EVIA 2010: The Third International Workshop on Evaluating Information Access

William Webber
University of Melbourne
wew@csse.unimelb.edu.au

Tetsuya Sakai
Microsoft Research, Asia
tesakai@microsoft.com

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

Abstract

The Third International Workshop on Evaluating Information Access (EVIA 2010) was held at the National Institute of Informatics, Tokyo, on June 15th, 2010. Themes were evaluation depth; the management of collaborative experiments; result diversity; and user studies.

1 Introduction

The remarkable effectiveness of today's web search engine springs not from the intelligence of retrieval algorithms alone, but also from the evaluation regimes in which these algorithms are tested and tuned. As retrieval adapts to the permanent revolution in the volume, nature, and provision of information on the web, so too must we renovate and reinvent the ways we evaluate information access.

The sesquiannual EVIA workshop provides a forum for researchers to report and discuss advances and innovations in the evaluation of information access. The term "information access" is chosen in deliberate preference to the more established "information retrieval". We want to reflect the diversity of information seeking, exploring, and processing that takes place on the contemporary web. We also want to prod researchers (and ourselves!) to go beyond conventional methods, and provide new domains of information access with the evaluation tools they need.

This year's workshop consisted of nine peer-reviewed papers and one invited talk. All papers are available online, as are many of the talk slides.[1] We briefly summarize these presentations in Section 2, and then look forward to the next workshop in Section 3.

[1] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/EVIA/

2 Summary of Presentations

The workshop was divided into four sessions: pool depth; CLEF, NTCIR, and TREC collections; diversity; and “we’re only human”, for user studies. The first session was preceded by an invited talk. We summarize the presentations in the following subsections.

2.1 Invited Talk

The invited talk was given by John Nave, Principal Development Manager of the Microsoft Search Technology Center, Japan [3]. John described new Bing features, focusing on result diversity and whole-of-page relevance. He also discussed the peculiarities of Japanese search and web usage. An example is the popularity of photo-tagging in Japan: taking GPS-located photos of places and objects and tagging them with comments. “Any public place in Tokyo,” he observed, “including probably the walls of this lecture theatre, is plastered online with tagged photos, super-imposing a virtual world on the physical one.” John also observed that down-town Tokyo’s three-dimensionality demands that locations be specified not just as latitute and longitude, but also as altitude. Japan has led the world by a decade in the uptake of mobile technology, and the Japanese experience has much to teach us about society’s adoption of and adaption to the mobile internet.

2.2 Evaluation and Pooling Depth

The opening session focused on a traditional topic of collection-based evaluation: the depth of pooling and of system evaluation. The first paper, presented by Sukomal Pal, of the Indian Statistical Institute, describes an adaptive method for deciding pool depth [4]. Whereas the standard method is to set the one, fixed pool depth for every topic, Pal et al. propose instead that the depth of the pool for each topic be separately determined at assessment time, from the topic’s observed density of relevant documents. The candidate pool is assessed in the order of each document’s maximum rank. When the rate at which new relevant documents are discovered in the pool falls below a certain threshold, processing of the pool stops. Thus, assessment effort is weighted towards topics with a greater number of relevant documents. Experiments show that, comparing the adaptive pooling scheme to full pooling, 40% of the assessment effort is required to find 80% of the relevant documents, and that the adaptive pool produces an almost identical system ranking to the full one.

The second paper in the session, presented by William Webber, of the University of Melbourne, investigates the relationship between evaluation depth and metric stability [11]. Previous studies have found normalized discounted cumulative gain (nDCG) to be a more stable and discriminative metric than rank-biased precision (RBP). The finding is interesting, because the two metrics are similar in construction, scoring by the rank-weighted sum over relevancies. The differences are, first, nDCG is normalized, RBP is not; second, RBP is typically evaluated to less depth; and third, the weights of RBP decline smoothly with rank, but nDCG is steep at the top and fat-tailed. The latter leads to the suspicion that nDCG is being overly influenced by system reinforcement beyond pooling depth, and that its stability is spurious. Webber et al., however, find (rather to their disappointment) that spurious reinforcement does not seem to have a major effect; but, surprisingly, neither does normalization. Evaluation depth is the most important ingredient for a stable metric; depth aside, DCG’s weight distribution does appear peculiarly conducive to stability.

2.3 Collaborative Evaluation

The second session examined various aspects of evaluation practice at the TREC, CLEF, and NTCIR efforts. The session began with the presentation, by Nicola Ferro of the University of Padua, of a paper describing the DIRECT tool for managing retrieval evaluation campaigns [1]. DIRECT is an online tool, informed by the DIKW (Data, Information, Knowledge, Wisdom) hierarchy. Collections and experiments are the data; effectiveness measurements present information; descriptive and analytical statistics provide knowledge; and models and theories impart wisdom. The DIRECT system currently supports the first three stages of the evaluation process, from the creation of topics through to the visualization and analysis of experimental results. The system has been used to manage CLEF campaigns since 2005, and it provides access to CLEF results going back to the effort's inception in 2000. Future work on the tool will seek to extend its capacity into the knowledge domain, by supporting community exploration, annotation, and discussion of results.

The second paper of the session, presented by Tetsuya Sakai, of Microsoft Research Asia, proposes the use of assessment-free evaluation as an interim analysis service for NTCIR participants [6]. In collaborative retrieval experiments such as NTCIR, there can be an extended interval between the submission of runs and the release of relevance assessments. During the interval, participating teams have little indication of how well their systems have performed. Sakai and Lin see assessment-free evaluation as filling this gap, providing participants with an immediate prediction of their performance. Numerous methods have been proposed for ranking systems, using their retrieval runs, without performing relevance assessments. Sakai and Lin compare five such methods, finding the only stable techniques to be those based upon the number of systems that return each document. Average Kendall's τ between predicted and assessed system rankings is around 0.73 for the more reliable methods, with NTCIR results being easier to predict than TREC ones.

The final presentation in the session was from Ian Soboroff of NIST [9]. Soboroff presents several recommendations for best-practice in test collection construction and analysis, based upon his extensive experience creating test collections for the TREC effort. Reliable pooling depends upon the diversity of participating runs; manual runs are particularly valuable in locating unique relevant documents. Where relevance assessments must be provided prior to run submission, iterative assessment and relevance feedback can be use to create the assessment set. At TREC, several heuristics are used to balance topics for subject matter and difficulty. The reusability of a pool can be diagnosed by the proportion of documents uniquely retrieved by one system or team, and by the effect on that system or team's score of removing the unique documents from the pool. Another useful diagnostic is the minimum delta test: the minimum score delta required between system pairs to assure that their ordering is not reversed on another set of topics. Finally, the titlestat statistic, namely the proportion of pooled documents containing topic title keywords, detects pools that are saturated with easily retrieved, keyword-rich documents, and which therefore are biased against alternative retrieval approaches.

2.4 Diversity

The third session of the workshop addressed the evaluation of result diversity. Tetsuya Sakai presented the first paper of the session, which proposed new evaluation metrics for diversified search results [7]. Unlike existing measures of diversity, the new metrics score for all query intents, weighted by intent probability, and incorporate graded relevance within each intent. Sakai et al. compare the behaviour of their new and of the existing metrics on the TREC 2009 Web Track Diversity Task data. The new metrics give similar system rankings to the existing ones, while producing more intuitive orderings on some individually analysed topics.

Diversity evaluation is a significant recent development, and there is an urgent need for more test collections to support diversity research. The second paper of the session, presented by Ruihua Song of Microsoft Research Asia, describes the construction of such a collection [10]. Existing collections, such as that of the TREC Web Track, derive a list of possible intents for a query from query log analysis, prior to assessment. Song et al. instead encourage assessors to add to the list of intents, seeded from Wikipedia disambiguation pages, by identifying new intents during document assessment. Furthermore, existing collections treat different intents as equally probable, in part because many existing metrics do not handle intent probability (as noted by Sakai et al. [7]). Song et al. propose two methods for estimating the probability of an intent: from the frequency of clicks on intent-identified documents in a query log; or from the frequency of intent-identified documents in the document corpus. In experiments, the addition of assessor-identified intents almost triples the number of intents found through Wikipedia disambiguation. The correlation between intent probabilities estimated by logs and by queries is surprisingly low; corpus frequency is a poor predictor of query popularity. Moreover, omitting the assessor-identified intents distorts scores both for probability-aware and probability-agnostic diversity metrics.

2.5 User studies

The final workshop session discussed user studies. The first paper of the session, presented by Katrin Lamm of the University of Hildesheim, investigates how user expectation of system performance affects their satisfaction with retrieval results [2]. In particular, the study tested the widely-established confirmation–disconfirmation hypothesis, that lower expectations are more easily satisfied. Human subjects were divided into four groups, by two dimensions. On one dimension, half were told they were using a professional retrieval system, the others that the system was a student project. On the other dimension, half the users were assigned to a system that was in fact of high performance, the other half of low performance, as simulated by artificially constructed result lists. Users performed three search tasks, and then completed a questionnaire on their satisfaction with the system. Although low-expectation users reported higher satisfaction, the difference was not significant, which Lamm et al. ascribe to insufficient manipulation of expectations. On the other hand, difference in performance did lead to a significant difference in satisfaction. Both findings contrast with previous results. In practice, though, users were able largely to compensate for the poorly performing system, locating relevant documents almost as reliably as for its high performing rival.

The final paper of the workshop, presented by Keun Chan Park of the Korea Advanced Instituted of Science and Technology, examines the use of games for relevance evaluation [5]. A well-designed game entices users to perform assessment tasks for fun, in a less artificial environment than that of formal assessment, reducing the cost of user evaluation. The task implemented as a game by Park et al. is that of matching advertisements to news articles. Users play in pairs, and must guess which advertisements the other user will select for a given article. Players are rewarded with scores based on how accurate their predictions are. Experiments indicate that player assessments are close to those of formal assessors, while being made with less effort and more enjoyment.

3 The future

Charlie Clarke, visiting EVIA for the first time this year, remarked, "Why hadn't I heard of this workshop before?", a question implying both praise and blame. We hope that this year's workshop has merited the praise, and our report has helped remedy the blame. We were fortunate to have many leading researchers and industry participants in attendance this year; not only were presentations of a high quality, but the discussion was extended and enlightening.

After this year's event, Tetsuya Sakai is stepping down as co-chair of EVIA, to take on the new responsibility of Evaluation Co-Chair for NTCIR, alongside Hideo Joho of the University of Tsukuba. Tetsuya's co-chairs thank him for the dedication and energy he has brought to making EVIA such a successful workshop.

The fourth EVIA workshop will be held on December 6th, 2011, in Tokyo, Japan. As usual, this coincides with the first day of the NTCIR evaluation forum. Several innovative evaluation tasks are being piloted and run in the current iteration of NTCIR, including a query intent task embedding a one-click, whole-of-page relevance subtask. The design of these tasks is sure to provoke interesting discussion at EVIA next year. We also look forward to the participation at EVIA of other researchers working on the evaluation of information access.

Acknowledgments

We were fortunate to have a highly qualified program committee[2]; we thank them for their contribution to making the workshop a success. We also thank Ian Soboroff for his comments on this report.

References

[1] Maristella Agosti, Giorgio Maria Di Nunzio, Marco Dussin, and Nicola Ferro. 10 years of CLEF data in DIRECT: Where we are and where we can go. In Sakai et al. [8], pages 16–24.

[2] Katrin Lamm, Thomas Mandl, Christa Womser-Hacker, and Werner Greve. The influence of expectation and system performance on user satisfaction with retrieval systems. In Sakai et al. [8], pages 60–68.

[3] John Nave. Microsoft's Bing and user behaviour evaluation. In Sakai et al. [8], pages 1–1.

[4] Sukomal Pal, Mandar Mitra, and Samaresh Maiti. Estimating pool-depth on per query basis. In Sakai et al. [8], pages 2–6.

[5] Keun Chan Park, Jihee Ryu, Kyung-min Kim, and Sung Hyon Myaeng. A game-based evaluation method for subjective tasks involving text content analysis. In Sakai et al. [8], pages 69–76.

[2] http://research.nii.ac.jp/ntcir/ntcir-ws8/EVIA-2010/

[6] Tetsuya Sakai and Chin-Yew Lin. Ranking retrieval systems without relevance assessments: revisited. In Sakai et al. [8], pages 25–33.

[7] Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin-Yew Lin. Simple evaluation metrics for diversified search results. In Sakai et al. [8], pages 42–50.

[8] Tetsuya Sakai, Mark Sanderson, and William Webber, editors. *Proceedings of the Third International Workshop on Evaluating Information Access*, Tokyo, Japan, June 2010. URL `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/EVIA/`.

[9] Ian Soboroff. Test collection diagnosis and treatment. In Sakai et al. [8], pages 34–41.

[10] Ruihua Song, Dongjie Qi, Hua Liu, Tetsuya Sakai, Jian-Yun Nie, Hsiao-Wuen Hon, and Yong Yu. Constructing a test collection with multi-intent queries. In Sakai et al. [8], pages 51–59.

[11] William Webber, Alistair Moffat, and Justin Zobel. The effect of pooling and evaluation depth on metric stability. In Sakai et al. [8], pages 7–15.