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.