Journal of Southeast University (English Edition) Vol. 23, No. 3, pp. 347 – 351 Sept. 2007 ISSN 1003—7985

User-oriented web search based on PLSA

Yu Fang[1] Chen Dongling[1,2] Wang Daling[1] Yu Ge[1] Bao Yubin[1]

([1]College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

([2]School of Information, Shenyang University, Shenyang 110044, China)

Abstract: In order to solve the problem that current search engines provide query-oriented searches rather than user-oriented ones, and that this improper orientation leads to the search engines' inability to meet the personalized requirements of users, a novel method based on probabilistic latent semantic analysis (PLSA) is proposed to convert query-oriented web search to user-oriented web search. First, a user profile represented as a user's topics of interest vector is created by analyzing the user's click through data based on PLSA, then the user's queries are mapped into categories based on the user's preferences, and finally the result list is re-ranked according to the user's interests based on the new proposed method named user-oriented PageRank (UOPR). Experiments on real life datasets show that the user-oriented search system that adopts PLSA takes considerable consideration of user preferences and better satisfies a user's personalized information needs.

Key words: user-oriented search; underlying search intention; probabilistic latent semantic analysis (PLSA); user profile; topics of interest

When different users submit the same query, typical search engines return the same result list regardless of who submits the query and what intention the user has. A common example is that user A issues query "apple" aiming to find some information on a laptop with the brand "apple", while user B issues the same query requiring some recipes of apple. Considered from this aspect, most of the current search engines are query-oriented, which means that the result list from a search engine is totally up to the query (within a certain period of time). But essentially, a search engine is a tool providing service for users rather than queries, and users are the target clients. From this point of view, user-oriented web search is in great need, which can be adaptable to different user information needs by taking account of the users' underlying intent of a query. In this paper, a novel method based on probabilistic latent semantic analysis (PLSA) is proposed to convert query-oriented search into user-oriented search.

1 Related Work

The most common method realizing personalization is utilizing user web usage information to construct user profile. Pretschner et al. [1] used ontology to create user profile. The user profile was represented as a hierarchical structure and the process was generated automatically, without explicit user feedback. Liu et al. [2] proposed a user profile and a general profile that were learned from the user's search history. The two profiles were combined to map a user query into a set of categories, which represented the user's search intention and served as a context to disambiguate the words in the user's query. Sugiyama et al. [3] proposed several approaches that were used to adapt search results according to each user's information requirement. The experiments showed that the one based on modified collaborative filtering outperformed the methods based on other techniques.

All of the above methods have their own deficiencies. In detail, some methods are based on ontology knowledge that requires predefining the ontology by dedicated professionals and checks those methods to be scalable. Some methods are based on collaborative filtering which requires facing up to sparse data and cold-start problems. Some methods are based on clustering algorithms which require that the users have multiple interest problems, and so on. Furthermore, those methods cannot describe a user's latent information requirements correctly. Moreover, they have a common shortcoming in that they do not reveal the underlying characteristics of a users' usage information and the latent relationships among the co-occurrence observation data.

2 Constructing User Profile Based on PLSA

A user's interest is an unobserved factor, which is concealed under the queries and the browsing behav-

Received 2007-05-18.

Foundation item: The National Natural Science Foundation of China (No. 60573090, 60673139).

Biographies: Yu Fang (1981—), female, graduate; Wang Daling (corresponding author), female, doctor, professor, dlwang@mail.neu.edu.cn.

ior. In order to discover the hidden semantic relationships between users and web objects, we incorporate PLSA to analyze user web usage information.

2.1 Probabilistic latent semantic analysis model

The PLSA model[4] is a statistical latent variable analysis model. It assumes that there exists a set of hidden factors underlying the co-occurrence among two sets of objects. The relationships between the hidden factors and the two sets of objects can be estimated by the expectation-maximization (EM) algorithm. PLSA have been successfully applied in information retrieval[5], collaborative filtering[6], co-citation analysis[7], and identifying user interest in commercial web sites[8]. This model has also been used in the web search field to compute preference predictions[9].

2.2 Representation of user profile

We assume that a user's query can be disambiguated in the context of a topic, to which the key words of the query belong. For example, with the topic constraint "cooking", "apple" cannot be misunderstood as "computer".

Definition 1 (topics of interests vector) $\boldsymbol{T} = \{t_1, t_2, \ldots, t_m\}$, where m is the number of considered topics.

Definition 2 (degree of preference) $\boldsymbol{P}_{\mathrm{T}} = \{P(t_1), P(t_2), \ldots, P(t_m)\}$, where $P(t_i)$ denotes the extent to which the user has preference for topic t_i, and $\boldsymbol{P}_{\mathrm{T}}$ is normalized to satisfy the equation $\sum_{i=1}^{m} P(t_i) = 1$.

When a user issues a query, she/he definitely has a topic in mind, while this topic is not explicit. She/he wants the web pages not only containing the keywords in the query but also correlating to the topic, while the pages having nothing to do with the topic are redundant. Fig. 1 depicts the status of topics in the web search process.

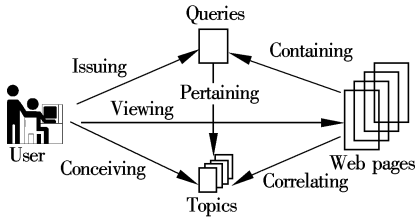

Fig. 1 The status of topic in web search

2.3 Discovering the relationship between query and web page

When a user visits a page for a query, she/he has some particular underlying intent. This underlying set of variants can be deduced through employing PLSA. After preprocessing the web log of a user, the set of queries $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_l\}$, and the sets of web pages $\boldsymbol{P} = \{p_1, p_2, \ldots, p_n\}$ are made available. The core of PLSA is a latent class statistical mixture model, which is a latent semantic class model for co-occurrence data which associates an unobserved class variable $z_i \in \boldsymbol{Z} = \{z_1, z_2, \ldots, z_m\}$ with each observation. In our study, (p_j, q_k) are co-occurrence objectives.

The probabilistic latent factor model can be described as the following generative model:

① $P(p_j)$ is the *a priori* probability that a user will visit a web page p_j;

② $P(z_i \mid p_j)$ is the conditional probability that a user has intent z_i when she/he visits p_j;

③ $P(q_k \mid z_i)$ is the conditional probability that a user issues a query q_k when she/he has intent z_i.

When a user issues a query q_k and browses web page p_j, she/he has an underlying intent z_i. According to the aspect model, an assumption should be made that $\boldsymbol{P}$ and $\boldsymbol{Q}$ are independent of each other. The relationships among unobserved factors and co-occurrence observed data can be derived from the Bayesian conditional probability as follows:

$$P(p_j, q_k) = P(p_j)P(q_k \mid p_j) = P(p_j)\sum_{i=1}^{m} P(q_k \mid z_i)P(z_i \mid p_j) \quad (1)$$

Using a full probability formula, we can obtain the following equation:

$$P(p_j, q_k) = \sum_{i=1}^{m} P(z_i)P(q_k \mid z_i)P(p_j \mid z_i) \quad (2)$$

Now, in order to explain a set of observations ($\boldsymbol{Q}$, $\boldsymbol{P}$), we need to estimate the parameters $P(z_i)$, $P(p_j \mid z_i)$ and $P(q_k \mid z_i)$, while maximizing the following likelihood $L(\boldsymbol{Q}, \boldsymbol{P})$.

$$L = \sum_{k=1}^{l}\sum_{j=1}^{n} \omega(p_j, q_k)\log(p_j, q_k) \quad (3)$$

Then, we resort to the alternation between the E-step and M-step of the iterative EM algorithm[10] to solve the problem. When the monotonically increasing log-likelihood function reaches a local optimal limit, we can obtain the final probabilities of observed data: $P(z_i)$, $P(p_j \mid z_i)$ and $P(q_k \mid z_i)$. In Eq. (3), $\omega(p_j, q_k)$ represents the weight of page p_j in terms of query q_k, which can be measured through observing the time the user spent on this page.

Maximal likelihood alternates between two steps.

In E-step:

$$P(z_i \mid p_j, q_k) = \frac{P(z_i)P(p_j \mid z_i)P(q_k \mid z_i)}{\sum_{i'=1}^{m} P(z_{i'})P(p_j \mid z_{i'})P(q_k \mid z_{i'})} \quad (4)$$

In M-step:

$$P(p_j \mid z_i) = \frac{\sum_{k=1}^{l} w(p_j, q_k) P(z_i \mid p_j, q_k)}{\sum_{k=1}^{l} \sum_{j'=1}^{n} w(p_{j'}, q_k) P(z_i \mid p_{j'}, q_k)} \quad (5)$$

$$P(q_k \mid z_i) = \frac{\sum_{j=1}^{n} w(p_j, q_k) P(z_i \mid p_j, q_k)}{\sum_{k'=1}^{l} \sum_{j=1}^{n} w(p_j, q_{k'}) P(z_i \mid p_j, q_{k'})} \quad (6)$$

$$P(z_i) = \frac{\sum_{k=1}^{l} \sum_{j=1}^{n} w(p_j, q_k) P(z_i \mid p_j, q_k)}{\sum_{k=1}^{l} \sum_{j=1}^{n} \sum_{i'=1}^{m} w(p_j, q_k) P(z_{i'} \mid p_j, q_k)} = \frac{\sum_{k=1}^{l} \sum_{j=1}^{n} w(p_j, q_k) P(z_i \mid p_j, q_k)}{\sum_{k=1}^{l} \sum_{j=1}^{n} w(p_j, q_k)} \quad (7)$$

2.4 Computing topic preference

Now we want to assign values to the topics of interest vector. Thus, a corresponding relationship should be found between T and Z. The following algorithm is based on two assumptions:

Assumption 1 If $P(q_k \mid z_i) = \max\{P(q_k \mid z_{i'}) \mid z_{i'} \in \mathbf{Z}\}$, then q_k is more representative in z_i.

Assumption 2 If q_k occurs more times in category c, then q_k is more pertaining to c.

Algorithm 1 Computing topic interest vector

Input: $P(q_k \mid z_i)$, $P(z_i)$ and the matches of q_k in category c in open directory project.

Output: $\boldsymbol{T} = \{t_1, t_2, \ldots, t_m\}$; $\boldsymbol{P}_T = \{P(t_1), P(t_2), \ldots, P(t_m)\}$.

Method:

$QS = \{QS_1, QS_2, \ldots, QS_m\}$, $QS_1 = QS_2 = , \ldots, = QS_m = \varnothing$;

For each q_k and each z_i

$q_k \in QS_i$ where $P(q_k \mid z_i) \geqslant P(q_k \mid z_{i'})$ $(i' = 1, 2, \ldots, n)$ but $i' \neq i$;

For $i = 1$ to m

For each q_k in QS_i and each category c in ODP,

freq(q_k, c) = matchings(q_k, c)/matchings (q_k);

freq$(QS_i, c,)$ + = freq (q_k, c);

For $i = 1$ to m

$t_i = c$ where freq (QS_i, c) > freq (QS_i, c');

$P(t_i) = P(z_i)$

2.5 Quantified probability of underlying topic for given query

One of the main advantages of the PLSA model in web usage mining is that it generates the probabilities which quantify relationships between web users and queries, as well as web pages and queries. We can derive a method to figure out the probabilities a query's underlying topics based on the Bayesian framework[11].

$$P(t_i \mid q) = \frac{P(q, t_i)}{P(q)} = \frac{P(t_i) P(q \mid t_i)}{P(q)} \propto P(t_i) P(q \mid t_i) \quad (8)$$

If query q is an familiar query, we can obtain the value of $P(q \mid t_i)$ directly from the above results of the EM iteration, else this probability may be computed as in Ref. [11] by counting the total number of occurrences of terms in query q in the web pages listed under the topic t_i in the open directory.

3 User-Oriented Ranking

Now we formulate the difference between query-oriented and user-oriented search in more formal expression. We deem the relevant documents to a query q as an unordered document set, which are the result of the "match" function with q as the parameter and regard the returned link list as a sorted array of the hit set of documents, with the PageRank (PR)[12] algorithm (the most far-reaching ranking algorithm) as the parameter of the "sort" function.

$$\text{list}(q) = \text{sort}(\text{match}(q), \text{PR}) \quad (9)$$

From the above equation, hyperlink structures of the web are focused on the mutual relationship of links and hardly any factor of user diversity are considered.

Our destination is to construct a mechanism that can convert the query-oriented search to a user-oriented search through re-ranking the result set for a user's query with respect to the user's specific topic interests. So the parameter of the "sort" function should include the user's preference, naming user-oriented PageRank (UOPR).

$$\text{list}(q) = \text{sort}(\text{match}(q), \text{UOPR}) \quad (10)$$

We design the UOPR re-ranking as follows:

$$\text{UOPR}(p) = \sum_{i=1}^{m} P(t_i) P(t_i \mid q) \cdot \text{ranking}(p)^{-\frac{1}{3}} \text{correlation}(p, t_i) \quad (11)$$

In Eq. (11), four factors contribute to the final ranking: ① $P(t_i)$ represents the user's preference on the topic; ② $P(t_i \mid q)$ is the probability that q belongs to t_i; ③ ranking(p) is the ordinal rank of the link returned by the search engine; ④ correlation(p, t_i) is the correlation of page p to topic t_i. As we do not expect the UOPR value to attenuate too fast with the ordinal rank increasing, we adopt the exponent $-1/3$. More importantly, user interests are embodied in the equation and so the re-ranked list can satisfy user requirements better.

4 Evaluation and Experiments

4.1 Data set

For the sake of persuasion, the experiments are done on real life data collected through Google API. User queries and browsing histories are recorded in the form of QuerySession (sessionID, username, query,

issue_time, querycategory1, querycategory2) and ClickedItem (clicked_time, sessionID, title, url, snippet, pagecategory1, pagecategory2, ordinal_rank). We collected 30 d of the English queries sent to Google from 32 students and faculties in our institute, amounting to 13×10^3 queries.

4.2 Metrics

● Accuracy of user's interest vector Since it is hard for users to specify to what extent she/he prefers a topic, we cannot calculate an accurate similarity by cosine metric between the interest vector we calculated and the vector user specified. But it is feasible for the users to rank their preferred topics in descending order, so we use the following method to evaluate:

$$\text{accuracy}(x) = 1 - \frac{y}{x} - \frac{\sum_{i=1}^{x-y} d_i}{x^2} \quad (12)$$

where x is the number of topics we take into consideration ($x < 15$); y is the number of topics that have not appeared in our recommendation; and $x - y$ is the number of topics matched; d_i means the distance of a matched topic's position posed by us to that posed by the user.

● Accuracy of re-ranking As Ref. [13] pointed out that recall and precision are based on the assumption that the set of relevant documents of a query is the same, independent of the user. However, different users might have a different interpretation of which document is relevant and which one is not. So we use $R_{\text{precision}}$ to evaluate the retrieval accuracy. If r denotes the number of relevant documents the user expected to find, and e denotes the number of documents examined in an attempt to find r relevant documents, then

$$R_{\text{precision}} = \frac{r}{e} \quad (13)$$

4.3 Experimental result

According to use usage data, we construct user profiles through identifying their latent intent. The detail of the user profile construction process is depicted in algorithm 1. Fig. 2 gives the 28th user's interest vector as an example.

It is clear to see that a user's preferences on different topics slant greatly. So detecting a user's interest vector is meaningful. We then ask the users of our system to rank their own interests with a corresponding ratio. We generate a top-n prediction candidate set of n topics, and then compare the predictions with the list the user ranked (see Fig. 3). Here we see that as n increases from 1 to 10, the precision decreases from 85% to 30% (using Eq. (12)).

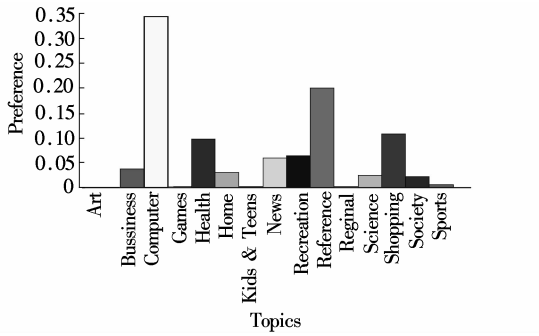

Fig. 2 Examples of user topics of interest vector(User_ID = 28)

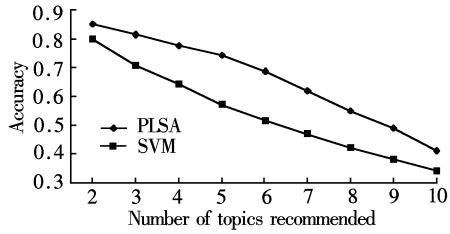

Fig. 3 PLSA vs. SVM on accuracy of interest vector

We also make a baseline method to compare. Here, we use the software package (Http://svmlight.joachims.org/) to implement an SVM. Fig. 3 shows the comparison of our method and the SVM.

We re-rank top n ($n > 30$) links based on our proposed UOPR method, and then take the top 30 links. From Fig. 4, we can see that the more links are re-ranked, the higher $R_{\text{precision}}$ is achieved. But the speed will slow when the number of candidate links exceeds 80.

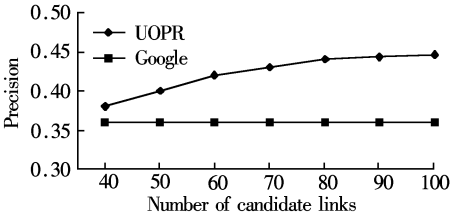

Fig. 4 UOPR vs. Google on ranking

5 Conclusion and Future Work

In this paper we pointed out that improper orientation is the intrinsic cause that leads to the low performance of current search engines. We proposed a framework to convert the query-oriented search to the user-oriented search by taking account of user's preference into the ranking. We employed the PLSA to detect the latent factor between the queries and the web pages. And we designed UOPR to re-rank result list based on the user's interests. In the real life data set experiments, we found that the user's interest vector estimated by our algorithm can express the user's interest well. And the re-ranked list is more adaptable to each

user's specific taste.

In the future, we plan to deepen the blueprint of our UOPR to build an applicable toolkit. We will combine a user's long-term interests with his/her short-term interests for updating a user profile more effectively. We also plan to incorporate more sophisticated learning and ranking algorithms, such as personalized search based on behavior, to further improve the performance of our system.

References

[1] Pretschner Alexander, Gauch Susan. Ontology based personalized search [C]//*Proc of the* 11*th IEEE International Conference on Tools with Artificial Intelligence*. Chicago, 1999: 391 - 398.

[2] Liu Fang, Yu Clement, Meng Weiyi. Personalized web search for improving retrieval effectiveness[J]. *IEEE Trans on Knowledge and Data Engineering*, 2004, **16**(1): 28 - 40.

[3] Sugiyama K, Hatano K, Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users[C]//*Proc of the* 13*th International World Wide Web Conference*. New York: ACM Press, 2004: 675 - 684.

[4] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis[J]. *Machine Learning*, 2001, **42**(1/2): 177 - 196.

[5] Hofmann T. Probabilistic latent semantic indexing [C]// *Proc of the* 22*nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, 1999: 50 - 57.

[6] Hofmann T. Latent semantic models for collaborative filtering [J]. *ACM Trans on Information Systems*, 2004, **22**(1): 89 - 115.

[7] Cohn D, Chang H. Learning to probabilistically identifying authoritative documents[C]//*Proc of the* 17*th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 2000: 167 - 174.

[8] Jin X, Zhou Y, Mobasher B. Web usage mining based on probabilistic latent semantic analysis[C]//*Proc of the* 10*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press, 2004: 197 - 205.

[9] Lin Chenxi, Xue Guirong, Zeng Huajun, et al. Using probabilistic latent semantic analysis for personalized web search [C]//*Proc of the Seventh Asia-Pacific Web Conference* (*APWeb*2005). Shanghai, 2005, **3399**: 707 - 717.

[10] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. *Journal of Royal Statistical Society B*, 1977, **39**(1): 1 - 38.

[11] Haveliwala T H. Topic-sensitive pagerank[C]//*Proc of the* 11*th International World Wide Web Conference*. New York: ACM Press, 2002: 517 - 526.

[12] Page Larry, Brin Sergey, Motwani R, et al. The PageRank citation ranking: bringing order to the web[R]. California: Stanford Digital Library Technologies Project, 1998.

[13] Baeza-Yates R, Ribeiro-Neto B. *Modern information retrieval*[M]. Beijing: China Machine Press, 2004: 82 - 84.

基于 PLSA 的面向用户的网络搜索

于 芳[1] 陈冬玲[1,2] 王大玲[1] 于 戈[1] 鲍玉斌[1]

([1] 东北大学信息科学与工程学院,沈阳 110004)

([2] 沈阳大学信息学院,沈阳 110044)

摘要:针对当前的搜索引擎提供面向查询、而非面向用户的服务,从而导致搜索引擎无法满足用户个性化的需求这一问题,提出了一种基于 PLSA 的新方法,将面向查询词的搜索转变成面向用户的搜索. 首先,通过分析用户查询历史和浏览记录建立代表用户模型的用户兴趣向量,在用户发出查询时用户的查询词根据用户兴趣向量被映射到兴趣分类上,最终根据面向用户排序算法将返回结果列表重新排序. 实验表明该面向用户搜索系统能够充分考虑用户的偏好,从而更好地满足不同用户的信息需求.

关键词:面向用户的搜索;潜在搜索意图;概率潜在语义分析(PLSA);用户模型;兴趣主题

中图分类号:TP311