Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments

Paul N. Bennett
Microsoft Research
paul.n.bennett@microsoft.com

Ben Carterette
University of Delaware
carteret@cis.udel.edu

Olivier Chapelle
Yahoo! Research
chap@yahoo-inc.com

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

1 Introduction

In most information retrieval or filtering applications, assessor judgments form the cornerstone of system evaluation. These judgments are critical when comparing systems or training ranking algorithms. Other "judgments" such as clicks, relevance feedback or ratings are also used for tuning and selection of ranking algorithms, and more broadly for user modeling, evaluating presentation techniques, *etc.* However, most judgment types (including binary or graded relevance judgments, ratings, and explicit feedback) are for an individual document independent of other documents. The past several years have seen a growing interest in the use of relative judgments or preferences (*Is document A better than document B?*), diversity judgments (*Is a retrieved set of results diverse?*), novelty judgments (*Is this document novel when added to a set?*), and implicit preferences from clicks that require considering multiple items. At the same time, there has been a surge in the development of machine learning methods for "structured learning". These methods are capable of optimizing more complex interactions in both input features and output values as well as directly optimizing for more complex performance measures (*e.g.*, MAP, ROC area, ranking, predicting parse trees, *etc.*).

The SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level judgments, held in conjunction with the 31st Annual International ACM SIGIR Conference, explored research challenges at the intersection of novel measures of relevance, novel learning methods, and core evaluation issues. The goal of the workshop was to examine how the type of response elicited from assessors and users influences the evaluation and analysis of retrieval and filtering applications. For example, research suggests that asking people which of two results they prefer is faster and more reliable than asking them to make absolute judgments about the relevance of each result. Similarly, many researchers are using implicit measures, such as clicks, to evaluate systems. New methods like preference judgments or usage data require learning methods,

evaluation measures, and collection procedures designed for them. This workshop tackled these and other related issues.

2 Workshop Structure

Reflecting the positioning of this workshop at the intersection of several fields, the call for papers resulted in submissions from a wide variety of viewpoints – less than half of which were accepted for publication. All papers were thoroughly reviewed by the program committee and external reviewers. The accepted papers were divided into two sessions: "Relevance and Evaluation" and "Diversity". We also organized two complementary sessions: one on "Position Statements and Early Work" to foster discussion on related research in early stages and another on "Algorithmics" with invited speakers to address relevant learning challenges and approaches. Additionally, the workshop and proceedings included a paper documenting a dataset and system baselines several of the organizing committee members collected to help promote research in this arena.

The workshop was structured to promote discussion by having very limited periods for focused questions immediately after a talk and reserving large time for general discussion within each of the sessions where issues raised in the session could be debated more generally. We felt this generally led to lively debates that focused on the broader challenges more than the specifics of one approach. The workshop drew 34 registered attendees at SIGIR many of whom took a very active role in the discussion periods.

3 Presentations

In this section, we give only highlights of the presentations made during the workshop. More details can be found in the papers and slide decks available from the workshop website: `http://research.microsoft.com/~pauben/bbr-workshop`. The presentations were organized into the following sessions (with the presenter highlighted in bold):

- Relevance and Evaluation
  - *Invited Talk: Empirical justification of the discount function for nDCG*
    Evangelos Kanoulas, **Javed A. Aslam** (Northeastern University)
  - *Learning the Gain Values and Discount Factors of DCG*
    Ke Zhou (Shanghai Jiao-Tong University), Hongyuan Zha (Georgia Tech), Gui-Rong Xue, Yong Yu (Shanghai Jiao-Tong University)
- Diversity
  - *Invited Talk: Promoting Diversity in Search Results*
    David Hawking (Funnelback)
  - *Creating a Test Collection to Evaluate Diversity in Image Retrieval*
    Thomas Arni, Jaiyu Tang, Mark Sanderson, Paul Clough (University of Sheffield)
  - *Nugget-based Framework to Support Graded Relevance Assessments*
    Maheedhar Kolla, Olga Vechtomova, Charles L.A. Clarke (University of Waterloo)

- Position Statements and Early Work
  - *Is Relevance the Right Criterion for Evaluating Interactive Information Retrieval?*
    Nicholas J. Belkin, Ralf Bierig, Michael Cole (Rutgers University)
  - *Beyond Relevance Judgments: Cognitive Shifts and Gratification*
    Amanda Spink, Frances Alvarado-Albertorio, Jia T. Du (Queensland University of Technology)
  - *Set Based Retrieval: The Potemkin Buffet Model*
    Soo-Yeon Hwang, **Paul Kantor**, Michael J. Pazzani (Rutgers University)
- Algorithmics
  - *Invited Talk: Learning Diverse Rankings by Minimizing Abandonment*
    Filip Radlinski (Cornell University)
  - *Invited Talk: Clicks-vs-Judgments and Optimization*
    Nick Craswell (Microsoft Research Cambridge)
  - *A Test Collection of Preference Judgments*
    Ben Carterette (University of Massachusetts Amherst), Paul N. Bennett (Microsoft Research), Olivier Chapelle (Yahoo! Research)

To start off the evaluation session, Javed Aslam presented work that demonstrated many evaluation functions can be rewritten as nDCG with an appropriate choice of gain and discount function. This then begs the question of what gain and discount should be used. The resulting analysis focused on issues of explaining the variance of results as the number of queries grows, and in particular, isolating the difference in systems to be truly due to the systems rather than complicating factors such as query variance.

Immediately following this, Ke Zhou presented work which offered a different view on choosing gain and discount. In their work, a user is presented with two ranked lists and is allowed to choose the one they prefer. Using a collection of these preferences, a learning problem can be formulated where the gains and discounts can be learned to predict the observed preferences. One issue raised in discussion was that the preferences elicited from the user may depend on properties like diversity or the number of duplicates, but the learning problem still decomposes the set independently – thus a potential alternative decomposition may be able to make progress here.

This led nicely to issues of diversity, and in David Hawking's invited talk, he highlighted that there are two broad types of diversification. The first can be thought of as diversification across users. That is, users have their own personal favorites as well as specific intents when they issue a query that change the criteria for relevance. Optimizing the diversity of results for the general population is one way of trying to address this when there is no knowledge of a particular user's intent (*i.e.*, personalization). However, the second type of diversity is quite distinct from personalization and is the diversity that exists within a query even when an intent is fixed. In either case, the diversity may be topical (*e.g.*, "broad coverage of responses" vs. "this portion of the population tends to be interested in computers"), geographical (*e.g.*, "I want a variety of local neighborhoods to be covered in the results when I search for restaurants" vs. "I tend to be interested in a particular neighborhood"), temporal, or along a variety of other dimensions.

Next, Thomas Arni presented one approach to building a diversty test collection using an image collection as a focal point. Here the work identified different aspects and dimensions along which diversity could exist. Throughout discussion, the attendees were very uncertain whether it would be as easy to identify aspects in documents as images.

Following this, Maheedhar Kolla presented work demonstrating how the judgments in a test collection can be augmented with "nugget" annotations for each document. In the study presented in the paper, a "nugget" roughly corresponds to a factoid, and graded relevance judgments are shown to correlate with having more nuggets. However, more interestingly a "nugget" can be thought of as any binary property and thus might be useful in creating datasets for diversity (*e.g.*, each nugget is a subtopic) or similar properties in an incremental fashion.

Next, in the position statement and early work session, Nicholas Belkin and Amanda Spink both presented work calling for a larger focus on measures of user satisfaction and task performance. The focus here was on user task performance, amount learned, general effort, *etc.* and how we can reliably ascertain these quantities. Following this, Paul Kantor presented one approach to how results can be diversified to explore the gain for an arbitrary set in an efficient way.

In the final session, the focus was on algorithmics. Filip Radlinski presented work demonstrating how a current ranking can be improved with low regret bounds by switching presentation order and using clicks. Following this, Nick Craswell talked about using click logs. The first portion of the talk focused both on the failings of relevance judgments where clicks are useful and where click logs fails (*e.g.*, did the user back-out, was the user satisfied, *etc.*). He wrapped up by arguing for viewing clicks as both evidence and truth. That is, clicks can be used as a feature to predict relevance judgments, but we can also try to predict the clicks (clicks as truth). Given both types of predictions, these can be combined into a utility estimate which is then used to perform the ranking.

Finally, the presentations were wrapped up by Ben Carterette who presented work and baseline learning results for a dataset of pairwise preference judgments of relevance. This dataset is now available for research (see pointers in the following section).

4 Resources and Data

Many of the presenters have made their slides available. The links to the presentations are available off of the workshop web page at `http://research.microsoft.com/~pauben/bbr-workshop/index_files/program.htm`. Additionally, an electronic version of the proceedings is also available on the page. An up-to-date list of data sets related to this problem can be found on the workshop web site at `http://research.microsoft.com/~pauben/bbr-workshop/index_files/data.htm`. Of particular note, one of the organizers collected a set of pairwise preferences that can be used in learning to rank experiments. Furthermore, baseline performance and a basic analysis of this dataset was written up by several of the organizers as a paper included in the workshop proceedings.

5 Conclusion

There were a variety of themes highlighted throughout the discussions. One of the primary points of discussion was on click logs. There was a general trend in the discussion that despite their flaws, click logs may be the best form of evaluation we have currently for studying a variety of issues including diversity – since presumably when diverse results are presented users will interact differently.

There was a general call for more formal methodology in the use of click logs and increased awareness or logging of what happens after the click – did the user reformulate the query, back out of the web site and click another link, *etc.* It was fairly broadly accepted that among the caveats are "no click does not imply not relevant" and that a click is, at best, *weak* evidence that the result clicked is more relevant than the results above it.

Also, one interesting discussion point was that for academics who may be limited in their access to click logs, building a deployed academic system (*e.g.* University Library Search) is both a nice potential interactive IR setting and a useful complement to static click logs. In particular, the resulting click logs can be analyzed but dynamic studies can also be performed where the system is changed on the fly or A/B testing is performed by routing users to different instantiations of the system.

In terms of evaluation, current approaches range from a Cranfield style evaluation to small design user studies. The traditional Cranfield-type studies have proven very useful in core relevance but currently are lacking extensions to set-level properties like novelty and diversity. Extending them involves formalizing the utility that matters as well as the cost of getting judgments (*e.g.* if costs must be made per set of results, different queries may have vastly different costs).

In terms of corpus building for diversity and similar types of judgments, some advances have been made that would allow more traditional evaluation such as the use of nuggets to augment the documents in a dataset with any arbitrary property. However, this is still early. Useful work would include extending these approaches or attempting to build corpora now. Corpora that might prove useful immediately could weight queries or different intents for a query differently in an attempt to explore diversification across users (subpopulations).

In discussion of user studies as an evaluation tool, there was significant recognition of the fact that relevance, especially topical relevance, does not fully capture whether the user's need has been satisfied or measure the user's performance on a task. In particular, there is interest in research on what components other than relevance play a role in measuring satisfaction and task performance and how one can reliably measure these quantities even in direct interviews.

Finally, regarding the need for diversity, there was general recognition that diversification can not only be achieved by diversifying a ranked list of results but also by altering how the results are presented. For example, site collapsing, presenting links that jump to various deeper locations within a site, advertising, and related searches can all be seen as ways of diversifying. However, evaluation methodologies are needed that reflect the savings afforded to the user from these different presentation styles.

In this conclusion, we have tried to reflect both the general sentiment of the discussions as well as identify the avenues where work is both needed and desired from the community. As in any discussion, it should not be assumed that these opinions were unanimous. Hopefully, they will

prove useful to the reader in obtaining the flavor of the workshop and aligning the community's vision on important research topics.

6 Acknowledgments

We would like to thank ACM SIGIR for its sponsorship and the SIGIR 2008 committee for its support. We are particularly thankful for the workshop chairs' (Peter Anick and Hwee Tou Ng) feedback and quick responses to questions and for the publication chair's (Ee-Peng Lim) template for workshop proceedings.

Finally we would like to thank the invited speakers, authors, program committee members, external reviewers, and participants for their roles in supporting the workshop and making it a success.