Personalized Web Search: User Modeling using Implicit Feedback from Click Through Data
Geetanjali Surange, Asso. Prof., ITM Gwalior

e-mail: geetanjalis@gmail.com
Suhel A. Khan
Lecturer, RJIT Tekanpur

ABSTRACT
As the Web is growing very fast and its structure is heterogeneous it has become very difficult for an individual to find some information of a particular relevance on web. Modern search Engines helps the user to organize their search to some extend by providing advance search features which requires the user to specify its personal interest regarding the search in advance or explicitly. Also as the search is based on the keywords provided in the query, most search engines return the same results even if the query is submitted by another person with some different intention. Therefore web search must result should adapt to users with different information needs. In other words the search technique must learn the user behavior and return the results accordingly. Current web search technologies return, large sets of results, based on a textual query, regard less of any special needs of specific type(s) of user(s). Because of this reason personalization of web search to retrieve documents as per the user’s specific interest without any explicit feedback from user has become necessary. In this paper we have proposed a theoretical framework for building a user model without explicit feedback form user and without any human assessment about the relevancy between the terms presented for query. The proposed framework also ensures that the click through data used is not biased towards the order of the result presented by the popular search engine.

KEYWORDS
Web mining, information retrieval, user modeling, web personalization.

INTRODUCTION
Current web search technologies return, large sets of results, based on a textual query, regard less of any special needs of specific type(s) of user(s). Because of this reason personalization of web search to retrieve documents as per the user’s specific interest without any explicit feedback from user has become necessary. For achieving this, a user model can be created by using information present in user’s browsing history and the relevance between the terms presented by user in the query. It was there for proposed that web mining can be done in three dimensions: 1.Web Content Mining 2.Web structure mining 3.Web usage mining. These techniques explore the different properties of the web. It was soon realized that using any one of the dimensions is not sufficient to personalize the search [1]. In this paper we have proposed a theoretical frame work for building a user model without explicit feedback form user and without any human assessment about the relevancy between the terms presented for query. The proposed framework automatically computes the relevance between the terms presented based on the previous browsing sessions of the user and uses the current browsing session data to modify user model. Once the user behavior is agreed upon, after “n” no. of user sessions, the frame work can return result of a search as per the user interest. The proposed framework also ensures that the click through data used is not biased towards the order of the result presented by the popular search engine.

The rest of the paper is organized as follows. Section 2 presents Personalization and user modeling task. Section 3 reviews the earlier efforts done in the field of personalization. Section 4 presents the model and methodology of the proposed frame work. Section 5 explains the algorithm and experimental setup. Section 6 describes future work and conclusion.

PERSONALIZATION
The web was designed as an information space with the goal that it should be useful not only for the human- human but also that machine would be able to participate and help in proper communication. This simply means that the machine should be intelligent enough to simulate human behavior. One of the major obstacles to achieve this is the fact that most of the information on the web has been designed for human consumption and documents are designed as per the creator’s perception about the subject matter of the document. Because of this feature of the web, currently it has become distributed but heterogeneous information space, which brings the people great trouble in finding needed information.

For the web to reach to its full potential, however, it is necessary to improve web services, make it more comprehensible and increase its usability. But comprehending and using the heterogeneous information stored on web poses a
great challenge to the researchers. Because of these issues data mining has become the most popular tool among the research community for finding the solution to the problem of delivering personalized results of web search. Some efforts have been made by using data mining techniques, to make efficient use of the web data. But doing this alone may not solve the problem as the data stored on the web is much more sophisticated than commercial data base. Besides the structure and the dynamic property of this data base, it is also very huge which is why designing techniques to improve the recall and precision values has become a challenging job.

Applying the mining techniques to the data available on web is known as web mining. Web mining techniques appears a suitable solution for minimizing the irrelevancy and redundancy in document retrieval and increasing the efficiency thereby increasing the precision and recall indices of a search engine. However when it comes to implementation there are many challenges posed by web itself on the mining:

- Structure of the web is far more complex than any of the commercial data base structures.
- Besides growing rapidly it is also highly dynamic information resource which receives continuous updates thereby affecting the linkage structure and access records.
- Serves a large community of the users hence a common level of searching methodology can not be fixed.
- Lack of relevance prevails in the access patterns of a user as most of the users normally access only small portions of the information source.

Hence web mining concept can be redefined as using data mining techniques to automatically discover and extract connotative and potentially useful patterns from the web documents and services. Thus it has been suggested [1] that the web mining can be done in three ways: web content mining web usage mining and web structure mining.

Web content mining is the technique where the documents are retrieved based on the contents of the documents. Most popular way of doing this is through the keyword / object search. This approach purely depends on how the creator of the web page has designed the document and what is his or her perception about the subject / keyword. Many efforts have been done to improvise this technique. One way of doing this is to ask the user about it’s preference regarding the domain where it would like to search the document [2]. A tree / hierarchical structure of the domain is stored and whenever a query is given only the desired sub tree is searched.

Another way is to improve the way the documentation of the web is done i.e. making Semantic Web [7]. A common feature of both the approaches is that they are static in nature and are non-adaptive. For a user to take advantage of the first approach, it is necessary to understand the domain classification as per the search engine, whereas most of the users do not understand such a classification. The second approach produces smart results but renovation of the structure of all the documents on web needs to be done, which requires a long transition period. Web pages do not only store the contents but also some hyperlinks to other relevant pages which are in some or the other way are related to the original document. Analysis of these link structures can reveal much useful and hidden information. By establishing association rules among these links or by forming clusters of similar web pages a user can be serviced with more accurate query results. The goal of web usage mining is to find the user’s access patterns quickly, such as the frequent access page set, frequent traversal path and users clustering. In this technique the record of user’s access to web pages, web server access log, the time a user spend, no. of times a particular link is clicked etc. are used as parameters for modeling a user category. Use of any one technique of the above does not seem sufficient as these are not the only parameter that models a user behavior of searching. Hence a user must be modeled by considering all of these aspects together, if possible.

RELATED WORK

Web personalization is understood in various dimensions. One way of doing this is categorization of users based on demographic information provided by the users at the time of selecting the style for personalization. An example of this is Google Personal search through igoogle. This approach requires that the user must exactly know what information is needed prior to searching. The research is also going on to modify the structure of the web documents and make it semantic so that the documents are then retrieved on the basis of the meaning of the query and not the terms present in the query [7, 8]. This approach seems very promising but is a long term project, the acceptability and usability of which depends on the user community. Another way to personalize the search is to classify users on the basis of pre-calculated classes. The classes may be pre-calculated through users browsing history. A classification of the on-line users to one of the predefined classes is typically based on similarity calculation between each predefined pattern and the current session. The current session is assigned to the most similar cluster [13, 14]. Further this approach is modified to accommodate fuzzy classification so as to prevent some users to become outliers [3].

Some authors have constructed user profiles on the basis of modified collaborative filtering with detailed analysis of user’s browsing history in one day [1]. They have pointed out that data obtained from server logs are not sufficient and reliable. Hence recommender system proposed by them works on two levels: 1. Using pure browsing history 2. Using modified collaborative filtering. The drawback of this approach is that a term is added to user profile as soon as the page is browsed where as other click through data can also be used to more accurately guess the user preferences. As per our understanding this is classification of users in already existing classes and not
typical personalization. User profiles are also constructed on the basis of ontology [8]. Some efforts have also been done to refine the search process by re-defining the queries and then submit it to the search engine. Refined queries are then clustered to form user’s profiles [13]. In this approach also only visiting a page makes it interesting enough to update user profile. Another method to personalize is to discover association among various links accessed by the user through its sessions [12].

Another interesting effort has been done in actual personalization of users’ interest in which they have considered that every user’s behavior is different on same search results obtained through same search query [12]. They have used two properties of a document for modeling users i.e. attractivity and perseverance. They have assumed that these properties depend on the popularity of the document among the similar user community and distance of that document from last selection. Normal user behavior suggests that after a certain no of unattractive documents the user stops navigating the search results. Efforts have also been done to construct user profiles using relevancy between the terms of the queries presented in current session and in earlier sessions [2].

MODEL AND METHODOLOGY
The model proposed here is a theoretical frame work for building a user model without explicit feedback from user and without any human assessment about the relevancy between the terms presented for query and the document. This relevancy is computed on the basis of past sessions by user and the contents of the documents. Then the documents with the higher relevance ratio are presented to the user. The current user session data is used to update the user’s profile for future reference. Two types of parameters are considered for constructing user model: static parameters and dynamic parameters. Static parameters are those which remain same for all the users belonging to the same category of users who are surfing the web, such as relevancy of documents with the specific category measured by the popularity of the document. Dynamic parameters vary from user to user and are generated through user’s navigation pattern.

Static parameter used in the model is: Term-document relevancy which is maintained in a 2-D matrix T. This matrix is similar to the term document matrix used by Jingqiu XU et.al.[2]. The relevancy calculation done here is based on the occurrence of terms in the document. We have tried to improve upon the modality of updating the matrix. The matrix is updated every with every user session with the browsing patterns of a user and for first ‘n’ sessions it keeps on constructing new columns with respect to the terms that relates to a document.

Dynamic parameters used are extracted from user logs obtained from the proxy server. The parameters are : no. of clicks on the page, time spent on the page, no of links followed recommended by that page, last time when page was visited (this is to be maintained to restrict the no. of days for which history will be recorded).

The proposed system works as follows: The results generated by the search engine are feed to the proposed system, where first it goes to relevancy calculation module which calculated the relevancy of the document with the terms presented. The newly calculated relevancy matrix is then passed to the filtering module. User’s profiles are also updated by the relevancy matrix for future use and search results are filtered using current relevancy matrix. The users sessions are maintained as vectors \( U = \{ p_1, p_2, p_3, \ldots, p_n \} \) where \( p_i \) indicates the pattern accessed by the user in \( i^{th} \) session. A pattern \( p_i \) stored as a record containing all dynamic parameters as mentioned above. Initially the users are initialized to null and user profiles start building as users navigates through web. After ‘n’ no sessions of a particular user, the system starts returning relevant results based on the learned profiles. The architecture of the system can be represented as follows:

4. EXPERIMENTAL SETUP AND ALGORITHM
The algorithm is designed with certain assumptions. First is that there are a fixed no of users taken as subjects. A proxy server is set up to record web logs of each user separately. The web logs format used is Common Log Format recommended by NCSE. The fields in the Common log file format are:

- \texttt{host}: The IP address or host/subdomain name of the HTTP client that made the HTTP resource request.
- \texttt{rfc931}: The identifier used to identify the client making the HTTP request. If no value is present, a "-" is substituted.
- \texttt{username}: The username, (or user ID) used by the client for authentication. If no value is present, a "-" is substituted.
- \texttt{date:time timezone}: The HTTP request. The request field contains three pieces of information. The main piece is the requested resource (index.html). The request field also contains the HTTP method (GET) and the HTTP protocol version (1.0).
- \texttt{statuscode}: The status is the numeric code indicating the
success or failure of the HTTP request.

bytes
The logs thus recorded are refined and cleaned so as to represent useful data about user’s browsing history. Second is that the browsing patterns of first ‘n’ sessions will be recorded as test data for learning module. Once the training data is available then cluster of users are created based upon similarity between various users based on access patterns of the user. These clusters are used to guide the user for future sessions. The similarity is calculated based on Jaccards Formula[3].

Algorithm:
1. Initialize $U_i$
2. Initialize Term-Document Relevancy Matrix $T$
3. Collect Training data for ‘n’ session for each user from web logs i.e. where $i$ is the no of users and $p_i$ is the pattern accessed by the user in a session
   \[ U_i = \{ p_j | j <=n \} \]
4. Update $T$ matrix on the basis of no. of times the term occur in the document and no of users accessing the document $d_k$ where $k$ is the no of documents viewed in user’s $i^{th}$ session
5. Update user profile with the help of updated matrix $T$ and current user browsing pattern
6. Create clusters $C' = \{ C^i | i <=m, m is the no of users under consideration \}$ of similar users to guide the user for the next session.
7. Re-rank the result obtained as output of search engine with ranks calculated on the basis of $U_i$ and $C'$. The ranks thus obtained are randomized in groups of 10 documents to avoid biasing of selection of a particular link because of the position of the document in the search results.

The process of collecting navigation patterns works as follows:
1. Take a snapshot of the web server log for every ‘t’ time.
2. Synthesize / clean the log to represent only useful information.
3. Separate the record of the each user from the collective log.
4. Create / Update data structure $U_i$ to include the current user session’s browsing pattern.

The clusters of users are created based on similarity calculated on the basis of the dynamic parameters of individual user’s profile.

CONCLUSION
The model proposed here is a theoretical frame work for building a user model without explicit feedback from user and without any human assessment about the relevancy between the terms presented for query. The relevance between the terms of the query is computed on the basis of past sessions by user and the contents of the documents. Then the documents with the higher relevance ratio are presented to the user. The current user session data is used to update the user’s profile.

5. FUTURE SCOPE
The frame work provided above is only a theoretical proposal of an innovative way of personalizing the search. The results are obtained only on dummy data and are producing better results in terms of personalization. The algorithm is to be developed and tested for real time data by setting up actual experimental setup as suggested above. Further the algorithm may be optimized as there are many iterations involved. The idea of proposing such a frame work is to achieve actual personalization.

REFERENCES
Personalized Web Search: User Modeling Using Implicit Feedback
From Click Through Data

International IIPWM’03, Advances in Soft Computing,
Springer Verlag
2003, 529-534.

Integrating Web Usage and Content Mining for More Effective
Personalization. LNCS 1875 Springer Verlag, 156-176