International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 5, Issue 2, Feb - Mar 2016; 123-130

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

APPROACHES TO PERSONALISED WEB SEARCH TO IMPROVE RETRIEVAL QUALITY

NAIF ALORFI

Monash University, Dandenong Road, Caulfield East, Victoria, Australia

ABSTRACT

As resources on the World Wide Web (WWW) are growing rapidly, search engines have become an essential tool for people to find what they need on the Web. Millions of users' queries are processed every day, but current Web search engines still have many disadvantages. Search engines serve all users in the same way, regardless of who submits the query, even though each user will have different information needs, associated with each query they submit. For that reason, search results should be adapted to users with different information needs. To solve this problem, a personalised web search is proposed that looks closely at each individual user to predict their intentions. This review focuses on two major tasks in developing a personalised Web search engine: user profile modelling and personalised query expansion, both of which can help to improve information retrieval quality. A user profile aims to find the best user model to help a system to predict user intentions or interests while searching the Web, without any additional activity from the user, such as explicit feedback. Personalised query expansion is widely used to decrease query ambiguity in information retrieval, expanding the user's query by, for instance, adding extra terms with statistical relations to a set of relevant documents or by adding terms with a similar meaning.

KEYWORDS: Personalized Search, Search Engine, Information Retrieval, Web Search

1. INTRODUCTION

Resources on the World Wide Web (WWW) are growing rapidly, and search engines process millions of queries every day. A user searching the Web for information of interest does so by typing a keyword query that describes the information desired. Search engines usually list many pages based on that keyword query, and these are subsequently displayed in order of higher page ranking. But these pages may not meet the searcher's needs—for example, a programmer may enter the query "Java" while developing an application; similarly, a coffee buyer could potentially use the same query ("Java") in searching for types of coffee to buy. Typically, a search engine will return the same results for both, regardless of who issues the query and the context of the particular query issued. The problem is that the user's keyword-based query is usually ambiguous to the search engine and does not describe exactly what is needed. Users commonly issue queries that are very short, making the process of extracting the most relevant pages among millions decidedly difficult. Based on Onstat.com analysis of their log files over a two–month period, Speretta and Gauch (2005) reported that 77.2% of keyword searches comprised three words or fewer while 32.6% of searches comprised only two words. Categorisation of Web pages can help to decrease the ambiguity of a user's query by associating query terms with a set of categories. Before issuing a query, the searcher can select an appropriate category from a hierarchy of categories (Liu, Yu, & Meng). For example, if a category "programming" is associated with the term "Java", the intended search becomes clearer. Categorisation of all vocabularies is usually substantial, and in consequence, a user may struggle to find the appropriate paths to identify the

appropriate category. A user profile consists of information about specific user interests that can be used to narrow down the number of retrieved pages, presenting those most relevant to the user in the current session. To improve retrieval quality, there are three areas of immediate relevance to this work: modelling user tasks by building a user profile for each individual user; personalization based on the user's historical behaviour, including short- and long-term user interests; and mining other users' search behaviours to find similar users and so complement and improve web search personalization(White et al.).

This review discusses several approaches to the development of a user profile-based personalised Web search to improve retrieval quality. It is organised as follows. Section 2 discusses several techniques of user modelling for personalised searches. Section 4 discusses a number of techniques used in recent studies on query expansion to improve personalised search retrieval quality. The concluding section presents a summary of findings.

2. USER PROFILE MODELLING

The main purpose of user profile modelling is to infer the user' sinterests, helping to envisage what information will meet the user's needs, to be employed in subsequent queries. Several available techniques can help to construct the user profile. The basic strategy in volvesa search engine asking the user to explicitly specify which web pages are relevant or irrelevant or to rate the results—for example: 1 (*very bad*) up to 5 (*very good*). This is called *explicit feedback*, where additional user action is required (Shen, Tan, & Zhai; Sugiyama, Hatano, & Yoshikawa, 2004). Some search engines ask searchers to setup their personal profile by registering details such as their interests, occupation, age and so on, which will then be used by the system to predict the searcher's level of interest in the retrieved Web pages (Sugiyama et al.). Such systems require additional actions from searchers to obtain the information needed, and searchers usually prefer a simpler method. A more efficient process is to deduce the information needs of searchers without any additional actions or interactions. This type of feedback is called *implicit feedback* (Shen et al., 2005). There are many studies of user modelling to support personalised Web searches that enhance retrieval quality and accuracy.

2.1 Personalisation based on User Search History

Many current search engines track and maintain a user's search history to learn about their interests and so construct an individual user model. Information items can be extracted from the user's history, including previous queries, relevant Web pages and categories that relate to the current query submitted by the user (Liu et al.). This can be modelled asa hierarchy tree with root node "Term" as shown in Figure 1. The child nodes of the root represent a set of categories. The leaves are documents associated with the parent category (Bounoy & Walairacht, 2010). For example, the word "Java" is related to both the "programming" and "coffee" categories. The search engine should implement effective design and analysis algorithms to generate the relationship between query terms and categories, extracting the best results to meet the individual user's preferences (Bounoy & Walairacht, 2010).

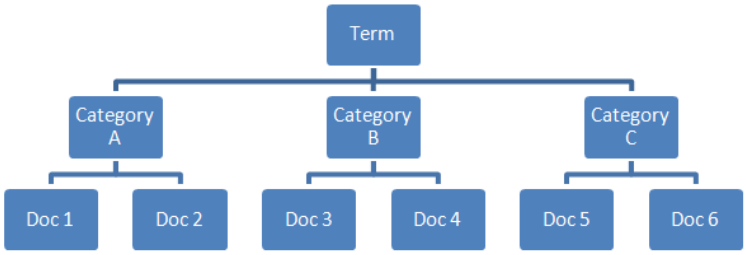

Figure 1: Model of Category Hierarchy in User Search History

Sugiyama et al. (2004) argued that each individual user's preferences consist of two phases: short-term and long-term preferences. In the case of short-term preferences, the user's profile is constructed only on the basis of the current session. For example, a user searches for a "used car" to buy, and eventually finds and purchases a suitable car. It follows that the user is no longer interested in such documents (Shen et al.). In the case of long-term preferences, the user's profile grows continuously over time with subsequent Web search sessions; indeed, it is likely that users perform different searches and different browsing behaviours on the same day (Sugiyama et al.).

Another technique, proposed by Cheqian, Kequan, Heshan, and Shoubin (2010)constructs the user profile based on user clicks history. The system records the page title and keywords used along with a summary of any clicked link for each user, and these a reindexed and scored using Lucene. Applying the *Naive Bayes* classification algorithm and the *support vector machine* (SVM) re-sorting algorithm, Cheqian et al. argued that the system improves retrieval quality in a short time.

Liu et al. (2004) proposed a two-step strategy to improve retrieval quality, based on constructing a user's profile by means of a weighted concept hierarchy assembled from the individual user's search history. They also discussed the construction of a general profile based on the open directory project (ODP) hierarchy of categories. Based on these two profiles, the system should automatically detect suitable categories related to each query. As a first step, for each query issued, the system automatically detects a small set of categories for each user, based on their search histories. As a second step, the system uses that set of categories to retrieve related web pages. Each category in the user profile consists of a set of query terms, and each term has a weight, calculated to represent a user's level of interest in a specific category(Liu et al.). The general profile is structured in the same way as the user's profile but for general knowledge (all users), constructed before the user's profile and obtained from the ODP category hierarchy (Bounoy & Walairacht). The terms in the submitted query are compared with terms stored in each category in the user profilec^u, as well as with those in the general profile c^gcreated by the cosine function as follows (Liu et al.):

$$Sim(q,c) = \max\left(Sim(q,c^u), Sim(q,c^g)\right)$$

Other experiments by (Liu et al.)and(Bounoy & Walairacht)indicated that using a user profile combined with the general profile achieves greater accuracy than using only the user profile or the general profile individually.

Speretta and Gauch (2005)implemented a Google Wrapper to monitor search history for a set of users; for each user, they collected two types of information: queries submitted(and at least one page examined) and snippets consisting of title and summary for pages examined by the user. The user's profile was constructed by classifying these two types of information into a concept hierarchy based on ODP, and then constructing two different profiles individually and comparing the

weights. After submitting a query to the Google Wrapper, the result snippets were also classified, using the same reference concept hierarchy. The similarity between result snippets and user profile concepts was calculated for re-ranking of results, where a higher page ranking represents higher user interest.

2.2 Collaborative Filtering for Similar User Profiles

Collaborative filtering is a set of popular algorithms that recommend items based on the preferences of similar users; if a set of users share similar interests, an item preferred by any user can be recommended to others in the set (Sun, Zeng, Liu, Lu, & Chen, 2005). Sugiyama et al. (2004) argued that this technique can usefully be implemented to predict auser's interests form other similar users. This can be represented as the missing value problem in the user-item weights matrix, where there are insufficient data on which to base predictions.

Figure 2 illustrates a simple example of a user-term weights matrix; when the user visit a new Web page, new terms are appended to the user's profile, but as other users may not have visited the same page, missing values occur in the user-term weight matrix. Collaborative filtering algorithms are used to complete these missing values (Sugiyama et al., 2004).

Table 1

	Term 1	Term2	…	Term C	…
User 1	0.754	0.805		0.543	
User 2	0.545	0.795		0.765	
…					
User A		0.645			
…					

Figure 2. User-term matrix for modified collaborative filtering.

Sugiyama et al. (2004)experiment evaluated retrieval accuracy based on three techniques: relevance feedback, pure search history and modified collaborative filtering. The evaluation indicates that modified collaborative filtering achieves the best accuracy of personalisation.

2.3 User Profile Based on Ontological User Profiles

Another approach is to construct ontological user profiles by allocating derived interest scores to current concepts in a domain ontology. Using a*sp reading activation algorithm*, user's interest scores (based on the user's ongoing behaviour) can be stored in their user profile, updating annotation for the current concept (Sieg, Mobasher, & Burke, 2007). The user profile is structured as a category hierarchy of Web pages. The relationship between concepts and categories plays an important role in constructing ontological user profiles (Sieg et al.). Interest scores for each concept for each user profile are initialized by 1; whenever the user is viewing a new Web page, the ontological profile is updated, and annotation of the current concept is redefined by separating activation; the interest score for the current concept is then updated incrementally (Sieg et al.). Figure 3 illustrates personalised Web search based on user profiles.

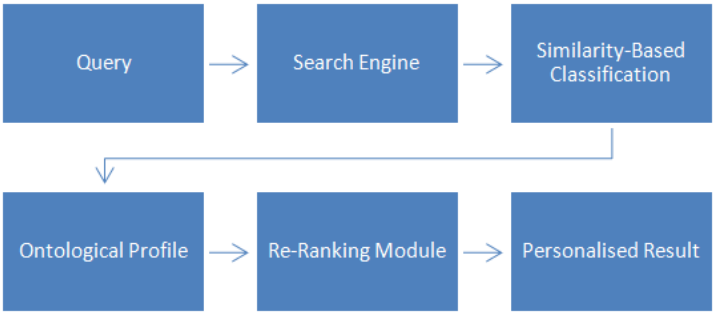

Figure 3: Personalised Web Search Based on Ontological User Profiles (Sieg Et Al.)

1. 3. Query Expansion Approaches

Query expansion is commonly used in information retrieval to reduce query ambiguity by expanding the user's query by adding extra terms with statistical relations to a set of relevant documents or by adding terms with similar meanings (Jayanthi, Jayakumar, & Akalya, 2011). Short queries are the most common; for any search engine, most queries consist of 1–3 words (Speretta & Gauch) and cannot describe abuser's needs. Another difficulty relates to the *dictionary problem*— for example, where two users have the same search intention, the probability that they will issue a similar query is less than 20%, and the search engine will therefore return different results (Jayanthi et al.).Shamim Khan and Khor (2004) proposed a *key phrase identification* algorithm, based on the documents retrieved by the original query. Although they reported that this algorithm can effectively expand the query string and retrieve more relevant documents, their study is not dealing with a personalisation web search.

In his article "Global analysis and local analysis", Aly (2008) describes two principal methods for query expansion. An example of global analysis is adding new terms from other external resources (such as a thesaurus) to the original query before searching. In contrast, local analysis involves the formulation of anew query from documents retrieved on the basis of the original query(Jayanthi et al.). For example, when users submit a search, the system returns the results and then collates the user's interest as implicit/explicit relevance feedback to retrieve relevant documents. The new query is then formulated on the basis of these documents to make the query more powerful. Such a system is known as personalised query expansion (Jayanthi et al.; Shamim Khan & Khor).

3.1 Query Expansion Based on Semantic Similarity of Phrases

Jayanthi et al. (2011) proposed a framework for personalised query expansion based on phrase similarity using a global analysis method. The process is in two stages: key phrase extraction and semantic similarity measured against phrases from the initial query (Jayanthi et al.). They also proposeda *profile-based phrase weight* algorithm, which gathers all relevant terms from Word Net and user interest from the user profile When the user query search is submitted, all relevant terms from Word Net are gathered to form a relevant phrase list. Candidate phrase sets are obtained by fixing the occurrence threshold. For each term in the relevant phrase list, term frequency is calculated from the titles and summaries in the initial result. If the term frequency is equal to or greater than the threshold, it will be appended to the candidate phrase set (Jayanthi et al.). The term weight is then calculated as follows:

$$\omega = \frac{df}{rp},$$

where$df$is the term frequency in all documents, and$rp$is the total number of relevant phrases in the list

(Jayanthi et al.). Candidate phrases are reweighted on the basis of user profiles. Next, related phrases are ranked, and those most similar to the query (from the top n phrases) are selected to form “set SP hraseExp and” to build a relevant link list; the most preferred links will be used (Jayanthi et al.). This work suggests that the framework improves retrieval by getting closer to the user’s intention.

3.2 Probabilistic Query Expansion

Palleti, Karnick, and Mitra (2007) proposed personalised Web search methods based on probabilistic query expansion. Where the data in a user profile are insufficient to make a prediction, their system performs collaborative filtering for automatic prediction of user interests from other similar users and then uses *pseudo query term selection* to enhance the user query. Their experiment uses the formula below to select a *pseudo query term*, based on the original query and search history (queries/documents):

Pseudo Query Term = argmax ∀qi∈QSP(qi|dj),

whereq_i represents the query term in the user’s query and previous queries’ *Query Space*, and d_j represents the user’s query term, which does not exist in previous queries (Palleti et al., 2007). The browsing history of each user is processed for the query session as follows:

$$QuerySession = < query > \ [clickedDocument, releventDocument] *$$

Similarity is calculated as follows:

$$P(q_i|d_j) = \frac{P(d_j|q_i) \times P(q_i)}{P(d_j)}$$

$$P(d_j|q_i) = \sum_{\forall D_k \in S} P(d_j|D_k) \times P(D_k|q_i),$$

where$P(q_i)$ represents the ratio of the number of query sessions in which the query term exists in the document to the total number of query session existing in the user’s query space; and $P(d_j)$ represents the ratio of the total term frequency of d_j in the user’s document space relative to the total term frequency of all document terms existing in the user’s document space (Palleti et al., 2007). Where *S* represents a set of documents in the user query session space, which contains q_i, Palleti et al. (2007) argued that performance is significantly improved and the system does not require explicit relevance feedback.

4. CONCLUSIONS

This review has discussed a number of approaches used in recent studies of personalised Web search to improve retrieval quality. The focus here was on two major tasks in personalised Web search engine development: user profile modelling and personalized query expansion. These approaches differ in their construction of user profiles to find the user model that will best help a system to predict the user’s intentions or interests while searching the Web, without any additional activity from the user. The present report has described methods that include the use of implicit feedback such as

mouse clicks and movement and search history, whether for long- or short-term user preferences. The ontological user profiles technique, and how it improves the user model, was also discussed. The discussed approaches based on personalised query expansion show that these improve information retrieval in respect of individual user intentions.

REFERENCES

1. Aly, A. A. (2008). Using a query expansion technique to improve document retrieval. *International Journal" Information Technologies and Knowledge, 2*, 343-348.
2. Bounoy, T., & Walairacht, A. (2010, 7-10 Feb. 2010). *User preference retrieval using semantic categorization for web search.* Paper presented at the Advanced Communication Technology (ICACT), 2010 The 12th International Conference on.
3. Cheqian, C., Kequan, L., Heshan, L., & Shoubin, D. (2010, 7-9 July 2010). *Personalized search based on learning user click history.* Paper presented at the Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on.
4. Jayanthi, J., Jayakumar, K. S., & Akalya, B. (2011, 8-10 April 2011). *Personalized Query Expansion based on phrases semantic similarity.* Paper presented at the Electronics Computer Technology (ICECT), 2011 3rd International Conference on.
5. Liu, F., Yu, C., & Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *Knowledge and Data Engineering, IEEE Transactions on, 16*(1), 28-40.
6. Palleti, P., Karnick, H., & Mitra, P. (2007). *Personalized web search using probabilistic query expansion.* Paper presented at the Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops.
7. Shamim Khan, M., & Khor, S. (2004). Enhanced web document retrieval using automatic query expansion. *Journal of the American Society for Information Science and Technology, 55*(1), 29-40.
8. Shen, X., Tan, B., & Zhai, C. (2005). *Implicit user modeling for personalized search.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany.
9. Sieg, A., Mobasher, B., & Burke, R. (2007). *Web search personalization with ontological user profiles.* Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.
10. Speretta, M., & Gauch, S. (2005, 19-22 Sept. 2005). *Personalized search based on user search histories.* Paper presented at the Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on.
11. Sugiyama, K., Hatano, K., & Yoshikawa, M. (2004). *Adaptive web search based on user profile constructed without any effort from users.* Paper presented at the Proceedings of the 13th international conference on World Wide Web.
12. Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y., & Chen, Z. (2005). *Cubesvd: a novel approach to personalized web search.* Paper presented at the Proceedings of the 14th international conference on World Wide Web.

13. White, R. W., Chu, W., Hassan, A., He, X., Song, Y., & Wang, H. (2013). *Enhancing personalized search by mining and modeling task behavior.* Paper presented at the Proceedings of the 22nd international conference on World Wide Web.