Personalized Web Search Results with Profile Comparisons

Jun Lai and Ben Soh
Department of Computer Science and Computer Engineering
La Trobe University Bundoora, VIC, Australia 3083
jun@cs.latrobe.edu.au ben@cs.latrobe.edu.au

Abstract

The information explosion on the Internet makes it hard for users to obtain required information from the web searched results in a more personalized way. For the same input word, most search engines return the same result to each user without taking into consideration user preference. For many users, it is no longer sufficient to get non-customized results.

It is crucial to analyze users' search and browsing behaviors based on searching keywords entered by users, the clicking rate of each link in the result and the time they spend on each site. To this end, we have proposed a method to derive user searching profiles. We have also proposed a mechanism to derive document profiles, based on similarity score of documents. In this paper, we discuss how to use our model to combine the user searching profiles and the document profile, with a view to presenting customized search results to the users.

1. Introduction

The tremendous growth in the amount of information available and the number of visitors to web sites in the recent years pose some key challenges for search engines. Search engine users not only expect high quality information, but also wish that the information presented were personalized. While search engines and information filtering tools are typically not personalized to individual users or their prevailing context, they tend not to deliver an appropriate amount of preferred information. To overcome these limitations, recommender systems have been advocated and some personalized approaches to web search have been proposed [1]. The user can specify number of web pages to be retrieved, domains (e.g. .gov, .edu, .com) to be included in the search results, number of Internet Spiders to be used, and so on. However, the search result is not personalized; different user will still get the same result as long as the query entered is the same. But different users have different preferences and interests. PHOAKS [2] recommends relevant, high-quality information on the web to users by automatically recognizing and redistribution recommendations of web resources based on group profiles instead of individual profile. The users from the same group are recommended the same search results. Some online commercial systems recommend music, movies and products using either content-based or collaborative recommendation. However, they are not personalizing documents in search result.

To alleviate these issues, we propose in this paper a personalized approach to web search which makes it possible to create personalized search results and take search from mass media to my media. This approach compares the accuracy between the personalized and non-personalized search results. The documents' similarity score is computed based on the weighting of keywords and the number of times of keyword appearance. The document profile is then derived. The customer profile is derived based on customer searching and browsing behaviors. Then document profile and customer profile are analyzed, and customized to present search results to users.

2. Research Methodologies

One of the aims of our research is to compare the document profile and customer (user) profile. From the comparisons, personalized search results are obtained. Customer profile is derived from customer searching and browsing behaviors. The document profile is generated by the algorithm based on similarity score of documents [14].

2.1 System Design

Our proposed system consists of three components: document profile component, customer profile component and search result component. The proposed system architecture is shown in figure 1.

Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)
0-7695-2316-1/05 $20.00 © 2005 IEEE

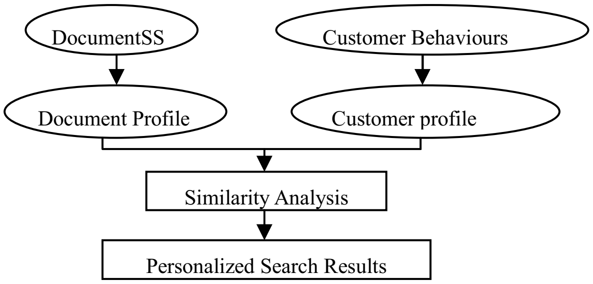

Figure 1. System architecture

For the customer profile component, customer behaviors - such as the searching keywords or phrases, a particular document the customer browsed and the time the customer spent on that document - are deemed as variables in the customer profile component. An average (generalized) customer profile will be derived from the overall customers currently profiled. This average profile will be used as the starting point for a new customer before their personalized profile is built up.

For the document profile component, the keywords, phrases and weight of the keywords are adopted for calculating the similarity score of documents. The document profile is then derived using keywords and similarity score as variables [14]. The document profile will be updated upon the arrival of any new document on the web.

Lastly, we make personalized search results based on the customer profile and document profile and carry out the analysis by adopting the accuracy comparison between the personalized search results and non-personalized search results.

2.2 Document Profile (DP)

Currently, there are a number of metadata standards proposed for web pages. Among them are two well-publicized, solid efforts: the Dublin Core Metadata standard and the Warwick frame-word. The Dublin Core is a 15-Element Metadata Element Set proposed to facilitate fast and accurate information retrieval on the Internet [3]. To compute the similarity score of documents, we initially select some keywords appearing in those documents. Each word is assigned a weight, ranging form 0 to 1(inclusive), depending on which element this keyword is in. Using sensitivity analysis, the weights are calibrated by the system designer, based on the importance and relevance of the keywords concerned. For example, the value of weight in the title should be higher than in the description, while value of weight in the description should be higher than in the content. The number of times that word appearing in the document also affects the value of relevance among documents.

The similarity score is computed as the sum of all product of keyword weight and the number of times that keyword appears in the document, using the following formula [14]:

$$SS_{(d,K)} = \sum_{j=1}^{n} (W_{K_j} * Count_{K_j}) \quad (1)$$

where:

- SS is the similarity score of a document based on the keyword K.
- d denotes a document ID.
- K_j is the keyword K appearing in the element j of the document ($1 \leq j \leq n$).
- W_{Kj} is the pre-defined weight of the keyword K_j appearing in the element j, determined by a system administrator via sensitivity analysis.
- $Count_{Kj}$ is the number of times that the keyword K appearing in the element j of the document.

Based on the similarity score, the document profile is derived as follows [14]:

$DP_{(d)} = \{ (K, SS_{(d,K)} \mid K \in C, SS_{(d,K)} \geq SS_{threshold} \}$ (2)

where:

- d denotes a document ID.
- K denotes a keyword.
- C is a set of all keywords to which the document can be related.
- $SS_{(d,K)}$ is the similarity score for document d for the keyword K.
- $SS_{threshold}$ is the minimum similarity score acceptable for a document to relate to that keyword.

If the similarity score of a document is above the threshold, then that document is relevant to that keyword. For instance, if the given threshold is 10 and the document No.61 has profile: DP(61) = {(filtering,12), (computing,13), (history,2), (internet technology,15)}

Then we consider that the document 61 is relevant to keyword filtering, computing and internet.

2.3 Customer Profile

In this paper, we define the customer profile as follows:

$CP_{(n)} = \{(K, D_i, T_{Di,Kj}) | K \in C, 0<i< N, 0 \leq T_{Di,Kj} < \infty)\}$ (3)

where:

- n denotes a customer ID.
- K denotes a keyword or phrase
- C is a set of keywords or phrases to which the document can related.
- D_i is the document of i and D_i has profile of $DP_{(i)}$ which is based on keywords and SS.
- $T_{Di,Kj}$ is the time of user spending on the document of D_i based on K_j as searching keyword.

Equation 3 shows that a customer's level of interest in a particular document is based on the customer's searching and browsing history. The history includes: (i) the keywords or phrases the customer usually searches for, (ii) the document the customer clicked in the search result, and (iii) the length of time the customer spends on a particular document. A particular customer's level of interest is quantified with respect to those of the general users in accordance with the average customer profile (shown in Equation 4). As mentioned before, for a new customer using our search engine, the average customer profile is adopted to him/her to start with.

$$\text{CPA}_{(n)} = \frac{CP_{(n)}}{1/n\sum_{n=1}^{N} CP_{(n)}(K, D_i, T_{D_i,K_j})},$$

$$n = 1,\ldots,\text{N}. \qquad (4)$$

2.4 Personalized Search Results

The personalized search results will be based on document profile and customer profile. Table 1 shows the customer A's profile.

Table 1. An example of customer profile

keyword	Document	Time
Filtering	http://sims.berkeley.edu/resources/collab/	50
	http://www/paulgraham.com/spam.html	10
	http://www.science.unitn.it/cirm/Annfiltering.html	0

Customer A has a profile
$\text{CP}_{(A)}$ = {(filtering, http://sims.berkeley.edu/resources/collab/, 50), (filtering, http://www/paulgraham.com/spam.html, 10), (filtering,http://www.science.unitn.it/cirm/Annfiltering.html, 0)}

From this profile, we infer that customer A searched the keyword "filtering" and browsed the document of http://ieeexplore.ieee.org/xpl/periodicals.jsp/InternetInformationFiltering.phf for 50 minutes. However the time of browsing the document is 0. That means this customer didn't click this search result and browse it at all. Therefore, those documents with less interest will be less recommended next time when customer A searches the keyword "filtering".

3. Evaluation

We adopt Accuracy (p) in this research as a measure to evaluate the effects of the personalized search results.

$$\text{Accuracy}_{(p)} = \frac{No.of\ visited\ documents}{No.of\ recommended\ documents}$$

The threshold of the time a customer spends on a document returned from a search result is used to recommend document to customers. If a given threshold, for instance, is 10 minutes, then the document the customer spent less than 10 minutes on browsing will not be recommended to this customer. The order of document recommended is also based on the customer profile.

This experiment is based on 40 users' search behaviors. Table 2 shows the accuracy comparison between personalized search results and non-personalized search results for different search query. The non-personalized search result is calculated as follows:

$$\text{Accuracy}_{(np)} = \frac{No.of\ visited\ documents}{No.of\ documents\ returned}$$

The given threshold of time is 10 minutes.

Table 2. Accuracy comparison between personalized and non-personalized search results.

keyword	Movies	Filtering	Cars	sports
Accuracy $_{(p)}$	0.75	0.72	0.74	0.75
Accuracy $_{(np)}$	0.53	0.56	0.52	0.51

The non-personalized search results are not based on customer profile. Whoever enters the search query, the search engine returns the same result. On the other hand, personalized search results returned are based on customer profile. As we can see from figure 2, it is about 20% more accurate than non-personalized search results. In particular, a keyword such as "movies", "cars" and "sports", is highly dependent on personal preferences. For example, someone who is a big comedy fan might like visiting the movie web site of comedy. The personalized search results know this fact from the customer profile and recommend most comedy web sites as search results for this customer. Therefore it is more accurate. However, non-personalized search results only return generic search result without knowing this customer's profile.

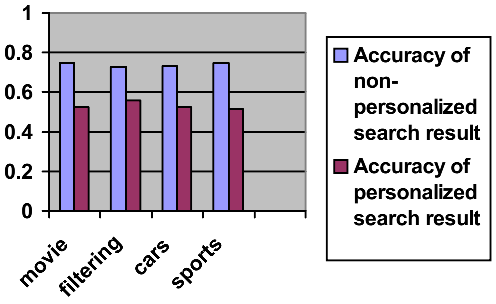

Figure 2. Accuracy comparison

4. Conclusions and Future Work

We believe that many other powerful techniques can possibly be implemented on search tools to improve the personalized search and effectiveness in web search as well as other information retrieval applications [15,16,17,18,19].

In this paper, we propose a personalized approach to web search. This approach applies an accuracy comparison between the personalized and non-personalized search results. The similarity scores of documents are computed based on the weighting of keywords and the number of times of keywords appearance. The document profile is then derived [14]. On the other hand, the customer profile is derived based on customer searching and browsing behaviors. Then document profile and customer profile are analyzed, customized search results are presented to users.

The customer's interest and preferences may change with time. Thus the recent searching and browsing behaviors can better reflect a customer's latest interests. Customer feedback can also be used to update the customer's preferences. In this regard, the customer profile component can be extended to include customer feedback, which is updating the customer profile component. Likewise, the document profile component can be extended to include a web crawler, which is crawling on web and updating document profile as documents on the web change with time.

5. References

[1] Meng, X and Chen, Z, "Personalized Web Search With Clusters", *International Conference on Internet Computing, IC'03*, 2003, pp. 46-52.

[2] Terveen, L., Hill, W., Amento, B., McDonald, D. & Creter, J. Phoaks: "A system for sharing recommendations". *Communications of the ACM*, 40(3), 1997

[3] Dublin Core – http://dublincore.org/documents/dces/

[4] J, A Hartigan. *Clustering Algorithms*. WILEY Publication, 1975.

[5] Hearst, M. TileBars, "Visualization of Term Distribution Information in Full Text Information Access", in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems(CHI'95)*, 1995, pp. 59-66.

[6] Veerasamy, A. and Belkin, N.J., "Evaluation of a Tool for Visualization of Information Retrieval Results", in *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96),* 1996, pp. 85-92.

[7] Hearst, M. and Pedersen, J. Reexamining the Cluster Hypothesis, "Scatter/Gather on Retrieval Results", in *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'96*), 1996, pp. 76-84.

[8] Rasmussen, E., "Clustering Algorithms", in *Information Retrieval Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates (eds.), Prentice Hall, N.J., 1992.

[9] Zamir, O. and Etzioni, O. Grouper, "A Dynamic Clustering Interface to Web Search Results", in *Proceedings of the 8th International World Wide Web Conference,* 1999.

[10] Chignell, M. H., Gwizdka, J. and Bodner, R. C., "Discriminating Meta-Search: A Framework for Evaluation", *Information Processing and Management*, 35, 1999.

[11] McBryan, O., "GENVL and WWW: Tools for Taming the Web", in *Proceedings of the 1st International World Wide Web Conference, Geneva, Switzerland,* 1994.

[12] Bowman, C., Danzig, P., Manber, U. and Schwarts, F., "Scalable Internet Resource Discovery: Research Problems and Approaches", *Communications of the ACM* 37(8), 1994, pp. 98-107.

[13] Cho, J., Garcia-Molina, H. and Page, L. "Efficient Crawling Through URL Ordering", in *Proceedings of the 7th World Wide Web Conference, Brisbane, Australia,* 1998.

[14] J, Lai and B, Soh. "Using Element And Document Profile For Information Clustering", in *Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service, TaiPei, TaiWan, 2004, pp.503-506.*

[15] S. Vrettos, A. Stafylopatis., "A Fuzzy Rule-Based Agent for Web Retrieval-Filtering", *Web Intelligence: Research and Development : First Asia-Pacific Conf.*, Springer-Verlag, October, 2001, pp. 448-453.

[16] Baeza-Yates, R. and Ribeiro_neto, B., *Modern Information Retrieval*, Addison-Wesley, Reading, MA, 1999.

[17] AI Mag, "Special issue on intelligent systems on the internet", *AI Mag. 18,*1997.

[18] Anick, P. G and Vaithyanathan, S., "Exploiting clustering and phrases for context-based information retrieval", *SIGIR Forum* 31, 1, 1997, pp. 314-323.

[19] Rasmussen, E., "Clustering algorithms", in *Information Retrieval: Data Structures and Algorithms*, W.B. Frakes and R. Baeza-Yates, (eds), Prentice-Hall, Inc., Upper Saddle River, 1992.