Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation

Shlomo Geva, Jaap Kamps, Carol Peters, Tetsuya Sakai, Andrew Trotman, Ellen Voorhees (editors)

July 23, 2009
Boston, Massachusetts

Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation. Held in Boston, Massachusetts. July 23, 2009.
Published by: IR Publications, Amsterdam. ISBN 978-90-814485-1-2.

Preface

These proceedings contain the invited talks and posters of the SIGIR 2009 Workshop on the Future of IR Evaluation, Boston, Massachusetts 23 July, 2009. The workshop will consist of three main parts:

– First, a set of keynotes by Chris Buckley, Susan Dumais, Georges Dupret, and Stephen Robertson that help frame the problems, and outline potential solutions.
– Second, a poster and discussion session with twenty papers selected by the program committee from 33 submissions (a 60% acceptance rate). Each paper was reviewed by at least two members of the program committee.
– Third, a final panel discussion where the ideas emerging from the workshop will be discussed with the four panelists Charles Clarke, David Evans, Donna Harman, and Dianne Kelly.

When reading this volume it is necessary to keep in mind that these papers represent the ideas and opinions of the authors (who are trying to stimulate debate). It is the combination of these papers and the debate that will make the workshop a success.

We would like to thank ACM and SIGIR for hosting the workshop, and Jay Aslam and James Allan for their outstanding support in the organization. Thanks also go to the program committee, the keynote speakers, the authors of the papers, and all the participants, for without these people there would be no workshop.

July 2009

Shlomo Geva
Jaap Kamps
Carol Peters
Tetsuya Sakai
Andrew Trotman
Ellen Voorhees

Organization

Program Chairs

Shlomo Geva
Jaap Kamps
Carol Peters
Tetsuya Sakai
Andrew Trotman
Ellen Voorhees

Program Committee

Omar Alonso
Chris Buckley
Charles Clarke
Nick Craswell
Susan Dumais
Georges Dupret
Nicola Ferro
Norbert Fuhr
Donna Harman
David Hawking
Gareth Jones
Noriko Kando
Gabriella Kazai
Mounia Lalmas
Stefano Mizzaro
Iadh Ounis
Stephen Robertson
Ian Ruthven
Anastasios Tombros
Stephen Tomlinson
Justin Zobel

Table of Contents

Invited Speakers

Human in the Loop

Social Data and Evaluation

Improving Cranfield

New Domains and Tasks

Towards Good Evaluation of Individual Topics

Chris Buckley
Sabir Research, Inc.

Test collection evaluation in information retrieval has necessarily focused on comparing systems over reasonably large sets of topics and averaging results—there are too many system-topic interactions to rely on just a couple of topics. This dependency on large numbers of topics has allowed us to sweep several non-flattering truths under the rug; the most important one being that our standard test collections really do a poor job at evaluating system performance on individual topics. Some of the reasons and important consequences of this are examined, and suggestions for improving individual topic evaluation are presented.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

Evaluating IR in Situ

Susan Dumais
Microsoft Research, Redmond

Information retrieval has a long and successful tradition of careful evaluation using shared testbeds of documents, queries, relevance assessments, and outcome measures. This paradigm has served us well for improving representations, matching and ranking algorithms, but it has limitations. Evaluations methodologies need to be extended to handle the scale, diversity, and user interaction that characterize information systems today. Previous research on interactive IR has focused on small-scale laboratory experiments. In contrast, Web search engines, e-commerce sites, and digital libraries all benefit tremendously from being able to study large numbers of searchers in situ as they interact with information resources using log data and/or more controlled experiments. Such data provide valuable insights about what users are doing, and how well current search systems are meeting those needs. There are important challenges in collecting and using interaction data (e.g., privacy of individual data, replicability of experiments in the face of changing content and queries, extracting signal from noisy behavioral data), but I believe that we must begin to address these issues and extend our evaluation methods to make continued progress in IR.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

User Models to Compare and Evaluate Web IR Metrics

Georges Dupret
Yahoo! Labs, Silicon Valley

In order to compare or evaluate the performance of two metrics objectively, we need to define a metric on the metrics. Because this new metric also needs to be evaluated, the problem seems to have no solution. In this work, we propose an alternative route: We argue that all Web IR metrics make assumptions on the user behavior, often implicitly. A Metric can then be judged on how realistic its associated assumptions are. The associated user model can also be evaluated against observations collected in clickthrough logs by search engines. If a model predicts better the user behavior on unseen data, then it is arguably more realistic, and the associated metric is superior. In this work we review some common metrics and propose a user model for each of them. We discuss the different assumptions to highlight their strength and weakness. In particular we illustrate these ideas by a discussion on Discounted Cumulated Gain (DCG) and its user model, we show how the discounting factor can be evaluated and suggest ways to improve it.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

Richer Theories, Richer Experiments

Stephen Robertson
Microsoft Research, Cambridge

The Cranfield approach to evaluation and that of its successors, including TREC, is oriented towards system effectiveness. The experimental paradigm is that we have a number of alternative systems, and the research question under investigation is: ‘Which system is best’. If we take seriously the notion that we are engaged in developing a *science* of search, then Cranfield would seem to fit with the idea of a scientific experiment, specifically a laboratory experiment, designed to test out ideas and to help in the development of models or theories. In fact, Cranfield would seem to give us the only notion that we have of a laboratory experiment in search. However, an analysis of the role of empirical knowledge in general and laboratory experiment in particular, in relation to models or theories, reveals some limitations of the Cranfield approach. Despite the huge advances in this experimental paradigm since Cranfield itself, due in large measure to TREC, I believe we are only scratching the surface of what experiments can tell us.

In the scientific approach, we would be looking for models or theories to explain and interpret the phenomena we see around us. In the case of information retrieval, we have some notion of what phenomena are of interest to us: people writing documents; other people (users) needing information in order to solve some problem or accomplish some task; these users undertaking search or information-seeking tasks; and the various mechanisms which might help them do this, by delivering or pointing at documents, or even by answering questions using information extracted from documents. Finally, we have a notion of success or failure, or perhaps degrees of success, in this process. This notion of success or failure we have taken to be central, exactly because we are trying (as engineers) to construct new and better mechanisms with a view to helping the users.

Again, in the scientific approach, we would be looking to the models or theories to tell us things about the phenomena that we did not know or understand before. We can see this as a process of *prediction* – a model might say, in effect, ‘if you do this [which we had not done before], or look at the phenomena in this way [which ditto], then this is what you will observe.’ In the IR case, because of our engineering emphasis on constructing mechanisms which work well, we have seen the function of models as telling us how to make them work better. Typically this is all we ask of a model in IR. We regard this as the only test we need to make of a model, that it gives us good retrieval effectiveness. Thus the function of experiment is (only) to tell us how well we are doing.

This feels like a major limitation. To be sure, the predictions about how to do things well are going to be the main *useful* predictions and applications of such models – although we might also ask if the same models are capable of making other useful predictions. But in any case, testing a model should not be restricted to testing its useful predictions. Less useful or even completely useless predictions may well tell us as much about the model and how to improve it as the useful predictions.

Furthermore, this seems to be one source of the (partial) standoff between the laboratory experimental tradition in IR and the user-oriented, often observational work on information seeking. While the user-oriented world may acknowledge the notions of success and failure (albeit with a somewhat broader notion of these qualities), there are many other aspects of information seeking processes, often orthogonal to the success/failure axis, that are of interest. In particular, user behaviours come to mind. In my view, one way to advance the field of IR would be to seek a much richer range of theories and models, and a correspondingly richer range of experimental and observational studies, with the primary aim of validating, or refuting, or deciding between, the models. I think we are in fact moving in this direction, but slowly.

I believe that what we need now is not so much better systems (though they are always welcome) as better understanding of the phenomena.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

New methods for creating testfiles: Tuning enterprise search with C-TEST

David Hawking,[1] Paul Thomas,[2] Tom Gedeon,[3] Timothy Jones,[3] Tom Rowlands[2]
[1]Funnelback [2]CSIRO [3]Australian National University
david.hawking@acm.org, paul.thomas@csiro.au, tom.gedeon@anu.edu.au, tim.jones@anu.edu.au, tom.rowlands@csiro.au

1. INTRODUCTION AND BACKGROUND

An evolving group of IR researchers based in Canberra, Australia has over the years tackled many IR evaluation issues. We have built and distributed collections for the TREC Web and Enterprise Tracks: VLC, VLC2, WT2g, WT10g, W3C, .GOV, .GOV2, and CERC. We have tackled evaluation problems in a range of scenarios: web search (topic research, topic distillation, homepage finding, named page finding), enterprise search (tuning for commercial purposes, key information resource finding and expertise finding), search for quality health information, automated bibliography generation, distributed information retrieval, personal metasearch and spam nullification.

We have found in-situ, in-context evaluations with real users using a side-by-side comparison tool [3] to be invaluable in A v. B (or even A v. B v. C) comparisons. When a uniform sample of a user population uses an n-panel search comparator instead of their regular search tool, we can be sure that the user needs considered in the evaluation are both real and representative and that judgments are made taking account the real utility of the answer sets. In this paradigm, users evaluate result sets rather than individual results in isolation.

But side-by-side comparisons have their drawbacks: they are inefficient when many systems must be compared and they are impractical for system tuning. Accordingly, we have developed the C-TEST toolkit for search evaluation,[1] based on XML testfile and result file formats designed for tuning and lab experiments. These testfiles can formally specify:

- The relative importance of one query to another.
- The relative utility of one result to another.
- The fact that certain groups of documents are near duplicates of each other.
- Different interpretations of the same query.
- The depth of result set which should be compared for this task.

C-TEST testfiles are potentially applicable in many search settings. Here, we focus on the specific problem of generating realistic testfiles for tuning an enterprise search system. Enterprise search is characterised by:

- Well-defined search engine workloads, which we can represent by sampling submitted queries.
- Great diversity, between organisations, in quantity and characteristics of documents to be searched.
- Financial motivation to tune for high performance. Enterprises sometimes spend large sums of money on enterprise search technology in order to boost productivity and competitiveness.

2. PROPOSED METHODS

We propose using a modified n-panel comparison tool (Figure 1). Assuming modest funding, we imagine supplying participants with large (30″ if practical), portrait-oriented, high-resolution screens. Care will be needed to position such a screen for usability. This is so that judging depth need not be arbitrarily restricted to ten and that many more results can be displayed without the need for scrolling or page-down actions. The use of two results set from two very different search engines is likely to promote a more thorough enumeration of the set of valuable results.

Logging: As in previous experiments with n-panel evaluation, we would log queries submitted, results clicked and judgments made. The testfile will comprise a sample of logged queries.

Utility tagging: Even with two deep result sets generated by different means, the list of correct answers may not be complete. Because searchers are assumed to be engaged in a real task, they are likely to continue to explore by browsing and further searching. We propose to provide them with a tagging interface in their browser toolbar which will enable them to tag an eventually-found document with the query they consider it to match (selected from a drop-down list based on their recent search history). Since users may not be motivated to tag answers in naturally-occurring searches, we could also use an instrumented browser to record their actions and attempt to detect when an information need is satisfied (e.g. at the end of a session).

Eye gaze tracking: In previous studies, we have looked at results users clicked on, and what features of clicking be-

[1] http://es.csiro.au/C-TEST/

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

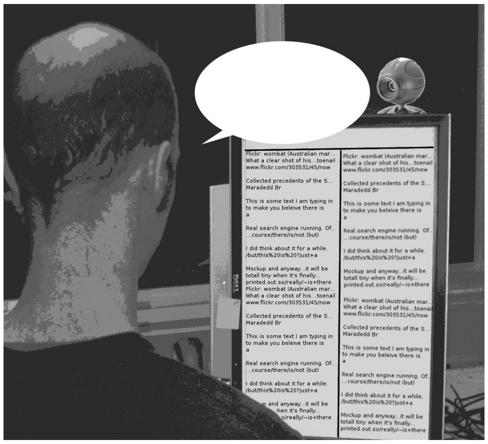


Figure 1: A possible interface for collecting query and judgement information. A two-panel configuration is shown, with logging; gaze tracking and expression recognition via a webcam; and audio feedback via built-in sound equipment.

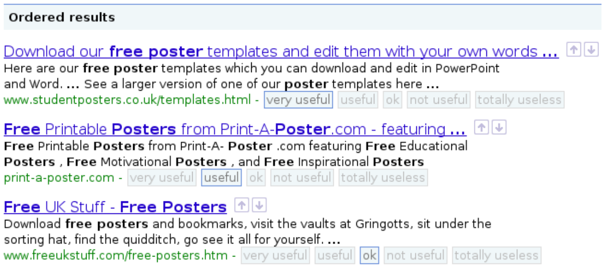
haviour most accurately predict the explicit judgment actually made. We now propose the use of an eye-gaze tracking facility built into the user's computer to observe which results are actually scanned by the user, detect some measure of attention from pupil diameter, and some indication of degree of cognitive processing from dwell times. Eye gaze reflects attention not selection, and needs to be fused with click data to differentiate between attention-getting bad results and results which are actually useful.

As well as indicating attention, knowing which results are scanned would allow us to choose an appropriate depth.

Audio commentary and feedback: We have previously used pop-up windows to elicit feedback ("You searched for 'IP policy' but so far you haven't clicked on any results. Is that because neither system gave you the answer you wanted?"). In the future we propose using speech generation and recording facilities to ask the user to describe what they are looking for (when they submit a query), and to comment on results they have clicked on. This could be used to enumerate interpretations and to assign utility values to results.

Face expression recognition: Human beings are used to expressing a lot of qualitative information about interactions via facial expressions. It is common to make facial expressions at the screen reflecting some judgements of the information provided, for example the match between expectation and result. The same cameras which are used to detect eye gaze could be used to identify facial expressions and gestures such as nodding or shaking the head.

Labelling and ordering documents: We are developing another approach to assigning utility values to query results. This approach asks subjects assign utility labels to documents, and to then rank them within those labels. Figure 2 shows a prototype interface to support this activity (to be demonstrated at the workshop). Obviously, the n-panel comparison tool would not be used in this activity, but labelling and ordering could be done in-situ and in-context, given cooperative subjects.



Ordered results

Download our **free poster** templates and edit them with your own words ... ↑ ↓
Here are our **free poster** templates which you can download and edit in PowerPoint and Word. ... See a larger version of one of our **poster** templates here ...
www.studentposters.co.uk/templates.html - very useful useful ok not useful totally useless

Free Printable **Posters** from Print-A-**Poster**.com - featuring ... ↑ ↓
Free Printable **Posters** from Print-A- **Poster** .com featuring **Free** Educational **Posters** , **Free** Motivational **Posters** , and **Free** Inspirational **Posters**
print-a-poster.com - very useful useful ok not useful totally useless

Free UK Stuff - **Free Posters** ↑ ↓
Download **free posters** and bookmarks, visit the vaults at Gringotts, sit under the sorting hat, find the quidditch, go see it all for yourself. ...
www.freeukstuff.com/free-posters.htm - very useful useful ok not useful totally useless

Figure 2: This prototype interface allows the result list to be arranged by usefulness, by a user clicking the up or down button for each result. The subject can also assign labels to results. Label sets can be used to indicate categories of relevance or to identify duplicates or spam.

3. DISCUSSION AND CONCLUSIONS

Any of these methods is of course subject to the bias inherent in selecting subjects. Those with the time and willingness to cooperate may not be representative of the full searching population. Obviously, we will provide the ability for participants to opt-out for particular queries, but this means that particularly important queries (e.g. 'employee retrenchment provisions') are not included.

Enterprise search testfiles are not likely to be made available for general distribution. Knowing what employees of a company are searching for and what documents they have access to, may be valuable competitive intelligence.

Like Cooper [1] we would like to evaluate search systems on the basis of the utility of the answers they provide. If considered appropriate, both "audio commentary and feedback" and "labelling and ordering documents" could be used to elicit utility values in dollars. Our approach replaces Cooper's human experimenter with much cheaper technological alternatives which are on-duty around the clock and arguably less likely to disrupt normal search behaviour.

Unlike Kelly and Belkin [2] our purpose is much narrower and more specific—we want to build testfiles capable of tuning search systems to maximise actual user satisfaction.

Our proposed method extends previous work in n-panel evaluation, by taking advantage of some newly available or newly affordable technology. It has many features in common with studies in a usability lab, but with the vital difference that the experiment is conducted in the workplace, using naturally occurring search needs and in-context judgments. Unlike logfile analysis, our method avoids the need to attempt to interpret or reverse engineer queries submitted and to deduce utility values from uncertain, incomplete, binary-only click data. As a result, we can obtain a representative sample of real workloads and use it to build a more realistic tuning testfile.

4. REFERENCES

[1] W. S. Cooper. On selecting a measure of retrieval effectiveness. *JASIS*, 24(2):87–100, 1973.

[2] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proc. ACM SIGIR*, pp. 377–384, 2004.

[3] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. CIKM*, pp. 94–101, 2006.

A Model for Evaluation of Interactive Information Retrieval

Nicholas J. Belkin, Michael Cole, and Jingjing Liu
School of Communication & Information
Rutgers University
4 Huntington Street, New Brunswick, NJ 08901, USA
{belkin, m.cole}@rutgers.edu, jingjing@eden.rutgers.edu

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems – *human information processing* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*

General Terms

Measurement, Performance, Experimentation, Human Factors

Keywords

Evaluation, Information seeking, Interaction, Usefulness

1.INTRODUCTION

Research in information retrieval (IR) has expanded to take a broader perspective of the information seeking process to explicitly include users, tasks, and contexts in a dynamic setting rather than treating information search as static or as a sequence of unrelated events. The traditional Cranfield/TREC IR system evaluation paradigm, using document relevance as a criterion, and evaluating single search results, is not appropriate for many circumstances considered in current research. Several alternatives to relevance have been proposed, including utility, and satisfaction. We suggest an evaluation model and methodology grounded in the nature of information seeking and centered on *usefulness*. We believe this model has broad applicability in current IR research.

2.INFORMATION SEEKING

As phenomenological sociologists note, people have their life-plans and their knowledge accumulates during the process of accomplishing their plans (or achieving their goals). When personal knowledge is insufficient to deal with a new experience, or to achieve a particular goal, a *problematic situation* arises for the individual and they seek information to resolve the problem [1]. Simply put, information seeking takes place in the circumstance of having some goal to achieve or task to complete.

We can then think of IR as an information seeking episode consisting of a sequence of interactions between the user and information object(s) [2]. Each interaction has an immediate goal, as well as a goal with respect to accomplishing the overall goal/task. Each interaction can itself be construed as a sequence of specific *information seeking strategies* (ISSs) [3].

We believe appropriate evaluation criteria for IR systems are determined by the system goal. The goal of IR systems is to support users in accomplishing the task/achieving the goal that led them to engage in information seeking. Therefore, IR evaluation should be modeled under the goal of information seeking and should measure a system's performance in fulfilling users' goals through its support of information seeking.

3.GOAL, TASK, SUB-GOAL & ISS

In accomplishing the general work task and achieving the general goal, a person engaged in information seeking goes through a sequence of information interactions (which are sub-tasks), each having its own short term goal that contributes to achieving the general goal. Figure 1 illustrates the relationships between the task/goal, sub-task/goal, information interaction, and an ISS.

Let us give an example. Suppose someone in need of a hybrid car wants to choose several car models as candidates for further inspection at local dealers. The *problematic situation* [1] here is that he lacks knowledge on hybrid cars. His general *work task* is seeking hybrid car information and deciding which models he should look at. He may go through a sequence of steps which have their own *short-term goals*: 1) locating hybrid car information, 2) learning hybrid car information, 3) comparing several car models, and 4) deciding which local dealers to visit. In each *information interaction* that has a short-term goal, he may go through a sequence of *ISSs*. For example, searching for hybrid car information can consist of querying, receiving search results, evaluating search results, and saving some of them.

There are several general comments about Figure 1. First, it shows only the simplest linear relations between the steps along the time line. In fact, the sequence of steps/sub-goals/ISSs could be non-linear. For instance, on the sub-goal level, after learning hybrid car information, the user may go back to an interaction of searching for more information. Another example on the ISS level is, after receiving search results, the user may go back to the querying step.

Second, the contribution of each sub-goal to the general goal may change over time. For instance, suppose in one information interaction, the user looks at information of car model 1 and decides to choose it as a final candidate. After he learns about car model 2, which outperforms car model 1 in all aspects, he removes model 1 from the candidate list. Therefore, some steps in the sequence (choosing car model 1) may contribute to the sub-goal positively, but it contributes to the final and overall goal negatively in that car model 1 is eventually removed.

Third, the leading goal of this task is, or can be taken to be, relatively stable over the course of the interaction. Different users can and will do different things to achieve similar leading goals. Some of differences in these sequences may be characteristics of classes of users, for example, high/low domain knowledge, cognitive capacities, and of task types, including task complexity.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

4.AN EVALUATION MODEL

We suggest IR evaluation should be conducted on three levels. First, it should evaluate the information seeking episode as a whole with respect to the accomplishment of the user's task/goal. Second, it should assess each interaction with respect to its contribution to the accomplishment of the overall goal/task. Third, it should assess each interaction, and each ISS, with respect to its specific goal. In this framework, an ideal system will support the task accomplishment by presenting resources and user support in an optimally-ordered minimum number of interaction steps.

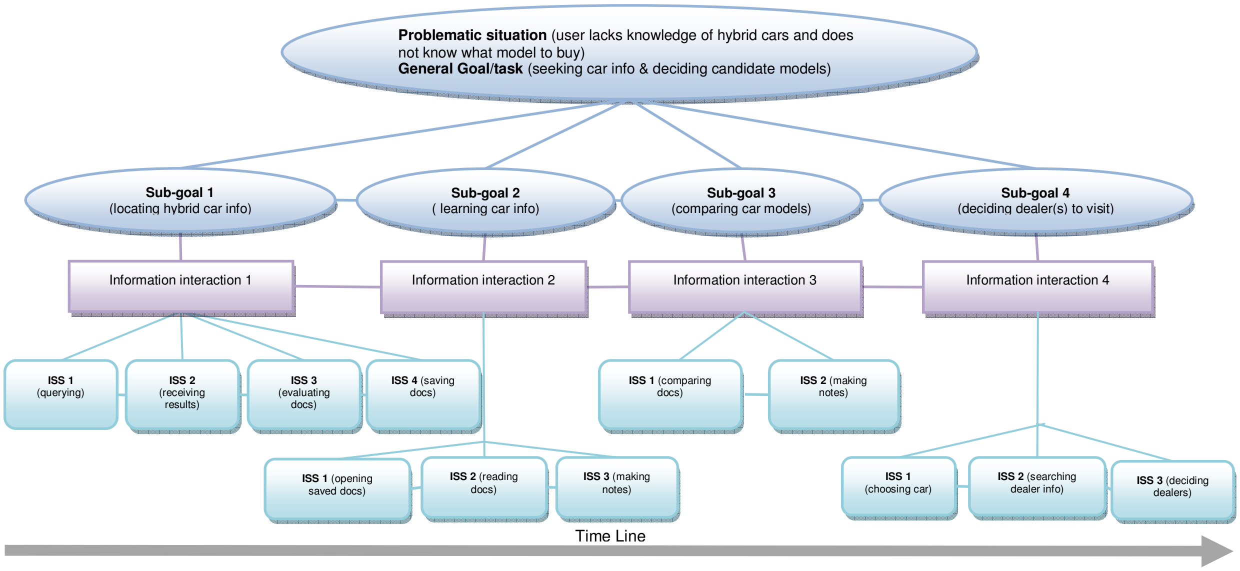
4.1Criterion: Usefulness

We suggest that *usefulness* is an appropriate criterion for IR evaluation. Usefulness should be applied both for the entire episode against the leading (work) task/goal and, independently, for each sub-task/interaction in the episode. Specifically, 1) How useful is the information seeking episode in accomplishing the leading task/goal? 2) How useful is each interaction in helping accomplish the leading task? 3) How well was the goal of the specific interaction accomplished? From the system perspective, evaluation should focus on: 1) How well does the system support the accomplishment of the overall task/goal? 2) How well does the system support the contribution of each interaction towards the achievement of the overall goal? 3) How well does the system support each interaction?

4.2Measurement

Operationalization of the criterion of usefulness will be specific to the user's task/goal, at the level of the IR episode; to the empirical relationship between each interaction and the search outcome, at the level of contribution to the outcome; and to the goals of each interaction/ISS at the third level.

Examples at each level might be: the perceived usefulness of the located documents in helping accomplish the whole task; task accomplishment itself; the extent to which documents seen in an interaction are used in the solution; the degree to which useful documents appear at the top of a results list; and the extent to which suggested query terms are used, and are useful. Identifying specific measures and how to achieve them are clearly difficult problems. However, we believe evaluation of IR systems should be grounded in the nature of the information seeking process that is the *raison d'etre* for these systems. Comments are welcome.


Problematic situation (user lacks knowledge of hybrid cars and does not know what model to buy)
General Goal/task (seeking car info & deciding candidate models)
Sub-goal 1 (locating hybrid car info)
Sub-goal 2 (learning car info)
Sub-goal 3 (comparing car models)
Sub-goal 4 (deciding dealer(s) to visit)
Information interaction 1
Information interaction 2
Information interaction 3
Information interaction 4
ISS 1 (querying)
ISS 2 (receiving results)
ISS 3 (evaluating docs)
ISS 4 (saving docs)
ISS 1 (opening saved docs)
ISS 2 (reading docs)
ISS 3 (making notes)
ISS 1 (comparing docs)
ISS 2 (making notes)
ISS 1 (choosing car)
ISS 2 (searching dealer info)
ISS 3 (deciding dealers)
Time Line

Evaluation based on the following three levels:

1. The usefulness of the entire information seeking episode with respect to accomplishment of the leading task;
2. The usefulness of each interaction with respect to its contribution to the accomplishment of the leading task;
3. The usefulness of system support toward the goal(s) of each interaction, and of each ISS.

Figure 1. An IR Evaluation Model

5.ACKNOWLEDGMENTS

These ideas have benefited from discussion at the 2009 Dagstuhl Seminar on Interactive Information Retrieval, especially the contributions of Pertti Vakkari, Kal Järvelin, and Norbert Fuhr.

This work is supported by IMLS grant LG-06-07-0105-07.

6.REFERENCES

[1] Schutz, A. & Luckmann, T. (1973). *The structures of the life-world.* Evanston, IL: Northwestern University Press.

[2] Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11: 251-265.

[3] Yuan, X.-J. & Belkin, N.J. (2007). Supporting multiple information-seeking strategies in a single system framework. In *SIGIR '08* (pp. 247-254). New York: ACM.

Stakeholders and their respective costs-benefits in IR evaluation

Cécile L. Paris[1], Nathalie F. Colineau[1], and Paul Thomas[2]

CSIRO ICT Centre
[1] PO Box 76, Epping NSW 1710, Australia
[2]GPO Box 664, Canberra ACT 2602, Australia
firstname.lastname@csiro.au

Ross G. Wilkinson

Australian National Data Service
700 Blackburn Rd.
Clayton, VIC 3700, Australia
ross.wilkinson@ands.org.au

ABSTRACT

Evaluations in Information Retrieval are dominated by measures of precision and recall. Is that enough? Probably not, as it somewhat assumes that all information seeking tasks are equal, and that everyone needs the same thing. In this position paper, we advocate a consumers' guide to systems that aim at supporting information seeking tasks. We propose a method that provides guidance in whole-of-system evaluations, explicitly considering all participants and both sides of the "bang for buck" equation.

1. INTRODUCTION

Evaluations of search engines have mainly focused on measuring their accuracy and completeness in returning relevant information, using metrics such as recall and precision. While important, these measures constitute in our view only a partial view of evaluation. First, accuracy and completeness are only one way to measure a system's impact on the user. Then, there are typically a number of stakeholders involved in any system aimed at supporting information seeking tasks, and we believe an evaluation may need to consider the goals of participants besides than the end-user, who is only one of the stakeholders. Finally, we argue that an evaluation should look at costs as well as benefits, for all parties involved.

Information Science also has a long tradition of evaluation: often taking a wider view, looking at a variety of factors such as the system quality (in terms of response time or data accuracy, for example), user satisfaction, individual impact and, interestingly, organisational impact (asking, for example, questions of cost, investment, return on investment, and productivity). Delone and McLean (1992) attempted to consolidate the work on evaluation in this field, and they introduced a comprehensive taxonomy with six major dimensions, placing previous work within that taxonomy (See also http://business.clemson.edu/IES/). In their work, there is a recognition that both benefits and costs have to be taken into account to decide on the success of an information system.

Inspired by Delone and McLean's work and drawing from ourown attempts both to evaluate systems and to choose an appropriate approach for a specific situation, we propose a method that provides guidance in whole-of-system evaluations, explicitly considering all participants and both sides of the "bang for buck" equation. The method we propose is akin to having to write a consumer's guide to a system.

In any consumer report, products are described with a set of attributes and evaluated along a variety of dimensions. These enable consumers to understand, compare and choose, given their own circumstances. A product appropriate for one person might not be appropriate for another. For example, a small two-door car might be appropriate for a single person, but not for a large family. There may also be preferences for some dimensions. For example, someone may put comfort over speed, while another individual will do the reverse. Or, there might be several concerns within the same family, with one member preferring one attribute and another member another feature. Finally, all benefits have to be balanced with costs: while someone might want a sport car, and that is absolutely their preference and desire, they might not want to pay the price it costs and will fall back on something they can afford.

The point here is that there is no such thing as one-size-fits-all, that benefits have to be considered in the context of costs, and that there might be more than one stakeholder to consider. Likewise, we argue that systems that support information seeking tasks must be evaluated along a number of dimensions. This view of evaluation is consistent with ISO 9000, a family of standards for quality management systems—and in particular ISO 9126, developed for software evaluation, which already accounts for attributes of a system such as reliability, usability, efficiency and maintainability. Finally, benefits must be balanced against costs, enabling people to choose what systems best suit their purposes, given their stakeholders.

We believe that one of the compelling attributes of our method is to allow researchers to characterise their system in terms of its strengths and weaknesses, its benefits, costs and impact on all affected stakeholders. Our method provides guidance to think explicitly about the different stakeholders involved in the construction, deployment, maintenance, funding and use of a system.

2. THE METHOD

Typically, a system that supports information seeking tasks involves different actors who have different goals. An evaluation must thus consider all the participants. We have identified four main participant roles:

- The *information seeker*, traditionally the end-user or consumer of the services offered by the system;

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

- The *information provider*, responsible for the content to be searched, explored and delivered;
- The *information intermediaries*. They can be categorised into two groups: resource builders and exploration partners;
- The *system provider*, responsible for the development and maintenance of the technology.

We realise that not all these roles are appropriate in all situations. For example, general search engines might not want to take into account the goals of all the information providers (i.e., anyone wishing to put content on the web). An enterprise search engine might, however, care about the goals of the enterprise. We believe it is important to think explicitly as to who the stakeholders are.

The costs and benefits of a system are likely to differ for each participant. The main benefits for the information seekers are related to the task effectiveness and their satisfaction in using the system. Their costs relate to the time needed to complete the task, the amount of effort required (i.e., the cognitive load) and, potentially, the necessary learning curve.

For the information provider, the benefits concern mostly the audience targeted – to what extent does the information reach a wide or desired audience? The costs here are the costs of providing the information in a form required by the system.

For the information intermediaries, we consider separately the resource builders from the exploration partners. The resource builders are responsible for creating the appropriate set of required resources (e.g., ontologies). Their benefits can be measured in terms of how easy it is to create the required resources, and their costs are related to the time needed to create them, include them in the system and maintain them if required. For the exploration partners, the benefits include those of the information seekers, i.e., related to the task performance and the quality of search and exploration support. Their costs include the time spent in capturing the information relevant to the information seekers' situation.

Finally, the benefits for the system provider are related to the system usage, with its possible corresponding revenue or corporate value, while costs are the cost of system implementation, maintenance and integration with other systems.

This explicit identification of what might constitute a benefit and a cost for whom (see Table 1) can guide researchers and developers in asking appropriate questions about a system and in identifying the relevant evaluation studies to conduct. This in turn helps understand where the technology fits in a larger picture and evaluate different approaches, characterising their strengths and weaknesses, thus allowing one to choose the approach (or system) best suited to one's needs. It also often becomes apparent that providing a benefit to one participant usually comes at a cost (sometimes to another participant). This is the key "bang for buck" equation. This can raise questions such as: to what extent can we trade the benefits of improved user experience with data and system provision costs?

3. CONCLUSIONS

We have briefly presented an evaluation method aiming at guiding researchers in evaluating their web-based information system, looking at benefits *and* costs for *all* participants. Our cost-benefit method provides the means to evaluate different approaches or systems to make an informed decision as to which costs we are willing to pay to obtain which benefits. We believe that our method also enables the framing of research questions that may not be immediately obvious otherwise. The interested reader is referred to Wu et al., (2009) and Paris et al., (2009) for case studies of this method.

4. REFERENCES

DeLone, W. H. and McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. In *Information Systems Research*, 3(1):60-96.

Paris, C., Colineau, N. and Wilkinson, R. Evaluating Web-Based Information Systems from a Cost-Benefit Perspective: A Case Study. CSIRO ICT Centre Technical Report 09/165. 2009.

Wu, M., Thom, J., Turpin, A., and Wilkinson, R., *Cost and Benefit Analysis of Mediated Enterprise Search*, Joint Conference on Digital Libraries, June 15-19, 2009, Austin, Texas.

Table 1. Cost-Benefit Assessment Method: identifying all participants, their benefits and costs

Participant	Information Seeker	Information Provider	Information Intermediaries	System Provider
Benefits	Task effectiveness Knowledge gained Accuracy of exploration Satisfaction	Audience reach Audience accuracy Message accuracy	Resource builders: Ease of knowledge creation & context modelling Exploration partners: Task effectiveness	System usage Reliability Response time Correctness
Costs	Time to complete task Cognitive load Learning time	Metadata provision Structured information Currency of Data	Resource builders: Time to create and integrate the resource Exploration partners: Time to capture contextual factors	Implementation hardware & software cost Syst. maintenance Syst. integration

A Plan for Making Information Retrieval Evaluation Synonymous with Human Performance Prediction

Mark D. Smucker
Department of Management Sciences
University of Waterloo
msmucker@uwaterloo.ca

ABSTRACT

Today human performance on search tasks and information retrieval evaluation metrics are loosely coupled. Instead, information retrieval evaluation should be a direct prediction of human performance rather than a related measurement of ranked list quality. We propose a TREC track or other group effort that will collect a large amount of human usage data on search tasks and then measure participating sites' ability to develop models that predict human performance given the usage data. With models capable of accurate human performance prediction, automated information retrieval evaluation should become an even better tool for driving the future of information retrieval research.

1. INTRODUCTION

In many respects, we believe that the future of information retrieval (IR) evaluation has already been written. In 1973, Cooper [8, 9] wrote a two-part paper outlining what he believed the evaluation of IR should be. In part 1, Cooper presented his "naive evaluation methodology" that held that IR effectiveness should be based on the users' personal utility gained from using an IR system. In part 2, Cooper put forth a possible plan of research that would establish ways to approximate this utility and in particular proposed *validation experiments* to measure the ability of an evaluation method to predict utility. With the rapid changes in computing and the fields of IR and human computer interaction (HCI) it is not too surprising that Cooper's vision was not quickly realized.

In 2009, we see a building momentum for adoption of these ideas but the majority of IR evaluations still focus only on measuring ranking quality with variants of precision and recall that are only loosely predictive of utility [2, 3, 4, 13, 26, 27]. In other words, today's IR researchers tend to evaluate IR systems much as was done prior to Cooper's proposal. In this paper, we renew Cooper's call for the future of IR evaluation and outline a plan to help the IR community move toward evaluation focused on *human performance prediction.*

2. REALIZING COOPER'S VISION

The Cranfield or "batch mode" style of evaluation has been a corner stone of IR progress for over 40 years and serves a complementary role to manual user studies. Cranfield style evaluation takes a ranked list of documents produced by a retrieval system in response to a query and evaluates the list by using a pre-existing set of relevance judgments.

A consistent criticism of the Cranfield style of evaluation is that it does not reflect the wide range of user behavior observed with interactive IR systems.

An important step towards realizing Cooper's vision was taken by Dunlop [11], who in 1997 made a case for the following ideas:

- Evaluation should be *predictive* of user performance.
- Evaluation should concern itself with both the user interface and the underlying retrieval engine.
- Evaluation should measure the time required for users to satisfy their information needs.

Whereas Cooper proposed to measure users' subjective utility, Dunlop examined performance with plots of time vs. number of relevant documents found — a measure inspired by Cooper's expected search length [7]. To make predictions of user performance, Dunlop built user models utilizing HCI methods developed in the decades following Cooper's proposal. Dunlop left as future work the validation of his predictions, i.e. a Cooper validation experiment.

While human performance is not always the same as users' subjective utility, we see Dunlop's ideas in combination with Cooper's validation experiments as the next step towards realizing Cooper's vision and the future of IR evaluation.

3. A BUILDING MOMENTUM

Dunlop's evaluation methodology is still a batch-mode evaluation that relies on a Cranfield style test collection. As Lin and Smucker [20] explain, the Cranfield style of evaluation can be seen as a form of automated usability [14] where the evaluation consists of some hypothetical user interface and a model of user behavior over that interface.

In the case of a Cranfield style evaluation, the hypothetical user interface allows for a query and display of a ranked list of results. The Cranfield style user model assumes the user will examine the results in rank order at a uniform rate and then stop at the end of the ranked list.

Dunlop extended the Cranfield style of evaluation to allow for different user interfaces and to utilize user models that predicted the time to examine the displayed ranked lists.

While not making time-based predictions and utilizing simple user models, several other researchers have also aimed to simulate the use of interactive IR systems [1, 19, 20, 25,

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

28]. Azzopardi [5] provides a useful discussion on the use of examined document sequences for evaluation of interactive IR as utilized by Aalbersberg [1] and others.

At the same time, work has been occurring that has in effect kept the hypothetical user interface fixed to the simple single query, single results paradigm but has aimed to incorporate different user models. Some of this work incorporates a user model into the retrieval metric with the focus on modeling when the user stops examining documents in the ranked list [7, 10, 15, 22].

Another body of work has utilized HCI user modeling techniques (c.f. Dunlop) to IR and IR-related tasks [6, 12, 17, 21, 23, 24]. In many of these cases, the simulations are compared to actual human studies to determine if the user model accurately reflects human performance.

Recently, Keskustalo et al. [18] have taken a significant step forward in simulating human search behavior with an evaluation methodology that examines and simulates query reformulation.

4. OUTLINE OF PLAN

We propose a TREC track or other group effort that defines a canonical search user interface (UI) and collects a large amount of user behavior on TREC-styled ad-hoc search topics. The aim of this effort is to evaluate different evaluation methods on their ability to predict actual human search behavior and performance.

We are only proposing to move IR evaluation in one direction: better prediction of human performance. There are many dimensions to IR evaluation and we do not aim to change the current accepted practices in these other dimensions. For example, we think the task should largely remain a searching of newswire documents, the saving of relevant documents, and the using of an interface that consists of a search box and 10 top ranked results with query-biased snippets.

This effort would in effect create an interaction pool [16] with possibly many participants plugging different retrieval engines into the canonical UI. An attempt would need to be made to collect as much relevant interaction data as possible (queries, clicks, keystrokes, mouse movement, eye tracking, server response times, time documents are saved, etc.).

In summary, we would collect real user data telling us when relevant information is discovered. This data will give us the means to train and test models of human performance prediction — a possible TREC track evaluation of evaluation methods.

5. CONCLUSION

Will it be easy to collect enough user interaction data to make it possible for new evaluation techniques to be created and tested on their ability to predict human search performance? No, but we believe it is preferable to directly predict human performance rather than continue in the current cycle of creating retrieval metrics and then post-hoc testing their predictive ability with expensive user studies.

6. REFERENCES

[1] I. J. Aalbersberg. Incremental relevance feedback. In *SIGIR'92*, pp. 11–22.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR'07*, pp. 773–774.

[3] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: does the test collection predict users' effectiveness? In *SIGIR'08*, pp. 59–66.

[4] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *SIGIR'05*, pp. 433–440.

[5] L. Azzopardi. Towards evaluating the user experience of interactive information access systems. In *SIGIR'07 Web Information-Seeking and Interaction Workshop.*

[6] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *CHI'01*, pp. 490–497.

[7] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Am. Doc.*, 19(1):30–41, Jan 1968.

[8] W. S. Cooper. On selecting a measure of retrieval effectiveness. *JASIS*, 24(2):87–100, Mar/Apr 1973.

[9] W. S. Cooper. On selecting a measure of retrieval effectiveness: Part ii. implementation of the philosophy. *JASIS*, 24(6):413–424, Nov/Dec 1973.

[10] A. P. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO'04*, pp. 463–473.

[11] M. D. Dunlop. Time, relevance and interaction modelling for information retrieval. In *SIGIR'97*, pp. 206–213.

[12] W.-T. Fu and P. Pirolli. Snif-act: a cognitive model of user navigation on the world wide web. *Hum.-Comput. Interact.*, 22(4):355–412, 2007.

[13] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR'00*, pp. 17–24.

[14] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516, 2001.

[15] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002.

[16] H. Joho, R. Villa, and J. Jose. Interaction pool: Towards a user-centered test collection. In *SIGIR'07 Web Information-Seeking and Interaction Workshop.*

[17] M. T. Keane, M. O'Brien, and B. Smyth. Are people biased in their use of search engines? *Commun. ACM*, 51(2):49–52, 2008.

[18] H. Keskustalo, K. Järvelin, T. Sharma, and M. L. Nielsen. Test collection-based IR evaluation needs extension toward sessions: A case of extremely short queries. In *AIRS'09.*

[19] J. Lin. User simulations for evaluating answers to question series. *IPM*, 43(3):717–729, 2007.

[20] J. Lin and M. D. Smucker. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. In *SIGIR'08*, pp. 19–26.

[21] C. S. Miller and R. W. Remington. Modeling information navigation: implications for information architecture. *Hum.-Comput. Interact.*, 19(3):225–271, 2004.

[22] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):1–27, 2008.

[23] V. A. Peck and B. E. John. Browser-soar: a computational model of a highly interactive task. In *CHI'92*, pp. 165–172.

[24] P. Pirolli. *Information Foraging Theory.* 2007.

[25] M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *SIGIR'06*, pp. 461–468.

[26] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR'06*, pp. 11–18.

[27] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR'01*, pp. 225–231.

[28] R. W. White, J. M. Jose, C. J. van Rijsbergen, and I. Ruthven. A simulated study of implicit feedback models. In *ECIR'04*, pp. 311–326.

Queries without Clicks: Successful or Failed Searches?

Sofia Stamou
Computer Engineering and Informatics Department
Patras University, GREECE
stamou@ceid.upatras.gr

Efthimis N. Efthimiadis
Information School, University of Washington
Seattle, WA, USA
efthimis@u.washington.edu

ABSTRACT

The critical aspect in the evaluation of retrieval effectiveness is the satisfaction of the user needs in the retrieved results. Current efforts for evaluating retrieval performance rely either on explicit user feedback or on the analysis of the search transaction logs in order to elicit the user needs and thus be able to infer their satisfaction in the retrieved results. In this paper, we propose a method for evaluating the user satisfaction from searches not followed by clickthrough activity on the retrieved results. To that end, we carried out a user study in order to identify the search intentions of queries without follow-up clicks. Our findings indicate that queries without clicks may pursue specific search goals that can be satisfied in the list of retrieved results the user views rather than in the contents of the documents the user visits for the query.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: search process

General Terms

Performance, Experimentation, Human Factors.

Keywords

Task-oriented search, queries without clickthrough.

1. INTRODUCTION

With the advent of the web and the proliferation of both information sources and information seekers, there has been a shift of interest from the retrieval of query-relevant documents to the retrieval of information that is relevant to the user needs. Automatically identifying the user needs is a challenging task that has mainly focused on the analysis of the user activity on the query results [6] [7]. Although clickthrough data can be perceived as an indicator of implicit user feedback on the relevance of retrieved results [5], it might generate biased relevance judgments unless we consider that users make click decisions based on a limited set of options, i.e. the displayed information on the results page [8].

Recently, researchers proposed that context of search, i.e. the task the user is trying to accomplish, should be the driving force in the quest for effective retrieval evaluation [4] [15]. In this respect, there have been proposed user-centric approaches to the evaluation of retrieval performance [13][14]. The commonality in the above approaches is that they rely on the analysis of the user interaction with the retrieved results for judging their usefulness in satisfying the user search intentions. One aspect that existing IR evaluation techniques do not systematically address is the user's perception of the usefulness of the results retrieved but not visited. Despite the acknowledgment that some queries are not followed by result clicks because the desired information is presented in the snippets (abstracts) of the results [12], to our knowledge no effort has been reported that investigates the contribution of retrieved but un-visited results in relation to users' tasks. In this paper, we investigate the impact that retrieved but not visited results might have on user satisfaction from retrieval effectiveness and examine whether and how these should be accounted in the retrieval evaluation process. First, we present the findings of a survey we carried out in order to identify the context of searches without clicks. In Section 3 we propose a model for evaluating the effectiveness for contextual searches not followed by result clicks.

2. SEARCHES WITHOUT CLICKS

The goals that lead people to engage in information seeking behavior affect their judgments of usefulness of the retrieved results [2]. This, coupled with the observation that nearly 50% of the searches do not result on a single click on the results [4], motivated our study on how to evaluate retrieval effectiveness for queries not followed by result clicks. To that end, we carried out a survey in order to identify the intentions associated with queries not followed by clicks. We recruited 38 postgraduate computer science students and asked them to answer four questions per search performed on their preferred search engine(s) in a single day. The questions, presented to our subjects via an online questionnaire, asked if they did or did not click on results and the reasons for it. Specifically, we instructed our participants to open the questionnaire in a new browser window while conducting their searches and answer the questions for each of their queries right after the submission of the query and the review of the retrieved results. Before conducting our survey we familiarized our subjects with the questions by giving them verbal explanations for every question. The collected user feedback was anonymous in the sense that neither the user identities nor their issued queries or preferred search engines were recorded. Table 1 reports selected results of our survey.

Table 1: Queries without clicks - Survey Results

Examined queries	908
Queries with clicks	87.22%
Queries w/o clicks (intentional-cause)	6.06%
Queries w/o clicks (unintentional-cause)	6.72%
Classification of unintentional queries w/o clicks	
No results retrieved	14.78%
Displayed results seemed irrelevant	62.29%
I have already seen these results for the query	13.11%
Search was interrupted	9.82%
Classification of intentional queries w/o clicks	
Check spelling/syntax of query term(s)	30.91%
See if there's a new page retrieved from the last time I issued the query	32.73%
Find out what the query is about by looking at the retrieved abstracts	21.82%
See if there's a web site about my query	14.54%

The study showed that the reasons for not clicking on the query results fall into two categories: intentional-cause and unintentional-cause. The unintentional cause for not clicking is encountered when the user submits a query, but the retrieved results are unexpected to the user, hence they decide not to click. These reasons (Table 1) are: nothing retrieved, seemed irrelevant, already seen, interrupted search. Conversely, the intentional cause for not clicking is encountered when the user issues a query with a predetermined intention to look for answers in the results' snippets and without following any link. According to our participants, searches without clicks are encountered when they want to accomplish the following types of tasks: (i) get an update or (ii) obtain instant information about the query. In particular, the information goal of users engaging in an *update*[1] search is to find out if there is *new* information retrieved since their last submission of the query. On the other hand, the goal of users performing an *instant* search is to obtain information about the query from the title or the snippets of the displayed results. In both cases, the information need of the user engaging in update or instant searches can be satisfied by the contents of the result list displayed (i.e. the snippets) without the need to follow any results per se. Therefore, retrieval effectiveness for update and instant searches that do not generate clickthrough activity could also be evaluated based on the results displayed to the user. We recognize that this is rather difficult; however a model that attempts this is discussed below.

[1] Update searches as determined by our users could be perceived as an instance of repeat searches [16] since they both concern queries the user has issued in the past.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

3. DISCUSSION

Given the findings of our study, we propose a retrieval evaluation framework for queries without clicks. Our evaluation relies on the observable user activity following a query submission in order to infer the user perception of the displayed results' usefulness. The idea of utilizing the searcher activity on the returned results as an indicator of implicit relevance judgments is not new. There exists a large body of work on how the different post-query activities can be interpreted as implicit feedback signals (for an overview see [11]). The searchers' behavior that researchers observed as implicit measures of interest are: time spent on a page combined with the amount of scrolling on a page [3], duration of search and number of result sets returned [5], click data on and beyond the search results [9], use of eye-tracking methods to capture the user's visual attention on the results [10], repetition of result clicks across user sessions [16]. Although, the above measures have been applied for inferring the user satisfaction from the results visited for some query, we propose their utilization towards capturing the user satisfaction from the results displayed for queries not followed by clickthrough events. From the above measures, we obviously exclude click data since we are dealing with searches not followed by result clicks.

Our proposed model examines the post-query user activity in order to firstly identify the user goals for queries without clicks and then based on the identified goals to infer the user satisfaction from search results. Our model first examines whether a query without clicks returned any results. If the query retrieved no documents, then it concludes that search failed to satisfy the user needs. On the other hand, if the query retrieved results that the user did not visit, our model tries to deduce the user satisfaction from retrieval effectiveness based on the examination of the following features (partially based on the proposal of [1]): (i) time spent on a results page combined with the amount of scrolling on the page (ii) terminological overlap between the query term(s) and the displayed result titles and/or snippets, (iii) terminological overlap between two consecutive queries, (iv) repetition of the query and (v) type of user activity on the displayed results (e.g. read, copy text from snippet, move to the next results page). The idea is that the features characterizing the post query user activity are valuable indicators of the query intentions. Thus, if the intention of the query is to obtain information in the snippets of the displayed results, then evaluation of retrieval performance should concentrate on the usefulness of the result snippets. The features that characterize the intentional cause of queries without clicks and which imply the user satisfaction from the search results can only be determined explicitly via user studies. Next, we discuss a probabilistic approach for capturing the query intention and the user satisfaction from searches not followed by clicks. Our approach relies on the combination of the following measures that are presented below: (a) query refinement probability, (b) query results usefulness, and (c) update search probability.

Query refinement probability, i.e. the probability that a query *q* which did not yield result clicks was refined in the search (q_i) that immediately followed. Formally, $p(q|q_i)$ can be determined proportionally to the number of overlapping terms between *q* and q_i. If $p(q|q_i)$ exceeds a threshold (to be empirically determined via user studies), then *q* was refined in its succeeding search (q_i) and we may conclude that the user did not satisfy her information needs in the results displayed for *q*. If $p(q|q_i)$ is below the threshold, i.e. *q* is not refined in the next search, we examine the following:

Query-results usefulness, i.e. the probability that *q* was not followed by result clicks because it was satisfied in the list of displayed results. To derive such probability, we rely on the terminological overlap between the query term(s) q_t and the terms s_t in the result titles and/or snippet, given by: $O(q, r) = |q_t \cap s_t| / |s_t|$. In addition, we estimate the amount of time the user spent on the results list as well as the type of the demonstrated user activity on the results. The combination of the above measures can serve as an approximation of the displayed results' usefulness to the query intention. Again, threshold weights should be empirically set via user studies before the deployment of our approach to a retrieval evaluation setting. Another factor we should examine is the:

Update search probability, i.e. the probability that the user intention is to obtain new information about a previous search. The probability *p*(*q*) that the query has been submitted before can be determined based on the frequency of *q* in the observable user searches. If *p(q)* exceeds a given threshold, then *q* probably represents an update search. User satisfaction from the results retrieved for an update query can be estimated based on: $p(r_n) \cdot O(q, r_n)$ where $p(\mathrm{r_n})$ is the probability that r_n is a new result not previously retrieved for *q* and $O(q, r_n)$ is the probability that *q* is satisfied in the information displayed for r_n. This probability combined with the amount of time spent on the results and the type of user activity on the results can give rough indications of the user satisfaction from update searches. Again, user studies need to be carried out in order to set the threshold values upon which conclusions can be drawn.

For queries without clicks that are not refined in their succeeding searches and do not represent update requests, as well as for queries without clicks that have low probability of being satisfied in the result snippets or they are the last searches in the user session, the only way to assess user satisfaction from displayed results is in terms of explicit user feedback.

4. CONCLUDING REMARKS

We have proposed the utilization of implicit feedback measures for inferring the user satisfaction from searches not followed by result clicks. The parameters of our approach need to be validated and fine-tuned via additional user studies. We hope that our approach will contribute towards the design of IR evaluation frameworks where search is seen holistically and incorporate multiple features for measuring retrieval quality.

5. REFERENCES

[1] Agichtein, E., Brill, E., Dumais, S. and Rango, R. 2006. Learning user interaction models for predicting search result preferences. In the 29th ACM SIGIR Conference.

[2] Belkin, N. 2008. Some(what) grand challenges for information retrieval. In ACM SIGIR Forum, 42 (1): 47-54.

[3] Claypool, M. Le, P., Waseda, M. and Brown, D. 2001. Implicit interest indicators. In Intl. Conference on Intelligent User Interfaces, pp. 33-40.

[4] Callan, J., Allan, J., Clarke, Ch.L.A., Dumais, S., Evans, D.A., Sanderson, M., and Zhai, Ch. 2007. Meeting of the MINDS: an information retrieval research agenda. In ACM SIGIR Forum, 41(2): 25-34.

[5] Fox, S., Karnawat, K., Mydland, M., Dumais, S. and White, T. 2005. Evaluating implicit measures to improve web search. ACM Transactions on Information Systems, 23(2): 147-168.

[6] Jansen, B.J. and Spink, A. 2006. How are we searching the www: a comparison of nine search engine transaction logs. Information Processing & Management 42(1):248-263.

[7] Jansen, B.J., Booth, D.L. and Spink, A. 2008. Determining the informational, navigational and transactional intent of web queries. Information Processing & Management 44:1251-1266.

[8] Joachims, T., Granka L., Pan, B., Hembrooke H., Padlinski, F. and Gay, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Transactions on Information Systems, 25(2):1-26.

[9] Jung, S., Herlocker, J.L. and Webster, J. 2007. Click data as implicit relevance feedback in web search. Information Processing & Management, 43(3):791-807.

[10] Granka, L.A., Joachims, T. and Gay, G, 2004. Eye-tracking analysis of user behaviour in www results. In ACM SIGIR Conference, pp. 478-479.

[11] Kelly, D. and Teevan, J. 2003. Implicit feedback for inferring user preference: a bibliography. In ACM SIGIR Forum, 37(2):18-28.

[12] Radlinski, F., Kurup, M. and Joachnims, T. 2008. How does clickthrough data reflect retrieval quality. In CIKM Conf.

[13] Sharma, H., and Jansen, B.J. 2005. Automated evaluation of search engine performance via implicit user feedback. In the 28th ACM SIGIR Conference, pp. 649-650.

[14] Spink, A. 2002. A user centered approach to evaluating human interaction with web search engines: an exploratory study. Information Processing & Management, 38(3):401-426.

[15] Taksa, I., Spink, A. and Goldberg, R. 2008. A task-oriented approach to search engine usability studies. Journal of Software, 3(1): 63-73.

[16] Teevan, J., Adar, E., Jones, R. and Potts, M. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In the 30th ACM SIGIR Conference.

Can we get rid of TREC assessors?
Using Mechanical Turk for relevance assessment

Omar Alonso
A9.com
Palo Alto, CA (USA)
oralonso@gmail.com

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
Udine (Italy)
mizzaro@dimi.uniud.it

ABSTRACT

Recently, Amazon Mechanical Turk has gained a lot of attention as a tool for conducting different kinds of relevance evaluations. In this paper we show a series of experiments on TREC data, evaluate the outcome, and discuss the results. Our position, supported by these preliminary experimental results, is that crowdsourcing is a viable alternative for relevance assessment.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software — performance evaluation

General Terms

Measurement, performance, experimentation

Keywords

IR evaluation, relevance, relevance assessment, user study

1. INTRODUCTION AND MOTIVATIONS

One issue in current TREC-like test collection initiatives is the cost related to relevance assessment: assessing requires resources (that cost time and even money) and does not scale up. Indeed, in recent years, there has been some trend on trying to save assessment resources: there is a vast body of literature on reducing the number of documents pooled and/or judged, and, more recently, on reducing the number of assessed topics [4] as well. Also, test collections are sometimes built in-house [3], and assessment effort is obviously a problematic issue when building your own test collection.

Stated briefly, our research question is: "Can we get rid of TREC assessors?" Our position is that crowdsourcing is a reliable alternative to "classical" assessors: in this paper we propose to use the Mechanical Turk crowdsourcing platform for relevance assessing; we also support this approach by some experimental data.

2. RELATED WORK

Amazon Mechanical Turk (MTurk, `www.mturk.com`) is a marketplace for work that requires human intelligence. The individual or organization who has work to be performed is known as the requester. A person who wants to sign up to perform work is described in the system as a turker. The unit of work to be performed is called a HIT (Human Intelligence Task). Each HIT has an associated payment and an allotted completion time. It is possible to control the quality of the work by using qualification tests. MTurk has already been used in some relevance related research [1,2,5], with good success.

E1	Graded relevance on a 4 point scale (3 = excellent, 2 = good, 1 = fair, 0 = not relevant) following closely TREC-7 guidelines. We summarized the main points from the TREC assessment guidelines as starting point.
E2	Graded relevance with modified instructions. Changes on the instructions, use more layman English (not so expert). We also included an input form in the task so turkers can provide feedback.
E3	Graded relevance with modified instructions II. Modified instructions using colors and examples of relevant content. Also included more documents in the test.
E4	Binary relevance without qualification test. Maintained same instructions but changed the answers to binary (1 = relevant and 0 = not relevant). Modified the feedback input to an optional entry for justifying answers. Passing grade was 80% of correct answers.
E5	Binary relevance with qualification test. Same as previous experiment but with a lower passing grade for the qualification test to 60%.

Table 1: The five experiments

Therefore, our research question can be framed as: "Is it possible to replace TREC-like relevance assessors with Mechanical turkers?". We think the answer is "Yes — at least to some extent"; we report in the next sections some preliminary experimental results that support our position.

3. EXPERIMENTS

We used the TREC topic about space program (number 011), in the domain of science and technology. We selected a subset of 29 FBIS documents (the first 14 not relevant, and the first 15 relevant). Each turker was given some instructions, a description of the topic, and one document, and he was asked to judge the relevance of the document to the topic. We decided to have each topic/document pair judged by 10 turkers, thus obtaining 290 judgments in total (when the task was 100% complete).

We performed 5 experiments, as shown in Table 1. We refined the experiments and methodology in each experiment run in an iterative way. By looking at the results data, we

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
.

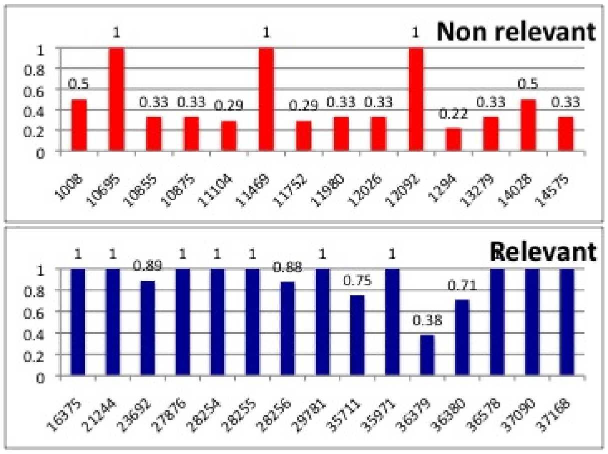


Figure 1: Turkers average relevance assessments

manually inspected the answers, adjusted the methodology accordingly, and tested again. This was done over several weeks as the completion for each experiment was set to 10 days. For each experiment we paid 0.02 cents per task.

The task design in MTurk can be framed as a user interface problem, so in every iteration we tweaked the language, instructions, and presentation. As the results looked closer to our initial hypothesis, we decided to use binary evaluation with qualification test. For this particular topic (space program), we felt that binary evaluation was more suitable given the content of the collection.

We measured the agreement between the turkers and TREC assessors as presented in Figure 1 (that shows the FBIS3 documents on the X axis and the average turkers score on the Y axis, with relevant = 1 and not relevant = 0). For the relevant documents the average across all turkers was 0.91 (relevant expert assessment was 1). In the case of not relevant document, the average was 0.49 (not relevant expert assessment was 0). There are 4 exception where turkers disagree with the experts, for documents: 10695, 11469, 12092, and 36379. We manually inspected the documents and concluded that, in three out of four cases, turkers were correct and TREC assessors were wrong: document FBIS3-10695 seems definitely relevant; 11469 is probably not relevant, but partially relevant; 12092 sounds relevant; and 36379 is not relevant.

Of all the assignments in E5, 40% contain turker's justifications for answers. This feedback field was not mandatory in the experiment. In most of the cases turkers provided a very good explanation. For example, concerning not relevant documents:

- This report is about the Russian economy, not the space program.
- The "MIR" in the article refers to a political group, not the Russian space station.
- This article is about Kashmir, not the space program.

And concerning relevant ones:

- This is about Japan's space program and even refers to a launch.
- On the Russian space program, not US, but comments about American interest in the program.
- The article is relevant, but it seems a non-native English speaker wrote it. For instance the article says the space shuttle will lift off from the "cosmodrome". NASA doesn't call the launch pad a "cosmodrome."

4. DISCUSSION AND OUTCOMES

As we can see from the data analysis, turkers not only are accurate in assessing relevance but in some cases were more precise than the original experts. Also, turkers tend to agree slightly more with the experts when the document is relevant, and less when it is not relevant.

It is important to design the experiments carefully. Mapping TREC assessment instructions [6] to MTurk is not trivial. The TREC-7 guidelines is a 4-page document that has to be summarized in a few sentences for reading online, since the turker sees a screen with instructions and task to be completed. It is important to be concise, precise, and clear about how to evaluate the relevance of a document. The usage of some basic usability design considerations for presentation is also important.

In our experience, all experiments without qualification tests were completed in less than 48 hours. Once qualification test was involved, the completion rate per turker was much higher. The number of turkers required to assess per document can have an impact on the duration.

5. CONCLUSIONS

Crowdsourcing-based relevance evaluation using MTurk is a feasible alternative to perform relevance evaluations. Using TREC data, we have demonstrated that the quality of the raters is as good as the experts. Our experience shows that it is extremely important to carefully design the experiment and collect feedback from turkers. Taking a TREC-like experiment and run it as is, would probably fail.

In the future, we plan to seek confirmation of these findings on more TREC topics, and also to study related issues like the effect of topics/documents of different kinds, the number of turkers needed to get reliable results, the importance of the qualification test etc.

6. REFERENCES

[1] O. Alonso and S. Mizzaro. Relevance Criteria for E-Commerce: A Crowdsourcing-based Experimental Analysis. In *Proceedings of SIGIR'09*, 2009. In press.

[2] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[3] J. Callan, J. Allan, C. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai. Meeting of the MINDS: An Information Retrieval Research Agenda. *SIGIR Forum*, 41(2):25–34, 2007.

[4] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM TOIS*, 2009. In press.

[5] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *CHI '08: Proceeding of the 26th SIGCHI*, pages 453–456, 2008.

[6] E. Voorhees. Personal communication.

Evaluating Network-Aware Retrieval in Social Networks

Tom Crecelius
MPI Informatik, Saarbrücken, Germany
tcrecel@mpi-inf.mpg.de

Ralf Schenkel
Saarland University, Saarbrücken, Germany
schenkel@mmci.uni-saarland.de

ABSTRACT

This paper discusses the problem of evaluating search and recommendation methods in social tagging networks that make use of the network's social structure. While many such methods have recently been introduced, they fall short of evaluating the quality of the results they produce in a systematic way, which is mostly caused by the lack of publicly available test collections.

1. INTRODUCTION

Collaborative recommendations and search in social networks has been a very active research topic in recent years, and there has been an increasing number of papers proposing different methods and algorithms in this area. A recently upcoming trend is keyword-based search in social tagging networks such as del.icio.us, Flickr, or Librarything, where users annotate their items with tags. While early works in this area focused on frequency-based methods to evaluate searches, more recent approaches such as [3, 5, 6, 8] take the connections of the querying user in the social network into account when computing results. However, as there is neither a standard evaluation methodology nor a standard collection of data sets and topics, the proposals evaluate their techniques in different ways, making it impossible in practice to compare the performance of different techniques without reimplementing and reevaluating them. This clearly shows that there is an increasing demand for a publicly available evaluation platform to compare the performance of different methods for searching social tagging networks. This paper first discusses existing evaluation methods and demonstrates their shortcomings. It then proposes a future community-based evaluation task for this scenario.

2. EVALUATION APPROACHES

Evaluating effectiveness of search methods in social tagging networks has been recognized as an important yet unsolved problem [2]. A number of different evaluation methodologies for assessing the quality of such search methods has been proposed, typically as a byproduct of proposing a novel search method [4, 6, 8, 9]. Due to the lack of publicly available, large-scale samples of social networks, each paper uses a different data set, either crawls from the Web sites of exsting social networks (del.icio.us for [6], del.icio.us, Flickr and Librarything for [8, 9]) or non-public snapshots of such networks (del.icio.us for [3], data from inside IBM for [4]). As snapshots and crawls had been done at different instances in time, the crawls had used different techniques, and each snapshot had been only a small fraction of the whole network, it is very likely that each evaluation used a largely different data set, limiting the possibility to compare the results. While all approaches use reasonably large sets of keyword queries and some notion of average result precision to evaluate result quality, they drastically differ in how they determine the set of ground truth results.

User-Independent Ground Truth. Exploiting that del.icio.us maintains bookmarks, Bao et al [6] used the DMOZ catalogue of Web sites to extract queries and ground truth. They first selected DMOZ categories with enough urls that were also present as bookmarks in their del.icio.us crawl. For each such category, a keyword query was defined based on the category label. The set of relevant results for this query was formed by the urls in that category that were also present in the crawl. While this yields a large test collection, it completely ignores the user who submits the query. Methods that determine user-specific results are therefore potentially penalized by this method.

Context-based Ground Truth. Our own previous work [8] generated a set of relevant answers which we assumed to be more relative to the querying user. We computed the set of relevant answers for a keyword query as the set of items from friends of the querying user that conjunctively match the keyword query. However, this method comes with some bias towards network-aware search methods because it gives priority to results in the neighborhood of the user. Additionally, there is no clear evidence if those results really satisfy the user's information need. An item that does not appear among the user's friends may as well be very relevant for the user.

Temporal Ground Truth. If not a single snapshot, but a series of snapshots of the same social network is available, a set of relevant answers to a query can be defined by exploiting the change of the network over time. Given a tag query and a user, the set of relevant answers is formed by the items with (at least) these tags that the user aded in the future.[1] While such an item will definitely be important for the user (or she would not have added it to her collection), there is no guarantee that it is also relevant for this specific query. Additionally, an item may not have been added by the user simply because she didn't know about it, not because she found it irrelevant.

User Study. We performed a small-scale user study with six users in [9] that were actual users of the LibraryThing social network. The experiment was done in a Cranfield style, with a set of topics that each user defined, results for each topic from different methods pooled, and each pool assessed by the user who defined the topic. However, while the queries have been made public, the snapshot of the social network is not available, making it difficult to reuse them to evaluate other approaches.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

[1] A temporal ground truth has been internally used by Yahoo! Research, but has not yet been published.

3. A SETUP FOR A COMMUNITY-DRIVEN EVALUATION TASK

3.1 Collection

Data from *Bibsonomy*[2] is available for research purposes, and is currently being used for the ECML challenge [3]. However, this data set does not provide information about how users are connected, it is limited to the narrow domain of scientific publications, and it is of rather limited size, so it is unclear how significant results derived from this corpus could be. More interesting candidates for social networks would be large-scale networks like del.icio.us or librarything, which combine rich annotations and complex friend networks with interesting and rich content. However, it is unclear to which extent the owners of these networks would be willing to supply data from these networks.

Maintaining such a publicly available collection of – potentially sensitive – data from private networks raises some privacy issues. However, we think that these issues can be dealt with through a combination of technical and legal means: First, attempts should be made to anonymize the users contained in the snapshot, for example by assigning them unique, but random identifiers. As experiences with other collections, for example with the AOL query log [1] and the NetFlix dataset [7], have show in the past, such an anonymization cannot make sure that anonymous users cannot be mapped to their real identity. The collection should therefore be made available only under a restricted license that allows its use only for (possibly limited) research. This has been successfully done in the past for several other collections. Finally, the data in the collection can be restricted to information that is already available on the Web, thereby making it of limited use to anybody wanting to break the anonymity of users.

3.2 Users, Queries and Assessments

Community-driven evaluation venues such as INEX have been successfully distributing the load of defining queries and assessing evaluation results among the participating organizations. We propose to use a similar community-driven approach for the evaluation of search in social tagging networks. Each participating organization needs to define several possible *topics* (including a description of the information need, a corresponding keyword query, and example results). Each such topic must come with a *user* from that organization that is actually a user in the social network from which we take the data. In the ideal case, this would be a user who already has a long history of activity in this network, but it could as well be someone who joins for the experiment (and, of course, needs to upload and tag items and make connections to other users). Once topics are fixed, a snapshot of the network can be taken that includes these users (or, alternatively, a large crawl of the network could be performed using these users as crawl seeds).

Once the data set is available, participating organizations can—just like in existing benchmarks such as INEX or TREC—submit their results, which will then be pooled per topic and assessed by the original topic author. The latter is necessary because we assume that most topics will be of a highly subjective nature, with results highly depending on the submitting user, so we think that they cannot easily be assessed by someone who did not define the topic.

3.3 Primary Measurements

Evaluation measurements can be similar to those currently used for evaluating text retrieval methods. More specifically, there should be at least one precision-based metric such as NDCG, and one recall-based metric such as MAP.

3.4 Additional Measurements

Given that the evaluation process that we described so far incurs a great deal of work for all participants, an important subgoal of this activity would be to understand if the automatic methods for ground truth building described in Section 2 yield results that are comparable to the results with manual assessments. More precisely, it should be examined if the automated methods result in similar ranking of the different participating systems (not necessarily in similar absolute precision for the different runs). If that was the case for one of the methods, future evaluations could be much easier.

4. CONCLUSION

This paper introduced the problem of evaluating search methods in social tagging networks, presented several evaluation approaches used by different publications, and showed why none of them is generally applicable. We proposed a novel community-based evaluation that successfully captures the pecularities of social networks. However, the success of such an initiative eventually depends on the cooperation of the companies and institutions owning social network data, and on the agreement of enough organizations to participate in such a project.

5. REFERENCES

[1] E. Adar. User 4xxxxx9: Anonymizing query logs. In *Query Log Analysis: Social and Technological Challenges*, 2007.

[2] S. Amer-Yahia, M. Benedikt, and P. Bohannon. Challenges in searching online communities. *IEEE Data Eng. Bull.*, 30(2):23–31, 2007.

[3] S. Amer-Yahia, M. Benedikt, L. V. S. Lakshmanan, and J. Stoyanovich. Efficient network aware search in collaborative tagging sites. *Proc. VLDB Endowment*, 1(1), 2008.

[4] E. Amitay, D. Carmel, N. Har'El, S. Ofek-Koifman, A. Soffer, S. Yogev, and N. Golbandi. Social search and discovery using a unified approach. In *18th International World Wide Web Conference*, pages 1211–1211, April 2009.

[5] I. Assent. Actively building private recommender networks for evolving reliable relationships. In *M3SN Workshop*, pages 1611–1614, 2009.

[6] S. Bao, G.-R. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, 2007.

[7] S. Greengard. Privacy matters. *Commun. ACM*, 51(9):17–18, 2008.

[8] R. Schenkel, T. Crecelius, , M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *SIGIR*, 2008.

[9] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. X. Parreira, M. Spaniol, and G. Weikum. Social wisdom for search and recommendation. *IEEE Data Engineering Bulletin*, 31(2):40–49, 2008.

[2] `http://www.bibsonomy.org`

[3] `http://www.kde.cs.uni-kassel.de/ws/dc09`

A Virtual Evaluation Forum for Cross Language Link Discovery

Wei Che (Darren) Huang
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
w2.huang@student.qut.edu.au

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

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

ABSTRACT

In this position paper we propose to extend the current INEX Link-the-Wiki track, based on the English Wikipedia, to a Cross Language Link Discovery (CLLD) track using the multi-lingual Wikipedia. We observe that the existing automatic evaluation methods used at INEX do not necessitate manual assessment as assessments are extracted directly from the collection and performance is measured relative to this ground-truth. The proposed track can therefore run online with continuous evaluation, free from the difficulties of scheduling and synchronizing groups of collaborating researchers. We also propose to continually publish peer-reviewed evaluation results online with open access.

Categories and Subject Descriptors

D.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval – *Search process.*

General Terms

Measurement, Performance, Experimentation.

Keywords

Link-Discovery, Cross Language Information Retrieval.

1. INTRODUCTION

Since the inception of TREC in 1992 interest in IR evaluation has increased rapidly and today there are numerous active and popular evaluation forums. It is now possible to evaluate a diverse range of information retrieval methods including: ad-hoc retrieval, passage retrieval, XML retrieval, multimedia retrieval, question answering, cross language retrieval, link discovery, and learning to rank, amongst others. Most forums are tied to a long evaluation cycle which includes a workshop, symposium, or conference at the end of each cycle.

In this position paper we propose a new *virtual evaluation track*: Cross Language Link Discovery (CLLD). The track will initially examine cross language linking of Wikipedia articles. This virtual track will not be tied to any one forum; instead we hope it can be tied to each of (at least): CLEF, NTCIR, and INEX as it will cover ground currently examined at each.

We suggest automatic as well as collaborative manual assessment of submissions. With automatic and manual assessment a continual evaluation and publication forum for CLLD is possible.

2. MOTIVATION

On the welcome page of the Wikipedia we see that different language versions of the Wikipedia have different numbers of articles. At the time of writing the Maori version has about 6,500 articles whereas the English version has about 2,800,000 articles. In the cases where an article exists in both languages a bilingual reader might prefer a Maori article to an English one. This preference should "travel" with the user as they navigate around the Wikipedia, and links to articles should be given in the user's preferred language. To achieve this it is necessary to support cross lingual links in the Wikipedia (and profiles, multiple links per anchor, and so on).

Overell [3] shows that the geographic coverage of the Wikipedia very much depends on the language version – places in the UK are best covered by the English language Wikipedia while places in Spain are best covered by the Spanish language version. Despite the geographic proximity of Spain to England, a search for the village *Wylam* in the Spanish version reports *No hay coincidencias de título de artículo.* The English language version, however, informs us that the village is the birth place of George Stephenson, the inventor of the Stephensonian locomotive (the "modern" steam train). Wylam is historically interesting to railway enthusiasts, regardless of nationality – so much so that the Spanish Wikipedia article on George Stephenson shows a link for Wylam in red (the page does not yet exist). Perhaps it should link to the English article.

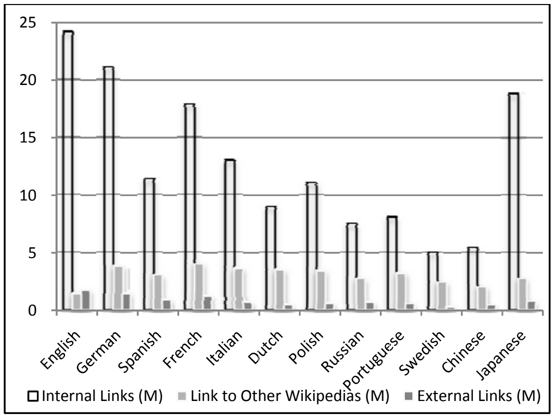


Figure 1: Cross-lingual link structure of the Wikipedia. MediaWiki provide English stats from Oct 2006 while others are from Dec 2008.

These two use-cases demonstrate a need for cross-language links within the Wikipedia. WikiMedia provide some statistics showing that there are already many cross language links. The statistics are summarized in Figure 1 where it can be seen that about a quarter of the Chinese links are to other languages (many Chinese articles

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

link to English pages, 诺森伯兰郡, for instance, links to “List of United Kingdom Parliament constituencies”). English articles, however, are not well linked to other languages.

3. TASK DEFINITION

We propose a Cross Language Link Discovery (CLLD) track run as a collaboration between INEX, CLEF, and NTCIR. Initially two linking experiments will be examined:

MULTILINGUAL topical linking is a form of document clustering – the aim is to identify (regardless of language) all the documents in all languages that are *on the same topic*. The Wikipedia currently shows these links in a box on the left hand side of a page.

BILINGUAL anchor linking is exemplified by the Chinese article 诺森伯兰郡, having a link from the anchor 国会选区 to the English article “List of United Kingdom Parliament constituencies”. The link discovery system must identify the anchor text in one language version of the Wikipedia and the destination article within any other language version of the Wikipedia.

4. STATIC EVALUATION

When Trotman & Geva [4] introduced the Link-the-Wiki track at INEX they noted that, technically at least, the evaluation required no human assessment. The same is true with cross-language link discovery.

Topics in the INEX Link-the-Wiki track are chosen directly from the document collection. All links in those documents are removed (the documents are orphaned). The task is to identify links for the orphans (both to and from the collection). Performance is measured relative to the pre-existing links.

For MULTILINGUAL linking the links on the left hand side of the Wikipedia page could be used as the ground truth. The performance could be measured relative to the alternate language versions of the page already known to exist.

BILINGUAL anchor linking from one document to another could also be automatically evaluated. Links from the pre-orphan to a destination page in an alternate language would be used as the ground truth – but there are unlikely to be many such links.

A same-language link from a pre-orphan to a target provides circumstantial evidence that should the target exist in multiple languages then the alternate language versions are relevant. This is essentially a triangulation: $A \overset{t}{\rightarrow} B \overset{l}{\rightarrow} C \Rightarrow A \overset{tl}{\rightarrow} C$ where *A*, *B*, and *C* are documents; and *t* designates a topical link, *l* a cross language link, and *tl* a topical cross language link.

Static assessment requires no human interaction. A web site with orphan sets, assessment sets (extracted from the pre-orphans), and evaluation software, can support a sound evaluation methodology which does not necessitate any fixed deadlines.

5. CONTINUAL EVALUATION

Huang *et al*. [1] question automatic evaluation. Their investigation suggests that many of the links in the Wikipedia are not topical, but are trivial (such as dates), and that users do not find them useful. Manual assessment is, consequently, necessary. This challenges cross language link discovery because finding assessors fluent in multiple languages is difficult – especially for a track with a relatively small number of participants but in a large number of languages (the Wikipedia has 266 languages).

We propose a novel form of evaluation called *continual evaluation* in which participants can download topics and submit runs at any time; and in which manual assessment is an on-going concern. The document collection will, initially, be static. Topics will either be chosen at random from the collection, or nominated by participants. For any given run a participant will download a selection of topics and submit a run. The evaluation will be based on metrics that consider the un-assessed document problem (such as a variant on rank-biased precision [2]), and comparative analysis will be relative to an incomplete, but growing, assessment set.

To collect manual assessments two methods are proposed: first, in order to submit a run the participant will be required to assess some anchor-target pairs in languages familiar to them; second, we will run an assessment Game With A Purpose (GWAP). Kazai et al. used a GWAP for the INEX Book track; Von Ahn & Dabbish [5] discuss GWAPS in other contexts (including the Google Image Labeler). Regardless of the method of assessment collection, we are trying to validate the minimum number of links necessary to disambiguate the relative rank order of the runs (within some known error).

6. PUBLICATION

Both automatic and manual assessment of cross language link discovery can be performed on a continual rolling basis; there is no need for topic submission deadlines, run deadlines, assessment deadlines, or even publication deadlines. At INEX the time difference between run-submission and the workshop paper submission date is long (6 July – 23 Nov). With automatic assessment it is possible to achieve a result, write, and publish a paper with a short turn around. As part of the virtual track we propose an open-access virtual CLLD workbook to which registered participants can submit their papers for peer review and publication.

7. CONCLUSIONS

We put the case for an online virtual track that examines Cross Language Link Discovery in the Wikipedia. Such a track can be *continual* because the assessments are drawn from the collection itself. To facilitate the exchange of results we propose a virtual open-access workbook to which participants can submit papers. We believe this virtual forum will better serve the link-discovery community than the existing calendar based evaluation forums.

REFERENCES

[1] Huang, W.C., A. Trotman, and S. Geva, *The Importance of Manual Assessment in Link Discovery*, in *SIGIR 2009*. 2009, ACM Press: Boston, USA.

[2] Moffat, A. and J. Zobel, *Rank-biased precision for measurement of retrieval effectiveness.* ACM Trans. Inf. Syst., 2008. 27(1):1-27.

[3] Overell, S.E., *Geographic Information Retrieval: Classification, Disambiguation and Modelling*, in *Department of Computing*. 2009, Imperial College London: London. p. 175.

[4] Trotman, A. and S. Geva. *Passage Retrieval and other XML-Retrieval Tasks*. in SIGIR 2006 Workshop on XML Element Retrieval Methodology. 2006. Seattle, USA.pp. 43-50

[5] von Ahn, L. and L. Dabbish, *Designing games with a purpose.* Commun. ACM, 2008. 51(8):58-67.

On the Evaluation of the Quality of Relevance Assessments Collected through Crowdsourcing

Gabriella Kazai
Microsoft Research
7 JJ Thomson Ave
Cambridge, UK
gabkaz@microsoft.com

Natasa Milic-Frayling
Microsoft Research
7 JJ Thomson Ave
Cambridge, UK
natasamf@microsoft.com

ABSTRACT

Established methods for evaluating information retrieval systems rely upon test collections that comprise document corpora, search topics, and relevance assessments. Building large test collections is, however, an expensive and increasingly challenging process. In particular, building a collection with a sufficient quantity and quality of relevance assessments is a major challenge. With the growing size of document corpora, it is inevitable that relevance assessments are increasingly incomplete, diminishing the value of the test collections. Recent initiatives aim to address this issue through crowdsourcing. Such techniques harness the problem-solving power of large groups of people who are compensated for their efforts monetarily, through community recognition, or by the entertaining experience. However, the diverse backgrounds of the assessors and the incentives of the crowdsourcing models directly influence the trustworthiness and the quality of the resulting data. Currently there are no established methods to measure the quality of the collected relevance assessments. In this paper, we discuss the components that could be used to devise such measures. Our recommendations are based on experiments with collecting relevance assessments for digitized books, conducted as part of the INEX Book Track in 2008.

Keywords

Test collection construction, relevance judgments, incentives, social game, quality assessment.

1. INTRODUCTION

The established approach to constructing a test collection involves employing a single judge, usually the topic author, to assess the relevance of documents to a topic. Recent practices are, however, diversifying the ways in which relevance judgments are collected and used. In Web search the tendency is to use explicit judgments from a sample of the user population or to analyze user logs to infer relevance. An increasingly popular strategy is to use crowdsourcing. Amazon’s Mechanical Turk service, for example, employs Internet users to complete ‘human intelligence tasks’, such as providing relevance labels, for micro-payments. Google’s Image Labeler game [7] works by entertaining its participants who label images for free. Community Question Answering (cQA) services, such as Yahoo! Answers, reward the members who provide the best answers with ‘points’ which leads to increased status in the community. Participants of the Initiative for the Evaluation of XML retrieval (INEX) [3] contribute relevance assessments of highlighted passages in Wikipedia documents [6] or digitized books [5] in order to gain access to the full test set.

Obtaining relevance judgments through a collective user effort, however, carries inherent risks regarding the quality of the collected data. For example, it has been shown that the different background knowledge of the assessors can lead to different conclusions in the evaluation [1]. A further critical factor is the incentive that motivates assessors to provide relevance judgments. For example, workers on Amazon’s Mechanical Turk benefit from completing more jobs per time unit. Thus, the quality of their output may not be of their concern unless tied to the potential loss of their income. Studies have also shown that some members of the cQA community ‘play the system’ by colluding in order to increase their status. Similar problems of user tactics have been reported in reputation systems like eBay.

This raises the question of how to estimate the trustworthiness of relevance labels provided by members of the ‘crowd’ and how to evaluate the quality of the collected relevance data set. In this paper, we make recommendations based on the experiments conducted at the INEX 2008 Book Track.

2. COLLECTIVE ASSESSMENTS AT INEX

In 2008, the INEX Book Track [4] experimented with a method for the collective gathering of relevance assessments using a social game model [5]. The Book Explorers’ game was designed to provide incentives for assessors to follow a predefined review procedure. It also made provisions for the quality control of the collected relevance judgments by facilitating the review and re-assessment of judgments and by enabling communication between judges. The game was based on two competing roles: explorers who discover and mark relevant content and reviewers who check the quality of the explorers’ work. Both were rewarded points for their efforts. Disagreements between explorers and reviewers led to point deductions which could be recovered by re-assessing the pages under conflict (though agreement was not necessary).

In two pilot runs, several types of relevance data were collected: text regions highlighted on a page, relevance labels assigned to a page, comments recorded for a page, and the relevance degree assigned to the books. In total, 17 assessors judged 3,478 books and 23,098 pages across 29 topics, and marked a total of 877 highlighted regions. The assessment system recorded 32,112 navigational events, 45,126 relevance judgment events, and 2,970 ‘search inside the book’ events.

In addition, as part of the assessment process assessors were asked to indicate their familiarity with their selected topics, as well as to record their familiarity with each book they judged before and after they browsed the book.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

3. TRUST AND QUALITY CONTROL

Given the assessors' diverse backgrounds and intentions, the question arises as to what degree relevance assessments can be trusted. For example, assessors' desire to win may influence their work, leading to more labels but of lower quality. To incorporate the notion of reliability, we may associate a *trust weight* with each assessment. The final assessments can then be derived as weighted averages of the individual opinions. However, how can such trust weights be derived without an established ground-truth to compare with? In the following sections we discuss possible sources of evidence for computing the trust score.

3.1 Assessor agreement

We hypothesize that *judgments agreed upon by multiple assessors can be trusted more*. Agreement can suggest that the topic is less ambiguous, that the interpretation of the document and the relevance criterion is similar across the judges, but it may also signal collusion. Judges may collude in order to increase their scores. Disagreement can indicate an ambiguous topic, possible differences in the assessors' knowledge or interpretation of the relevance criterion, or may reflect their intention to reduce each others' scores. The trust weight will depend on being able to differentiate between these reasons.

In our data set, a total of 239 books were judged by multiple assessors (between 2-4) across 18 topics. The level of pairwise agreement between judges, based on binary relevance, was relatively high, around 80.7%. Out of 239 books, judges only disagreed on the relevance of 24 books. Their opinion differed only on the degree of relevance for 34 relevant books (71% by 1 degree, 20% by 2 degrees, 6% by 3 degrees and 3% by 4 degrees). At the page level, 4,622 pages were judged by multiple assessors with an agreement level of 57%.

Suggestive influence. The observed levels of agreement are relatively high compared to those reported elsewhere (i.e., around 33-49% for documents at TREC, and 27-57% for documents and 16-24% for elements at INEX). This high level of agreement could suggest collusion between judges or could simply reflect bias in their work. Since reviewers were shown the relevance labels that explorers assigned to a page, their own judgments could have been influenced by these opinions. Indeed, the majority of the multiple judgments were results of reviewers checking the explorers' work (74%). However, the reviewers were not aware of the relevance labels that explorers assigned to books.

Topic familiarity. The average difference between assessors' familiarity with the topics for books on which they agreed on (based on binary relevance) was 1.95 while for books on which they disagreed was 3.36. This shows that background knowledge does contribute to differences of opinions.

Collusion. Possible collusions may involve judges from the same institution who agree with each other. In the collected data, 6 books and 606 pages were judged by different members of the same group. Judges agreed on the relevance of all 6 books at the binary level and disagreed on the degree of relevance for 4 of the books. They also agreed on the (binary) relevance of all, expect 5, pages. This agreement may, however, be genuine rather than a result of collusion. The amount of time spent on assessing a page (dwell time) could provide a clue: it may be reasonable to expect that judges with similar levels of topic and book familiarity spend similar lengths of time assessing the same page. Collusion could thus be detected when judges consistently agree whilst having different averages for time spent on a page or book.

3.2 Annotations

Annotations, i.e., comments added to pages by assessors, could be used when considering the trustworthiness of the assessments. We hypothesize that *the judgments of annotated pages may be more trustworthy* since judges spent extra time and effort.

Comments were added by 9 of the 17 assessors to 227 pages in 98 books. The distribution of comments varied greatly, with an average of 25 comments per judge (σ=37, min=1, max=102). Two judges, in particular, made frequent use of this feature, adding 102 and 75 comments, respectively. This reflects commitment on their part and suggests that their judgments may be more trustworthy.

Most comments were explanations of relevance decisions or short summaries (76%), or qualitative statements about the relevance of the content (15%). We suspect that the comments may have acted as indirect messages, purposefully added by explorers to preempt possible challenges and thus penalty from reviewers. The presence of comments may thus also signal ambiguous content or questions about relevance. Furthermore, comments can also provide clues on the user background and the user task.

3.3 Learning effect

At the start of the assessment process, judges indicated their familiarity with their selected topics. However, although initially unfamiliar with a topic, a judge may learn about it during the review process. One way to assess this is to examine changes in the length of time that judges spend on assessing pages for a given topic. Assuming that judges learn, we expect them to become faster in assessing pages over time. This should be combined with their self-declared familiarity with the book and incorporated into the trust weight.

4. CONCLUSIONS

In this paper we draw attention to potential issues with collective relevance assessments through crowdsourcing, where judges with diverse backgrounds and intentions contribute data with varied levels of reliability. We discuss several sources of evidence that could be used to derive a trust weight for the judgments: topic familiarity and familiarity with the content being assessed, dwell time and changes in the patterns of dwell time, agreement between judges, and the presence and length of comments. However, other factors, such as the incentives that influence judges' behavior, also need to be considered. How to define the trust weight function based on these factors, taking into account their complex relationship, is the subject of our further research.

5. REFERENCES

[1] Alonso, O., Rose, D. and Stewart, B.. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., and Yilmaz, E. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proc. of SIGIR 2008*, 667-674.

[3] Fuhr, N., Kamps, J., Lalmas, M., Malik, S., and Trotman, A. 2007. Overview of the INEX 2007 ad hoc track. In *Proc. of INEX'07*.

[4] Kazai, G., Doucet, A., Landoni, M. 2009. Overview of the INEX 2008 Book Track. In *Proc. of INEX'08*. LNCS Vol. 5613, Springer.

[5] Kazai, G., Milic-Frayling, N., Costello, J. 2009. Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments. In *Proc. of SIGIR 2009*.

[6] Trotman, A. and Jenkinson, D. 2007. IR evaluation using multiple assessors per topic. In *Proc. of ADCS*.

[7] von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In *Proc. of SIGCHI 2004*, 319-326.

CiteEval for Evaluating Personalized Social Web Search

Zhen Yue[1], Abhay Harpale[2,] Daqing He[1], Jonathan Grady[1], Yiling Lin[1], Jon Walker[1], Siddharth Gopal[2], Yiming Yang[2]

[1]School of Information Sciences,
University of Pittsburgh
{zhy18,dah44,jpg14,yil54,jdw8}@pitt.edu

[2]Language Technology Institute,
Carnegie Mellon University
{aharpale,sgopal1,yiming}@cs.cmu.edu

Categories and Subject Descriptors: H.3 Information Storage and Retrieval; H.3.4 Systems and Software: Performance Evaluation
General Terms: Design, Experimentation, Reliability
Keywords: Personalization, Search, Evaluation, Dataset.
.

1. DEVELOPING CITEEVAL

The technologies and the ideas of Web 2.0 have significantly changed users in thinking and using Web information in their work and other aspects of daily life. More and more Web users, from sophisticated to naïve, are more willing to share online their own ideas, readings, documents, and many other materials. As a result, there is much more potential relevant information in social Web setting for users to search on, at the same time, by knowing more about individual users' interests, knowledge and preference, it is possible to build personalized search systems to support users' searches. Personalization has attracted researchers from information retrieval, user modeling, machine learning communities, and has generated many interesting results. However, no reasonable large test collection is yet available for researchers to compare their personalization algorithms.

The rapid development of modern information retrieval technologies owns great debt to TREC and other benchmark evaluation frameworks. Although Cranfield inspired evaluation frameworks still have many limitations, they are the best available test beds for examining the effectiveness of retrieval algorithms across different sites, different platforms, and even different time periods. Researchers in IR and related areas, such as text classification and information extraction, all understand the importance of having standard benchmark evaluation datasets.

In this position paper, we will present a new dataset called CiteEval for benchmark evaluation of personalized algorithms in social Web searches. However, before we talk in detail about the construction of CiteEval, we want to discuss the key features that such benchmark datasets should have:

- Currently most personalization algorithms still work on text. Therefore, the documents in the dataset should be primarily textual social web content. Ideally, the documents should have full text information, but the reality is that maybe only basic bibliographic information such as author, title, abstracts and keywords is available.
- The dataset should explicitly contain users and their search tasks for evaluating personalization. Since many personalization algorithms rely on users' past behaviors and results for adaptation, the tasks and the queries associated with the tasks should provide rich history. To obtain true personalization, the relevance annotations should only be done by the person who proposed the search task.
- The dataset should include as many extra features about the documents as possible. The preferable minimum set should have hyperlinks, tags, categories/topic labels, and virtual communities. Past personalization algorithms have utilized lots extra information than the basic document content. For example, Hyperlinks have been combined with user profiles to provide personalized PageRank among documents; categories of topics have been used to identify users' interests and document similarities; and social tags and online communities are among the newly applied social Web features in identifying users' expertise and interests.

CiteEval contains academic articles extracted from CiteULike and CiteSeer repositories, with multiple features such as bibliographic information, tags, topic categories, and citation information.

CiteULike (www.citeulike.org) is a social Web site designed for scholars to store, organize and share the papers that they are reading. CiteULike papers are organized around individual CiteULike users, of which there is a private library to store all the papers the users have read, the tags that the users have entered, and the virtual communities (called groups) that the users have subscribed to. However, as an open free access environment, CiteULike suffers from spam contamination, unintentional human errors and inaccurate information. We, therefore, used CiteSeer (http://citeseer.ist.psu.edu/) to extract critical document metadata such as document abstract, authors, publication year, and keywords. CiteSeer is another popular repository, but it is widely accepted as an authoritative source for academic publication. To obtain the citation/link relationships among documents, all CiteSeer papers cited by at least one selected paper in CiteULike is included into the final CiteEval collection.

To obtain focused user-tasks and personalized relevance judgments, we solicited experts who have at least several years research experience in the areas of Computer Science and Information Systems. The selection of the right experts for our annotation was balanced with the availability of related documents and users in CiteULike. Our goal is to make sure that the proposed search tasks have enough relevant documents and similar users in CiteULike, and at the same time our experts can develop tasks according to their own research interests for true personalization. To achieve this, we identified potential topics by looking at relevant CiteULike groups that contain at least 10 users and more than 500 articles. Then we selected the groups whose topics fit to the research areas of the recruited experts.

Each expert was asked to develop a full topic statement to describe his/her search task, and he or she then searched the

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

collection with four to six search queries that are related to the search task. This not only gave the experts opportunities to review and examine the search tasks against the collection, but also helped us to collect their relevance annotations. Figure 1 shows an example of the search tasks.

UserID	**network03**
Topic	**Information Network Security**
Topic Statement	Access control is the process in which a request to a data resource or service is mediated to determine whether the access should be granted or denied. Access control mechanism is managed by an authorization policy which generally states which subjects can perform what operations or have what rights on which objects. Different access control models have been proposed to address specific environmental requirements and challenges or provide more powerful and expressive policies.
Query1	role based access control
Query2	workflow access control
Query3	authorization delegation
Query4	distributed access control
Query5	XML access control

Figure 1: Search Task "Information Network Security"

Task ID	# Queries	# Highly relevant	# Slighly relevant	# Not relevant
blog01	5	49	310	1611
education01	4	166	148	1178
education02	5	110	241	1829
network01	5	67	17	1861
network03	5	73	58	1699
p2p01	6	396	326	1546
statistic01	5	9	54	1827
web02	5	231	84	1610
web03	5	27	76	1822
Average	**5**	**125**	**146**	**1665**

Table 1: Relevance Annotations of Some CiteEval Tasks

During the annotation process, the expert judged the relevance of the top 500 returned documents for each query. However, considering the possible limitation of CiteULike search engine, we used two additional resources to enhance the annotation coverage. First, by assuming that all documents in the corresponding CiteULike group(s) could have higher chance to be relevant, each document in the group library was judged by the expert for relevance to one of the queries. The second resource come from a well studied relevant annotation strategy -- pooling method used in TREC experiments [2]. We used seven different retrieval algorithms to return from CiteEval a pool of articles for each query and asked our experts to annotate every article in the pool. Through this complex relevant annotation process, we built a comprehensive ground truth annotation for our test collection.

In total, CiteEval contains 81433 documents, of which 39327 were extracted from CiteULike initially. 42106 were added from CiteSeers. We recruit 20 experts who developed 20 different tasks that belong to 13 groups. Table 1 shows the statistics of the annotations for nine out of the 20 search tasks. In average, each search task has 5 queries. The average number of highly relevant documents identified for each task is 125, and that of somewhat relevant documents is 146. But to obtian this amount of relevance annotation, our experts in average annotated 1936 documents.

2. DISSCUSSIONS

As the initial study of the usages of CiteEval dataset [3], we conducted searches on the dataset using our implementations of several personalized and unpersonalized algorithms. We used Indri search engine as the representative unpersonalized system. Indri results were personalized using three different strategies. One method called TDS (Topic Distribution Search) re-ranked documents based on the user's topical interest distribution. Another method was based on the popular Personalized PageRank (PPR) to re-rank Indri results based on a weighted combination of PPR scores and Indri-based relevance scores. Finally, another method, which we call PCF, used the probabilistic Latent Semantic Analysis (pLSA) to estimate user's topical interests based in a collaborative filtering setting. MPS (Meta Personalized Search) used a weighted combination of TDS, PPR and PCF for generating the final ranked-list. In our experiments, we observe a significant improvement of personalized search approaches over the unpersonalized ones. Using these results, we ran Cronbach's alpha, which is a reliability value based on the classical test theory [1]. The alpha value is 0.97, which indicates that results obtained by testing on CiteEval are reliable. Therefore, CiteEval dataset is useful for researchers to test their personalized search algorithms. Because of the rich features in the dataset, the personalized algorithms to be tested can utilize any combination of links among documents, document categories, social tags, online communities and other user related information.

One of the major challenges in creation of a personalized search dataset is the issue of quality control. The users' relevance annotation completely depends on that particular user. Although it helps us establish the true personalization in relevance, it is difficult to guarantee that the annotation is in fact correct for a particular search task. How to reassure the quality and still maintain valid personalization is an interesting challenge that we would like to focus on for future work.

3. ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation under Grant No. 0704628.

4. REFERENCES

[1] D. Bodoff and P. Li. Test theory for assessing ir test collections. In *SIGIR 2007.* pp:367-374. 2007.

[2] D. K. Harman. The TREC test collections. *TREC: Experiment and evaluation in information retrieval.* E. M. Voorhees and D. K. Harman (Eds). The MIT Press. pp:21-52. 2005.

[3] A. Harpale, Y. Yang, Z. Yue and D. He. Citeeval: A new multi-faceted dataset for evaluating personalized search performance. Submitted to *the 18th ACM Conference on Information and Knowledge Management (CIKM2009).* 2009.

Relative Significance is Insufficient: Baselines Matter Too

Timothy G. Armstrong, Justin Zobel, William Webber, Alistair Moffat

Computer Science and Software Engineering
The University of Melbourne, Victoria 3010, Australia
{tgar,jz,wew,alistair}@csse.unimelb.edu.au

ABSTRACT

We have tabulated retrieval effectiveness claims from a large number of information retrieval research papers from 1998–2008, a period that has seen many innovations. The results of our analysis are not encouraging. Over this period, although a great many papers claimed significant effectiveness improvements, there has been no overall gain in absolute retrieval effectiveness on TREC ad hoc collections. A decade of development has not, it appears, led to better systems.

To promote verifiable improvement, reporting practices that allow rigorous comparison with prior results are needed. We propose several measures: ongoing longitudinal surveys; better reporting of baselines and use of standard systems; and use of resources such as our `evaluatIR.org`, an accessible database of test results.

1. INTRODUCTION

A core goal of information retrieval (IR) research is to make ongoing improvements in retrieval system effectiveness. A tenet of our community is that – through incremental improvement, and innovations such as language models and query expansion – we have gradually improved the effectiveness of search systems. To verify claimed improvements, we create standard test collections, in particular through the TREC mechanism; and we carry out "before" and "after" trials, measuring performance using a standard metric such as mean average precision (MAP). We also use the literature to argue the details of test collection creation and of effectiveness measures, but are confident that their systematic adoption has let us measure progress in the field.

However, a careful tabulation of the last decade of IR literature reveals a picture that for ad-hoc retrieval is far from encouraging. The reported effectiveness results show no pattern of improvement in MAP at all, and even in 2008 many new results that were validated via experiments using old collections were below the median results of a decade ago. Furthermore, these "improved" results are often worse than those available from the publicly available Terrier system. It seems that over a decade or more, authors have published and referees have approved work that, taken collectively, has done little to advance the effectiveness of IR systems.

We see this problem as a broad failure of experimental method. There are straightforward mechanisms that could lead to better outcomes, but adopting them will require determination on the part of the community, as at face value they would mean that many current papers would not be publishable.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston
.

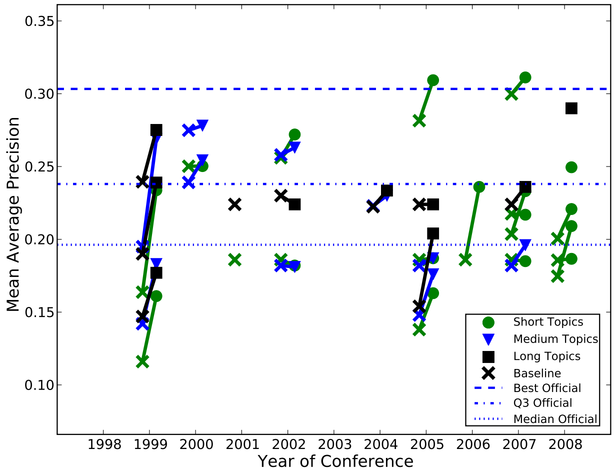


Figure 1: Published MAP scores for the TREC 7 Ad-Hoc collection. The connections show before-after pairs.

2. METHODOLOGY

All papers published at the annual ACM SIGIR conference for the period 1998–2008, and at the ACM CIKM conference for 2004–2008, were scrutinized for experimental effectiveness results. A large proportion of new IR techniques are first presented in SIGIR, so it is where we expect to find results that are indicative of the overall state of IR research. In recent years the CIKM conference has also become a significant forum for IR research.

Results were tabulated for papers that presented effectiveness scores for ad-hoc style retrieval on TREC collections, meaning TREC Ad-Hoc, Robust, Web, and Terabyte collections, and subsets thereof. Note was made of all MAP and P@10 effectiveness scores, as these are the most commonly reported metrics and the only ones used regularly enough in the period surveyed to permit a longitudinal analysis. Careful attention was paid to the distinction between "baseline" and "improved" (or "before" and "after") values. The analysis identified 87 SIGIR papers and 21 CIKM papers. Of these, 90 were focused on retrieval effectiveness; 8 on efficiency; 5 on distributed retrieval; and 5 reported scores but did not make clear claims. The set of papers studied included four that had authors in common with this abstract.

Results for a representative test collection and measure (TREC 7, using MAP), are shown in Figure 1. The trend visible in this plot is typical of what we found for all of the Ad-Hoc retrieval tasks, including the Robust track, and the Web tracks in TREC 9 and TREC 2001. There is no clear upward or downward trend in retrieval effectiveness, and since 1998 the vast majority of scores

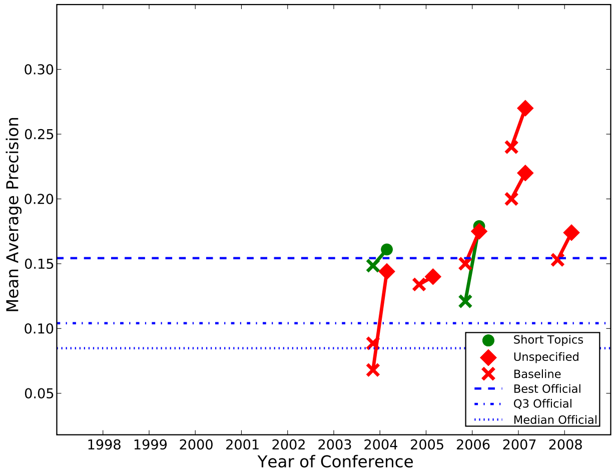


Figure 2: Published MAP scores for the TREC 2003 web track topic distillation task in SIGIR and CIKM Proceedings.

fluctuated in the range of the upper 50% of official TREC run scores. Baselines show a similar trend: the relationship between baseline score and the claimed score in each paper is stronger than any incremental performance improvements over time.

With the exception of the TREC 6, TREC 7 and TREC Robust 2004 and 2005 scores reported in two papers [Liu et al., 2005, Zhang et al., 2007], and a TREC 4 score reported in a 1998 paper [Mitra et al., 1998], we found no ad-hoc retrieval results that exceeded the scores of the best corresponding automatic TREC run. It might be argued that the maximum TREC scores are an unstable and unfair baseline, and that because of per-topic variation in system performance we would expect some outlier systems in a big pool such as the TREC competitions, purely by chance. However, given the time that has elapsed and the number of publications claiming significant (and sometimes substantial) effectiveness improvements, it is surprising that the original best systems are so rarely bettered – especially given the fact that the original runs were the only ones conducted without the benefit of hindsight. As a contrast to the ad-hoc retrieval tasks, Figure 2 shows that there have been ongoing performance improvements for the web topic distillation task of TREC 2003.

Another finding of our analysis was the large number of variant test collections used, despite the survey's restriction to 11 base collections. In 108 publications, 83 different test collections were used, with variants derived by subsetting or combining topics and corpuses from different base collections. There was also little use of standard retrieval systems, even though public domain systems are competitive with published results, and are natural baseline candidates. For instance, Terrier achieves a MAP of 0.248 on the TREC 7 Ad-Hoc collection[1], beating all but four results published since Terrier's 2005 release.

3. PROPOSALS

Future IR evaluations will need to consider the issues raised by our analysis, including the lack of gains overall, the apparent readiness of reviewers to accept papers that have results that are demonstrably weak, and the lack of transparency in many retrieval experiments. It is our view that even significant improvements on a poor baseline should not in themselves merit publication, as such results do not prove that the method being tested would be effective when added to a more competitive baseline. Yet many papers report experimental results using non-standard test-collections, make poor baseline choices, do not report best prior results, and do not provide sufficient experimental detail that would allow their claims to be independently reproduced.

Having an expectation of thorough and consistent reporting of past results would go some way to addressing these concerns, but in our view more is required. We have created a resource for researchers that can bring together all relevant effectiveness results in a way that permits easy comparisons and benchmarking, namely `evaluatIR.org` [Armstrong et al., 2009]. We see several uses for the system: as a resource for analysis of a researcher's own runs against a large database of existing results; as a repository for use by readers and reviewers of papers who wish to evaluate published claims; and as a database that allows the IR community to perform longitudinal and other comparative analyses. Use of this resource is, however, a challenging step: few new methods appear to be competitive with established benchmark systems, and the papers describing them would thus be at risk of summary rejection.

As a related step, we should expect researchers to use multiple test collections, and, more significantly, multiple retrieval systems, to demonstrate that new techniques provide verifiable improvements in combination with a range of configurations.

4. CONCLUSION

Our longitudinal survey of published IR results in SIGIR and CIKM proceedings from 1998–2008 has revealed that ad-hoc retrieval does not appear to have measurably improved. There are many possible explanations for this apparent stagnation, but it is troubling that it appears to have gone largely unremarked within the IR community. It is also paradoxical that the stream of incremental "significant" effectiveness improvements in the literature has not resulted in any apparent cumulative improvements.

Whatever the future direction of IR evaluation, there are fundamental issues with reporting practices that must be addressed. Our evaluation suggests that current methods for measuring improvement are not adequate, and that unless we adopt rigorous strategies for identifying which techniques in the field are of genuine value, we risk remaining on a treadmill of inconclusive experimentation.

Acknowledgement: This work was supported by the Australian Research Council, and the Australian Government through NICTA.

References

T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. EvaluatIR: An online tool for evaluating and comparing IR systems. In *Proc. 32nd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Boston, USA, 2009. Demo.

S. Liu, C. Yu, and W. Meng. Word sense disambiguation in queries. In *Proc. 14th ACM Int. Conf. on Information and Knowledge Management*, pages 525–532, Bremen, Germany, 2005.

M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 206–214, Melbourne, Australia, 1998.

W. Zhang, S. Liu, C. Yu, C. Sun, F. Liu, and W. Meng. Recognition and classification of noun phrases in queries for effective retrieval. In *Proc. 16th ACM Conf. on Information and Knowledge Management*, pages 711–720, Lisbon, Portugal, 2007.

[1] From `ir.dcs.gla.ac.uk/terrier`, specifically Terrier 2.2 with BM25 similarity ($b = 0.3$) and query expansion (Bose-Einstein 1 term weighting model with 3 documents and 10 terms) using Title+Description queries.

Accounting for Stability of Retrieval Algorithms using Risk-Reward Curves

Kevyn Collins-Thompson
Microsoft Research
1 Microsoft Way
Redmond, WA 98052-6399 U.S.A.
kevynct@microsoft.com

ABSTRACT

Past evaluation of information retrieval algorithms has focused largely on achieving good *average* performance, without much regard for the *stability* or variance of retrieval results across queries. In fact, two algorithms that superficially appear to have equally desirable average precision performance can have very different stability or *risk profiles*. A prime example comes from query expansion, where current techniques typically give good average improvements in mean average precision, but are also unstable and have high variance across individual queries [3]. We propose the use of *risk-reward curves* and related statistics to characterize the tradeoff an algorithm exhibits between a reward property such as mean average precision and a risk property such as the variance of the algorithm – particularly the downside variance, when the algorithm fails or makes performance worse. Such evaluation methods are broadly applicable beyond query expansion to other retrieval operations that must balance risk and reward, such as personalization, document ranking, resource selection, and others.

Categories and Subject Descriptors: H.3.3 [**Information Retrieval**]: Evaluation
General Terms: Experimentation, Measurement
Keywords: Algorithm risk, stability, query expansion

1 Risk-reward tradeoff curves

We observe that many IR scenarios have a risk-reward tradeoff. In query expansion, for example, when interpolating a feedback model with the original query model using a parameter α, giving more weight to the original query model (lower α) reduces the potential harm of a noisy expansion model, but also reduces the potential gains when the feedback model is effective, and vice versa. By plotting risk and reward jointly as α is varied from $\alpha = 0.0$ (original query only) to $\alpha = 1.0$ (all feedback), we obtain a *risk profile* in the form of a risk-reward tradeoff curve that gives a more complete picture of algorithm quality. As Fig. 1 shows, two algorithms that appear identical in terms of mean average precision (MAP) gain may have very different risk profiles.

In general, to compute a risk-reward tradeoff curve for an information retrieval algorithm, we must first decide on how to quantify risk and reward. The appropriate measures will vary with the retrieval task: a good 'reward' measure for Web search, for example, may be precision in the top-20 documents (P20); legal IR applications may focus on recall; and general IR evaluations may use mean average precision (MAP). We generally will focus on risk-reward curves using relative or absolute MAP or P20 gain as the 'reward' measure, and this is plotted on the y-axis of the chart.

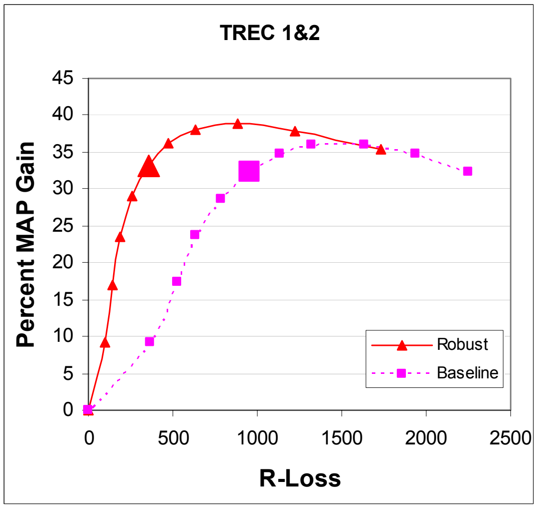


Figure 1: One form of risk-reward tradeoff curve for query expansion, showing how two algorithms that give almost identical MAP gain (33%) at a typical operational setting ($\alpha = 0.5$, shown as the enlarged points) can have very different risk profiles: the 'robust' version of the expansion algorithm is much more stable and has a much smaller net loss of relevant documents for expansion failures. The downside risk/variance (R-Loss) and MAP improvement change together as the feedback interpolation parameter α is increased from 0. (original query, no expansion) to 1.0 (all feedback model, no original query). Curves that are *higher* and *to the left* give a better tradeoff. This example is an actual experiment result (TREC 1&2 topics) taken from [2].

The key aspects of the 'risk' measure are: 1) that it captures *variance* or a related negative aspect of retrieval performance across queries, and 2) this variance/risk is based on the corresponding reward measure chosen. We are particularly interested in the *downside risk* of an expansion algorithm: the reduction in reward due to expansion failures,

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

which are defined as cases where applying expansion gives worse results than the initial query. The risk measure is assigned to the x-axis of the risk-reward curve.

As one example, we can choose the reward measure to be 'percent gain in precision at k ($P@k$)' compared to using no expansion, and the risk measure, which we call *R-Loss at k* as the *net loss of relevant documents from the top k due to expansion failure*. R-Loss at k is an appropriate risk measure because it both reflects the downside variance of the reward measure and net loss in relevant documents is a concrete and important measure for users. When we use MAP gain as the reward measure instead of $P@k$, we refer to the risk measures simply as *R-Loss*, setting k to the size of the retrieved document set (typically $k = 1000$). Because R-Loss is a document count, queries with more relevant documents have greater influence on the measure. Alternatively, we could consider normalizing R-Loss over the number of relevant documents, to give each query equal weight.

A number of potentially useful concepts and extensions follow from exploiting connections to computational finance. We say that one algorithm's tradeoff curve A *dominates* another curve B if the reward achieved by A for any given risk level is always at least as high as achieved by B at the same risk level. For example, in Figure 1 the robust algorithm dominates the baseline expansion method. The *efficient frontier* on a risk-reward graph is the boundary of the convex hull of points produced by (in theory) all possible parameter settings and represents the best performance that an algorithm can achieve at any given level of risk, for any choice of parameters. Typically, the efficient frontier can be approximated, although at considerable computational cost, by broad sampling of the parameter space.

The *risk-reward ratio* $\rho(P) = G(P)/F(P)$ of a point P that achieves MAP gain $G(P)$ and R-Loss $F(P)$ is the *slope* of the line joining P to the origin. The *midpoint risk-reward tradeoff* at $\alpha = 0.5$ gives a single value that could be used to compare with other algorithms on the same collection. The *Sharpe ratio* is the optimal $\rho^\star = \rho(P^\star)$ at the point $P^\star$ of maximum slope on the (approximate) efficient frontier, identifying the *best achieved tradeoff* of an algorithm. These are just a few examples of how investigating risk-aware versions of standard retrieval statistics like MAP or P20 may be a fruitful direction for future research.

2 Related work

Risk/reward tradeoff curves were introduced by Markowitz [4] as part of his pioneering finance work on portfolio selection. Risk-aware algorithms and analysis methods are well-developed in the computational finance community but we have seen little work in IR fully exploit this connection. The downside risk of query expansion has been noted for decades [6], but only recently has this gotten more extended attention in evaluations. An early version of risk-reward curves was used by the author for query expansion robustness evaluation [3]. The connection between Markowitz-type mean-variance models and risk and reward for retrieval algorithms was first noted in a study that applied this idea to reduce the downside risk of existing query expansion methods [1]. A greatly extended exploration of risk and reward, including extensive refinement and employment of risk-reward curves for evaluation, may be found the author's doctoral dissertation [2]. Recently, a similar mean-variance paradigm was described for document ranking [7]. *Robustness histograms*[1][5], shown in Fig. 2 are another useful evaluation approach that captures variance at a single choice of risk parameter α but not the entire risk profile across all values of α. Precision-recall curves can also present a limited form of risk-reward tradeoff, but assume a binary good/bad label for the objects of interest (e.g. an expanded query), which gives only a crude approximation of variance since it ignores the magnitude of the retrieval failure or result. Risk-reward curves, in contrast, can make more effective distinctions between systems by observing the magnitude of changes in the reward measure and not merely whether gains were positive or negative.

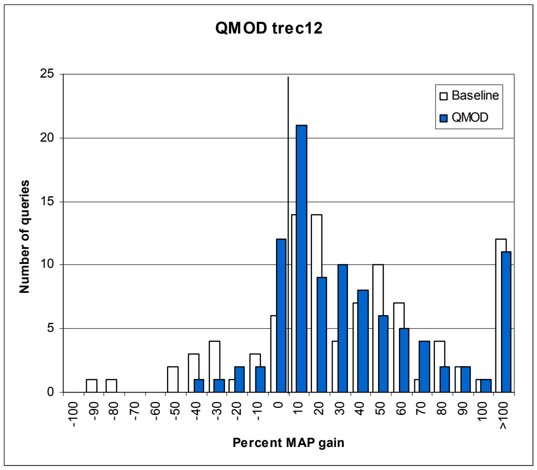


Figure 2: A *robustness histogram*, showing the variance in MAP gain/loss across queries for two different expansion algorithms at a single choice of $\alpha = 0.5$. The 'baseline' expansion method has higher downside variance than the QMOD algorithm [1], as shown by the increased left-hand tail (queries hurt by expansion).

3 Conclusion

We propose the joint analysis of risk and reward behavior for retrieval algorithms using risk-reward curves, which can capture the tradeoff between algorithm risk or variance, and a reward measure such as average-case performance. We believe risk-reward tradeoff curves are a highly useful evaluation method not only for query expansion, but also personalization, document ranking, resource selection and other risk-sensitive scenarios.

4 References

[1] K. Collins-Thompson. Estimating robust query models using convex optimization. In *Advances in NIPS 21*, 2008.

[2] K. Collins-Thompson. *Robust model estimation methods for information retrieval*. PhD thesis, Carnegie Mellon Univ., 2008.

[3] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of SIGIR 2007*, pages 303–310, 2007.

[4] H. M. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[5] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR 2007*, pages 311–318.

[6] A. Smeaton and C. J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[7] J. Wang. Mean-variance analysis: A new document ranking theory in information retrieval. In *ECIR 2009*, pages 4–16, 2009.

Toward Automated Component-Level Evaluation

Allan Hanbury
Information Retrieval Facility
Operngasse 20b
1040 Vienna, Austria
a.hanbury@ir-facility.org

Henning Müller
University of Applied Sciences Western
Switzerland (HES SO)
TechnoArk 3, 3960 Sierre, Switzerland
henning.mueller@sim.hcuge.ch

ABSTRACT

Automated component-level evaluation of information retrieval is discussed. The advantages of such an approach are considered, as well as the requirements for implementing it. Acceptance of such systems by researchers is discussed.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Performance, Measurement

Keywords

Information retrieval evaluation, future benchmarking

1. INTRODUCTION

The majority of information retrieval evaluation campaigns today are run based on the TREC (Text REtrieval Conference) organisation model. This consists of a yearly cycle in which participating groups are sent data and queries by the organisers, and subsequently submit retrieval results obtained by their system for evaluation. The evaluation produces a set of performance measures, quantifying how each participating group's system performed on the queries.

This approach has a number of disadvantages [2]. One of the main disadvantages is the evaluation at system level only. As each system contains many components (e.g. stemmer, tokeniser, feature extractor, indexer), it is difficult to judge the effect of each component on the final result returned for a query. For this reason, when reviewing a number of years of an evaluation task, it is often difficult to go beyond superficial conclusions based on complete system performance and textual descriptions of the systems. Little information on where to concentrate effort so as to best improve results can be obtained. A further disadvantage of the system-level approach, where the result of an evaluation is a ranked list of participants, is the potential to view the evaluation as a competition. This can lead to a focus on tuning systems to the evaluation tasks, rather than the scientific goal of determining how and why systems perform as they do.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

A solution that has been proposed is a component-level evaluation of systems. An example is the MediaMill Challenge [3] in the area of video semantic concept detection. A concept detection system, data and ground truth are provided, where the concept detection system is broken down into feature extraction, fusion and machine learning components. Researchers can replace any of these components with their own components to test the effect on the final results. However, browsing the papers that cite [3] gives the idea that while many researchers make use of the data and ground truth, few use the system framework.

The Grid@CLEF initiative[1] is implementing a component-level evaluation within an evaluation campaign. A basic linear framework consisting of tokeniser, stop list, word decompounder, stemmer and weighting/scoring engine components is specified. Each component should use as input and output XML data in a specified format. This design is an intermediate step between traditional evaluation methodologies and component-based evaluation — participants run their own experiments, but are required to submit intermediate output from each component.

In this paper, we discuss moving towards a fully automated component-level evaluation. Participation in such an evaluation would consist of registering a number of components at a central server for access over the web. The components would then be called as needed for experiments by the server. Such an idea has already been proposed for CBIR in 2001 [1], in which a communication framework (MRML) was specified, and a web server for running the evaluation by communicating in MRML over a specified port was provided. This system did not receive much use.

In the following sections, we discuss the requirements for an automated evaluation system. As use by researchers of the already proposed systems is often lacking, we pay particular attention to the problem of motivating participants.

2. AUTOMATED EVALUATION

The basic framework for a fully automated component-level evaluation framework follows. An information retrieval system built out of a set of components will be specified (as e.g. for Grid@CLEF and the MediaMill Challenge). Participating groups in the evaluation may choose which components they wish to submit. These components should be written so as to run on the participants' computers, callable through a web interface. Participants register their components on a central server. The central server then runs

[1] `http://ims.dei.unipd.it/websites/gridclef/`

the experiments using a large number of combinations of components, accessed through their web interfaces. This approach has the following advantages: (1) A large number of experiments can be done. Each participant makes available online components, which are then called from a central server. This reduces the amount of work for each participant in running complete information retrieval experiments. (2) The best performing combination(s) of components can be identified, where components making up this best performing combination could be from different groups. Different search tasks will also possibly be best performed by different constellations of components. (3) Significantly less emphasis will be placed on the final ranking of complete systems. The results will be in the form of which constellations of which components are best suited for which tasks. This will allow participants to concentrate on developing and improving specific components. It also reduces the perceived competitiveness by removing the ranked list of participants.

An alternative to the web service approach is to require the participants to submit code or executables to the organisers, although this variant leaves the onerous and time-consuming task of system integration to the organisers.

2.1 System Requirements

To create such a system, the following are needed:

- Software and a central server to run the evaluation.
- Protocols for interfacing with programs over the web, exchanging data and exchanging results.
- As for any IR evaluation: large amounts of data, realistic queries and relevance judgements.

The protocol design is the key challenge. The participants' task will shift from performing the experiments to adapting their code to conform to the protocols. In order to make this attractive to participants, the protocols should be designed to have the following properties:

Stability: The protocols should be comprehensively designed to change little over time — After an initial effort to get their systems compliant, little further "interface work" would have to be done by participants.

Simplicity: The initial effort by participants to get their systems compliant should not be high, as a large initial hurdle could discourage participation. In addition to a specification, code implementing key interface components should be provided.

Wide Applicability: Implementing the protocols should enable groups to achieve more than participation in a single evaluation campaign. Standardising the protocols for different evaluation campaigns and potentially for other uses is therefore important.

These properties can be contradictory. For example, a stable protocol that covers all possible eventualities is less simple. Wide applicability can be obtained through the use of a common web service protocol, however many of these protocols do not meet the requirement for simplicity.

For the control software, as the amount of participation increases and the number of components included in the IR system specification increases, the potential number of component combinations will explode. It will therefore not be feasible to test all possible combinations. Algorithms for selecting potentially good component combinations based on previous experimental results and the processing speeds of components, but with low probability of missing good combinations, will have to be designed. Further difficulties to be considered are the remote processing of large amounts of data, where participants with slower Internet connections may be disadvantaged (an initial solution may be to continue distributing the data to be installed locally). It will also have to be considered how to ensure that participants with less computing capacity are not at a disadvantage.

A current problem in IR evaluation that is not addressed at all in this framework is the provision of sufficient data, queries and relevance judgements. With the potential for more efficient experiments, this problem might become worse.

2.2 Participation

It is important to design the system so that it is accepted and used by the targeted researchers. The system should be designed so that there are clear benefits to be obtained by using it, even though an initial effort is required to adopt it. These benefits should be made clear through a "publicity campaign". Potential benefits include: more extensive experimental results on component performance, the opportunity for each research group to concentrate on research and development of those components matching their expertise, and the reuse of components by other researchers to build a working system. It is expected that web service-based systems will become common and thus many researchers might have an interest in such an interface anyway. With having other research group's components available, the building of systems can become easier.

3. LONG-TERM CONSIDERATIONS

Given the additional experimental data that will become available through such a framework, a long-term aim can be to design a search engine that can be built from components based on the task that a user is carrying out and analysis of his/her behaviour (targeted search, browsing, etc.).

The problem of obtaining a sufficient number of queries and relevance judgements in order to allow large scale experiments should be considered. Innovative approaches to harnessing Internet users for continuously increasing the number of relevance judgements should be examined, such as games with a purpose [5], or remunerated tasks [4].

4. REFERENCES

[1] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[2] S. Robertson. On the history of evaluation in IR. *Journal of Information Science*, 34(4):439–456, 2008.

[3] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. ACM Multimedia*, pages 421–430, 2006.

[4] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Proc. CVPR Workshop on Internet Vision*, 2008.

[5] L. von Ahn. Games with a purpose. *IEEE Computer Magazine*, pages 96–98, June 2006.

Building a Test Collection for Evaluating Search Result Diversity: A Preliminary Study

Hua Liu[1]* , Ruihua Song[2,3], Jian-Yun Nie[4], Ji-Rong Wen[3]
[1]Xi'an Jiao Tong University, [2]Shanghai Jiaotong University
[3]Microsoft Research Asia, [4]University of Montreal
doudoulh@gmail.com, {rsong,jrwen}@microsoft.com, nie@iro.umontreal.ca

ABSTRACT

Users often issue vague queries. When we cannot predict users' intentions, a natural solution is to improve user satisfaction by diversifying search results. Such an area, usually called "result diversification", lacks a systematic approach to construct a test collection, by which we can evaluate how search systems perform. In this paper, we propose leveraging the user contributed data in Wikipedia[1] to build up a test collection for ambiguous queries. A preliminary experiment shows promising results.

1. INTRODUCTION

Queries issued by Web users often have multiple meanings or intentions. For such queries, it is important for search engines to retrieve documents covering different requirements. Sanderson [2] has surveyed previous research work on ambiguity and the effort taken to diversify search results. Although there is a long history of research on addressing ranking problems for ambiguous queries, little work done to build test collections has hampered research of this type. This motivates us to construct a test collection that has ambiguous topics and a range of relevance judgments with regard to more than one interpretation.

It is challenging to sample representative ambiguous queries and enumerate their different intentions. First, a set of ambiguous queries proposed by a few people tend to be biased by individual experiences. Second, it is costly to sample ambiguous queries from query logs manually because it is difficult for humans to judge whether a query is ambiguous. Third, even if we have ambiguous queries sampled, there are still difficulties in listing all major intentions of a query.

Fortunately, thousands of people contribute a huge amount of knowledge to Wikipedia. For an ambiguous entry, Wikipedia provides a disambiguation page to allow users to choose their interested interpretations. We propose the idea of leveraging such data to sample queries, pool documents, and labeling the intentions that a document is relevant to. In a preliminary experiment, we build a test collection containing 50 representative queries for evaluating result diversification.

2. BUILDING A TEST COLLECTION

In general, an IR test collection is comprised of queries, documents, and judgments for query document pairs. For ambiguous queries, the intentions that a document is relevant to are also required for evaluating diversity. In this section, we describe how we leverage Wikipedia to achieve these goals.

2.1 Sampling Queries

We make use of disambiguation pages to identify ambiguous entries as Sanderson does in [2]. Then we filter the ambiguous entries from Wikipedia by checking whether it is in a half-a-year query log from a commercial search engine. This is to make it sure that our sampled ambiguous entries are real web queries. Finally, we obtain 38,606 candidate queries.

By observing the candidate queries, we find some ambiguous queries have more diverse meanings than others. For example, "TREC"[2] refers to Text Retrieval Conference, Texas Real Estate Commission, the Trans-Mediterranean Renewable Energy Cooperation, etc., which are quite different from each other. In contrast, "A Beautiful Mind"[3] tends to have more similar meanings, such as A Beautiful Mind (book), A Beautiful Mind (film), and A Beautiful Mind (soundtrack).

Therefore, to compose a set of representative ambiguous queries, we propose sampling the queries with different levels of Similarity of Intentions (SI). For each distinct meaning of an ambiguous query Q, denoted as QM_1, QM_2, …, QM_n, we use their corresponding Wikipedia entry pages Wiki(QM_1), Wiki(QM_2), …, Wiki(QM_n) to calculate SI as the average of cosine similarities between pairs of pages:

$$SI(Q) = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} \text{cos\_sim}(\text{Wiki}(QM_i), \text{Wiki}(QM_j))}{n \cdot (n-1)/2}$$

where, $SI(Q)$ is in the range of 0 to 1. The larger $SI(Q)$ is, the less diverse meanings the ambiguous query Q covers.

We calculate SI for all the candidate queries and show the distribution in Figure 1. Among 38,606 ambiguous queries, 7,454 queries have SI values less than 1.0×10^{-8}, which means these ambiguous queries have quite distinct intentions. For example, "TREC" is in this group. Different from "TREC", "A Beautiful Mind" gets a medium SI value because its interpretations are related to each other. Furthermore, some examples of the queries with high SI values are "dream" (0.0835), "Hercules" (0.0509), "David Copperfield" (0.0441) and "Saint Mary's" (0.0295).

*Work was done when the author was visiting Microsoft Research Asia

[1] www.wikipedia.org

[2] http://en.wikipedia.org/wiki/TREC

[3] http://en.wikipedia.org/wiki/A_beautiful_mind

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

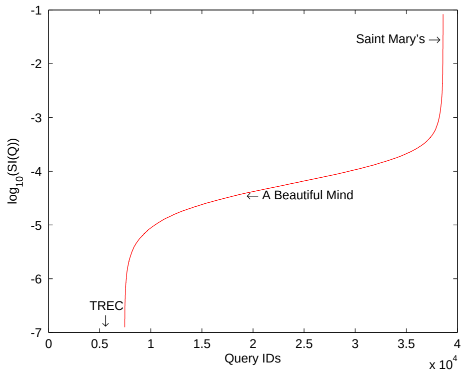


Figure 1: Distribution of $\log_{10}(SI(Q))$ among the candidate ambiguous queries

In our test collection, we randomly select 30 ambiguous queries with low SI values, 10 queries with medium SI values, and 10 queries with high SI values.

2.2 Pooling Documents

An ambiguous query alone may be not enough to retrieve the documents that are relevant to its main intentions, because some unpopular meanings may be overwhelmed by the documents on popular meanings. Thus, we create additional queries that are related to the different meanings in Wikipedia. For example, in terms of the meanings at the disambiguation page of “A Beautiful Mind”, we create three additional queries: “A Beautiful Mind book”, “A Beautiful Mind film”, and “A Beautiful Mind soundtrack”. Then we submit the query and its additional queries respectively to two commercial search engines and retrieve the top 20 returned documents for each query. Finally, by merging the retrieved documents and removing duplicates, we make a pool of documents for each sampled query.

2.3 Labeling Relevance and Topics

To evaluate result diversification, we develop a labeling tool to judge whether a document is relevant to a query as well as which main intentions the page covers. The frame on the right displays the page with keywords highlighted. On the left questionnaire frame, an annotator can mark a page as “Not Found”, if the page fails to be loaded; or “Irrelevant”, which means the page content is not relevant to the query at all; or “Relevant”, which means the page content is relevant to the query. If “Relevant” is clicked, the annotator is also asked to choose one or more relevant intentions from a list of “candidate intentions” that are extracted from the Wikipedia disambiguation page. In addition, the annotator is allowed to input other intentions that are not covered by the list if necessary.

3. EXPERIMENTS

We set up a test collection of 50 queries in a preliminary experiment. On average, there are 5.98 intentions provided and about 213 pages judged per query. In the labeled data, annotators input new interpretations for only about 3.45% of pages. This indicates that the candidate intentions from Wikipedia can cover the meanings of ambiguous queries well. In addition, annotators select multiple intentions for 7.1 pages per query on average. Most of the pages come from dictionary-type websites, such as thefreedictionary.com and britannica.com. These websites usually have a page that shows all the meanings of an ambiguous query.

Table 1: Evaluating two search engines by using a test collection containing 50 ambiguous queries

	MAP-IA@3		MAP-IA@10	
	SE1	SE2	SE1	SE2
Low	0.401	**0.422**	0.427	**0.448**
Medium	**0.335**	0.296	**0.383**	0.337
High	**0.471**	0.437	**0.484**	0.463
All	0.402	0.400	0.429	0.429

We evaluate the performance of result diversification of two commercial search engines by using the test collection. To preserve anonymity, we refer to them as SE1 and SE2. MAP-IA proposed in [1] is used as the measure. Results are shown in Table 1.

We can see that there is no significant difference between two search engines in terms of the overall MAP-IA. However, when looking closely into different types of queries, we find the two engines are obviously different: SE2 outperforms SE1 on the ambiguous queries with clearly different intentions, whereas it performs worse than SE1 on the queries with medium and high SI values. This verifies that query sampling strategies do affect evaluation results.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we present a simple approach to build a test collection by leveraging disambiguation pages from Wikipedia. First, Similarity of Intentions (SI) is proposed to measure how different the meanings of an ambiguous query are. Based on SI, we can sample the representative queries with different properties. Second, in pooling documents, we expand an ambiguous query by additional queries from the disambiguation page. Third, we design a labeling tool that allows annotators to judge both relevance and the topics that a document is relevant to. A preliminary experiment shows that our proposed approach is feasible to construct a test collection for evaluating search result diversity.

In this preliminary study, we use Similarity of Intentions to measure how diverse the intentions of an ambiguous query are. However, there are some alternative measures, such as the number of intentions and the number of categories. Our next step is to investigate the methods and compare their performance in sampling representative queries. In addition, the set of 50 queries is too small to infer statistically sound conclusions. Is it possible to construct a large-scale dataset with minimal human effort? For example, can we label only a few documents and then employ supervised learning approaches to learn classifiers and get more labeled documents further? These interesting research problems await our future research work.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[2] M. Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR*, pages 499–506, 2008.

Are Evaluation Metrics Identical With Binary Judgements?

Milad Shokouhi Emine Yilmaz Nick Craswell Stephen Robertson
Microsoft Research Cambridge
{milads,eminey,nickcr,ser}@microsoft.com

ABSTRACT

Many information retrieval (IR) metrics are top-heavy, and some even have parameters for adjusting their discount curve. By choosing the right metric and parameters, the experimenter can arrive at a discount curve that is appropriate for their setting. However, in many cases changing the discount curve may not change the outcome of an experiment. This poster considers query-level directional agreement between DCG, AP, P@10, RBP($p = 0.5$) and RBP($p = 0.8$), in the case of binary relevance judgments. Results show that directional disagreements are rare, for both top-10 and top-1000 rankings. In many cases we considered, a change of discount is likely to have no effect on experimental outcomes.

1. INTRODUCTION

In the field of information retrieval, many different evaluation metrics have been proposed and used. Each of these metrics is believed to evaluate different aspects of retrieval effectiveness. Hence, much research has been devoted to identifying what constitutes a good metric and which metric to use for evaluation [1, 2].

Since users care more about the top end of the ranking, most evaluation metrics employ a discount function that aims at modelling how much users care about each ranking. The discount functions used by some metrics are parametric, and different methods have been used to learn the optimal values of these parameters [5, 6].

In this poster, we consider four of the most commonly used metrics in IR, precision at 10 (P@10), DCG [3], rank biased precision (RBP) [4], and average precision (AP). When the binary versions of these metrics are considered, the difference between these metrics is the discount function.

Precision at 10 (P@10), for example, assumes that users equally care for the top 10 documents and completely ignore the rest of the ranking.

Even though the discount function used in DCG [3] is not completely specified, most commonly used discounts are the $\frac{1}{\log_b(r+1)}$ (b specified depending on the persistence of the user) and the Zipfian $1/r$ discount, where r is the rank at which document is retrieved.

RBP assumes that the users scan the ranked list of documents from top to bottom and at each step may continue scanning the ranked list with some probability p or stop with probability $1-p$. Hence, the discount function used by RBP follows a geometric distribution.

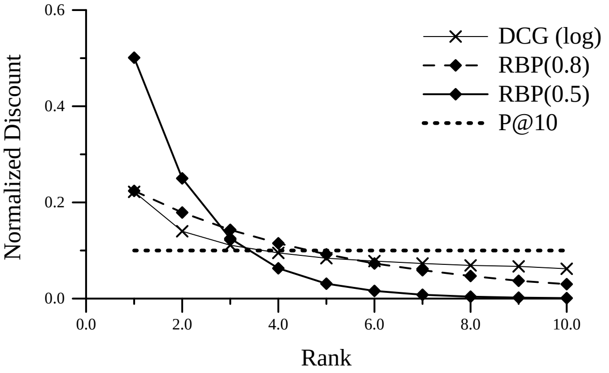


Figure 1: Normalized discount functions for different evaluation metrics. The discount function for AP is adaptive and not shown.

Average precision is defined as the average of the precisions at relevant documents. Therefore, discount function used by AP is adaptive; i.e., the discount of a document retrieved at rank r depends on the relevance of documents retrieved above rank r.

Figure 1 depicts the discount functions of different evaluation metrics. The numbers are normalized so that the area under each curve is equal to one. That is, each data point represents the importance of each rank according to a discount function. Even though these metrics seem quite different when the discount function is considered, what is important for evaluation purposes is whether they agree with each other on the relative quality of two different ranked lists. In this poster, we focus on the case of binary relevance judgments and we analyze whether the difference in the discount functions leads to different conclusions on the relative quality of rankings. In particular, we show that especially when real rankings are considered, most metrics agree on what is a better ranking. We conclude that using different discount functions (i.e., different evaluation metrics) actually leads to similar outcomes when judgments are binary.

2. EXPERIMENTS AND ANALYSIS

We measure the *agreement* rates of different metrics, by comparing their pair-wise preferences for various pairs of rankings. For a given pair of metrics M_a and M_b, and a given pair of ranked lists l_i and l_j, the metrics are in *agreement* if they both prefer the same list. That is, if $M_a(l_i) > M_a(l_j)$,

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

Table 1: The agreement rate between different metrics over several ranked lists (with different distribution of relevant and nonrelevant documents). In the *uniform* experiments, all ranked lists are considered equally likely. In the *sampled* experiments, the likelihood of ranked lists are approximated by using the previous TREC runs. Parameter N denotes the size of the ranked lists, and Δ is the *Fuzziness* value.

Metric Pairs	uniform	sampled	uniform	sampled	uniform	sampled	uniform	sampled
	$N = 10, \Delta = 0$		$N = 1000, \Delta = 0$		$N = 10, \Delta = 0.01$		$N = 1000, \Delta = 0.01$	
DCG/AP	0.95	0.98	0.90	0.92	0.96	0.99	0.99	0.99
DCG/P@10	0.75	0.92	0.54	0.70	0.93	0.97	0.76	0.81
DCG/RBP(0.5)	0.83	0.93	0.61	0.71	0.84	0.94	0.69	0.76
DCG/RBP(0.8)	0.94	0.98	0.64	0.74	0.96	0.98	0.72	0.78
P@10/AP	0.76	0.92	0.51	0.76	0.94	0.98	0.73	0.91
RBP(0.5)/AP	0.82	0.93	0.56	0.76	0.83	0.94	0.63	0.87
RBP(0.5)/P@10	0.60	0.86	0.59	0.81	0.77	0.92	0.76	0.90
RBP(0.8)/AP	0.94	0.98	0.59	0.80	0.96	0.98	0.67	0.89
RBP(0.8)/P@10	0.72	0.90	0.71	0.87	0.90	0.96	0.88	0.96
RBP(0.8)/RBP(0.5)	0.85	0.94	0.85	0.93	0.87	0.95	0.86	0.93

then $M_b(l_i) > M_b(l_j)$ and vice versa. In our experiments, we compare AP, DCG (logarithmic discount), P@10 and two variants of RBP with $p \in \{0.5, 0.8\}$. We use binary judgements for relevance, and consider the top-N documents in rankings to measure the agreement rates ($N \in \{10, 1000\}$).

The first five columns in Table 1, include the pairs of metrics, and their agreement rates for short ($N = 10$), and long ($N = 1000$) ranked lists. For short rankings ($N = 10$), there is a total possible of $\binom{1024}{2}$ ranking pairs that can be generated by varying the number of relevant documents in the top N. For each of these possible permutations, we calculate the value of each metric on both lists, and measure the ratio of *inter-metric* agreement accordingly. The agreement ratios computed this way, assume *uniform* likelihood for each pair of ranked lists. However, IR systems are biased towards returning more relevant documents on top of the ranked lists. Therefore, we also report the *sampled* version of agreement rates, by approximating the likelihood of each ranking according to previous TREC runs.[1] For long ranked lists ($N = 1000$), it is not feasible to try all the possible $\binom{2^{1000}}{2}$ permutations. Therefore, we generated about 5×10^7 random ranking pairs, where the probability of visiting a relevant document at each position is always 0.5.

The numbers in Table 1 suggest strong agreement rates between all the tested metrics for $N = 10$. In general, P@10 has the lowest agreement with the other metrics, which is not surprising given its shallow cutoff. For $N = 1000$, P@10 shows higher agreement rates with the two variants of RBP. This can be explained by aggressive discount function of RBP (Figure 1) that does not noticeably reward relevant documents at lower ranks. Furthermore, the agreement rates between the sampled lists are consistently higher than the uniform sample case. This shows that when metrics disagree, the disagreement is usually between the lists that are unlikely to appear in practice, and metrics mostly agree on the relative quality of reasonable ranked lists.

Fuzziness value (Δ). Buckley and Voorhees [2], defined the *Fuzziness value*, as the "the percentage difference between scores such that if the difference is smaller than the fuzziness value the two scores are deemed equivalent". The last four columns in Table 1 include the results for $\Delta = 0.01$. Here, the metrics M_1 and M_2 are in *disagreement* for a ranking pair l_i, l_j, *iff* (a) they prefer opposite rankings, and (b) $|M_a(l_i) - M_a(l_j)| > \Delta$, and $|M_b(l_i) - M_b(l_j)| > \Delta$. As was expected, employing a fuzziness threshold consistently boosts the agreement rates across all experiments.[2]

3. CONCLUSIONS

We compared four of most commonly used evaluation metrics in information retrieval over millions of pairs of ranked lists. When all lists are considered equally likely, the metrics may look different than each other. However, in reality, not all lists are equally likely. In most cases, the probability of relevance decreases by rank. In order to identify whether metrics are different when reasonable ranked lists are considered, we used TREC runs to approximate the likelihood of each ranked list. When such a background distribution is employed, metrics seem highly correlated with each other, substantially more than uniform scenario. The agreement increases further by considering even small fuzziness intervals (e.g. $\Delta = 0.01$), to the extent that many metrics become almost identical (e.g. AP versus DCG). This suggests that most metrics agree on reasonable lists, and the most disagreements between metrics are only on the lists that are very unlikely to be real search results.

4. REFERENCES

[1] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proceedings of the ACM SIGIR conference*, pages 27–34, Salvador, Brazil, 2005.

[2] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the ACM SIGIR conference*, pages 33–40, Athens, Greece, 2000.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[4] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008.

[5] L. A. Park and Y. Zhang. On the distribution of user persistence for rank-biased precision. In *Proceedings of the Twelfth Australian Document Computing Symposium*, pages 17–24, Melbourne, Australia, 2007. School of CS and IT, RMIT University.

[6] K. Zhou, H. Zha, G.-R. Xue, and Y. Yu. Learning the gain values and discount factors of dcg. In *Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments.*, Singapore, 2008.

[1] We employed all the runs submitted to TREC7 and TREC8 ad hoc tracks. In total, there were 232 systems, each returned rankings for 50 queries.

[2] Increasing the value of Δ leads to further increase in agreement ratios. For example, when $\Delta = 0.05$ and $N = 10$, the agreement is always above 95%, except for RBP(0.5)/P@10 where it is 92%.

Enhanced Web Retrieval Task

M. S. Ali
University of Toronto
sali@cs.toronto.edu

Mariano P. Consens
University of Toronto
consens@cs.toronto.edu

ABSTRACT

This paper presents the Enhanced Web Retrieval Task to model how enhanced web search engines serve the information needs of users. To evaluate the task, we model enhanced results as trees that users navigate to locate relevant information and we propose suitable measures.

1. INTRODUCTION

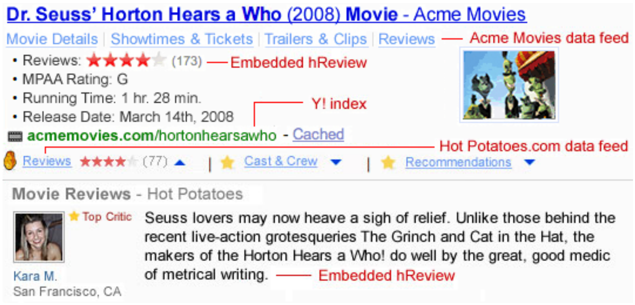
State-of-the-art commercial web search engines retrieve links to web pages annotated with information facets such as a text summary of key phrases in the page [5], folksonomic tags that categorize the page or site [8], links to relevant related pages [7], semantic web relationships to retrieve reviews and ratings [10], and other information to inform or entice users to review sponsored content [4]. These are aggregated search results [9] where the search engine retrieves a main link which is annotated with information facets from other sources. We refer to these type of results as *enhanced results*. Enhanced results have been shown to improve the accuracy of search results [7, 3], and improve user satisfaction of systems [6, 4].

Enhanced results are typically composed of information retrieved from across pages and sites on the web. In this paper, we propose that this retrieval paradigm can be represented as the retrieval of trees of information from the web. In the next section, we present an example and show how trees provide a basis for this paradigm. In Section 3, we propose the Enhanced Web Retrieval Task and conclude in Section 4.

2. ENHANCED RESULTS

Figure 1 shows an enhanced result from an example online movie search application based on the Yahoo! Search Monkey service [2] (the recently announced Google Rich Snippets provides a similar service). The presentation of the example result includes the main link to a retrieved movie, a summary description with details of the movie, embedded reviews of it (`hReview`'s), supporting links to provide the user with show times and ticketing information, and opinion ratings of the movie from other people on the web.

This example demonstrates how an enhanced result can satisfy the information need of users who pose the same query but have very different needs. In this work, we can model the enhanced result as a set of web links. The example includes 8 links to pages on the web; (1) more details about the movie, (2) show times and ticketing information, (3) trailers and video clips for the movie, (4,5) links to two different sites where the movie was reviewed, (6) a link to see the cast and crew who made the movie, (7) a link to recommendations to see other similar movies, and (8) the main link to the web page of the movie. It should be noted that enhanced results (such as in Figure 1) may not be optimal for all users.



Dr. Seuss' Horton Hears a Who (2008) Movie - Acme Movies

Movie Details | Showtimes & Tickets | Trailers & Clips | Reviews

- Reviews: ★★★★☆ (173)
- MPAA Rating: G
- Running Time: 1 hr. 28 min.
- Release Date: March 14th, 2008

acmemovies.com/hortonhearsawho - Cached

Reviews ★★★★☆ (77) | ★ Cast & Crew | ★ Recommendations

Movie Reviews - Hot Potatoes

★ Top Critic

Kara M.
San Francisco, CA

Seuss lovers may now heave a sigh of relief. Unlike those behind the recent live-action grotesqueries The Grinch and Cat in the Hat, the makers of the Horton Hears a Who! do well by the great, good medic of metrical writing.

Figure 1: Enhanced web search result

The effectiveness of the search engine can be measured via inferring classical precision-recall based on the click-through rates mined from weblogs of the main link to the movie [3, 10], inferred relevance of the different information facets from click-through rates on them mined from weblogs [4], and user studies to determine user satisfaction of the retrieved information and its presentation [6]. Researchers have also considered how search results help users locate relevant information on the web via navigation. This has led to the need to also evaluate issues such as redundancy and the effort that users expend to navigate [3, 4, 7].

It is challenging to evaluate enhanced results because each facet of a result can be assessed as to whether it represents relevant information for the user. In addition, the amalgam of the facets can be assessed to determine whether they together represent a relevant answer to the user. Moreover, the web is a vast, non-homogeneous collection that spans the gamut of human knowledge in a format that is not neatly organized. The number of possible combinations of facets in a result makes it impractical to utilize pooling without introducing system bias into assessments. For instance, if two search engines retrieve the same answer but use different facets to enhance the primary part of the answer (i.e., the main link), then should this affect the relative performance

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

measure of these systems? We contend that it should affect performance based on how users navigate to locate relevant information from enhanced results.

We propose to model the retrieval of enhanced search results as trees of information from the web that are used to form a single answer that is structured analogously to a sitemap of the relevant links across the web. A sitemap is typically a single web page in a web site that contains a set of links to the pertinent pages of general interest to the audience of the website.

3. ENHANCED WEB RETRIEVAL TASK

We define the Enhanced Web Retrieval Task as the retrieval of a ranked list of trees of information where each contains a main link and ancillary links that answer a priori known facets of the users' information need(s). An effective system for this task helps the user to navigate to different parts of an answer that are interspersed across the web.

Tree retrieval has been proposed in [1] as a search task for retrieving trees of information from structured documents (such as XML). A key differentiator of tree retrieval from other ad-hoc structured retrieval paradigms (such as passage or element retrieval) is that the purpose of the tree is meant to improve how users navigate to relevant information and to improve how complex information (such as, in this case, enhanced results) can be encoded. Specifically, in [1], it is noted that *the task of returning trees to satisfy an information need builds on a more complex notion of relevance that extends beyond the classical content-based criterion. The relevance of a tree depends on both its content and its context. Tree retrieval involves not only finding relevant information, but also finding trees that afford users access to this information.*

For instance, the result shown in Figure 1 can be represented as a tree as shown in Figure 2. The representation of a movie in Figure 2 suggests that the user seeks a single answer to combine information facets such as whether the movie is highly rated, how to go and see the movie, and to find details that might further entice the user to go see the movie such as who are the stars in the cast. In general, any movie retrieved from the web could be encoded in this way. Enhanced Web Retrieval provides a general way to consider the retrieval of enhanced results, particularly, as in this case, where search is embedded into a focused task (such as searching the web for movies).

Tree retrieval provides a basis for representing this search task, but important questions remain. The most significant is the question of the user's information need given enhanced results. Preliminary work in aggregated search [9] addresses issues such as defining the user's core information need, aggregating information from multiple sources, presenting enhanced results, and exploring how users will interact with systems that retrieve enhanced results. In short, the key challenge will be to assess the relevance of complex enhanced results in a way that is practical and effective.

We propose the following steps to evaluate Enhanced Web Retrieval Task. First, determine a suitable way to infer relevance from web query logs [3, 10]. Second, adapt evaluation measures that consider relevance and user navigation such as structural relevance [1]. Third, utilize appropriate user navigation models, such as user browsing graphs [7].

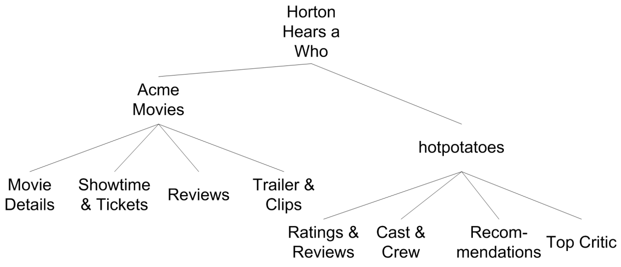


Figure 2: Tree model for enhanced result

4. CONCLUSION

To our knowledge, the Enhanced Web Retrieval Task outlined above is the first proposal in the literature for modelling enhanced search results. It can be applied to numerous, active areas in web IR including semantic relationships, opinions, sponsored content (i.e., advertising), geo-spatially localized results, personalization of search, and multilingual support in search results. A user study should be conducted to determine users' information needs and to validate whether users consider enhanced results as trees.

5. REFERENCES

[1] M. S. Ali, M. P. Consens, G. Kazai, and M. Lalmas. Structural relevance: a common basis for the evaluation of structured document retrieval. In *CIKM '08*, pages 1153–1162, 2008.

[2] R. Baeza-Yates, M. Ciaramita, P. Mika, and H. Zaragoza. Towards semantic search. In *NLDB '08*, pages 4–11. 2008. (See also `http://developer.yahoo.com/searchmonkey`).

[3] M. Bilenko and R. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *WWW '08*, pages 51–60, 2008.

[4] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *WWW '08*, pages 61–70, 2008.

[5] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *SIGIR '99*, pages 121–128, 1999.

[6] M.-Y. Kan, K. McKeown, and J. Klavans. Domain-specific informative and indicative summarization for information retrieval. In *DUC '01*, pages 19–26, 2001.

[7] Y. Liu, M. Zhang, S. Ma, and L. Ru. User browsing graph: Structure, evolution and application. In *WSDM 2009 (Late Breaking-Results)*, 2009.

[8] M. Melenhorst, M. Grootveld, M. van Setten, and M. Veenstra. Tag-based information retrieval of video content. In *UXTV '08*, pages 31–40, 2008.

[9] V. Murdock and M. Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2):80–83, 2008.

[10] U. Shah, T. Finin, A. Joshi, R. Cost, and J. Matfield. Information retrieval on the semantic web. In *CIKM '02*, pages 461–468, 2002.

Towards Information Retrieval Evaluation over Web Archives

Miguel Costa
Foundation for National Scientific Computing
Lisboa, Portugal
miguel.costa@fccn.pt

Mário Silva
University of Lisbon, Faculty of Sciences
LaSIGE
Lisboa, Portugal
mjs@di.fc.ul.pt

ABSTRACT

We present the first overview of a web archive user profile and the searching technology that supports it. Most web archives only support URL search and just a few provide full-text search in response to users' expectations. Their technology is essentially based on web search engines, which ignore the temporal dimension of collections. As consequence, the quality of results is poor. We suggest the creation of an initiative for information retrieval evaluation, meeting the needs of web archives. We believe this initiative would foster research in web archives, in resemblance with what other initiatives achieved in their domains.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation, Measurement, Design

Keywords

web archives, ranking, evaluation

1. INTRODUCTION

All kinds of information are published on the world wide web. Part of this information is unique and historically valuable. However, since the web is too dynamic, a large amount of information is lost everyday. Several initiatives started to archive parts of the web, mainly to preserve their web heritage (see `http://www.nla.gov.au/padi/topics/92.html`). The Internet Archive is the most ambitious initiative with 150 billion documents archived since 1996. As time passes, more and more documents will be archived and their historic interest increased with age. These collections of web data offer a great potential to understand the past, but that requires the development of mechanisms to access this information in areas so diverse as sociology, history, anthropology, culture, politics or journalism.

The prevalent access in web archives is based on the search over automatically extracted metadata from web documents, specially their URLs. A URL search returns a list of the versions of that URL chronologically ordered, such as in the Internet Archive's Wayback Machine (see `http://www.archive.org/web/web.php`). However, the requirement of the user having to know the URL limits its use. A web archiving user survey indicates that full-text search is the most desired web archive functionality [6]. Users expect an interface similar to the one offered by web search engines. In conformity with this idea, a few web archives have implemented full-text search. However, all are based on the Lucene search engine, which is the core of NutchWAX, an extension of the Nutch search engine with Web Archive eXtensions. All the institutions managing these web archives are members of the International Internet Preservation Consortium (IIPC), which has the goal of aggregating efforts to produce common tools and standards. This explains the convergence to NutchWAX and was also the reason for us to adopt it in the developing of the Portuguese web archive [3]. We have indexed until now more than 200 million documents. To the best of our knowledge, the Internet Archive performed the largest indexing over parts of its collection that have close to a billion documents.

This general tendency of adapting web search engines technology to provide full-text search for web archives raises several questions. Does the technology provide good results? Cohen et al. showed that the out-of-the-box Lucene produces low quality results, a MAP of 0.154, which is less than half when compared with the best systems participating in the TREC Terabyte track [2]. We believe that the specific characteristics of web archive collections that are not handled by Lucene, degrade even more the quality of results. Being time present in all the processes and foreseen solutions over a web archive, shouldn't time be present in the ranking model to provide better results for the users? If so, which combination of temporal attributes should be used: the crawl date, creation date, last-modified date or temporal expressions extracted from text with the help of NLP and information extraction technology? Temporal information retrieval (IR) uses temporal data embedded in documents and queries, implicitly or explicitly, to improve search results. Can the rich time-based characteristics of web archive collections be explored with temporal IR? Can we take advantage from the several versions of a document or from the evolution of its links? How should the results of successive crawls from the web be fused? How many versions of a document should be returned to the user? All these questions and others require a dedicated testbed to be studied.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

2. USERS' INFORMATION NEEDS

A clear understanding of what users search is fundamental for the development of web archives search functionalities and to evaluate their performance. A shallow analysis over the top queries at PANDORA's web archive (see `http://pandora.nla.gov.au/search-trends/`) indicates that web archive queries are short like web search engines queries, which contain on average around 2 terms [4]. Unexpectedly, there isn't almost any mention to dates or temporal expressions in web archive queries. This is in conformity with Nunes et al. analysis over the AOL logs that showed that only 1.5% of the queries mention temporal expressions [5]. Our preliminary experiments with users using the Portuguese web archive revealed that they also type short queries without temporal expressions. This may be due to the dominant use of web search engines that today influences the way how users search in other systems. On the other hand, users sometimes use a date range filter incorporated in the interface to narrow the search to a specific period. This filter exists in most web archives and in some cases serves to disambiguate queries. For instance, searching for 'Iraq war' can return documents about three different wars occurring in different periods. When the documents were published during each war, the 'Iraq war' query identified unequivocally the conflict. With the accumulation of all these documents, the query is insufficient to do so.

Users try to find specific pages to see them as they were published in the past. Sometimes they browse their archived versions after that to see for instance, the oldest or youngest version. This search for specific pages is a navigational need. Users also search information about a topic, such as in a topic distillation task. The difference is that web archive users want to see what was known and written about the topic in the past, recreating an historical period. For instance, a user can find what political leaders said about the invasion of Iraq led by the U.S. when it happened in 2003.

Besides navigational and informational queries, Broder classified another query type as transactional, when the query intent is to obtain a resource available via the web (e.g. download a file or buy a product) [1]. Despite the fact that this type is significant in web search engines, we did not detect transactional queries submitted by web archive users. One of the reasons why this occurred is that the web services supporting products purchasing are mostly discontinued when trying to access these services through archived pages. However, we envision that users will use web archives to download old files, for instance, an old manual.

3. TEST COLLECTION

Web archive collections are distinct due to their temporal dimension, so time must be present in the criteria to select the test collection elements: corpus, topics and relevance judgments. The corpus should follow the same diversity of subjects, literary styles and lengths, the same heterogeneity of formats and contents, and a similar word, language and link distribution. Web archives crawl and store different snapshots of the web from different periods. Some crawls are selective, for instance focusing in one sub-domain or topic (e.g. elections). These snapshots are narrower but deeper, trying to crawl all about the topic. More general snapshots, such as country codes top-level domains (e.g. pt), are wider, but more shallow. Another aspect is that some documents, such as newspapers, have a higher change rate, while others, such as scientific articles, tend to be static for long periods. Due to this heterogeneity in crawling frequency, the number of versions of a document can be highly variable. The versions can be very similar or even duplicates, while others are totally different. These characteristics affect the ranking algorithms. For instance, link-based algorithms such as PageRank would have to handle more sparse and versioned web graphs derived from these collections.

The topics must reflect the web archive users' information needs, as described in Section 2. Despite simplistic, the general web archive user profile portrays the user performing navigational or informational queries, some times restricted with a date range or a domain name. We are presently preparing a user survey and a study over the query logs to understand this profile better. We believe that there are at least two types of users: the casual user, whose behaviour and expectations are those of a web search engine user, and the researcher, who needs to explore a topic exhaustively over a timeline. We also want to understand the taxonomy and distribution of the various types of queries to see how different they are from the web search engines queries, analyse the search trends and all critical aspects to engineer effective searching systems and representative test sets.

4. CONCLUSION

The technology used to enable search in web archives provides unsatisfactory results to web search engines and was never evaluated over web archives. Time is the main feature of web archive collections and is completely ignored. Other problems were also raised in this paper that require investigation. Being IR mostly an empirical discipline, joint evaluation initiatives are undeniably important to foster IR research and technology. The elaboration of an initiative towards the evaluation of IR over web archive collections, seems like the natural next step to study the search technology under a set of controlled conditions. It is essential to demonstrate the superior effectiveness and robustness of some retrieval approaches and to produce sustainable knowledge for future development cycles.

5. REFERENCES

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] D. Cohen, E. Amitay, and D. Carmel. Lucene and Juru at Trec 2007: 1-million queries track. In *Proc. of the 16th Text REtrieval Conference*, 2007.

[3] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the Portuguese web archive initiative. In *Proc. of the 8th International Web Archiving Workshop*, 2008.

[4] B. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[5] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. In *Proc. of the Advances in Information Retrieval, 30th European Conference on IR Research*, pages 580–584, 2008.

[6] M. Ras and S. van Bussel. Web archiving user survey. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek), 2007.

Building Pseudo-Desktop Collections

Jinyoung Kim and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
{jykim,croft}@cs.umass.edu

ABSTRACT

Research on the desktop search has been constrained by the lack of reusable test collections. This led to a high entry barrier for new researchers and difficulty in the comparative evaluation of existing methods. To address this point, we introduce a method for creating reusable pseudo-desktop collections by gathering documents and generating queries that have similar characteristics to actual collections. Our method involves a new query generation method and a technique for evaluating the similarity of generated queries with user-generated queries.

Categories and Subject Descriptors

H.4 [**Database Management**]; D.2.8 [**Information Storage and Retrieval**]: [Information Search and Retrieval]

Keywords

Desktop Search, Test Collection Generation

1. INTRODUCTION

Although desktop search plays an important role in personal information management, past research has been limited by the lack of availability of shareable test collections. For instance, desktop search prototypes such as Stuff I've Seen [2] and Connections [4] employ evaluation methods based on real users' desktop collections and queries. Based on actual use cases, this type of evaluation is certainly valuable. Yet this approach requires a fully functional desktop search engine and the lack of reusability makes it difficult, if not impossible, to repeat experiments and make comparisons to alternative search techniques.

In this paper, we suggest a methodology for automatically building reusable pseudo-desktop collections, consisting of document gathering and query generation. The resulting collections have many of the characteristics of typical desktop collections and, importantly, are free from the privacy concerns that are common with personal data.

While we cannot claim that a generated test collection is an ideal substitute for a real desktop environment with actual user queries, we tried to make the collection generation procedure as realistic as possible, and verify the validity of the resulting test collection for retrieval experiments by comparison to actual instances of desktops and user queries.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

2. GENERATING A PSEUDO-DESKTOP

2.1 Collecting Documents

As a first step, we need a collection of documents that has the characteristics of a typical desktop. The criteria that we used for the documents in a desktop were that the documents should be related to a particular person, there should be of a variety of document types, the different document types should have metadata or fields. The privacy of the target individual was another concern.

Given these conditions, our choice of a document collection method was to focus on people mentioned in the email collection from the TREC Enterprise track (crawl of the W3C website) and fetch a variety of publicly-available documents on the web related to those people. More details will be provided in Section 3.1.

2.2 Generating Known-Item Queries

Azzopardi et al. [1] suggested a set of methods for generating a known-item query in a multilingual web collection by algorithmically selecting a set of terms from a target document, based on a observation that an user may formulate query by taking whatever terms she can remember from the document.

However, since we assume that a user's querying behavior would be somewhat different in desktop search, we adapted their generation method by incorporating the selection of fields in the generation process, which results in the following algorithm:

```
1. Initialize an empty query $q = ()$ and select the query length $s$ with probability $P_{length}(s)$
2. Select document $d_i$ to be the known-item with probability $P_{doc}(d_i)$
3. Repeat $s$ times:
   3-1. Select the field $f_j \in d_i$ with probability $P_{field}(f_j)$
   3-2. Select the term $t_k$ from field language model of $f_j$ $P_{term}(t_k|f_j)$ and add $t_k$ to the query $q$
4. Record $d_k$ and $q$ to define a known-item/query pair
```

The only step added here is step 3.1, where we choose the field from which the query term is selected. We call this modification field-based generation method to contrast with document-based generation method suggested in previous work [1]. For P_{term}, we use random selection, TF-based selection, IDF-based selection and TF*IDF-based selection, as suggested in Azzopardi et al. [1].

Table 1: Number and average length of documents for each pseudo-desktop collection.

Type	Jack	Tom	Kate
email	6067 (555)	6930 (558)	1669 (935)
html	953 (3554)	950 (3098)	957 (3995)
pdf	1025 (8024)	1008 (8699)	1004 (10278)
doc	938 (6394)	984 (7374)	940 (7828)
ppt	905 (1808)	911 (1801)	729 (1859)

2.3 Evaluating Equivalence to Manual Queries

Azzopardi et al. [1] introduced the notion of predictive and replicative validity to show that generated queries are equivalent to hand-built queries. Predictive validity means whether the data (e.g., query terms) produced by the model is similar to real queries, while replicative validity indicates the similarity in terms of the output (e.g., retrieval scores).

2.3.1 Verifying Predictive Validity

In verifying predictive validity, we need to evaluate how close the generated queries are to hand-built queries. While previous work [1] introduced only the idea of predictive validity, we suggest using the generation probability $P_{term}(Q)$ of the manual query Q with the term distribution P_{term} from the given query generation method, as follows:

$$P_{term}(Q) = \prod_{q_i \in Q} P_{term}(q_i) \qquad (1)$$

For document-based query generation method [1], we can just use the simple maximum-likelihood estimates for each word. For the field-based query generation method, since every field has different P_{term}, we need to take the linear interpolation of P_{term} for all fields.

2.3.2 Verifying Replicative Validity

Azzopardi et al. [1] measured replicative validity by the two-sided Kolmogorov-Smirnov test (KS-test) using the score samples of real and generated queries as input. Since KS-test determines whether two samples are from the same distribution, we can conclude that two distributions are equivalent if resulting p-value is greater than a certain threshold.

3. EXPERIMENTS

3.1 Building a Pseudo-desktop Collection

As described in Section 2, we built each pseudo-desktop collection so that it may contain typical file types in desktop like *email*, webpage (*html*) and office document (*pdf*, *doc* and *ppt*) related to specific individuals. Table 1 lists the statistics from the resulting pseudo-desktop collections corresponding to three pseudo-users – "Jack", "Tom" and "Kate".

To get the emails related to a person, we filtered the W3C mailing list collection where the name occurrence of each person was tagged, which enabled us to identify several individuals whose activities in W3C were prominent. For other document types, using the Yahoo! search API with the combination of name, organization and speciality (provided by TREC expert search track) of each pseudo-user as query words, we collected up to 1,000 documents for each individual and document type, which roughly matches the statistics of previously used desktop collections [3].

3.2 Generated Queries

Table 2: P-values of Kolmogorov-Smirnov test for different query generation methods.

Extent	P_{term}	DLM	PRM-S	PRM-D
Document	Uniform	0.068	**0.417**	**0.129**
	TF	0.058	**0.619**	**0.244**
	IDF	0.000	**0.116**	0.003
	TF*IDF	0.000	**0.266**	0.007
Field	Uniform	**0.621**	**0.299**	**0.406**
	TF	**0.456**	**0.207**	**0.605**
	IDF	**0.110**	0.027	0.061
	TF*IDF	**0.227**	0.030	0.066

We generated queries using methods described in Section 2.2 and verified its predictive and replicative validity using three pseudo-desktops each with 50 queries written by three people for random sample of documents. For predictive validity, the field-based generation method showed higher generation probability (-13.7 in log scale) than the document-based generation method (-13.9 in log scale). We also verified the replicative validity using three retrieval models – document query likelihood (DLM), PRM-S [3] and the interpolation of DLM and PRM-S (PRM-D). The result in Table 2 confirms the replicative validity of field-based generation methods, especially when query-terms were selected randomly or based on term frequency. All document-based generation methods show replicative validity only for some of the retrieval models.

4. CONCLUSION

In this paper, we described a method for generating a reusable test collection for desktop search experiments and showed that pseudo-desktop collections generated with the field-based method are valid based on various criteria. For future work, we can refine the generation procedures using more sophisticated query generation models or scale the collection by adding more file types and metadata fields. We are also working on verifying the result in pseudo-desktops with the actual desktops.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-0707801, and NSF grant #IIS-0711348. Any opinions, findings and conclusions or recommendations expressed are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07*, pages 455–462, New York, NY, USA, 2007. ACM.

[2] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03*, pages 72–79, New York, NY, USA, 2003. ACM.

[3] J. Kim and W. B. Croft. *Retreival Experiments using Pseudo-Desktop Collections*. CIIR Technical Report. 2009.

[4] S. Shah, C. A. N. Soules, G. R. Ganger, and B. D. Noble. Using provenance to aid in personal file search. In *ATC'07: 2007 USENIX*, pages 1–14, Berkeley, CA, USA, 2007. USENIX Association.

Evaluating Collaborative Filtering Over Time

Neal Lathia
Dept. of Computer Science
University College London
London, WC1E 6BT, UK
n.lathia@cs.ucl.ac.uk

Stephen Hailes
Dept. of Computer Science
University College London
London, WC1E 6BT, UK
s.hailes@cs.ucl.ac.uk

Licia Capra
Dept. of Computer Science
University College London
London, WC1E 6BT, UK
l.capra@cs.ucl.ac.uk

ABSTRACT

Collaborative Filtering (CF) evaluation centres on accuracy: researchers validate improvements over state of the art algorithms by showing that they reduce the mean error on predicted ratings. However, this evaluation method fails to reflect the reality of deployed recommender systems, which operate algorithms that have to be iteratively updated as new users join the system and more ratings are input. In this work we outline a method for evaluating CF over time, and elaborate on work done exploring the temporal qualities of CF algorithms and recommendations.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: :Information Filtering

General Terms

Algorithms

Keywords

Temporal Collaborative Filtering, Time-Averaged Error

1. INTRODUCTION

Collaborative filtering (CF) [1] fuels the success of online recommender systems; in fact, the benefits of filtering information collaboratively are so compelling that facets of CF are now making their way into search engines [2]. The crux of CF algorithm evaluation has become accuracy [3]: a plethora of research in this field focuses on methods that reduce the error between the *predictions* an algorithm makes of user-ratings and the ratings themselves. In other words, to measure the performance of a CF algorithm, a user-rating dataset is split into training/test sets and error is measured on test set predictions after the algorithm has been fed the training ratings. Improvements are then measured by repeating this process, with the same data and modified algorithms. This methodology is reflected in the ongoing Netflix prize[1]. The use of accuracy in and of itself has been questioned before [4]; however, more importantly, the *method* used to test CF algorithms fails to address an important aspect of recommender systems: time. Deployed recommender systems will be iteratively updated as users input ratings in order to update the recommendations that each user is offered [5]. The underlying rating dataset will grow, and any summary statistics derived from it will be subject to change. Experiments on an unchanging dataset do not reflect the reality of a deployed recommender system, and the effect that users will experience as a result of updated recommendations cannot be explored with any static method.

In this paper, we outline a method for evaluating collaborative filtering over time (Section 2), and elaborate on two aspects of CF: how user-similarity changes with time (Section 2.1) and how the system's time-averaged accuracy fluctuates (Section 2.2). We then argue that a broader range of characteristics of recommendations (beyond mere accuracy) are yet to be investigated, and briefly summarise our current work in this area.

2. TEMPORAL CF

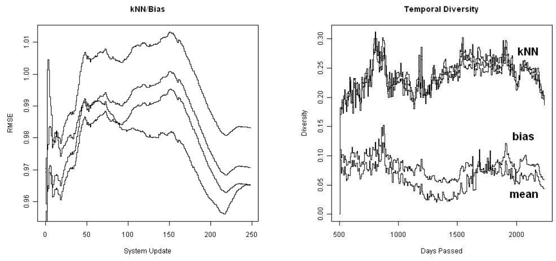
In order to incorporate time into CF experiments, we sort user ratings according to when they were input and then simulate a system that is iteratively updated (every μ days). Beginning at time ($t = \epsilon$), we use all ratings input before ϵ to train the algorithm and test on all ratings input before the next update, at time ($\epsilon + \mu$). We then repeat this process for each time t, incrementing by μ at each step. At each step, what was previously tested on becomes incorporated into the training set; we thus mimick the actual operation of deployed recommender systems by augmenting training sets with ratings in the order that users input them and only testing on ratings that users will make before the next round of recommendation updates. Altering CF experiments in this way highlights a number of hidden characteristics of recommender systems: in the next sections, we briefly summarise some key findings observed to date.

2.1 Similarity Over Time

The basic assumption of CF is that users who have been like-minded in the past are likely to be like-minded in the future. This assumption leads to the intuitive use of the k-Nearest Neighbour (kNN) algorithm for CF [1]: given a user (or item), the ratings of similar users (items) can be used to predict the former's ratings. The focus thus shifts toward the problem of finding *like-minded* neighbours, by measuring the similarity between users or items. In this context, a range of similarity measures have been adopted, including the Pearson Correlation, Cosine Similarity, and many oth-

[1] http://www.netflixprize.com

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.



(a) TA-RMSE (b) Temporal Diversity

Figure 1: Time-Averaged RMSE Of a Series of CF Algorithms and Temporal Top-N Diversity of Netflix Data

ers. However, once similarity is examined on the temporal scale, there is no guarantee that users who were measurably similar at a previous update will continue to be deemed similar. In [6], we found that the similarity between pairs of users (and thus likelihood that they repeatedly be each other's kNN neighbours) highly fluctuates over time, and depends more on *how* similarity is being measured rather than *what* the users are rating. In other words, CF algorithms do not necessarily reflect their founding assumption in the way that they manipulate data over time.

2.2 Accuracy Over Time

To measure the temporal accuracy of a CF system that is iteratively updated, we applied the time-averaged root mean squared error (TA-RMSE) metric. If we define R_t as the set of predictions made up to time t, then the time-averaged error is simply the RMSE achieved between the predictions $\hat{r}_{u,i}$ and ratings $r_{u,i}$ made so far:

$$\text{TA-RMSE}_t = \sqrt{\frac{\sum_{\hat{r}_{u,i} \in R_t}^{N} (\hat{r}_{u,i} - r_{u,i})^2}{|R_t|}} \tag{1}$$

Figure 1(a) shows the TA-RMSE results of the kNN algorithm (with a variety of k values) and Potter's bias model [7] over a sequence of updates on Netflix data subsets. The results highlight that there is no single algorithm that dominates over all others over time. In fact, in [8] we explored how techniques that *improve* accuracy in static experiments actually *degrade* time-averaged accuracy during iterative experiments; furthermore, techniques that produce the best results at the *global* level do not produce similar results when analysing the per-user performance.

2.3 Temporal Recommendations

Observing how CF operates over a sequence of updates also paves the way for exploring a broader range of recommendation characteristics. Given a method of evaluating CF over time, let us focus on the metrics. Since minimal improvements to accuracy bear little meaning to the end user [4], other metrics are worth considering, like temporal diversity. While diversity has been explored in the static case [9], one may be interested in measuring the extent that users are recommended the same items repeatedly over time [10]. To explore this facet of recommendations, we defined the *diversity* between two top-N lists, $L_{u,a}$ and $L_{u,b}$, generated for user u at times a and b, by looking at the proportion of items that appear in both lists, using the Jaccard distance:

$$div(L_{u,a}, L_{u,b}) = 1 - \frac{|L_{u,a} \cap L_{u,b}|}{|L_{u,a} \cup L_{u,b}|} \tag{2}$$

Figure 1(b) plots the temporal diversity in the recommendation rankings when three different algorithms are applied to the Netflix data. From these, we observe that of the methods explored, those that are more *accurate* produce lower *diversity* over time: researchers must therefore question what characteristics they aim to achieve with their recommendations, and prioritise accordingly.

3. CONCLUSIONS

In this paper we have outlined a method for evaluating CF over time, and introduced a number of metrics that relate to temporal evaluations: the time-averaged RMSE measures how prediction accuracy varies over time, while the temporal diversity metric measures the extent that users are being recommended the same items over a number of updates. A number of further metrics are possible. For example, researchers may be interested in the *novelty* of recommendations: how quickly items are recommended after being first rated.

More generally, evaluating any information system requires a notion of what *good* results are: in this work, we argue that an awareness of the temporal nature of recommender systems not only better reflects how CF algorithms are deployed online, but broadens the set of qualities that can be explored when examining the dynamics of recommendations.

4. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE TKDE*, 17(6), June 2005.

[2] J. Pujol, R. Sanguesa, and J. Bermudez. Porqpine: A Distributed And Collaborative Search Engine. In *Proc. 12th WWW*, Budapest, Hungaryt, 2003.

[3] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM TOIS*, 22(1):5–53, 2004.

[4] S. M. McNee, J. Riedl, and J. A. Konstan. Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *Extended Abstracts of the ACM CHI Conference*, Montreal, Canada, 2006.

[5] M. Mull. Characteristics of High-Volume Recommender Systems. In *Proceedings of Recommenders '06*, Bilbao, Spain, September 2006.

[6] N. Lathia, S. Hailes, and L. Capra. kNN CF: A Temporal Social Network. In *Proceedings of Recommender Systems (ACM RecSys '08)*, Lausanne, Switzerland, 2008.

[7] G. Potter. Putting the Collaborator Back Into Collaborative Filtering. In *Proceedings of the* 2^{nd} *Netflix-KDD Workshop*, Las Vegas, USA, August 2008.

[8] N. Lathia, S. Hailes, and L. Capra. Temporal Collaborative Filtering With Adaptive Neighbourhoods. In *Proceedings of ACM SIGIR*, Boston, Massachusetts, 2009.

[9] J. A. Konstan G. Lausen C.N. Ziegler, S. M. McNee. Improving Recommendation Lists Through Topic Diversification. In *Proceedings of WWW 2005*, Chiba, Japan, May 2005.

[10] N. Lathia, S. Hailes, and L. Capra. Tuning Temporal Recommendations With Adaptive Neighbourhoods. In *Under Submission*, June 2009.

How long can you wait for your QA system?

Fernando Llopis
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
llopis@dlsi.ua.es

Alberto Escapa
Departamento de Matemática Aplicada
Escuela Politécnica Superior
University of Alicante, Spain
alberto.escapa@ua.es

Antonio Ferrández
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
antonio@dlsi.ua.es

Sergio Navarro
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
snavarro@dlsi.ua.es

Elisa Noguera
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
elisa@dlsi.ua.es

ABSTRACT

Common approaches to evaluate Question Answering (QA) systems consider exclusively the accuracy of the answers. It ignores an essential feature of all the computational procedures: the efficiency. In this note, we explore new evaluation measures that take into account, in addition to the accuracy, the efficiency, which is incorporated through the magnitude of the answer time of QA systems. In particular, we have developed a family of metrics where the signification of the efficiency can be balanced. By applying this metric to a real time experiment performed in CLEF 2006, it is showed different possibilities to evaluate in a more realistic way the performance of QA systems.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Question Answering, Performance, Evaluation Measures

1. INTRODUCTION

The main evaluation measures used for QA systems are *accuracy*, or some related metrics such as *Mean Reciprocal Rank (MRR)*, *K1 measure* and *Confident Weighted Score (CWS)*. In any case, the answer time of the QA systems is not considered, that is to say, it is neglected the efficiency of the systems. By so doing, we face two main difficulties: some systems can have a good performance being extremely slow in obtaining the right answers; and the comparison among QA systems is not realistic when they had employed different answer times. Therefore, a realist performance analysis requires to take into account the accuracy of the answers and the time needed to obtain them. The aim of this note is to develop a metric that considers these two properties of QA systems, in such a way that the user can balance the dependence of the metric on the efficiency of the system.

2. NEW EVALUATION MEASURES BASED ON ANSWER TIME

One simple possibility to define a metric depending on the accuracy and the efficiency of a system is to associate two real numbers, x and t, to each of these characteristics. Then, we can construct a real function f of two independent real variables and order the systems accordingly the values obtained when evaluating $f(x,t)$. We refer to f as a ranking function, since it allows ranking the different systems depending on their accuracy and answer time. This approach also provides a graphical view of the ordering procedure of the systems through the level curves of f, which we will call iso-ranking curves. Mathematically all the systems that are tied in the classification belong to the same level curve. In the case of accuracy based metrics the level curves are vertical straight lines increasing from left to right, but when the metric also considers the efficiency this is not longer true. We can view an example for one particular metric $MRRT_{E,1}$ (see the next section) in figure 1.

It is important to note that this procedure is of an ordinal type. This means that the relevant information to classify the systems is the relative difference of the numerical values of the ranking function, being meaningless the concrete value of the ranking function for a single system. On the other hand, the ranking functions are not completely arbitrarily but must fulfill some mathematical requirements ([1]).

Within this framework there have been considered different kinds of ranking functions ([1]), in such a way that the efficiency has less weight than the accuracy, since by no means a completely inaccuracy system is preferred over a very efficient one. Anyway, it is possible to modulate the weight of the efficiency in the evaluation of QA systems. To this end, we have introduced a family of ranking functions of the same type controlled by a parameter. By so doing, the value of the parameter could be adjusted in any QA task allowing to design different evaluation measures, accordingly some prefixed criteria. In particular, we have constructed a

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

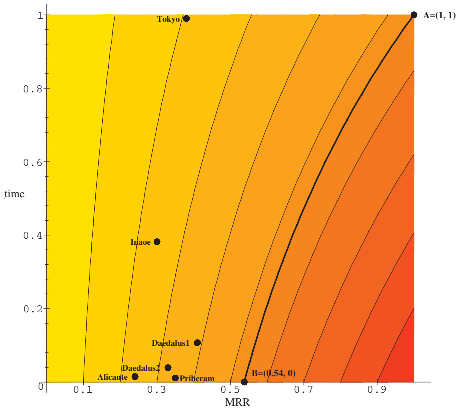


Figure 1: Iso-ranking curves for CLEF-2006 results with metric $MRRT_{E,\,1}$.

family of ranking functions of the form

$$MRRT_{E,\,r}(x,t) = \frac{2x}{1+e^{rt}}. \quad (1)$$

Here, the accuracy of the system x is given the mean reciprocal rank (MRR), so $x \in [0,1]$. The efficiency is measured by considering the answer time of each system, in such a way that a smaller time to answer means a better efficiency of a system. Anyway, to obtain a more suitable scale of representation, we have considered the effective time resulting from dividing the answer time by the maximum answer time obtained in the QA task under consideration, hence we will have that this effective time, denoted as t, belongs to the interval $(0,1]$. Finally, r denotes the parameter that controls the efficiency dependence.

If we take $r = 0$ we recover the MRR measure, which only takes into account the accuracy of the system. In general, the real parameter r can only take values in the interval $[0,+\infty)$. When the value of r increases from 0 to $+\infty$ the weight of the efficiency is also increased. In this way, a ranking function with a small value of the parameter r takes into account very little the efficiency of the systems. This is clear if we observe the functional form of the ranking function family, where the MRR value is multiplied by a function that only depends on time and always take positive values equal or smaller than 1. For higher values of r the value of MRR is more and more penalized as the time grows up.

3. DISCUSSION

Next, we analyze an application of the above designed metric to a real evaluation scenery. In accordance with CLEF organization, we carried out a pilot task at CLEF-2006 whose aim was to evaluate the ability of QA systems to answer within a time constraint, in others words, to consider the efficiency as a relevant part in the evaluation. This experiment followed the same procedure that the main task at QA@CLEF-2006, but the main difference was the consideration of the answer time. The participating groups were: *daedalus* (Spain), *tokyo* (Japan), *priberam* (Portugal), *alicante* (Spain) and *inaoe* (Mexico) (for further information about the realtime experiment see [2]). In table 1, the results of the competition are displayed. We have evaluated the performance of these teams with the uniparametric family of evaluation measures $MRRT_{E,\,r}$. In this way, it is possible to obtain different classification of the systems determined by the values of the parameter r (see table 2). For example, daedalus1 and tokyo obtain the best results of $MRR = MRRT_{E,\,0}$ (0.41 and 0.38 respectively). But, the position of tokyo goes down in the ranking accordingly we increase the values of r, that is to say, when the answer time becomes more important. On the contrary, alicante obtains the worst value of MRR (0.24), as a consequence it is the last one in the ranking if we take only the MRR into account, but it goes up if we increase the parameter r. The teams daedalus1 and priberam do not change practically their position in the ranking, although if we increase the parameter r their values bring near, because priberam has a shorter answer time than daedalus1.

Table 1: CLEF-2006 results

Team	MRR	t (s)	Ef. time
daedalus1	0.41	549	0.10
tokyo	0.38	5141	1.00
priberam	0.35	56	0.01
daedalus2	0.33	198	0.03
inaoe	0.3	1966	0.38
alicante	0.24	76	0.02

Table 2: Accuracy-efficiency evaluation

Participant	r=0	r=0.51	r=0.99	r=1.95
daedalus1	**0.41 (1°)**	0.40 (1°)	0.39 (1°)	0.37 (1°)
tokyo	0.38 (2°)	**0.28 (4°)**	**0.19 (6°)**	0.09 (6°)
priberam	0.35 (3°)	**0.35 (2°)**	0.35 (2°)	0.35 (2°)
daedalus2	0.33 (4°)	**0.33 (3°)**	0.32 (3°)	0.32 (3°)
inaoe	0.30 (5°)	0.27 (5°)	0.23 (5°)	0.19 (5°)
alicante	0.24 (6°)	0.24 (6°)	**0.24 (4°)**	0.24 (4°)

Summarizing up, we have proposed a procedure to define different metrics that consider both the accuracy and efficiency of QA systems and that allows to control the weight of the efficiency on the metric. It opens a new line beyond the traditional evaluation paradigm, since efficiency of QA systems should not be longer ignored.

4. ACKNOWLEDGEMENTS

This paper has been partially supported by the Spanish government, project TIN-2006-15265-C06-01, and by the framework of the project QALL-ME, which is a Sixth Framework Research Programme of the European Union (EU), contract number: FP6-IST-033860.

5. REFERENCES

[1] Noguera, E., Llopis, F., Ferrández, A. and Escapa, A. New Measures for Open-Domain Question Answering Evaluation Within a Time Constraint. Lecture Notes in Computer Science, 4629, 540–547, 2007.

[2] Magnini, B., *et al.*: Overview of the CLEF 2006 Multilingual Question Answering Track. In: WORKING NOTES CLEF 2006 Workshop, 2006.

Author Index

ISBN 978-90-814485-1-2

90000 >

9 789081 448512