



A Review on Personalized Search Engine

Snehal D. Jadhav¹, Vaishali P. Suryawanshi²

Department of Information Technology, MITCOE, Pune, India

SavitribaiPhule Pune University, Pune

MITCOE, Paud Road, Kothrud, Pune- 411038

INDIA

Abstract: Trillions of pages are available on the World Wide Web which can provide users with effective means of information source. Web Search is the most frequent activity on the Internet but is still a difficulty when a user uses mobile devices with small screens and default keypad. A major problem of current web search is that the queries are short and obscure and the results of such queries are not specific to user needs. To relive this problem some of the search engines suggest users the semantically related terms to the user submitted queries so that user can select the terms as per need. In this paper, we provide review of various personalization techniques so that users are provided with more relevant data. The prior step for personalization process includes techniques to extract contents of user query. The next step involves applying personalization algorithms to give user more relevant search results.

Keywords: Ontology, Personalization, concept-based clustering, clickthrough, user profiling, concept.

I. Introduction

As the internet expand the indexed pages related to them also increases. With such a large volume of data, it becomes more difficult to find the relevant information in order to satisfy user queries using a simple web search engine. The queries submitted by users in search engines are limited and inexplicable. These short queries are not explicitly able to express what user needs exactly. As a result, lots of query results are retrieved which may be irrelevant to user query. To improve users' search many of the search engines provide query suggestions helping users to formulate their queries. When a user submits query, the terms relevant to the queries are provided to help user identify the terms he/she wants and therefore increasing effectiveness of retrieval. In this way, search engine will provide the same semantically related word to the user queries without considering user's personal specific interests.

Wilfred Ng, et al, proposed a method that provides personalized query suggestions based on personalized concept based clustering technique. This approach used clickthrough data to estimate user's preferences and then provide personalized query suggestions. DikLun Lee, et al proposed a personalization technique that captures the user's preferences in the form of concepts by mining their clickthrough data. An ontology-based user profiling is used to capture the user preferences. Location concepts are also given importance in order to give more specific results related to user profile. Three major functions are done in this personalization process: Capturing user preferences, reranking them and updating user profile. User preferences are captured by using the user clickthrough data and are treated as positive preferences.

II. Literature Survey

Query clustering techniques have been developed in various ways. The very first query clustering technique comes from information retrieval studies. Similarity between queries was measured based on overlapping keywords or phrases in the queries. Each query is represented as a keyword vector. Similarity functions such as cosine similarity or Jaccard similarity were used to measure the distance between two queries. One major limitation of the approach is that common keywords also exist in unrelated queries. For example, the queries, "apple iPod" and "apple pie", are very similar since they both contain the keyword "apple."

DikLun Lee et al [2], proposed a personalized concept-based clustering technique on the basis of variations of the queries submitted to the search engines with different meanings. For example, depending on the user, the query "apple" may refer to a fruit, the company Apple Computer or the name of a person, and so forth. Thus, providing personalized query suggestion (e.g., users interested in "apple" as a fruit get suggestions about fruit, while users interested in "apple" as a company get suggestions about the company's products) certainly helps users formulate more effective queries according to their needs. Proposed approach consists of the following four major steps.

First, when a user submits a query, concepts (i.e., important terms or phrases in web-snippets) and their relations are mined online from web-snippets to build a concept relationship graph. Second, clickthroughs are collected to predict user's conceptual preferences. Third, the concept relationship graph together with the user's conceptual preferences is used as input to a concept-based clustering algorithm that finds conceptually close queries. Finally, the most similar queries are suggested to the user for search refinement.

A. Personalized Concept-Based Clustering:

Algorithm:

Input: A Query-Concept Bipartite Graph G

Output: A Personalized Clustered Query-Concept

Bipartite Graph G_p

// Initial Clustering

- 1: Obtain the similarity scores in G for all possible pairs of queries using the noise-tolerant similarity function.
- 2: Merge the pair of most similar queries $q_i; q_j$ that does not contain the same queries from different users.
- 3: Obtain the similarity scores in G for all possible pairs of concepts using the noise-tolerant similarity function
- 4: Merge the pair of concepts $c_i; c_j$ having highest similarity score.
- 5: Unless termination is reached, repeat steps 1-4.

// Community Merging

- 6: Obtain the similarity scores in G for all possible pairs of queries using the noise-tolerant similarity function.
- 7: Merge the pair of most similar queries $q_i; q_j$ that contains the same queries from different users.
- 8: Unless termination is reached, repeat steps 6 and 7.

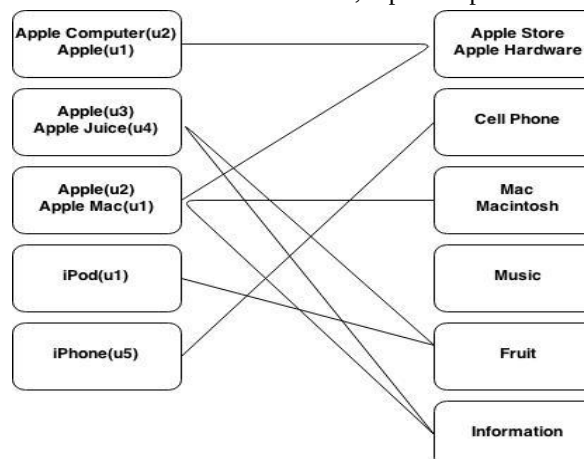


Figure. 1 Initial clustering

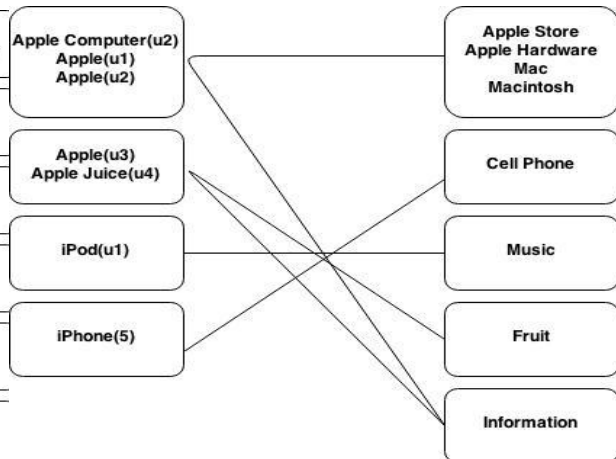


Figure 2. Community Merging

The main consequence of this technique is timing to start community merging is important for the success of the algorithm which affects the values of recall and precision which in turn affect the accuracy of the algorithm.

B. Personalization Based On Ontology

The personalization approach [3] is based on “concepts” to profile the interests and preferences of a user. Therefore, an issue we have to address is how to *extract* and *represent* concepts from search results of the user. In this paper an *ontology-based, multi-facet (OMF)* profiling method is proposed in which concepts can be further classified into different types, such as content concepts, location concepts, name entities, dates etc. As an important first step, we focus on two major types of concepts, **content concepts** and **location concepts**. A content concept, for example a keyword or key-phrase in a web page, defines the content of the page, whereas a location concept refers to a physical location related to the page.

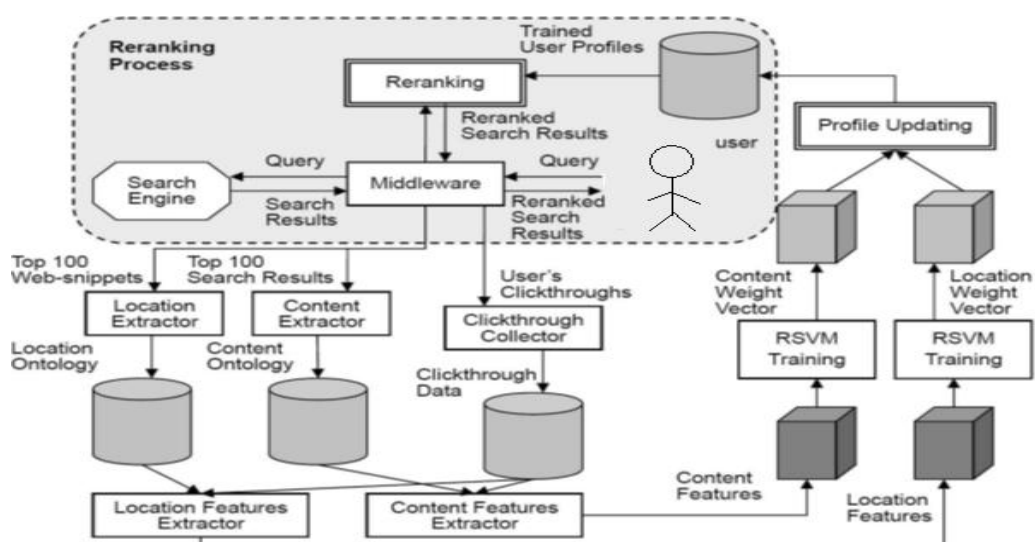


Figure. 3. Personalization Process

1. Joachim's Method

In Joachim method scanning of results is done from top to bottom. The document which has higher rank is displayed first and other results are displayed after that in descending order of ranking. If a user does not read a

document which is at a higher rank and reads on document at lower rank, it means that the user is not interested in document at the higher rank. Thus, Joachim's method concludes that the user prefers d_i to document d_j (denoted as $d_j <_{r_0} d_i$, where r_0 is the user's preference order of the documents in the search result list).

2. SpyNB Method

Similar to Joachim's method, SpyNB learns user behavior models from preferences extracted from click through data. SpyNB assumes that users would only click on documents that are of interest to them. Then we can treat the clicked documents as positive samples. However, documents that are not clicked are treated as unlabeled samples because they could be either relevant or irrelevant to the user.

3. RSVM

Ranking SVM is an application of SVM aimed to solve certain ranking problems. The main purpose is to improve the performance of the internet search engine. Click through data can be used as the input; RSVM aims at finding a linear ranking function. It is valid for as many document preference pairs as possible. It maps the similarities between queries and the clicked pages onto certain space. Calculates the weights between any two of the vectors obtained. Re ranks the search results based on the weights. If a user has visited the GPS location l_r , the weight of the location concept is incremented. Hence, we assume that the location that the user has visited a long time ago is less important than the location that the user has recently visited. Thus GPS plays an important role in location information.

C. Personalization Technique With Privacy

1. Concept and Entropy:

The *content and location entropies* [4] are introduced for measuring the diversity of content and location information from the search results of a query. In addition, the *click content and location entropies* were introduced to determine how much a user is interested in the content and location information associated with a query

$$H_{\overline{C}}(q, u) = - \sum_{i=1}^t p(\overline{c}_{i_u}) \log p(\overline{c}_{i_u})$$

$$H_{\overline{L}}(q, u) = - \sum_{i=1}^t p(\overline{l}_{i_u}) \log p(\overline{l}_{i_u})$$

2. Personalization Effectiveness:

A query result set with high content/location entropy indicates that it has a high degree of ambiguity. Thus, applying personalization on the search results helps the user to find out the relevant information.

$$e_C(q, u) = \frac{H_C(q)}{H_{\overline{C}}(q, u)} \quad e_L(q, u) = \frac{H_L(q)}{H_{\overline{L}}(q, u)}$$

Chen et al. studied the problem of efficient query processing in location-based search systems [7]. A query is assigned with a query footprint that specifies the geographical area of interest to the user. Several algorithms are employed to rank the search results as a combination of a textual and a geographic score.

3. Personalized Ranking Function

Once the user preferences are received, Ranking SVM is applied to get a personalized ranking function for rank adaptation of the search results according to user content and location preferences.

4. User Preferences Extraction and Privacy Preservation

Two feature vectors were introduced namely Content feature vector and Location feature vector to represent the content and location information associated with the document and to be used for training.

PMSE [4] incorporates a user's physical locations in the personalization process. We conduct experiments to study the influence of a user's GPS locations in personalization. The results show that GPS locations helps improve retrieval effectiveness for location queries (i.e., queries that retrieve lots of location information).

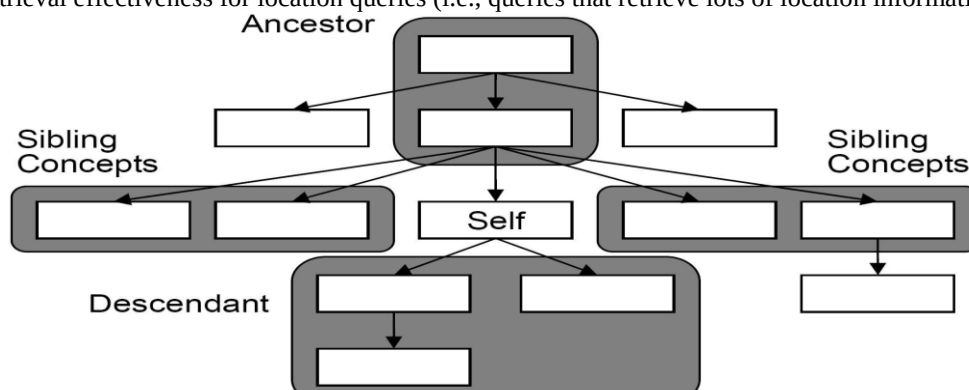


Figure.4. Parent-child relationships, 1) Ancestors, 2) Descendants, and 3) Sibling Concepts, in a concept ontology.**a) Content Feature Vector**

If content concept C_i is in web snippet S_k the value of content feature vector will be incremented by 1

$$\forall c_i \in s_k, \phi_C(q, d_k)[c_i] = \phi_C(q, d_k)[c_i] + 1.$$

Related Concepts are then give as:

$$\begin{aligned} \forall c_i \in s_k, \phi_C(q, d_k)[c_j] &= \phi_C(q, d_k)[c_j] \\ &+ sim_R(c_i, c_j) + ancestor(c_i, c_j) \\ &+ descendant(c_i, c_j) + sibling(c_i, c_j). \end{aligned}$$

b) Location Feature Vector

If location concept L_i is in a web snippet D_k , value of location feature vector is incremented by 1

$$\forall l_i \in d_k, \phi_L(q, d_k)[l_i] = \phi_L(q, d_k)[l_i] + 1.$$

Related Location Concepts are:

$$\begin{aligned} \forall l_i \in d_i, \phi_L(q, d_k)[l_j] &= \phi_L(q, d_k)[l_j] + ancestor(l_i, l_j) \\ &+ descendant(l_i, l_j) + sibling(l_i, l_j). \end{aligned}$$

III. Conclusion

The concept-based clustering technique used for personalization gives greater personalization to the user queries. The main critical factor affecting the effectiveness of the techniques was the time at which the agglomerative algorithm should be applied so as to get personalized clustering. Later the ontology based multi facet profiling technique was introduced which used content and location concept in ontology tree. Here the privacy preservation was not provided to the user profiles and to the user queries. The privacy of users was preserved by parameters such as minimum distance of the leaf node from the root in ontology giving user more accurate n secured results.

The future scope of these personalization techniques may involve relating the queries with the user name. By recording the historical places the user visited in near past, frequent travelling patterns can also be found out in order to give more specific results according to location concept.

References

- [1] Dr. C. R .Rene Robin and R Divya, "Onto-Search: An Ontology Based Personalized Mobile Search Engine", 2014
- [2] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept-Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008.
- [3] K.W.-T. Leung, D.L. Lee, and W.-C. Lee, "Personalized Web Search with Location Preferences," Proc. IEEE Int'l Conf. Data Mining (ICDE), 2010.
- [4] K.W.-T. Leung, D.L. Lee, and W.-C. Lee, "PMSE: A Personalized Mobile Search Engine", IEEE Trans. Knowledge and Data Eng., vol. 25, no. 4, April 2013
- [5] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006
- [6] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning User Interaction Models for Predicting Web Search Result Preferences," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
- [7] Y.-Y. Chen, T. Suel, and A. Markowetz, "Efficient Query Processing in Geographic Web Search Engines," Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
- [8] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002
- [9] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search," Proc. Int'l Conf. World Wide Web (WWW), 2007
- [10] S. Yokoji, "Kokono Search: A Location Based Search Engine," Proc. Int'l Conf. World Wide Web (WWW), 2001.