EVIA 2008: The Second International Workshop on Evaluating Information Access

Tetsuya Sakai
NewsWatch, Inc.
tetsuyasakai@acm.org

Mark Sanderson
University of Sheffield
m.sanderson@sheffield.ac.uk

Noriko Kando
National Institute of Informatics
Noriko.Kando@nii.ac.jp

Abstract

The Second International Workshop on Evaluating Information Access (EVIA 2008) was held at the National Institute of Informatics, Tokyo, Japan on December 16th, 2008. It was composed of sessions covering various aspects of information access evaluation and featured eleven refereed regular and short papers and two unrefereed "very short" papers.

1 Introduction

Evaluation of Information Retrieval, Question Answering and Text Summarisation systems has been central to Information Access research for decades. As retrieval becomes more pervasive and diverse, the need for effective and efficient evaluation has never been more important. Following the success of Open Submission Sessions at NTCIR-4 (June 2004), NTCIR-5 (December 2005) and the First International Workshop on Evaluating Information Access (May 2007), the Second International Workshop on Evaluating Information Access (EVIA 2008) was held on Day 1 of the NTCIR-7 Workshop Meeting at the National Institute of Informatics, in Tokyo, Japan.

The workshop consisted of oral presentation of eleven refereed papers and two unrefereed "very short" papers. We welcomed two categories of refereed papers – "regular" and "short," and each of the submitted papers was reviewed by at least three members of the EVIA 2008 Program Committee.

We are proud that the contributions came from diverse evaluation communities such as NTCIR, TREC, CLEF, INEX and the EU-based multimedia coordination action CHORUS. Authors came from different countries and regions – Japan, China, Hong Kong, India, Australia, Spain, Sweden, UK, Ireland and USA. The papers were organised into the following sessions: Asian test collections, question answering, evaluation metrics, users, patent search

and multimedia. The online proceedings is available at:
`http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/`

Below, we briefly touch upon the topics covered by the eleven refereed papers only.

2 Refereed Regular Papers

The EVIA 2008 Program Commitee accepted six regular papers.

Álvaro Rodrigo, Anselmo Peñas and Felisa Verdejo reported on their work on *answer validation* – the task of selecting the best answer given a set of different question answering (QA) systems. This evaluation framework is important, especially given that the traditional pipeline approach to QA often leads to error propagation, as Kui-Lam Kwok reports (See Section 3). They use the Answer Validation Excercise data from CLEF 2007.

Sukomal Pal, Mandar Mitra and Arnab Chakraborty reported on their comparative study of existing evaluation metrics for content-oriented XML passage/element retrieval. According to their experiments using the INEX 2007 data, Mean Average Interpolated Precision (MAiP), which is similar to Mean Average Precision (MAP), is more stable than other metrics.

Tetsuya Sakai and Stephen Robertson proposed a family of information retrieval evaluation metrics called Normalised Cumulative Utility (NCU), which subsume Average Precision (AP) and *Q-measure* – a graded-relevance version of AP. They constructed NCU by considering a population of users stopping at different ranks in the search result, and the utility of the result given the stopping point.

Falk Scholer, Andrew Turpin and Mingfang Wu tackled the *relevance threshold mismatch* problem – the fact that the threshold for judging whether a document is relevant or not varies widely from user to user. They employ 40 users to re-assess documents for three topics from TREC, and show that the number of users who are "TREC-like" depends heavily on how the agreement between the TREC judge and the user is quantified.

Erik Graf and Leif Azzopardi reported on their project of building a European patent test collection for prior art search. In order to avoid manual relevance assessments, they considered the use of *inferred* relevance assessments – references automatically extracted from patent texts were treated as relevant documents. This was probably a good approach, as the NTCIR-5 and -6 patent test collections were also built this way, yielding thousands of topics.

Gareth Jones, Cathal Gurrin, Liadh Kelly, Daragh Byrne and Yi Chen described their ongoing project on building personal lifelog collections at Dublin City University. They discussed the challenges of retrieving information from mixed-media lifelogs, and of generating digital narratives from them, by contrasting the tasks with more traditional ones such as IR and summarisation.

3 Refereed Short Papers

The EVIA 2008 Program Commitee accepted five short papers.

Guanglai Gao, Wei Jin, Fei Long and Hongxu Hou reported on their work on building a Traditional Mongolian IR test collection, following the general methodology of TREC. Lemur and Lucene were used for making pools for relevance assessments. Mongolian

is a language that is to the family of asian language test collections, which so far covered have Japanese, Korean, Simplified and Traditional Chinese as well as English.

Kui-Lam Kwok reported on his work on component-by-component evaluation of Chinese monolingual and English-Chinese crosslingual QA systems. He used the factoid QA test collections from NTCIR-5 and NTCIR-6 and showed that the answer selection module can be a bottleneck. A similar component-by-component evaluation methodology has been tried in the NTCIR ACLIA (Advanced Crosslingual Information Access) Task Cluster, which handles non-factoid QA.

Yuka Egusa, Masao Takaku, Hitoshi Terai, Hitomi Saito, Noriko Kando and Makiko Miwa proposed a method for visualising how users examine a search result for the purpose of studying user behaviours. Both eye movement data and clickthrough data were used. Such a technique may help us classify different queries, tasks and users for applying different IR techniques.

Fredric Gey and Ray Larson proposed an alternative evaluation method for the NTCIR-7 patent classification task, which required systems to assign international patent classification (IPC) codes to research papers. While the task officially used Average Precision by treating each IPC code as a document ID, they proposed a more relaxed evaluation method by utilising the hierarchical nature of the IPC code scheme.

Jussi Karlgren talked about the ongoing research project called CHORUS, funded by the European commission for multimedia information access. Some of the questions addressed were: how can we go beyond text, beyond topical relevance, and evaluate multimedia information access which may be interactive and exploratory in nature?

4 Conclusions

According to the feedback we received from the EVIA 2008 attendees, holding a focussed workshop on information access evaluation on Day 1 of NTCIR is very useful – it makes everyone "evaluation aware" before the main NTCIR task sessions start, and provides a common background for discussing how to evaluate each task. We therefore plan to hold the Third EVIA Workshop on Day 1 of next NTCIR (NTCIR-8), which will be held in June 2010.

Acknowledgments

We would like to thank the Program Committee members (See `http://ntcir.nii.ac.jp/index.php/EVIA-2008/`) for their punctual and high-quality reviews, the authors for their high-quality papers, and the attendees for their discussions.