aspects of geographic search using two distinct forms of relevance, one for each aspect. Cardoso used experimental evidence to question the orthodoxy. Further work, however, was needed.

2.1.4 Dynamic Focused Retrieval of XML Documents and Its Evaluation

The retrieval of XML data has been studied for a number of years by the INEX evaluation campaign and a wide range of evaluation measures have been proposed: (Kazai, Lalmas & de Vries 2004), (Kazai & Lalmas 2006). Shimizu and Yoshikawa proposed two new measures for a particular form of XML retrieval, dynamic retrieval of focused elements of a document. The measures they proposed were "benefit" and "effort". They described how to calculate the measures and detailed their properties.

2.1.5 How Many Experts?

Demartini proposed a new search task for the expert search track of TREC, suggesting a task to calculate the number of experts known to be skilled in a particular topic and a task to determine "highly expert" people in an organization. Evaluation measures for both tasks were also proposed.

2.1.6 Angle Seeking as a Scenario for Task-Based Evaluation of Information Access Technologies

Barker et al proposed the task of "angle seeking" as an experimentally rich task to examine. The scenario was constructed around the work of journalists who seek to find diverse background information on a subject of current interest. They described their work in evaluating an angle seeking system.

2.2 Changing evaluation campaigns

The effectiveness of aspects of the large scale evaluation campaigns was also addressed in the following paper.

2.2.1 Large-Scale Interactive Evaluation of Multilingual Information Access Systems – the iCLEF Flickr Challenge

While TREC, CLEF and NTCIR can all claim to have a wide range of participants involved in their test collection based activities, it has proved much harder to get broad involvement in more interactive user focused evaluation campaigns. Clough, Gonzalo, et al described their latest effort to get more researchers involved in this important aspect of search evaluation. Their focus of interest was studying interaction in cross language image search. Inspired by (von Ahn & Dabbish 2004) they described a searching game constructed on top of Flickr. A wide range of users would be encouraged to play the game and logs from the interactions with it would be distributed to interested research parties. The paper described their plans for running the challenge in the summer of 2008; the working notes of CLEF 2008 will describe the outcome of their work.

2.3 Alternative approaches to evaluation

Moving beyond particular tasks or evaluation measures, two papers described work that provided an alternative perspective to evaluation.

2.3.1 VisualVectora: An Interactive Visualization Tool for Cumulated Gain based Retrieval Experiments

Järvelin and his team in Tampere have built a number of tools over the years to help experimenters conduct better evaluations. Their latest system "VisualVectora" allows the studying of the cumulative gain family of evaluation measures across a set of search results. By visualizing results, the tool allows researchers to move beyond simple quantitative views of data and to examine a broader range of experimental configurations and to better understand common search behavior across groups of topics.

2.3.2 Document Accessibility: Evaluating the access afforded to a document by the retrieval system

Azzopardi and Vinay introduced their preliminary work on a totally new approach to evaluation: examining how easily a document can be retrieved from a collection. Here notions of relevance are ignored, instead the chances of a document being retrieved in the top N for a wide range of queries is the question that is addressed. The authors showed how this simple analysis revealed differences in the way that alternate ranking algorithms operated. The work presented in the workshop was preliminary, but more results will be expected from this novel approach to evaluation.

2.4 Evaluating beyond web search

Two papers on other aspects of a web search engine evaluation were presented.

2.4.1 A Large Time Aware Web Graph

Boldi, Santini and Vigna presented their work on building a suite of highly compressed snap shots of the link structure of the ".uk" domain. These collections would allow others to study the evolution over time of the link structure. This paper described the crawling strategy for building the collection and the attributes of the collection.

2.4.2 Compressed Collections for Simulated Crawling

In a "sister paper" Orlandi and Vigna presented their work on compressing these collections so

as to enable rapid processing of the collection and facilitate easy distribution of it.

3 Acknowledgements

The workshop was supported by the TrebleCLEF project: a coordinated action funded under ICT-1-4-1 Digital libraries and technology-enhanced learning; grant agreement 215231. More information about the TrebleCLEF can be found at http://www.trebleclef.eu/.

4 References

von Ahn, L. & Dabbish, L., 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press New York, NY, USA, pp. 319-326.

Gey, F. et al., 2006. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *Accessing Multilingual Information Repositories*. Lecture Notes in Computer Science. pp. 908-919.

Kazai, G. & Lalmas, M., 2006. INEX 2005 Evaluation Measures. In *Advances in XML Information Retrieval and Evaluation*. Lecture Notes in Computer Science. pp. 16-29.

Kazai, G., Lalmas, M. & de Vries, A.P., 2004. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 72-79.