CiteData: A new multi-faceted dataset for evaluating personalized search performance

Abhay Harpale
Carnegie Mellon University
Pittsburgh, PA - 15217
aharpale@cs.cmu.edu

Yiming Yang
Carnegie Mellon University
Pittsburgh, PA - 15217
yiming@cs.cmu.edu

Siddharth Gopal
Carnegie Mellon University
Pittsburgh, PA - 15217
sgopal1@cs.cmu.edu

Daqing He
University of Pittsburgh
Pittsburgh, PA - 15260
dah44@pitt.edu

Zhen Yue
University of Pittsburgh
Pittsburgh, PA - 15260
zhy18@pitt.edu

ABSTRACT

Personalized search systems have evolved to utilize heterogeneous features including document hyperlinks, category labels in various taxonomies and social tags in addition to free-text of the documents. Consequently, classifiers, PageRank algorithms and Collaborative Filtering methods are often used as intermediate steps in such personalized retrieval systems. Thorough comparative evaluation of such complex systems has been difficult due to the lack of appropriate publicly available datasets that provide such diverse feature sets. To remedy the situation, we have created CiteData, a new dataset for benchmark evaluations of personalized search performance, that will be made publicly accessible. CiteData is a collection of academic articles extracted from CiteULike and CiteSeer repositories, with rich feature sets such as authors, author-affiliations, topic labels, social tags and citation information. We further supplement it with personalized queries and relevance judgments which were obtained from volunteer users. This paper starts with a discussion of the design criteria and characteristics of the CiteData dataset in comparison with current benchmark datasets, followed by a set of task-oriented empirical evaluations of popular algorithms in statistical classification, collaborative filtering and link analysis as intermediate steps for personalized search. Our results show significant performance improvement of personalized approaches, over that of unpersonalized approaches. We also observe that a meta personalized search engine that leverages information from multiple sources of features performs better than algorithms that use only one of the constituent source of features.

Categories and Subject Descriptors

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

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

General Terms

Algorithms,Experimentation,Human Factors

Keywords

Personalization, Search, Evaluation, Dataset, Social Data

1. INTRODUCTION

Personalized search has become an increasingly important topic in IR (information retrieval) research in the recent years. Personalized search systems have been evolved to not only focus on keyword-based search, but also utilize diverse information sources as possible, such as hyperlinks among documents, category labels of documents and queries, social tags, and user preferences in various forms. For example, various Personalized PageRank algorithms [1][2] have been developed for applying link-analysis with user profiles, producing authority-based ranking of documents with respect to each individual user. Topical distribution in user's search history, as another example, has also been used to construct personalized user profiles, and to rank documents based on their topical match to user's interests in addition to keyword-based similarity score with respect to queries [3] [4]. Query categorization has also been studied for improving personalized search performance [5]. Social tagging or *Folksonomy* is another important source of information from which the interests of individual users and groups can be learned using Collaborative Filtering algorithms and utilized in personalized search [6]. It is also possible to use combinations of these strategies to further improve personalized search performance.

While various approaches have been studied for personalized search, comparative evaluation across current methods has been difficult, primarily due to the lack of a common benchmark dataset that offers a rich set of diverse features so that different personalization strategies can be tested and compared in a controlled manner. For example, personalized PageRank algorithms have been compared amongst each other and against non-personalized PageRank algorithms on document collections with hyperlinks (e.g., using the Stanford WebBase [1][2] dataset), but not compared with any methods using social tags or Folksonomy information, because the dataset lacks social tagging information. As another example, the Web Track [7] and Relevance

Feedback Track [8] in TREC are popular evaluation datasets in the information retrieval community; these datasets provide inter-document hyperlinks but lack social tagging information and topical assignment of documents. Similarly, popular text categorization datasets such as RCV1 [11] and Reuters21578 [12] lack relevance judgments and social tagging information, thus they can only be used for comparing classifiers, but are insufficient for evaluating the impact of document categorization on the ultimate goal, i.e., personalized retrieval performance. On the other hand, user preference information is available in popular Collaborative Filtering datasets such as Netflix, EachMovie and MovieLens [13] but these datasets lack textual content, and hyperlinks. Social tagging websites such as Digg, Del.icio.us, and CiteULike provide information about user-generated tags for websites and articles, but lack document categorization information. Personalized search evaluation requires availability of personalized queries and relevance judgments (qrels). Unlike conventional TREC-style relevance judgments, personalized qrels are not provided by a group of annotators, because that will defeat the purpose of a personalized dataset. Personalized relevance of document should be judged only by the user issuing the query, and not by a group of annotators. None of the popular datasets described above provide personalized queries and relevance judgments.

Clearly, having a multi-faceted benchmark dataset is crucial for facilitating personalized retrieval research and evaluation, but current benchmark datasets for evaluations of retrieval, classification, collaborative filtering and social tagging forums do not offer such a solution. To remedy the issue of dataset unavailability, we have created a new dataset which we call CiteData. This dataset was collated from information extracted from the Citeseer and CiteULike websites and supplemented with personalized queries and relevance judgments that we obtained from volunteer users. The dataset will be made publicly available for download at our research website[1]. With this dataset, we showcase the desirable characteristics of benchmark datasets for evaluation of personalized retrieval systems. In this paper, we also present a comparative evaluation of popular personalization strategies that utilize the different facets of CiteData, such as document hyperlinks, category labels and social tags, including variants of Personalized PageRank algorithms, and classification and collaborative filtering methods as intermediate steps in support of multi-faceted personalized search.

The rest of the paper is organized as follows. In Section 2, we present the idea of a multi-faceted wholesome benchmark dataset for comparing complex personalized search systems. Following up on this idea, we describe the creation of the new CiteData dataset. In Section 3, we present the intrinsic analysis of the dataset to describe the annotation statistics. We also present results of a test to ensure the reliability of the annotations for evaluation of search systems. In Section 4, we present the empirical comparison of some of the popular personalized search strategies based on the CiteData dataset. In Section 5, we present ideas for the potential usage of CiteData for other tasks beyond personalized search. Finally, in Section 6, we conclude with directions for future enhancements and potential uses of the CiteData dataset.

[1]http://nyc.lti.cs.cmu.edu/datasets/citedata

2. CITEDATA

As described earlier, none of the existing publicly available benchmark datasets satisfy all the characteristics of a rich personalized search evaluation dataset. However, existing publicly available datasets can be enriched to add all the desired characteristics. It is infeasible to add social tags-based information to existing IR datasets such as TREC, TDT, RCV1 or Reuters21578 because that will require setting up a large scale user-study. Hence, we choose one of the social tagging websites, CiteULike, as the foundation for the creation of the new benchmark collection.

The CiteULike website allows users to bookmark academic articles matching their interests, by associating them with appropriate user-chosen tags. Academic articles are inherently textual in nature, and the citations/references between academic articles are akin to document hyperlinks. Thus, the data extracted from the CiteULike website can readily provide social tags, textual content and document hyperlinks. The two important lacking features are document categorization information and personalized queries and relevance judgments. We add these additional features to the dataset using information sifted from CiteSeer, and from annotations obtained from volunteers, respectively. It is a challenge to obtain near-exhaustive annotations for each user-query for a large repository like CiteULike, as the user will have to judge the relevance of each document. Hence, to facilitate the annotation effort, we instead select a much smaller subset of articles (about 81433 out of 800k) from CiteULike which are constrained to several research areas in Computer Science. This constraint also helped us in inviting a focused group of volunteers, who are graduate researchers in related fields, to maintain the quality of annotations in the dataset. In the rest of this section, we describe the creation of various features of the CiteData dataset.

2.1 Obtaining Document text, meta-data and hyperlinks data from CiteSeer

CiteULike website is publicly editable and consequently suffers from spam contamination, hence unsuitable for extraction of crucial document meta-data such as document text, authors and conference information. As a result, we used an alternative source, CiteSeer, as the canonical source of information about academic articles such as the document abstracts, and meta-data information such as authors and year of publication. CiteSeer is a popular repository of academic articles, majority of which belong to Computer Science research and is widely accepted as a authoritative source for academic publications. Additionally, we also extracted the affiliation of most authors listed in the dataset.

We extracted the citation for each of the academic articles in the dataset to create a graph of academic articles for facilitating research in link-analysis based algorithms such as Personalized PageRank. Figure 1 shows the distribution of in-links in the dataset. It can be seen that the CiteULike link-structure follows a power-law distribution, similar to the nature of in-link distribution of web-pages on the internet. Intuitively, a few seminal and authoritative papers are highly cited by most other papers, just as, on the internet, a few popular websites receive many in-links from most of the other websites.

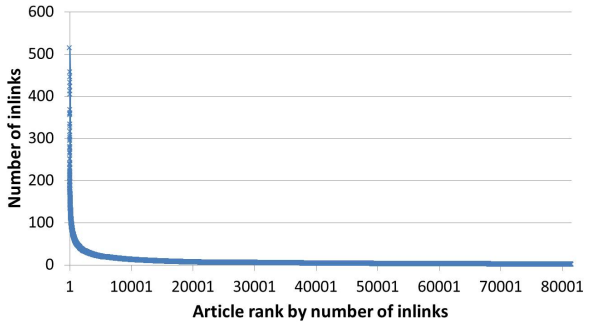

Figure 1: Inlink distribution of the articles in the CiteULike dataset

2.2 Obtaining Social Tagging information from CiteULike

The CiteULike website follows the del.icio.us model of tagging for academic articles. Social tagging information is publicly available for download from their website[2] in a 4-tuple format $< a, u, s, t >$, where t is the tag assigned by user u to an article a at time s.

The data available from the CiteULike website is not directly usable due to spam contamination and automated postings by robots. We have filtered the original dataset to remove spam and automatic postings by setting heuristic selection criteria over what constitutes legitimate users, articles and tags. For example, articles that were bookmarked by less than 4 genuine users were removed, where genuine users are those that have marked more than 4 and less than 500 articles on the website. Such strategies are common in the Collaborative Filtering community for creating usable benchmark evaluation datasets [13]. We could obtain social tagging information for only about 39327 articles (out of 81433) as other articles are not tagged on CiteULike at the moment. We will be updating this information as tags become available for more articles on CiteULike.

2.3 Automatic Document Categorization

In an internet setting, due to the large volume of the web corpus, it is infeasible to obtain true class labels for each web-page. In such a situation, a search engine may solicit labels for a smaller subset of webpages from the users. Alternatively, sample labels may also be extracted from online topic ontologies like the Yahoo topic hierarchy[3] or the Open Directory Project (ODP)[4], that provide a manually labeled taxonomy of websites into user-defined categories. The labels for the remaining documents are usually estimated using automatic classification algorithms.

Along similar lines, in the case of academic articles, we could obtain classification information for only a limited set of academic articles (about 6630 out of the total 81433) from the publicly available Citeseer classification hierarchy[5]. The remaining articles were automatically categorized by training a classifier on the available classification information. We analyzed the performance of several popular text classification algorithms such as K-Nearest Neighbors (KNN), Logistic Regression (LR), and two variants of Support Vector Machines (SVM), namely the linear SVM and polynomial SVM (poly-SVM) of degree 2. CiteData is a multi-labeled dataset, i.e. each document can be assigned to more than one categories. Multi-labeled classification was achieved by using S-Cut [18] thresholding strategy, that discovers optimal thresholds for classifying a document into more than one category, based on the scores that the different classes receive for that document. In Figure 2, we present a comparison of the various algorithms we tried in terms of the Micro-F1 and Macro-F1 classification performance for the multi-labeled dataset. The results in Figure 2 have been averaged over 5-fold cross-validation based runs over the dataset.

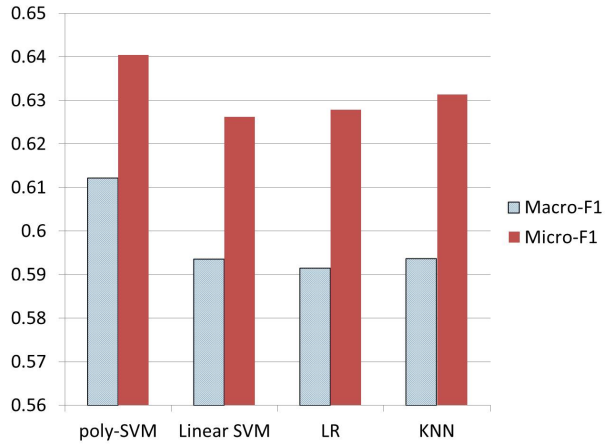

Figure 2: Classification performance of various classifiers on the explicitly labeled subset of the CiteData dataset

Based on the superior performance of the polynomial SVM kernel of degree 2, we have chosen it for classifying the remaining 74803 documents for which we do not have explicit classification information from CiteSeer. We have used the popular implementation of SVM available from the SVM-Light[6] project for this purpose. In Figure 3, we present the distribution of articles per topic in the dataset after the SVM-based categorization step. The categories in the CiteData dataset are listed in Table 1. We observed that on an average, each document has been assigned to 1.3 categories in this multi-labeled task.

2.4 User-tasks, and Personalized Queries and Relevance Judgments

To obtain focused user-tasks and personalized relevance judgments, we solicited experts who can provide such annotations. Our experts consisted of graduate and PhD students who have several years of research experience in the areas of Computer Science and Information Systems. Selecting the right experts for our annotation was not a straightforward task. This is because, on the one hand, we wanted to make sure that the proposed search tasks have enough relevant documents in the collection, and there are similar users in CiteULike who could also be interested in the tasks; on the other hand, we also wanted our experts to de-

[2]http://www.citeulike.org/faq/data.adp
[3]http:/www.yahoo.com
[4]http://dmoz.org
[5]http://citeseer.ist.psu.edu/directory.html
[6]http://svmlight.joachims.org/

Table 1: List of categories available in the CiteULike dataset, sorted by the number of documents in each topic

ID	Category Title
1	Computer Programming
2	Machine Learning and AI
3	Networking and Security
4	Computer Architecture
5	Agents and Applications
6	Computer Theory
7	Databases
8	Human Computer Interaction
9	Digital Libraries
10	Web and Information
11	Natural Language Processing
12	Other research areas in Computer Science

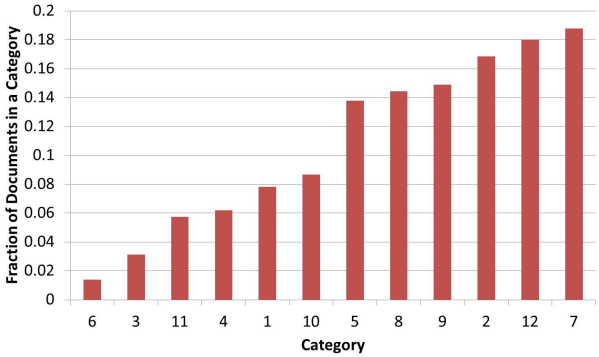

Figure 3: Topic distribution of the CiteData dataset

velop tasks according to their own research interests so that we can make sure that the tasks they developed are valid, genuine, and personalized. To help us identifying the potential candidates, we used the groups information on the CiteULike website. CiteULike allows users to form groups to share articles in common areas of interests. Groups can be very specific such as *Boosting for Support Vector Machines* or very broad such as *Information Retrieval*. First, we used CiteULike groups to identify potential topics that have groups containing at least 10 users and more than 500 articles, to gauge the nature of topical documents available on CiteULike. Then considering the expertise areas of the potential experts to be recruited, we selected those CiteULike groups whose topic fits in the research areas of PhD students in Computer Science and Information Systems. Once the groups and the experts were selected, we asked the experts to describe his/her search task in the form of a *Task statement* according to his/her own expertise. The task descriptions are similar to those available with the TDT 4 [9] dataset. The experts would then search the collection with four to six search queries that are related to their self-designed search task. This controlled study imitates the real-life situation where a computer science researcher sets out to identify interesting papers that are relevant to his/her research problem. Table 2 shows an example search task with corresponding queries and task description.

Table 2: Search Task "Information Network Security"

UserID	network03
Task	Information Network Security
Task 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....
Query1	role based access control
Query2	workflow access control
Query3	authorization delegation
Query4	distributed access control
Query5	XML access control

During the annotation phase, the experts searched for articles using four to six queries to provide relevance judgments. We realized that the CiteULike search engine (on their website) is still in its infancy, and does not retrieve the correct set of documents. This could have affected the annotation process, as the volunteers seldom browse the entire ranked list to judge relevance of each document. To enhance the coverage of annotations obtained from the volunteers, we followed a two-fold strategy. First, by assuming that all documents in the corresponding group(s) could have higher chance to be relevant, the experts were ask to judge each document in the group library and link the relevant documents to each of their queries. The second strategy comes from a well studied annotation strategy prevalent at TREC [8], i.e. pooling [19] from several different search engines to present a wider array of results to the annotators, without biasing towards a particular search engine. We used 7 different retrieval algorithms to generate a pool of articles for each query and ask our experts to annotate the relevance of each article in the pool, and to link each relevant document to a specific query. The 7 algorithms include Indri based retrieval, 3 from the Lemur toolkit, namely, KL-divergence, Okapi and Tf-IDF Cosine based retrieval, and 3 variants of PageRank for link-analysis. (We describe query-specific ranking using PageRank in Section 4). Through this complex annotation process, we built up a comprehensive ground truth annotation for the CiteData test collection.

3. INTRINSIC ANALYSIS OF CITEDATA

3.1 Basic Statistics of the Annotation

To date, we have recruited nine experts who developed nine search tasks across six different CiteULike groups. There are 45 queries associated with these nine search tasks. All these tasks are related to areas of Computer and Information Science, such as Blogging, Computer Networks, Web 2.0, and Information Network Security.

Table 3 shows the statistics of the relevance annotations for each search task. On an average, each search task has

Table 3: Characteristics of various tasks in CiteData (Rel: Relevant)

Task ID	# Queries	# High Rel	# Low Rel	# Not Rel
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

5 queries, the only exceptions are education01 which has only 4 and p2p01 that has 6. The average number of highly relevant documents identified for each task is 125, and that of somewhat relevant documents is 146. But in order to obtain this amount of relevance annotations, our experts annotated 1936 documents on an average.

3.2 Testing the reliability of CiteData as an evaluation dataset

A test collection with good-quality relevance annotation should be reliable as it will be used to predict the effectiveness of retrieval algorithms. We apply the Classical test theory [14] [20] to test the reliability of the CiteData collection. Classical test theory has been widely used in educational field to estimate the reliability of tests, such as standardized college entrance exams. When applying classic test theory to the information retrieval field, we treat each search algorithm as a student facing an exam [14]. In our case, we have 7 retrieval algorithms that could be viewed as 7 students participating this exam of providing relevant articles for queries. In the exam, there are 45 test items (45 queries) and the Mean Average Precision (MAP) score of the 45 queries is the test score for each retrieval algorithms. The reliability coefficient can be estimated by analyzing the variance of individual test items and total test scores. Cronbach's alpha is the best-known measure that can be used to estimate reliability coefficient and is calculated as:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_i \hat{\sigma}_i^2}{\hat{\sigma}^2}\right) \quad (1)$$

where k is the number of items on the exam (45 in this case), $\hat{\sigma}_i^2$ is the estimated variance for item i, and $\hat{\sigma}^2$ is the estimated variance of the total MAP scores. α scores above 0.7 indicate reliable test collections that are effective at comparing performance of various algorithms.

The Cronbach's alpha for CiteData collection is α=0.9717, which is above 0.7, indicating the reliability of the CiteData dataset according to the Classical test theory.

4. EMPIRICAL ANALYSIS OF PERSONALIZED SEARCH ALGORITHMS

As described earlier, unavailability of a rich multi-faceted benchmark evaluation dataset has presented a challenge in comparing personalized search systems that leverage diverse sources of information such as hyperlinks and social tags. In this section, we present one of the first empirical comparisons of such diverse personalized search systems.

4.1 Personalized search by matching user's topical interest to document categories

Some of the earliest personalized search systems present search results that closely match the user's topics of interest. Intuitively, a sports enthusiast is probably searching for sports-related documents, while a stock investor is searching for financial and investment reports. The user's topical interests can be discovered based on the user's search history and bookmarks. For example, the user's topical interests can be discovered based on the documents the user has marked relevant his for past queries.

$$\pi_c^{(u)} = \frac{\#\text{ relevant articles in topic } c \text{ for user } u}{\#\text{ relevant articles for user } u} \quad (2)$$

where, $\pi_c^{(u)}$ denotes the level of interest the user u has in topic $c \in 1, \ldots, C$. Consequently, the user's interest at the document level can be computed as a linear combination of the user's topical distribution based on the categorization of that particular document.

$$d_i^{(u)} = \sum_{c=1}^{C} \pi_c^{(u)} I(d_i, c) \quad (3)$$

where, $d_i^{(u)}$ denotees a measure of the interest of user u in the document d_i. $I(d_i, c)$ is an indicator whether document d_i belongs to the cateogry c. Note that the user-specific $d^{(u)}$ scores are not query sensitive. Query-sensitive personalized scores $\Psi_i^{(u)}$ for a document fd_i can be obtained by combining the user-specific scores $d^{(u)}$ with query-specific retrieval scores q_i. A simple implementation can be a weighted combination of query-specific retrieval scores provided by a search engine like Indri and the corresponding user-specific interest scores for the document as shown below:

$$\Psi_i^{(u)} = q_i + w d_i^{(u)} \quad (4)$$

The approach in Equation 4 has been shown to perform reasonable well in IR literature [15]. In our experiments, we will be referring to this approach as TDS to denote topical distribution based search, where the topical distribution is specific to a particular user, and hence personalized.

4.2 Personalized search by PageRank based link-analysis

Link-analysis based approaches such as PageRank [10] have gained immense academic and commercial interest for discovery of authoritative documents in a collection. The general premise of such algorithms is that authoritative documents are usually highly cited by other documents, and that the users are usually more interested in such authoritative documents than other documents. The PageRank

scores are usually estimated by simulating a random walk over the linked graph of documents, and each document receives scores proportional to the number of times it will be visited if this simulation was carried out infinitely. At each document along the walk, the surfer is faced with a choice: follow a randomly chosen link from the current document or randomly teleport to a document from the collection. Mathematically, it can be expressed as:

$$\vec{r} = (1-\alpha)M\vec{r} + \alpha\vec{t} \tag{5}$$

where the matrix M encodes the transition probability from each page to each of its hyperlinks, and the vector $\vec{t}$ denotes the random teleportation vector. The parameter α, called the dampening factor, is the probability that the random surfer will choose to teleport to a random page, instead of following a link, for which the probability is $(1-\alpha)$. The vector $\vec{r}$ denotes the PageRank scores of each of the articles in the network.

If $\vec{t}$ is a uniform vector, meaning, the user is equally likely to teleport to any page in the network, then we call this approach Global PageRank (GPR) as it is not biased towards a particular user or topic. Some variants of the GPR algorithm are specifically tailored for personalized and topic-sensitive search. Topic-sensitive PageRank (TSPR) ranks documents based on their importance within a particular topic, while Personalized PageRank ranks documents based on their importance to a particular user. For calculation of PPR, the PageRank equation is tweaked to accommodate user-specific preferential treatment visiting documents of interest. This preferential treatment is achieved by replacing the uniform teleportation vector $\vec{t}$ from Equation 5 with a personalized teleportation vector $\vec{t}^{(u)}$ which reflects the users interests in those pages.

$$\vec{r}^{(u)} = (1-\alpha)M\vec{r} + \alpha\vec{t}^{(u)} \tag{6}$$

Several optimization strategies have been proposed for improving the scalability of the personalized approach in Equation 6 to millions of users, a realistic situation on the internet. A popular approach by Jeh et. al. [1] computes the topic sensitive pagerank vectors for a canonical set of topics $c \in 1, \ldots, C$, and then builds personalized pagerank vectors by a linear combination of these TSPR vectors weighted by the user's interest in those particular topics. Mathematically,

$$\vec{r}^{(c)} = (1-\alpha)M\vec{r} + \alpha\vec{t}^{(c)} \tag{7}$$

for each category $c \in 1, \ldots, C$.

$$\vec{r}^{(u)} = \sum_{c=1}^{C} \vec{r}^{(c)}\pi_c^{(u)} \tag{8}$$

In our experiments, we have implemented this variant of PPR. PPR results are query-insensitive. To generate the final rank list, we combined the relevance scores from Indri with those from the respective PageRank algorithms. We have used the linear weighted-log combination approach proposed in [15]. Specifically, we compute the final combined score Ψ_i for each document i retrieved by Indri.

$$\Psi_i^{(u)} = q_i + w \log r_i^{(u)} \tag{9}$$

where q_i is the query-specific relevance score provided by Indri for a document i. $r_i^{(u)}$ is the PageRank score for document i for the personalized PageRank algorithm.

As a comparative unpersonalized baseline that uses link-analysis, we will also compare the performance of GPR by ranking documents based on the weighted combination of GPR and query-specific retrieval scores, similar to Equation 9, by replacing $r_i^{(u)}$ with r_i, i.e. unbiased PageRank.

4.3 Personalized search using Collaborative Filtering over social tags

With the advent of social bookmarking websites such as Digg and Del.icio.us, a new possibility has emerged for discovering users with similar interests and then personalizing search based on the shared interests of users. A user's act of tagging an article depicts an implicit interest of the user in the particular article, because the user is bookmarking the article for later retrieval. Numerous approaches [6] have been proposed for utilizing such social tagging information to improve personalized search performance. Comparison of all such approaches is beyond the scope of this paper and is left for future exploration. Our approach is based on a popular Collaborative Filtering (CF) algorithm called Probabilistic Latent Semantic Analysis (pLSA) [16, 17] that can be used to discover users with similar interests based on the similarity of their tagging patterns and then recommend articles to users based on their shared interests. pLSA is a probabilistic model in which users are considered to be a mixture of multiple interests or aspects. Thus each user $u \in U$ has a probabilistic membership in each of the aspects, $z \in Z$. If m is a binary random variable indicting interest in document d, then the probability of each tuple in the dataset can be computed as follows:

$$P(m|u,d) = \sum_{z \in Z} p(m|d,z)P(z|u) \tag{10}$$

Equation 10 consists of two parts, $p(m|d,z)$ and $P(z|u)$. It can be observed that the first term $p(m|d,z)$ does not depend on the user and represents the aspect-specific model. The second term $P(z|u)$ is the user-personalization term. pLSA works in three steps: 1) discover latent topics that best explain the available data by using an Expectation Maximization approach, i.e. discover $p(m|d,z)$ 2) discover the user's topical interests $P(z|u)$ by matching the items that the user has already shown interest in with the latent topics discovered in step 1. 3) Finally, recommend new items to the user based on scores obtained by Equation 10.

The CF scores $P(m|u,d)$ obtained for each of the documents estimate the user's interest in a particular document. Along the lines of TDS, these CF scores can be combined with query-specific retrieval from a search engine like Indri to create a novel collaborative personalized search solution. In our experiments we will call this approach PCF, which ranks the documents based on the scoring function:

$$\Psi_i^{(u)} = q_i + wP(m|u,d_i) \tag{11}$$

4.4 Meta Personalized Search

It is also possible to combine diverse scores such as TDS, PCF and PPR to generate a meta personalized search engine. Specifically, for each query, the documents can be ranked based on the function:

$$\Psi_i^{(u)} = q_i + w_{\text{ppr}} \log r_i^{(u)} + w_{\text{tds}} d_i^{(u)} + w_{\text{pcf}} P(m|d,u) \quad (12)$$

We call this approach MPS, short for Meta Personalized Search.

4.5 Experimental Setup

For completeness of the exploration, we also evaluate the benefit of using personalized approaches over not using any personalization on the CiteData dataset. Our chosen unpersonalized baselines include query-specific Indri retrieval and GPR, as a representative link-analysis based approach. For the Indri retrieval, we used the default inbuilt #combine operator which mimics a probabilistic OR function of the query terms.

User's topical interest distribution $\pi_c^{(u)}$ is required for the TDS and PPR approaches. As mentioned earlier, this topical interest distribution can be estimated by calculating the topical distribution of documents in the user's search history. To simulate the user's search history, for each test query from the user, we consider documents marked as relevant by that user for other queries as the corresponding search history for that search query.

For each of the weighted combination based approaches (UDIST, PPR, PCF, and MPS), the weights were tuned using 5 fold cross-validation over the user-query pairs. Cross-validation was also used for tuning other parameters such as the dampening factor α for PageRank approaches, and the number of latent factor Z for the pLSA based Collaborative Filtering approach.

4.6 Results

Before presenting the main results comparing all the aforementioned personalized search approaches on a common evaluation benchmark, we would like to present results for the each of the intermediate tasks such as user-distribution estimation, and PageRank computation.

4.6.1 *Quality of user interest distribution estimation*

Table 4 shows the estimated distribution of top topics, estimated according to Equation 2, for a few example users. It can be observed that this simple user interest estimation strategy works reasonably well. It is crucial that this approach work well because it will affect the performance of two of the compared approaches, PPR and TDS.

4.6.2 *Quality of PPR estimation*

As explained earlier in Equation 8, we compute PPR as a linear combination of TSPR vectors weighted by the user's topical interest distribution. Performance of estimating a user's topical interest distribution is evident from the Table 4. To demonstrate the qualitative performance of the second crucial factor for PPR computation, i.e. TSPR, in Table 5 we list the titles of top 5 articles ranked according to TSPR for 3 exemplary topics. It can be observed that TSPR performs quite well by ranking on-topic articles higher in the list.

4.6.3 *Personalized search comparison*

In Figure 4, we compare the performance of the various aforementioned approaches, namely the personalized approaches (TDS, PPR, PCF, MPS) and the unpersonalized approaches (Indri and GPR). It can be observed that there is a significant benefit of using personalized search algorithms on the CiteData dataset, as the representative personalized approaches significantly outperform the chosen unpersonalized approaches. It can also be observed that the combined meta search engine MPS performs better than each of the constituent scoring functions. This is encouraging for future research in identifying methods to leverage information from multiple diverse sources simultaneously for personalized search.

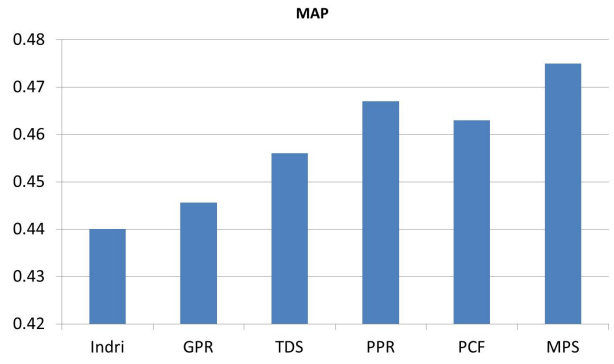

Figure 4: Comparison of representative Personalized and unpersonalized approaches on the CiteData dataset. The evaluations are based on Mean Average Precision (MAP)

5. CITEDATA USAGE BEYOND PERSONALIZED SEARCH EVALUATION

CiteData is a rich dataset with several diverse features and is therefore amenable to evaluations beyond just personalized search. From our experiments, it is evident that certain common tasks such as text classification, collaborative filtering for item recommendation, and link-analysis for discovery of authoritative documents can be evaluated on the collection, without personalized search as the ultimate goal.

CiteData documents provide multiple heterogeneous fields such as authors, hyperlinks, and conference information. Additional fields such as tags associated with each document and information about users interested in a particular document are also available. Owing to this, CiteData can be used to evaluate classification performance of algorithms that can benefit from treating such heterogenous features preferentially or by leveraging relationships between those features. CiteData can also be used for evaluation of content-based Collaborative Filtering algorithms that can leverage additional information about users, and items and combine them in novel ways. For example Basilico et. al [21] learn feature relationship kernels and combine them using a tensor product to improve the task of item recommendation in Collaborative Filtering. CiteData can also be used to evaluate some of the latest Graphical Model approaches such as Correspondence-LDA [22] or Correlated Topic-Models [23] that guide the inference of topic models based on correlation between various features.

Many Collaborative Filtering algorithms [16] [25] rely on the discovery of latent aspects. Such latent aspects typically represent automatically discovered groups of users with

Table 4: Example users with corresponding queries. Also shown are the top estimated topics for each user, and some selected articles that were marked relevant by the user.The figures in parenthesis indicate the distribution $\pi_c^{(u)}$ for that particular topic c and the user u

User	User 1	User 2	User 3
Top Estimated Topic (score)	Networking and Security (0.68)	Human Computer Interaction (0.50)	Networking and Security (0.40)
Queries submitted by the user	SQL injection	blog using behaviour	P2P File-sharing
	logic bomb software attack	wiki usability	P2P Network Theory
	user authentication	wikis collaboration	p2p algorithm
Documents marked relevant by the user	Denial of Service Resilience in Ad Hoc Networks	Software Architecture Analysis of Usability	Making Gnutella-like P2P Systems Scalable
	Toward Acceptable Metrics of Authentication	The Reengineering Wiki	Scalable Application Layer Multicast
	Exchange-based Incentive Mechanisms for Peer-to-Peer File Sharing	Fractal Behaviour Analysis	Building Low-Diameter P2P Networks
	Linear Logic Proof Games and Optimization	Timewarp: Techniques for Autonomous Collaboration	A Peer-To-Peer Approach To Resource Location In Grid Environments

similar interests or groups of items with shared patronage. CiteULike allows users to form user groups and create a corresponding library of articles that are interesting to that group. CiteData data can be easily supplemented with this information that is readily available from the CiteULike website. Availability of such user-group and item-library information can help in two ways. Firstly, it will help compare different CF algorithms directly on the quality of discovered aspects by comparing them to the known groups in the collection. So far, CF algorithms have been compared only on the ultimate goal of item recommendation, and crucial intermediate steps such as latent aspect discovery has not been evaluated explicitly. Directly comparing explicit user groups with discovered aspects may provide more insight into topic discovery methods. This can lead to performance improvements of topic discovery methods and consequently, improve item recommnedation. Secondly, in another research direction, the explicit information about user-groups and item-libraries can also be used to guide the discovery of the latent aspects in a supervised fashion.

CiteData is also suitable for evaluating Adaptive Filternig (AF) [24] algorithms. As the name suggests, AF approaches filter out documents from a stream in an online fashion based on the relevance feedback available from the user on documents that were recommended to the user in the past. The documents in the CiteData dataset can be easily ordered chronologically based on their year of publication for simulating such a document stream.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented CiteData, a new multi-faceted dataset for the primary task of evaluating personalized search. This dataset will help in bridging the evaluation gap between diverse personalized search systems that have so far been compared only their counterparts that use similar sources of information, but never with methods that leverage information from other features. To validate and demonstrate the usability of the dataset we presented an empirical comparison of a rich set of representative personalized search approaches that utilize topic discovery, link-analysis and collaborative filtering. Our experiments show strong evidence for effectively utilizing a rich sources of information for personalized search. Besides personalized search, we also discussed other important potential uses of the CiteData dataset for evaluation of diverse tasks such as classification, topic-discovery, adaptive filtering, content-based collaborative filtering. In the future, we would like to explore approaches for leveraging such heterogeneous features for the aforementioned array of tasks.

7. REFERENCES

[1] Glen Jeh and Jennifer Widom. Scaling Personalized Web Search. In Proceedings of the International World Wide Web Conference, WWW 2003.

[2] Taher Haveliwala. Topic-Sensitive PageRank. In Proceedings of the International World Wide Web Conference, WWW, 2002.

[3] Gauch, S., Chafee, J. and Pretschner, A. (2004). Ontology-based personalized search and browsing. Web Intelligence and Agent Systems, 1(3-4): 219-234

[4] Speretta, M. and Gauch, S. (2004). Personalizing search based on user search history. CIKM Ś04.

[5] Liu, F., Yu, C. and Meng, W. (2002). Personalized Web search by mapping user queries to categories. In Proceedings of CIKM Ś02, 558-565.

[6] Shengliang Xu, Shengua Bao, and Ben Fei. Exploring Folksonomy for Personalized Search. SIGIR 2008

[7] Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. Overview of the TREC 2003 Web Track. TREC Report 2004

[8] Yuanhua Lv and ChengXiang Zhai. A study of Adaptive Relevance Feedback Ű UIUC 2008 Relevance Feedback Experiments. TREC Report 2008

[9] The Linguistic Data Consortium. http://www.ldc.upenn.edu/

Table 5: List of top 5 articles ranked according to their TSPR score for 3 sample topics

Natural Language Processing	Programming	Network and Security
Building a Large Annotated Corpus of English: The Penn Treebank	Revised 4 Report on the Algorithmic Language Scheme	New Directions in Cryptography
A Practical Part-of-Speech Tagger	Automatic Translation of FORTRAN Programs to Vector Form	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems
A Simple Rule-Based Part Of Speech Tagger	A Semantics of Multiple Inheritance	Congestion Avoidance and Control
A Statistical Approach To Machine Translation	Implementing Mathematics with The Nuprl Proof Development System	A Scheme for Real-Time Channel Establishment in Wide-Area Networks
Finding Structure in Time	Interprocedural Slicing Using Dependence Graphs	A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998.

[11] David Lewis, Yiming Yang, T. Rose and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research 5 (2004) 361-397.

[12] Shantanu Godbole, Abhay Harpale, Sunita Sarawagi, and Soumen Chakrabarti. Document Classification through interactive supervision of document and term labels. ECML-PKDD 2004.

[13] Rong Jin and Luo Si. (2004). "A Bayesian Approach toward Active Learning for Collaborative Filtering" In Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence. Banff, Alberta. (UAI)

[14] David Bodoff, Pu Li: Test theory for assessing IR test collections. SIGIR 2007: 367-374

[15] Nick Craswell, Stephen Robertson, Hugo Zaragoza and Michael Taylor. Relevance Weighting for Query Independent Evidence. In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2005.

[16] Hoffman, T. Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis SIGIR 2003

[17] Hofmann, T., and Puzicha, J. Latent class models for collaborative filtering. In Proceedings of International Joint Conference on Artificial Intelligence, 1999.

[18] Yiming Yang. A study on thresholding strategies for text categorization. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), pp 137-145 2001.

[19] 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.

[20] Crocker, L. and J. Algina, Introduction to Classical and Modern Test Theory. 1986: Holt, Rinehart, and Winston

[21] Justin Basilico, and Thomas Hofmann. Unifying Collaborative and Content-based Filtering. Proceedings of the 21 st International Conference on Machine Learning, Banff, Canada, 2004.

[22] David Blei and Michael Jordan. Modeling Annotated Data. SIGIR 2003.

[23] David Blei and John D. Lafferty. Correlated Topic Models. Advances in Neural Information Processing Systems 18 (2005).

[24] Yi Zhang, Jamie Callan and Thomas Minka. Novelty and redundancy detection in adaptive filtering. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002

[25] Si, L. and Jin, R. Flexible Mixture model for collaborative filtering. In Proceedings of the Twentieth International Conference on Machine Learning, 2003

[26] Abhay Harpale, Yiming Yang. Personalized Active Learning for Collaborative Filtering. SIGIR 2008