International Journal of Research

International Journal of Research (IJR)

e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 09, September 2015

Available at http://internationaljournalofresearch.org

Better Metrics to Predict the Performance of Personalized Web Search

Repalle Tejaswi[1]& Md.Asim[2]

[1]Student,Dr.K.V.Subba Reddy College of Engg for Women, Kurnool, Andhra Pradesh

[2]Asst.Professor, Dr.K.V.Subba Reddy College of Engg for Women, Kurnool, A.P

Abstract:

Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. We study privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two greedy algorithms, namely Greedy DP and Greedy IL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that Greedy IL significantly outperforms Greedy DP in terms of efficiency..

Keywords: Privacy protection; personalized web search; utility; risk; profile

1. INTRODUCTION

THE web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query. The solutions to PWS can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods are straightforward— they simply impose bias to clicked pages in the user's query history. Although this strategy has been demonstrated to perform consistently and considerably well [1], it can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. Profile-based methods can be potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances [1]. Although there are pros and cons for both types of PWS techniques, the profile-based PWS has demonstrated more effectiveness in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users, which is usually gathered implicitly from query history [2], [3], [4], browsing history [5], [6], click-through data [7], [8], [1] bookmarks

[9], user documents [2], [10], and so forth. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life. Privacy issues rising from the lack of protection for such data, for instance the AOL query logs scandal [11], not only raise panic among individual users, but also dampen the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services.

2. RELATED WORK

There are several prior attempts on personalizing web search. One approach is to ask users to specify general interests. The user interests are then used to filter search results by checking content similarity between returned web pages and user interests [22, 6]. For example, [6] used ODP2 entries to implement personalized search based on user profiles corresponding to topic vectors from the ODP hierarchy. Unfortunately, studies have also shown that the vast majority of users are reluctant to provide any explicit feedback on search results and their interests [4]. Many later works on personalized web search focused on how to automatically learn user preferences without any user efforts [22, 19, 29, 26]. User profiles are built in the forms of user interest categories or term lists/vectors. In [19], user profiles were represented by a hierarchical category tree based on ODP and corresponding keywords associated with each category. User profiles were automatically learned from search history. In [29], user preferences were built as vectors of distinct terms and constructed by accumulating past preferences, including both long-term and short-term preferences. Tan et al. [31] used the methods of statistical language modeling to mine contextual information from long-term search history. In this paper, user profiles are represented as weighted topic categories, similar with those given in [28, 6, 22], and these profiles are also automatically learned from users' past clicked web pages. Many personalized web search strategies based on hyperlink structure of web have also been investigated. Personalized PageRank, which is a modification of the global Page Rank algorithm, was first proposed for personalized web search in [20]. In [10], multiple Personalized Page Rank scores, one for each main topic of ODP, were used to enable "topic sensitive" web search. Jeh and Widom [14] gave an approach that could scale well with the size of hub vectors to realize personalized search based on Topic-Sensitive PageRank. The authors of [32] extended the well-known HITS algorithm by artificially increasing the authority and hub scores of the pages marked relevant by the user in previous searches. Most recently, [17] developed a method to automatically estimate user hidden inter ests based on Topic- Sensitive PageRank scores of the user's past clicked pages. In most of above personalized search strategies, only the information provided by user himself/herself is used to create user profiles. These are also some strategies which incorporate the preferences of a group of users to accomplish personalized search. In these approaches, the search histories of users who have similar interest with test user are used to refine the search. Collaborative filtering is a typical group-based personalization method and has been used in personalized search in [29] and [30]. In [29], users' profiles can be constructed based on the modified collaborative filtering algorithm [15]. In [30], the authors proposed a novel method CubeSVD to apply personalized web search by analyzing the correlation among users, queries, and web pages contained in click-through data. In this paper, we also introduce a method which incorporates click histories of a group of users to personalize web search. Some people have also found that personalization has variant effectiveness on different queries. For instance, Teevan et al. [34] suggested that not all queries should be handled in he same manner. For less ambiguous queries, current web search ranking might be sufficient and thus personalization is unnecessary. In [6] and [5], test queries were divided into three types: clear queries, semi-ambiguous queries, and ambiguous queries. The authors also concluded that personalization significantly increased output quality for ambiguous and semi-ambiguous queries, but for clear queries, one should prefer common web search. In [31], queries were divided into fresh queries and recurring queries. The authors found that recent history tended to be much more useful than remote history especially for fresh queries while the entire history was helpful for improving the search accuracy of recurring queries. This also gave us a sense that not all queries should be personalized in the same way. These conclusions inspired our detailed analysis.

3. PERSONALIZED SEARCH BASED ON USER PREFERENCE

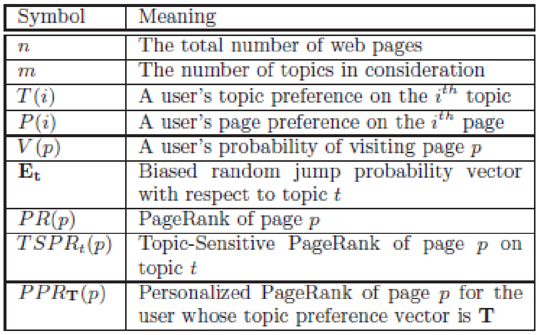
3.1 User Preference Representation

Given the billions of pages available on the web and their diverse subject areas, it is reasonable to assume that an aver-age web user is interested in a limited subset of web pages. In addition, we often observe that a user typically has a small number of topics that she is primarily interested in and her preference to a page is often affected by her general interest in the topic of the page. For example, a physicist who is mainly interested in topics such as science may find a page on video games not very interesting, even if the page is considered to be of high quality by a video-game enthu- siast. Given these observations, we may represent a user's preference at the granularity of either topics or individual web pages as follows:

Definition 1 (Topic Preference Vector) A user's topic preference vector is defined as an m-tuple T = [T(1), . . . , T(m)], in which m is the number of topics in consideration and T(i) represents the user's degree of interest in the ith topic (say, "Computers"). The vecP tor T is normalized such that m i=1 T(i) = 1.

Definition 2 (Page Preference Vector) A user's page preference vector is defined as an n-tuple P = [P(1), . . . , P(n)], in which n is the total number of web pages and P(i) repre-sents the user's degree of interest in the ith page. The vector P is normalized such that Pn i=1 P(i) = 1.

In principle, the page preference vector may capture a user's interest better than the topic preference vector, be-cause her interest is represented in more detail. However, we note that our goal is to learn the user's interest through the analysis of her past click history. Given the billions of pages available on the web, a user can click on only a very small fraction of them (at most hundreds of thousands), making the task of learning the page preference vector very diffi-cult; we have to learn the values of a billion-dimension vector from hundreds of thousands data points, which is bound to be inaccurate. Due to this practical reason, we use the topic preference vector as our representation of user interest in the rest of this paper. We note that the this choice of preference representation is valid only if a user's interest in a page is mainly driven by the topic of the page. We will try to check the validity of this assumption later in the experiment section — even though it is indirect — by measuring the effectiveness of our search personalization method based on topic preference vectors. In Table 1, we summarize the symbols that we use through- out this paper. The meaning of some of the symbols will be clear as we introduce our user model.



Symbol	Meaning
n	The total number of web pages
m	The number of topics in consideration
$T(i)$	A user's topic preference on the i^{th} topic
$P(i)$	A user's page preference on the i^{th} page
$V(p)$	A user's probability of visiting page p
$\mathbf{E_t}$	Biased random jump probability vector with respect to topic t
$PR(p)$	PageRank of page p
$TSPR_t(p)$	Topic-Sensitive PageRank of page p on topic t
$PPR_{\mathbf{T}}(p)$	Personalized PageRank of page p for the user whose topic preference vector is $\mathbf{T}$

Table 1: Symbols used throughout this paper and their meanings

3.2 User Model

To learn the topic preference vector of a user from her past click history, we need to understand how the user's clicks are related to her preference. In this section, we describe our user model that captures this relationship. As a starting point, we first describe the topic-driven random surfer model.

Definition 3 (Topic-Driven Random Surfer Model)

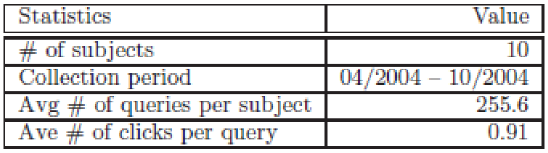
Consider a user with topic preference vector T. Under the topic-driven random surfer model, the user browses the web in a two-step process. First, the user chooses a topic of interest t for the ensuing sequence of random walks with probability T(t) (i.e., her degree of interest in topic t). Then with equal probability, she jumps to one of the pages on topic t (i.e, pages whose Et(p) values are non-zero). Starting from this page, the user then performs a random walk, such that at each step, with probability d, she randomly follows an out-link on the current page; with the remaining probability 1−d she gets bored and picks a new topic of interest for the next sequence of random walks based on T and jumps to a page on the chosen topic. This process is repeated forever.

4. EXPERIMENTS

In this section we discuss various experiments we have done to evaluate our proposed methods and show the results. We first describe our experimental setup in Section 4.1. Then in Section 4.2 we describe a simulation-based experiment to measure the accuracy of our learning method. Fi-nally in Section 4.3 we present the results from our user survey that measures the perceived quality of our personalized ranking method.

4.1 Experimental Setup

In order to apply the three evaluation metrics described in Section 3.5, we need the following three datasets: (1) users' click history, (2) the set of pages that are deemed relevant to the queries that they issue, and (3) the Topic-Sensitive PageRank values for each page. To collect these data, we have contacted 10 subjects in the UCLA Computer Science Department and collected 6 months of their search history by recording all the queries they issued to Google and the search results that they clicked on. Table 2 shows some high-level statistics on this query trace.



Statistics	Value
# of subjects	10
Collection period	04/2004 – 10/2004
Avg # of queries per subject	255.6
Ave # of clicks per query	0.91

Table 2: Statistics on our dataset

To identify the set of pages that are relevant to queries, we carried out a human survey. In this survey, we first picked the most frequent 10 queries in our query trace, and for each query, each of the 10 subjects were shown 10 randomly selected pages in our repository that match the query. Then the subjects were asked to select the pages they found relevant to their own information need for that query. On average 3.1 (out of 10) search results are considered relevant to each query by each user in our survey. Finally, we computed TSPR values from 500 million web pages collected in a large scale crawl in 2005. That is, based on the link structure captured in the snapshot, we computed the original PageRank and the Topic-Sensitive PageRank values for each of the 16 first-level topic listed in the Open Directory. The computation of these values was performed on a workstation equipped with a 2.4GHz Pentium 4 CPU and 1GB of RAM. The computation of 500 million TSPR values for each topic roughly took 10 hours to finish on the workstation.

4.2 Accuracy of Learning Method

In this section we first try to measure the accuracy of our learning method. Here, we are concerned with both the accuracy of our method and the size of the click history necessary for accurate estimation. Even if a user's preference can be learned accurately in principle, it may not be possible in practice if it requires a sample size significantly larger than what we can actually collect. The best way of measuring the accuracy of our method is to estimate the users' topic preferences from the real-life data we have collected, and ask the users how accurate our results are. The problem with this method is that, although users could tell which are the topics they are most interested in, it tends to be very difficult for them to assign an accurate weight to each of these topics. For example, if a user is interested in “Computers” and “News,” is her topic preference vector [0.5, 0.5] or [0.4, 0.6]? This innate inaccuracy in users’ topic preference estimations makes it difficult to investigate the accuracy of our method using real-life data. Thus, we will use a synthetic dataset generated by simulation based on our topic-driven searcher model:

1. Generation of topic preference vector. In our implementation, the number of topics the user is in-terested in is fixed to K as an experimental parameter. Then we randomly choose K topics and assign random weights to each selected topic. The weights for other topics are set to 0. The vector is normalized to sum up to one.

2. Generation of click history. Once we generate a user’s topic preference vector, we generate a sequence of L clicks done by the user using the visit probability distribution dictated by our model

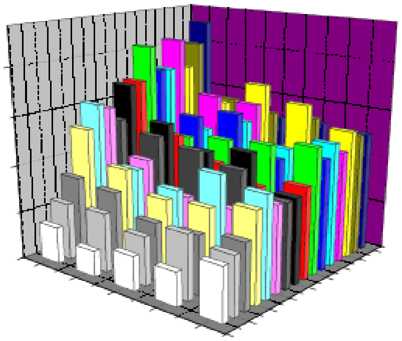


Figure 1: Relative errors in estimated weights

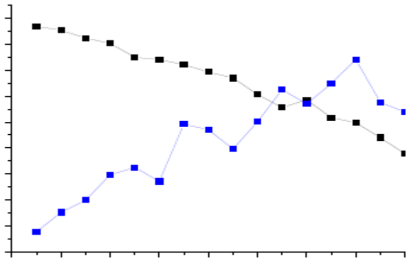


Figure 2: Comparison of relative errors in estimated weights on sample size 100

4.3 Quality of Personalized Search

We now try to measure how much our ersonalization method improves the overall quality of search results based on our user survey. To measure this improvement we compare the following four ranking mechanisms:

- PageRank: Given a query, we rank the pages that match the query based on their global PageRank values.
- Topic-Sensitive PageRank: We rank pages assuming that the user is interested in all topics. That is, we rank pages based on PPRT(p) of Equation 12, but assuming that T(i) = 1 16 for i = 1, . . . , 16. This represents a ranking method that does not take the user’s preference into account.
- Personalized PageRank: We rank pages based on Equation 12 using the estimated topic preference vector, but excluding the second term Pr(q|T(i)). That is, we rank pages by Pm i=1 T(i) · TSPRi(p). This represents a ranking method that uses the user preference, but not the query in identifying the likely topic of the query.
- Query-Biased Personalized PageRank: We rank pages based on Equation 12 without omitting any terms. This represents a ranking method that uses both the user preference and the query to identify the likely topic of the query.

5. RELATEDWORK

Researchers have also proposed ways to personalize web search based on ideas other than

PageRank [16, 17, 18]. For example, [16] extends the well-known HITS algorithm by artificially increasing the authority and hub scores of the pages marked "relevant" by the user in previous searches. [17] explores ways to consider the topic category of a page during ranking using user-specified topics of interest. [18] does a sophisticated analysis on the correlation between users, their queries and search results clicked to model user preferences, but due to the complexity of the analysis, we believe this method is difficult to scale to general search engines. There also exist much research on learning a user's preference from pages she visited [19, 20, 21]. This body of work, however, mainly relies on content analysis of the visited pages, differently from our work. In [19], for example, multiple TF-IDF vectors are generated, each representing the user's interests in one area. In [20] pages visited by the user is categorized by their similarities compared to a set of pre-categorized pages, and user preferences are represented by the topic categories of pages in her browsing history. In [21] the user's preferences are learned from both pages she visited and those visited by users similar to her (collaborative filtering). Our work differs from these studies in that pages are characterized by their Topic-Sensitive PageRank's, which are based on the web link structure. It will be an interesting future work to develop an effective mechanism to combine both the content and the web link structure for personalized search. Finally, Google7 has started a beta-testing of a new personalized search service8, which seems to estimate a searcher's interests from her past queries. Unfortunately, the details on the algorithm is not known at this point.

6. CONCLUSION

In this paper we have proposed a framework to investigatethe problem of personalizing web search based on users' past search histories without user efforts. In particular, we first proposed a user model to formalize users' interests on webpages and correlate them with users' clicks on search results. Then, based on this correlation, we described an intuitive algorithm to actually learn users' interests. Finally, we proposed two different methods, based on different assumeptions on user behaviors, to rank search results based on the user's interests we have learned.

7. REFERENCES

[1] B.J. Jansen, A. Spink, and T Saracevic. Real life, realusers, and real needs: A study and analysis of user queries on the Web. Information Processing and Management, 36(2):207 – 227, 2000.

[2] Robert Krovetz and W. Bruce Croft. Lexical ambiguity and information retrieval. Information Systems, 10(2):115–141, 1992.

[3] Nielsen netratings search engine ratings report.http://searchenginewatch.com/reports/article.php/2156461, 2003.

[4] J. Carroll and M. Rosson. The paradox of the active user. Interfacing Thought: Cognitive Aspects of Human-Computer Interaction, 1987.

[5] T. Haveliwala. Topic-sensitive pagerank. In Proceedings of the Eleventh Int'l World Wide Web Conf., 2002.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In Proc. of WWW '98, 1998.

[7] F. Qiu and J. Cho. Automatic identification of user preferences for personalized search. Technical report, UCLA technical report, 2005.

[8] J. Cho and S. Roy. Impact of Web search engines on page popularity. In Proc. of WWW '04, 2004. [9] Dimitri P. Bertsekas and John N. Tsitsiklis. Introduction to Probability. Athena Scientific, 2002.

[10] M. Kendall and J. Gibbons. Rank Correlation Methods. Edward Arnold, London, 1990.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web.
Technical report, Stanford Digital Library Technologies Project, 1998.

[12] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In Advances in Neural Information Processing Systems 14. MIT Press, 2002.

[13] G. Jeh and J. Widom. Scaling personalized web search. In Proceedings of the Twelfth Int'l World Wide Web Conf., 2003.

[14] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank. Technical report, Stanford University, 2003.

[15] M. Aktas, M. Nacar, and F. Menczer. Personalizing pagerank based on domain profiles. In Proc. Of WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis, 2004.

[16] F. Tanudjaja and L. Mui. Persona: A contextualized and personalized web search. In Proc. of the 35th Annual Hawaii International Conference on System Sciences, 2002.

[17] P. Chirita, W. Nejdl, R. Paiu, and C. Kohlschuetter. Using odp metadata to personalize search. In Proceedings of ACM SIGIR '05, 2005.

[18] J. Sun, H. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: A novel approach to personalized web search. In Proceedings of the Fourteenth Int'l World Wide Web Conf., 2005.

[19] L. Chen and K. Sycara. Webmate: a personal agent for browsing and searching. Proc. 2nd Intl. Conf. on Autonomous Agents and Multiagent Systems, pages 132–139, 1998.

[20] A. Pretschner and S. Gauch. Ontology based personalized search. In ICTAI, pages 391–398, 1999.

[21] K. Sugiyama, K. Hatano, , and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In Proceedings of the Thirteenth Int'l World Wide Web Conf., 2004.