Content and Location based Information Retrieval System

J Swathi,
Asst. Professor,
Department of Computer Science and Engineering,
PSG College of Technology,
Coimbatore, India.

G Seethalakshmi,
Asst. Professor,
Department of Computer Science and Engineering,
PSG College of Technology,
Coimbatore, India

ABSTRACT
Personalized Web search is an effective means of providing precise results to different users when they submit the same query. As the amount of web information grows rapidly an efficient personalization approach that modifies the appearance of a website's content to satisfy a specific user's instructions or preferences is required. It is also essential to keep track of the change of interest of the user from time to time. An approach which involves a concept based user profiling strategy, along with the click-through data and keyword-based search, is developed. Concepts are split into content and location concepts and are maintained separately for monitoring the gradual transition in the interest of a user over the time. The user's interest is captured from the click-through information. Depending upon the links clicked and the concepts returned users' information access behavior is analyzed and re-ranking is performed to obtain the relevant results.

Keywords
Personalization, User-profiling, Click-through, tf-idf, Content and location concept.

1. INTRODUCTION
The availability of high-speed internet, high capacity networks and advanced websites has increased the amount of information over the Web at a quick rate. It has become much easier for the internet users to publish data over the Web using interactive websites like Facebook, YouTube and blogging,. With this information flooding, it has become difficult for a user to find out the right information over the Web. Search engines act as a tool in retrieving relevant information from the web. Traditional search engines like Google, Altavista, Yahoo, Bing, etc. bring thousands of search results for a given search query. It is almost impossible for a human user to go through the content of all the search results manually. Also the retrieval of documents is based on calculation of similarity measures showing how close each document is to the desired results [1]. The results produced by search engines are totally dependent on the keywords typed by the user. Users who are ignorant about the mechanism of information retrieval many times remain unsatisfied due to improper keyword selection in the query. Search engines that are presently available do not have any mechanism to categorize the output based on variation in user requirement.

An information retrieval system must make sure that everybody it is meant to serve finds the information needed to accomplish tasks, solve problems, and make decisions, no matter where that information is available. Personalized content retrieval aims at improving the retrieval process by taking into account the particular interests of individual users. However, not all user preferences are relevant in all situations. Personalized search is an important research area that aims to resolve this ambiguity. It provides exact results to different users when they submit the same query. In personalization, the previous search behavior of a user is used for an efficient and focused search. By knowing the history of different search behavior of many users, such personalized search could offer customized search solutions for each user. To increase the relevance of search results, personalized search engines create user profiles to capture the users' personal preferences and as such identify the actual goal of the input query. Since users are usually reluctant to explicitly provide their preferences due to the extra manual effort involved, recent research has focused on the automatic learning of user preferences from users' search histories or browsed documents and the development of personalized systems based on the learned user preferences.

The main objective of this paper is to develop a concept based user profiling strategy. The concepts are split into content and location concepts. Content concept refers to a word that is closely related to the query term whereas location concept refers to the geographic location in which the user is interested. Both the concepts are extracted and maintained separately. The keyword based user profiling strategy is also included here. This is done by capturing the term frequency-inverse document frequency of the pages. The user's interest is captured from the click-through information. Depending upon the links clicked and the concepts returned users' information access behavior is analyzed. To perform re-ranking, the results from the backend search engine are combined with the user profile. This gives relevant information to the user.

The rest of the paper is organized as follows: In section 2 the existing work is discussed. Section 3 presents the proposed work. Section 4 gives the conclusion for the paper.

2. EXISTING WORK
Literature gives some specific attempts of personalization of web search. A personalized web search approach based on probabilistic query expansion and collaborative filtering aimed at exploiting browsing history of a user to get probabilistic correlations among the query terms and the documents terms. It performs a collaborative filtering and a pseudo query term selection for better query expansion [2]. An adaptive personalized approach based on context was proposed. It works on each user's need in different situations by using a process of a context-based adaptive personalized search. Then, it uses three technologies semantic indexing for web resources, modeling and acquiring user context and semantic similarity matching between web resources and user context [3].

A preference model that provides an algorithm for the selection of preferences related to a query was projected. It also incorporates an algorithm for the progressive generation

of personalized results, which are ranked based on user interest. [4]

A strategy for personalization of Web search in which a user's search history is collected without direct user involvement was described. The user's profile is constructed automatically from the user's search history and is augmented by a general profile which is extracted automatically from a common category hierarchy. The categories that are likely to be of interest to the user are deduced based on his/her query. These categories were used as a context of the query to improve retrieval effectiveness of Web search [5].

Click-through data is important for tracking user actions on a search engine. A method to learn users' clicking and browsing behaviors from the click-through data using a scalable implementation of neural networks called RankNet was proposed [6].A method to use document preference mining and machine learning to rank search results according to user's preferences was developed [7].

Search queries can be classified into two types, content and location**.** A classifier was built to classify geo and non-geo queries, and the properties of geo queries were studied in detail [8]. It was found that a significant number of queries were location queries focusing on location information. Hence, a number of location-based search systems designed for geo queries have been proposed.

A parser was employed to extract location information from web documents, which was converted into latitude-longitude pairs or polygons. When a user submits a query together with the location information specified in a latitude-longitude pair, the system creates a search circle centered at the specified latitude-longitude pair and retrieves documents containing location information within the search circle [9].

A hybrid index structure to handle both content and location-aware queries was proposed. The system first detects geographical scopes from web documents and represents the geographical scopes as multiple minimum bounding rectangles (MBRs) based on geographical coordinates. A hybrid index structure is used to index the content and location information of the web documents. A user is required to present their content and location interest in their search queries. A ranker is then employed to rank the search results according to the content and location relevance using the hybrid index [10].The major drawback of [8], [9] and [10] is that it requires the user's to specify the location preference explicitly. An approach that considers the dynamic interest of the users based on ontology-driven dynamic representation of the semantic context was developed. It exploits the contextual information and integrates it with the retrieval system. [11]

3. PROPOSED WORK

Personalized content retrieval aims at improving the retrieval process by taking into account the particular interests of individual users. There are a number of approaches for web personalization. Each of it differs in the way the user profiling strategy is performed. Figure 1 shows the generalized personalization approach.

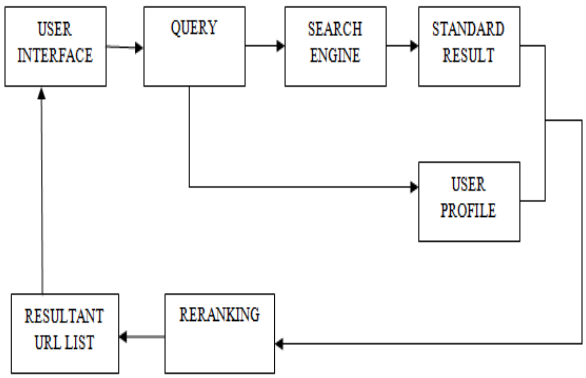

Figure 1: General personalization approach

A user interface is created to receive the search query and return the results to the user. When a user submits a query, the search results are obtained from the backend search engines (e.g., Google, MSN Search, and Yahoo). The search results are combined and re-ranked according to the user's profile. The change in the interest of the user can be monitored using the user profile.

The user profiling used in our approach is a combination of 1) keyword based user profiling strategy and 2) concept based user profiling.

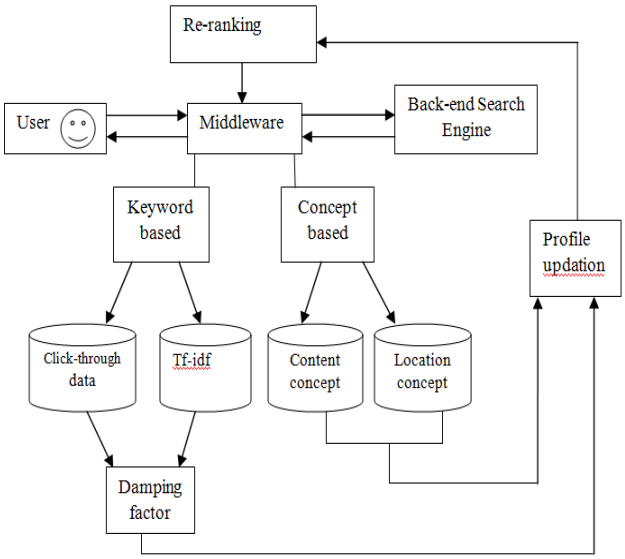

Figure 2: Proposed approach

Figure 2 shows an approach that considers the traditional keyword based retrieval system along with the concept based user profiling to return results relevant to the user's preference.

3.1 Keyword Based User Profiling

The most common approach of user profiling used in the traditional search engine is the keyword based strategy. This approach involves computing the term frequency-inverse document frequency. The term count ($tf_{i,j}$) in the given document is the number of times a given term appears in that document. The inverse document frequency (idf_i) is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. Then the tf-idf of the document is computed as follows

$$(tf\text{-}idf)_{i,j} = tf_{i,j} \times idf_i \qquad (1)$$

The tf-idf value for a term will always be greater than or equal to zero.

The user's interest is captured from the click-through information. The pseudo code for re-ranking based on click-through information is given below.

1. User Login
2. If User Profile exist then
   - If Query keyword exists then
     - Take URL which has the highest frequency count in user profile.
     - Add this URL list with the standard set of URL list from search engine.
     - Display the resultant URL list.
   - Else Add query keyword into user profile
3. Else Display result from back-end search engine
   - Create a user profile
4. Add user preference into user profile
5. End

In our keyword based user profiling strategy we combine the click-through information with the tf-idf approach to obtain relevant results

Algorithm 1: Keyword based user profiling

Input: Search Query

Output: Re-ranked URL list

1. Create a user profile for each user.
2. Obtain the top 10 results(links) for each search query given by the user
3. For each link clicked by the user the count value is incremented.
4. The tf-idf of the clicked document is captured
5. Re-ranking is done based on the damping factor formula(4) given below

$$Damping\ factor = \frac{(0.65 \times \text{count}) + (0.35 \times \text{tf} - \text{idf})}{2} \quad (2)$$

The major drawback of the traditional keyword-based approaches for user profiling is that it is unable to capture the semantics of user interests. It just matches the words in the document with the search query and produces the result.

3.2 Concept Based User Profiling

A combination of the concept based user profiling strategy with our keyword based approach would give better results. The concepts are split into content and location concepts and are maintained separately. A content concept, like 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.

3.2.1 Content Concept Extraction

The content concept is extracted from the web snippets according to user preferences. Web-snippet denotes the title, summary and URL of a Web page returned by search engines. Extraction of concepts is similar to the problem of finding frequent item sets in data mining. When a user submits a query to the search engine, a set of web –snippets are returned to the user for identifying the relevant items. We assume that if a keyword or a phrase appears frequently in the web-snippets of a particular query, it represents an important concept related to the query because it coexists in close proximity with the query in the top documents. The following support formula is used for measuring the interestingness of a particular keyword t_i with respect to the returned web snippets arising from a query q:

$$support(\text{ci}) = \frac{sf(\text{ci})}{n} \times |\text{ci}| \quad (3)$$

Where n is the total number of web-snippets returned, sf (c_i) is the snippet frequency of the keyword/phrase c_i (i.e., the number of web-snippets containing t_i), and | c_i | is the number of terms in the keyword/phrase c_i.

The related concepts are grouped using the similarity value of Church and Hanks' formula given as follows:

$$sim(t1, t2) = \frac{n.df(t1 \cup t2)}{df(t1) \times df(t2)} \Big/ \log n \quad (4)$$

where n is the number of documents in the corpus, df(t_1U t_2) is the joint document frequency of t1 and t2, and df(t) is the document frequency of the term t. Click-through data is used to capture the analysis behavior of the user [12] [13].

3.2.2. Location Concept Extraction

Location concept extraction method is composed of three steps.1) capturing the entire document of the web page clicked, 2) Preprocessing the documents, and 3) Matching the terms present in the document with the pre-defined ontology.

A document usually embodies only a few location concepts. As a result, very few of them co-occur with the query terms in web snippets. Hence location concepts are extracted from the full documents. Pre-processing of the documents involves the stop word removal and stemming. Porter stemmer algorithm can be used in this case. The preprocessed documents are then matched with the pre-defined ontology. The geographical relationships among the locations are already been captured as facts. For this purpose, predefined location ontology can be created using protégé tool.

An ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. It describes the concepts and relationships that are important in a particular domain. Figure 3 shows the location ontology created using protégé tool for India domain.

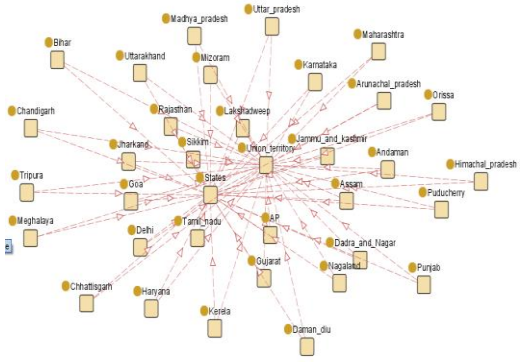

Figure 3: Ontology for location domain

Table 1 shows the set of links and the concepts returned for the search query "university" and the links that are clicked by the user.

Table 1: Search query "university"

Doc	Links Returned	Content	Location
d1	en.wikipedia.org/wiki/ University	research, faculty	-
d2	www.mu.ac.in	faculty, science, department	Mumbai
d3	www.du.ac.in	science, faculty	Delhi
d4	www.unipune.ac.in	department	Pune
d5	**annamalaiuniversity.ac.in**	**faculty, engineering**	**Tamilnadu**
d6	www.ignou.ac.in	research, science	Delhi
d7	**www.annauniv.edu**	**engineering**	**Tamilnadu**
d8	www.indiauniversity.me	engineering	India
d9	www.puchd.ac.in	engineering, science	Punjab
d10	www.caluniv.ac.in	Research	Calcutta

Depending upon the links clicked it is identified that the user is interested in "Engineering colleges in Tamilnadu". So the location concepts related to Tamilnadu are explored from the ontology created. Re-ranking is then done based on these concepts and the click-through information to return relevant results to the user.

4. CONCLUSION

As the amount of information on the Web is increasing at a very high pace there are thousands of results returned for a single search query. This is because the traditional backend search engines consider the keyword/query to retrieve the documents without considering the dynamic change in interests of the user. So an efficient personalized intelligent information retrieval system that returns result based on the interest of the user was proposed. The approach is a combination of the keyword based and content based user profiling strategy. In the keyword based user profiling technique the click-through information of the user is maintained which helps to identify the interest of the user. Based on this information and the traditional tf-idf technique the damping factor is calculated. To monitor the interest of the user the content and location concepts are extracted. Content concepts are extracted from the web snippets and the location concepts are extracted from the ontology created. Finally the user profile is updated using the damping factor calculated and the concepts extracted which in turn is used for re-ranking the links that are relevant to the user. However this approach does not completely consider the implicit interest of the users but can reduce the inaccuracy to certain extent.

As a future work few other concepts such as people name, profession and their interests can be added to the user profile. Also the search queries can be clustered and used to provide recommendations to users. Also the time taken for personalization can be reduced by optimizing the algorithm even further.

5. REFERENCE

[1] Margaret H. Dunham, "Data Mining Introductory and Advanced Topic", Delhi, Pearson Education, 2003.

[2] Pallavi Palleti, Harish Karnick, Pabitra Mitra, "Personalized Web Search using Probabilistic Query Expansion", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshop,2007.

[3] Xuwei Pan, Zhengcheng Wang, Xinjian Gu, "Context-Based Adaptive Personalized Web Search for Improving

[4] Information Retrieval Effectiveness", IEEE, 2007.

[5] Georgia Koutrika, Yannis Ioannidis, "Personalized Queries under a Generalized Preference Model", Proceedings of the 21st International Conference on Data Engineering, 2005.

[6] Fang Liu, Weiyi Meng, "Personalized Web Search for Improving Retrieval Effectiveness", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 1, January 2004.

[7] E.Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information", in Proc. of ACM SIGIR Conference, 2006.

[8] T.Joachims, "Optimizing search engines using click through data," In Proc. of ACM SIGKDD Conference, 2002.

[9] Q.Gan, J.Attenberg, A. Markowetz, and T. Suel,"Analysis of geographic queries in a search engine log", in Proc. of the International Workshop on Location and the Web, 2008.

[10] S. Yokoji, .Kokono, "A location based search engine", in Proc. of WWW Conference, 2001.

[11] Y.Zhou, X.Xie, C. Wang, Y. Gong, and W.Y.Ma, "Hybrid index structures for location-based Web search", in Proc. of CIKM Conference, 2005.

[12] David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas,Yannis Avrithis, "Personalized Content Retrieval in Context Using Ontological Knowledge", IEEE Transactions On Circuits And Systems For Video Technology, Vol. 17, No. 3, March 2007.

[13] K. W.-T. Leung, W. Ng, and D. L. Lee, "Personalized concept-based clustering of search engine queries," IEEE TKDE, vol. 20, no. 11, 2008.

[14] V.K.Priyanka Kolluri, A.Bala Ram, "Efficient Personalized Search using Ranking SVM", International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012.