

Personalized web pages ranking algorithm based on user preferences

Zhu Rongbo

(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract: In order to rank searching results according to the user preferences, a new personalized web pages ranking algorithm called PWPR (personalized web page ranking) with the idea of adjusting the ranking scores of web pages in accordance with user preferences is proposed. PWPR assigns the initial weights based on user interests and creates the virtual links and hubs according to user interests. By measuring user click streams, PWPR incrementally reflects users' favors for the personalized ranking. To improve the accuracy of ranking, PWPR also takes collaborative filtering into consideration when the query with similar keywords is submitted by users who have similar user interests. Detailed simulation results and comparison with other algorithms prove that the proposed PWPR can adaptively provide personalized ranking and truly relevant information to user preferences.

Key words: web page; user preference; ranking algorithm; personalization

In order to reduce the time spent on browsing search results, personalized web search^[1-2] schemes are proposed. To provide personalized ranking for search results, one should first extract user interests^[3-4]. However, the drawback of this approach is that an ordinary user is not always able to satisfy the user preferences all the time. Moreover, the user interests may change as time goes on. Research effort, such as PageRank^[3], explores the link structure among web pages, which builds a topic-oriented PageRank and a set of PageRank vectors based on 16 main topics of an open directory project are computed^[5]. The authors in Ref. [3] exploited a personal PageRank vector (PPV) to speed up the calculation of PageRank, where PPV is a personalized view of the importance of pages on the web^[6-8].

Although the research works mentioned above have explored user interest extraction, none of these research works take the personalization into consideration when ranking the Web pages in the search list. Consequently, in this paper, we explore data mining techniques to automatically extract user interests and devise a personalized web pages ranking algorithm called PWPR.

1 Proposed PWPR Algorithm

1.1 PWPR algorithm

The procedure of PWPR (personalized web page ranking) is divided into four steps. The first step performs the regular HITS analysis and re-weight for each page in the search re-

sults. Then, we assign larger weights to those web pages whose categories are similar to user interests. The second step is the virtual hubs and virtual links addition step, which further assigns the weights to those web pages whose categories satisfy the user interests and the query term. Furthermore, user interests may vary with time. In order to capture the change of user interests, algorithm PWPR monitors the user behavior to observe whether the user interests are changing or not in the third step. The final step is the user group history adjustment step in which collaborative filtering is used to generate a new rank recommended by others having similar interests.

1.2 Assigning initial weights

In order to reflect user interests in search results, the initial value of each page should be modified. By the user interest vector UIV_c , the search result cluster RC_c and the query q , we later perform a weight function to find out the relationship between the UIV_c and the RC_c . When the search engine receives a query q , we can find out which cluster may satisfy the user interests. There are some issues involves in determining the cluster. The weight of each category in the search result clusters RC_c is the summation of the similarity to each category in user interests UIV_c . The weight of category i in the search result clusters UIV_c can be defined as

$$weight_{RC_{c,i}} = \sum_{\forall \text{ category } j \text{ in user cluster } UIV_c} \frac{|UIV_{c,j} \cap RC_{c,i}|}{|UIV_{c,j} \cup RC_{c,i}|} \quad (1)$$

where $UIV_{c,j}$ is denoted as category j in the user interests vector UIV_c . After the calculation in each category in RC_c , the initial weight of the page is the weight of the category it belongs to.

1.3 Creating virtual hubs and virtual links

To assign different weights to web pages, we further adjust the scores so as to emphasize the importance of those web pages similar to user interests. Since the link analysis is used in PageRank, we come up with the idea of adding virtual links and virtual hubs go to the pages that are very similar to user interests. When a user submits a query term q , we compute the similarity between q and each category i in the user interests vector UIV_c . Only top k categories similar to query q are selected, where the number of k is a controlled parameter to show the personalized degrees. With the smaller values of k , only very similar categories are chosen. Once we have selected the top k categories, we check whether the search result categories are in the top k categories of user interests. If the search result category is exactly matched with the top k categories, the virtual hubs and links are added to the web pages whose categories belong to $RC_{c,i}$ with the weights of virtual links. The weights of virtual links are set to the values of the categories in the user interest vector.

Received 2008-04-15.

Biography: Zhu Rongbo (1978—), male, doctor, associate professor, rongbozhu@gmail.com.

Foundation item: The Natural Science Foundation of South-Central University for Nationalities (No. YZZ07006).

Citation: Zhu Rongbo. Personalized web pages ranking algorithm based on user preferences[J]. Journal of Southeast University (English Edition), 2008, 24(3): 351 – 353.

1.4 Measuring feedbacks from user click streams

The first two steps of PWPR can be viewed as user long term interest adjustment. But if the user changes his interests, the first two steps cannot make real-time adjustment, so we need a real-time adjustment mechanism to reflect user short term interests. Based on the mechanism, the re-ranking algorithm will not always perform the same prediction mechanism by the feedback of the user. User clicking streams are logged to observe the browsing behavior of users. From the user clicking streams, we can clearly justify whether the search results have fulfilled the requirements of the users or not. When the user submits a query q , it falls into four categories. We log the category which the user clicks the most to represent the query.

When query q is sent to the server, we try to find the query history and find out which category the user really chooses when facing the queries in the same category. The query history can be transformed by a click stream history. Thus, the adjustment function can be designated as

$$AF_p(i) = \frac{n_i \sum_{j \in \text{category } i} \text{weight}(j)}{n \sum_{j=1}^n \text{weight}(j)} \quad (2)$$

1.5 Collaborative adjustment

In order to achieve user group collaborative adjustment, grouping users is the first step. Because each user has own preference, and the preference can be viewed as a point in an m -dimensional space where m is the number of global categories. In the m -dimensional space, each axis represents a different category. The density-based clustering can be used in clustering users. The distance function used in density-based clustering can be the Euclidean distance. If the distance is less than a threshold σ , two users can be in the same cluster. Before calculating the group preferences, we should check the members in the group and remove the members who may change their interests to another category in order to provide more precise ranking. The contribution for a user i in a category j can be defined as

$$\text{contribution}_u(i, j) = \frac{n_{i,q} n_i C_j}{\sum n_{i,q}} \quad (3)$$

where $n_{i,q}$ represents the number of total queries for user i ; $n_i C_j$ represents the number of users i clicking the category j in the query history.

After removing the users with lower contributions, we have to calculate the total contribution value for every category in the query history. The weight of user i with category j can be defined as

$$\text{weight}_c(i, j) = \sum_{j=1}^m n_{i,q} n_i C_j \quad (4)$$

The contribution of the category k can be defined as

$$\text{contribution}_c(k) = \frac{\sum_{i=1}^n n_{i,q} n_i C_j}{\sum_{j=1}^m \sum_{i=1}^n n_{i,q} n_i C_j} \quad (5)$$

where n is the number of selected users, and m is the number of categories.

The weight of a category is based on the weight, and the weight is based on the number of queries and the number of hits. The weight of each category can be saved in the vector Vect_g . We insert two personalized vectors, Vect_p and Vect_g , to compute the personalized ranking for different users and different groups. Vect_p will return the users preferences to the page and Vect_g will return the groups adjustment to the category the page is in. Finally, the PWPR can be written as

$$\text{PWPR}(v) = \text{weight}_{\text{global}} \left[(1 - c) \sum_{u \text{ link to } v} \frac{\text{PWPR}(u)}{O(u)} + c \right] + \text{weight}_p \text{Vect}_p p(v) + \text{weight}_{\text{group}} \text{Vect}_g p(v) \quad (6)$$

where $p(v)$ is m by 1 vector, and m is the number of categories, and the sum of the three parameters satisfies:

$$\text{weight}_{\text{global}} + \text{weight}_p + \text{weight}_{\text{group}} = 1 \quad (7)$$

2 Numerical and Simulation Results

A data set provided by the webKB project^[9] is used to model the web environment. This data set contains five major categories: “Cornell”, “Music”, “Texas”, “Washington” and “Wisconsin”. The user preferences are simulated by randomly selecting the interest of the user and then generating the access patterns randomly. In this simulation we generate four different user preferences. Each user has his user interests randomly generated. In order to measure the ranking performance in the search results, we use coefficients named importance and discrepancy in Ref. [7]. In the simulation, we compare the query results in the same keyword but ranking in different algorithms with PageRank^[21] and VIPAS^[17].

Figs. 1 (a) to (d) show parts of the simulation results. Here the results are computed by three different algorithms: PageRank (PR), VIPAS, and PWPR, and the importance of each category is the average of each page in the category. In order to obtain precise results, we ignore the page that has the lowest value in each category. By ignoring these pages, we can obtain more important pages. According to the results, we can easily observe the difference between the regular PageRank and the proposed PWPR. In PWPR, it is clear that the pages that users are interested in have higher ranks and they can satisfy user preferences. The results in Figs. 1 (a) to (d) also show the change of rank score in the proposed PWPR, which proves that the category that the user is interested in has a higher importance than that of PR. The reason is that, to improve the accuracy of ranking, the PWPR takes collaborative filtering into consideration when the query with similar keywords is submitted by users having similar user interests. In VIPAS, the importance of the category that the user is interested in is higher than the importance of the category in PWPR. It is because VIPAS just adds weights to the pages. While in the proposed PWPR, user click streams are measured, and it incrementally reflects users' favors for the personalized ranking, which decreases the weights of virtual hubs and virtual links but increases the validity of ranking.

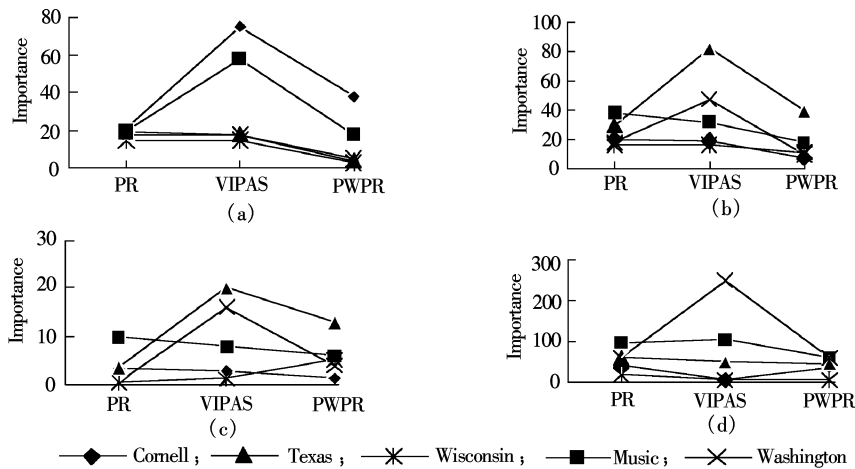


Fig. 1 Importance of same keyword ranking in different algorithms. (a) Queries “coding”; (b) Queries “hotlist”; (c) Queries “informatics”; (d) Queries “administrative”

3 Conclusion

In this paper, a new personalized web page ranking algorithm called PWPR with the idea of adjusting the ranking scores of web pages is proposed. The adjustments are in accordance with user preferences mined from user browsing behavior. Specifically, PWPR includes four steps. The first step assigns the initial weights based on user interests. In the second step, the virtual links and hubs are created according to user interests. By observing user click streams, PWPR will incrementally reflect user favors for the personalized ranking in the third step. To improve the accuracy of ranking, collaborative filtering is taken into consideration when the queries with similar keywords are submitted by users who have similar user interests. Detailed simulation experiments and comparison with other algorithms show that the proposed PWPR algorithm is very effective and it is very adaptive in providing personalized ranking.

References

[1] Geng Liqiang, Hamilton Howard J. Interestingness measures for data mining: a survey [J]. *ACM Computing Survey*, 2006, 38(3): 1 – 32.
[2] Castells Pablo, Fernández Miriam, Vallet David, et al. Self-

tuning personalized information retrieval in an ontology-based framework [C]//*Proceedings of OTM Workshops*. Berlin: Springer-Verlag, 2005: 977 – 986.
[3] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine [C]//*Proceedings of the 7th International World Wide Web Conference*. Netherlands: Elsevier Science Press, 1998: 107 – 117.
[4] Chakrabarti S, Dom B, Raghavan P, et al. Automatic resource compilation by analyzing hyperlink structure and associated text [C]//*Proceedings of 7th International World Wide Web Conference*. Netherlands: Elsevier Science Press, 1998: 65 – 74.
[5] Haveliwala T. Topic-sensitive PageRank [C]//*Proceedings of the 11th International World Wide Web Conference*. New York: ACM press, 2002: 517 – 526.
[6] Jeh G, Widom J. Scaling personalized web search [C]//*Proceedings of the 12th International World Wide Web Conference*. New York: ACM press, 2003: 271 – 279.
[7] Lin C C, Chen M S. VIPAS: virtual link powered authority search in the web [C]//*Proceedings of the 29th International Conf on Very Large Data Bases*. New York: ACM press, 2003: 1 – 12.
[8] Law Effie Lai-Chong, Klobučar Tomaž, Pipan Matic. User effect in evaluating personalized information retrieval systems [C]//*Proceedings of EC-TEL*. Berlin: Springer-Verlag, 2006: 257 – 271.
[9] Webkb project [EB/OL]. (2005-03-01) [2008-04-01]. <http://www.webkb.org/>.

基于用户偏好的个性化网页排序算法

朱容波

(中南民族大学计算机科学学院, 武汉 430074)

摘要: 为了按用户偏好对搜索结果进行排序, 提出了一种新的个性化网页排序算法 PWPR. PWPR 基于按照用户偏好调整网页排序的思想, 根据用户兴趣为网页分配初始权值, 并建立虚连接, 通过测量用户的点击流实现用户喜好的区分. 对于具有相似兴趣的用户提交的相似关键词查询, PWPR 采用协作过滤方式提高排序精确性. 仿真结果及与其他算法的比较证明 PWPR 算法能自适应地实现个性化排序, 并根据用户偏好提供相关查询信息.

关键词: 网页; 用户偏好; 排序算法; 个性化

中图分类号: TP311