Improving the Information Retrieval System through Effective Evaluation of Web Page in Client Side Analysis

Rekha.C
Assistant Professor,
R. D. Government Arts College,
Sivagangai,
Tamil Nadu, India.

Usharani.J
Assistant Professor,
M. K .U. College,
Madurai,
Tamil Nadu, India

Dr. K. Iyakutti
CSIR Emeritus Scientist,
School of Physics,
Madurai Kamaraj University,
Madurai,
Tamil Nadu, India.

ABSTRACT
To improve the information retrieval system for user, programmers have to learn a user's preferences accurately. In order to optimize retrieval accuracy, modeling the users appropriately based on their preferences and personalizing search according to each individual user are important. Implicit feedback information improves the user modeling process. The advantage of implicit modeling is effectively improving the user model without extra effort of user. Several implicit feedback features are used to develop user modeling process. We present a new model to find a user's preferences from click through behavior and using the exposed preferences to adapt the search engine's ranking function for improving search service. In this proposed model, the combination of viewed and stored document summaries is used.

1. INTRODUCTION
Today, World Wide Web has turned to be the largest resource of information available in this world and plays an important role as an information channel in our daily life. Moreover, the integer of Web pages is still rising rapidly and the types of information are varying so that users visit Web pages for different purposes. In order to consent people to get their target information effectively, many researchers are paying attention to the improvement of intelligent information delivery.

Each user has a specific goal while searching for information through entering keyword queries into a search engine. Keyword queries are inherently ambiguous but often formulated while the user is occupied in some larger task. A center part of intelligent information delivery is to build user profiles based on the content of visited Web pages. Due to the overload of information on the web, the need to automatically construct a user profile to assist the user in everyday browsing and searching tasks is steadily escalating, primarily for web personalization purposes. Building a user profile that adapts to a user's daily interests is a demanding task.

Personalization is a course of action, by which it is feasible to give the user optimal support in accessing, retrieving and storing information, where solutions are built so as to fit the preferences, their characteristics and the taste of the individuals [1]. Effective personalization of information access involves two significant challenges: perfectly identifying the user context, and organizing the information in such a way that matches the particular context. Since the understanding of user interests and preferences is vital element in identifying the user context, most personalized search systems employ a user modeling component to identify the user preference. Server side analyses cannot observe a user continuously if the user leaves a site to visit other Websites. On the other hand, through a client-side analysis, a user's interest can be analyzed using various sites, and a user model can be constructed using a wealth of information [2].

Many approaches and techniques have been used to generate a personalized information system. Information retrieval system is critical for overcoming information overload A foremost inconsistency of existing retrieval system is that they generally lack user modeling and are not adaptive to individual users, resulting in inherently non-optimal retrieval performance [3]. To get better information retrieval system for user, programmers have to learn a user's interest and information needs accurately based on personal search history information. Therefore, the information of personal search history should be appropriately stored, managed and exploited. The user modeling process is the critical process for user personalization techniques.

In order to optimize retrieval accuracy, modeling the users appropriately based on their interest and personalizing search according to each individual user are important. The key goal of user modeling for information retrieval is to exactly identify a user's information need, which is fatefully a very difficult task. Originally based on keyword matching between query and documents, the user model is constructed, and then this user model is effectively updated based on interest which is gathered explicitly or implicitly from user. An effective way to improve user modeling in information retrieval is to ask the user specify which documents are relevant. But unfortunately, in real world applications, users are usually unwilling to make the additional effort to afford relevant examples for feedback [4]. Another way to improve user modeling is based on implicit relevance feedback information, in his way, without any extra user effort, one can improve the user modeling from the user's browsing behavior. Indeed several Preview studies have shown that implicit user modeling can improve retrieval accuracy.

Implicit relevance feedback for ranking and personalization has become an active area of research. Implicit relevance measures have been studied by several research groups. Incorporating implicit feedback can augment other features improving the accuracy of competitive web search ranking algorithms by as much as 31% relative to the original performance [5].

2. USER MODELING
To present perfect personalization for any particular users, the system needs to be aware of the user's goal, preferences, as well as the overall contest relevant to the user [6]. User

model represents knowledge about the user that can be anything from general information of the user such as name, sex, age etc. to specify user characteristics. User modeling is the core of each personalization information system [1].

User modeling is concerned with the representation of the user's knowledge and interaction within a system to adjust the system to the needs of the user. The gain of utilizing a dynamic user model within a system is to let that system to adapt over time to a specific user's preferences, work flow, goals, etc. To realize this benefit, the user model must effectively represent the user's knowledge and intent within the system to accurately predict how to adapt the system.

There can be a several types of user model, and models can be classified along the four major important issues listed below [7].

- Whether the user model is constructed for Individual user or Canonical user?
- Based on the Source of modeling information the model can be constructed in two ways.
  in explicit way, model constructed using explicit information which is provided by user.
  in another way model constructed by the system on the basis of user's behavior.
- The time sensitivity of the model: the model can consist of short-term information, highly specific information or Longer-term information, more general information.
- Based on the updating method, the model can be static or dynamic.

User modeling is not a compulsory part of software but it improves the information retrieval system. Any information stored about the user or usage pattern is not a user model unless it is used to get some explicit assumption about the user [8].

There are various approaches and methods being used to create the user model [1] outline. Two major user modeling approaches are as follows:

1) Overlay modeling: Overlay modeling based on adaptive hypermedia research is used to find the initial knowledge estimation.
2) Stereotype user modeling: here the model is constructed by observing the user's behavior and it is used in web mining domain.

According to Perugini et al [9], user modeling can also be categorized based on the way data are acquired by explicit feedback from user or implicit feedback which is gathered from user's behavior. In explicit user modeling, the user has to supply the personal information or user's interest to the system. In contrast, implicit user modeling collects the data from user's behavior or task implicitly. Many personalization approaches are based on some type of a user profile. The user profile is a data instance of a user model that captured based on the user's interaction. User modeling using short-term personal search history, particularly the clicked document summaries, can improve information retrieval process significantly [10]. Long-term search history contains helpful information that can assist to get better results for the present query, the history also has a lot of noise and it is not instantly understandable. Thus one can extract the most useful information and at the same time avoid introducing noise or information [11].

Gasparetli and Micarelli propose a user model which tries to represent human memory. This user model based on user profile consists of two keyword vectors. One vector is used to represent the short-term interests whereas the other represents long-term interest [12]. Ahu Sieg et al [13] differ from these approaches, since they utilize a concept based model as opposed to representing the profile as keyword vectors.

User profile or usage log analysis for user modeling can be done not only at the server side but also at the client side. Server side analyses have exposed fine performances in consumer analyses of business websites. But server side analyses have some critical limitation, and the contents of a server are not adequate to construct a universal user model. On the other hand through a client side analysis, a user's interest can be analyzed using various sites, and a used model can be constructed using a wealth of information. However client side analysis also has some drawbacks in that prior knowledge of websites cannot be used , and hence the analysis should be done mostly based on collected usage logs[2]

3. IMPLICIT FEEDBACK AND FEATURES

In general an effective way to improve the used modeling process, first one has to identify the user type and their interest. To enhance that process, the information about the user in two ways can be gathered. (i)In explicit way: Getting information\Interest of user directly from user by asking some questions. (ii)An implicit way: Analyzing the user's behavior, the interest of users is identified. In this way, without any extra user effort, the user modeling from the user's browsing behavior can be improved. Several user behavior features are used for implicit feedback

Modern web search engines revise the user model in implicit way based on large number of features. These features are used to represent post-search navigation history for a given query and search result URL. The features used to represent user interaction with web search result may be classified into three groups.

(i) Click through features
(ii) Browsing features and
(iii) Query text features

Ranking the web pages is the fundamental problem in information retrieval. The implicit feedback is the action that users take when interested with the search engine can be used to improve the ranking Fox et al [9] explored the association between implicit and explicit measures in web search, and created Bayesian models to bond implicit measures and explicit relevance judgments for both individual queries and search sessions. In this work, modeling effort was aimed at predicting explicit relevance judgments from implicit user action like dwell time, scroll time and click through features and not specifying at learning ranking functions.

According to Bayesian decision theory, the optimal decision at time t is to choose a response that minimizes the Bayes risk

$$rt^*=\text{argmin}_{r\in} \; RatMLat,r,mtPmtU,D,At,Rt\text{-}1dmt \quad (1)$$

Here the space of all possible user models is M, L(a,r,m) is the loss function where $a\in A$ is a user action $r\in R(a)$ is system

response and m∈ M is a user model. PmtU,D,At,Rt-1 Is the posterior Probability of the user model m_t given all the observation about the user U we have made up to time t

Joachims et al [14] presented an empirical evaluation of interpreting click through evidence. By performing eye tracking studies and correlating predictions of their strategies with explicit ratings, the authors showed that it is possible to accurately interpret click through in a controlled laboratory setting. Unfortunately, the extent to which previous research applies to real world web search is unclear. But the recent work [15] on using click through information for improving web search ranking is promising. it captures only one aspect of the user interaction with web search engines.

Table taken from [5]. Some features used to represent post-search navigation history for a given query & search result URL

Query-text features	
TitleOverlap	Fraction of shared works between query and title
Summary overlap	Fraction of shared works between query and summary
Query URLOverlap	Fraction of shared works between query and URL
QueryDomainOverlap	Fraction of shared works between query and domain
QueryLength	Numbers of tokens in query
QueryNextOverlap	Average fraction of words shared with next query
Browsing features	
TimeOnpage	Page dwell time
Cumulative TimeOnpage	Cumulative Time for all subsequent pages after search
TimeOnDomain	Cumulative dwell Time for this domain
TimeOnShortUrl	Cumulative Time on URL prefix, dropping parameters
IsFollowLink	1 if followed link to result,0 otherwise
IsRedirected	1 if initial URL same as final URL, 0 otherwise
IspathFromSearch	1 if only followed links after query.0 otherwise
ClicksFromSearch	Number of hops to reach page from query
AverageDwellTime	Average time on page for this query
DwellTimeDeviation	Deviation from over all average dwell time on page
Cumulative Deviation	Deviation from average cumulative time on page
Domain Deviation	Deviation from average time on domain
Short URL Deviation	Deviation from average time on short URL
Clickthrough features	
Position	Position of the URL in current ranking
ClickFrequency	Numbers of clicks for this query,URL,pair
ClickRelativeFrequency	Relative frequency of a click for this query and URL
ClickDeviation	Deviation from expected click frequency
IsNextClicked	1 if there is a click on next position,0 otherwise
IspreviousClicked	1 if there is a click on previous position,0 otherwise
IsClickAbove	1 if there is a click above,0 otherwise
IsClickBelow	1 if there is a click Below,0 otherwise

Web search performance can be improved when user feedback is incorporated directly with popular content and link – based features are showed in Eugene Agichlein et al [5] experiments. Jinhyuk Choi et al [2] showed that the possibility of performing required task for user modeling successfully at the client side without any inherent limitation of server side analysis.

According to Ganeshan Velayathan et al [16] the information which is gathered implicitly from the user's behavior can also be used to help automatically evaluate web pages that the user has interest. In their experiment ,they developed a client-side analyzing tool do not focus on clicking ,scrolling , navigation or duration of visit alone , but they planned integrating user patterns of interaction to recognize and evaluate a user's response to a given web page using machine learning method.

According to palakorn Achananuparp[6], the aim of Implicit user modeling approach is creating a user model by requiring a minimum effort from users. This can be done by monitoring and analyzing the user's behaviors. The implicit user modeling process can be improved, if the system is able to understand the types of content on the visited page at a semantic level. The semantic web can contribute to implicit user modeling method is the use of ontology that provides semantic information describing the content of the web pages. Claypool et al [17] found that the time spent on a page, the amount of scrolling on a page, and the combination of time and scrolling have a strong positive relationship with explicit interest while individual scrolling methods and mouse clicks were not correlated with implicit interest.

Fox et al [18] found that click through was the most important individual variable but that predictive accuracy could be improved by using additional variables, notably dwell time on a page. In this approach, the user model is generated by two components of user behavior. 1. A relevance component (query) specific behavior and 2.Background component (users clicking indiscriminately).

4. PROPOSED USER MODELING

It is concluded that when user is having more interest on the document definitely he should store the document by using save operation or partial document by using cut or copy operation in his system for future reference. According to this conclusion, The combination of viewed document summaries and stored document summaries which are implicit feedback factors for effective user modeling in personalization process is provided.

Initially, the documents are ranked by using viewed summaries, when a user stored a entire document or the portion of document, the rate of the rank is increased. For the security issue, the proposed method can be used in client –

side system. In this method, the gathered information about the user or the summary details resides on client side system forever, thus the user does not need to release any personal information of the user to the outside. Performing personalized search on the client-side in more scalable that on the server side.

In general, client–side personalized search process has 3 steps.

(1) The user modeling module captures and analyzes the user's search history information and their behavior based on this analysis the user model is created or updated.

(2) The query which is submitted by user is customized based on current user model.

(3) Rerank the document whenever the user model is updated

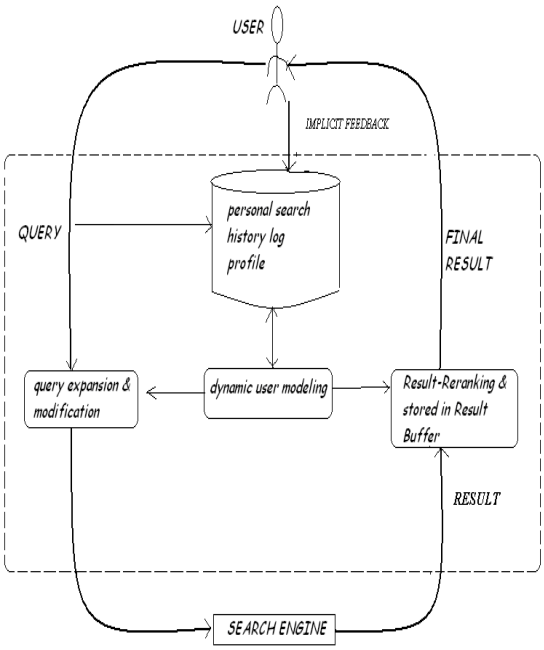

Figure 1: ARCHITECTURE OF PROPOSED MODEL

The user model is updated based on implicit feedback factors. As described in [3], Query expansion based in previous queries and immediate result re ranking based on implicit feedback information. Specific techniques to capture and exploit three types of implicit feedback information are proposed.

1. Decide whether the previous query is related to the current query and if so expand the current query with useful terms from the previous query or the results of the previous query,
2. Exploiting viewed documents summaries to immediately re rank and documents that have not yet been seen by the user. The viewed documents summaries are based on the browsing behavior when viewing a document, such as dwelling time, scrolling time.
3. 3. Exploiting stored documents summaries model is updated. The stored documents summaries are based on the behavior such as mouse click for save, cut and copy on documents

Browsing behavior, when viewing a page, like dwelling time, mouse click, mouse movement and scrolling time may be increased because of user interest on that page as well as some other external factors like user understanding capability or surroundings of user. Sometimes user takes long time to read the document for understanding the content of the document. So automatically the dwell time and scrolling time can be large. After understanding the content of the document, the user may decide that the document is not needed. But the document may get high score because of long dwell time and scrolling time.

5. A DECISION THEORETIC FRAMEWORK FOR THE PROPOSED MODEL

In order to utilize user's personal search history to improve search exactness in a general way, we view information retrieval (IR) as a decision optimization problem and propose a formal decision theoretic

Framework based on Bayesian decision theory for optimizing interactive retrieval [19].

Let A be the set of all user actions and R(α) be the set of all possible system responses to a user action α € A. At any time, let $A_t = (a1_{,\ldots\ldots,} a_t)$ be the observed sequence of user actions so far (up to time point t) and $R_{t-1} = (r_{1,\ldots\ldots} r_{t-1})$ be the responses that the system has made responding to the user actions. The system's goal is to choose an optimal response $r \in R(a_t)$

for the current user action a_t.

Let M be the space of all possible user models. A loss function is defined as L(α, r, m) € R, where a € A is a user action, r € R(α) is a system response, and m € M is a user model. L(α, r, m) encodes our decision preferences and assesses the optimality of responding with r when the current user model is m and the current user action is a. According to Bayesian decision theory, the optimal decision at time t is to choose a response that minimizes the Bayes risk, i.e.,

rt*=argminr∈ RatMLat,r,mtPmtU,D,At,Rt-1dmt (1)

where PmtU,D,At,Rt-1 is the posterior probability of the user model m_t given all the observations about the user U we have made up to time t.

To simplify the computation of Equation 1, let us assume that the posterior probability mass PmtU,D,At,Rt-1 is mostly concentrated on the model mt*=argmaxmt PmtU,D,At,Rt-1

We can then approximate the integral with the value of the loss function at m^*_t

That is,

rt*=argminr∈ RatL(at,r, mt*) (2)

Where mt*=argmaxmt PmtU,D,At,Rt-1

Leaving aside how to define and estimate these probabilistic models and the loss function, such a decision-theoretic formulation suggests that, in order to choose the optimal response to α_t, the system should perform two tasks:

(1) Compute the current user model and obtain m^*_t based on all the useful information.

(2) Choose a response r_t to minimize the loss function value L (a_t, r_t ,m^*_t). When α_t does not affect our belief about m^*_t, the first step can be omitted and one may reuse m^*_{t-1} for m^*_t.

Note that the framework is quite general since one can potentially model any kind of user actions and system responses. In most cases, as we may expect, the system's

response is some ranking of documents, i.e., for most actions a, R(a) consists of all the possible rankings of the unseen documents, and the decision problem boils down to choosing the best ranking of unseen documents based on the most current user model.

In our method let A_t =($a_{1,\ldots\ldots,}$ a_t) be the observed sequence of user actions may consist of three major actions.

1. Query expansion
2. Page dwell time.
3. Save/cut and copy operation on the documents.

R_{t-1} = ($r_{1,\ldots\ldots,}$ r_{t-1}) be the responses that the system has made responding to the above user actions, that is the system response is ranking of document for the above action

6. CONCLUSION

One of the important advantages of the proposed model is by which automatically detect the user interest from their behavior at the client side. Implicit feedback information improves the user modeling process. The advantage of implicit modeling is effectively improving the user model without extra effort of user. Several implicit feedback features are used to develop user modeling process. User modeling based on click through features is proposed. User model is constructed by using viewed and stored document summaries. Viewed document summaries like dwell time may be influenced by some other reasons like slow reading or understanding capability of user, working environment of user etc. So, viewed document summary is not alone used in the proposed model. Along with viewed document summary, stored document summary is also considered here. If user has more interest on document, definitely dwell time is long and user can store the entire or portion of the document for future reference. Initially, the documents are ranked based on viewed document summaries, and then the document should be re ranked according to stored document summaries. Further we plan to analyze the benefits of the integration the user modeling process with semantic content.

7. REFERENCES

[1] M. Baldoni, C. Barigkui and N. Henze, " Personalization for the semantic web", Reasoning web, First international summer school, LNCS, 2005.

[2] Jintiyuk cho and Geehyak Lec, "New Techniques for data preprocessing Based on usage Logs for efficient web uses profiling at client size", IEEE / WIC /ACM international conference on web intelligence and intelligent Agent Technology – Workshops, 2009.

[3] Xuehua Shen, Bin Tan, Chengziang Zhai, " Implicit user modeling for personalized search, ACM 1 -59593 - 140 -6/05/0010, 2005.

[4] D. Kelly and J. Teavan, "Implicit feedback for Inferring user preference", A bibliography SIGIR forum 37(2); 18-28, 2003

[5] Eugene Agichtein, Eric Bril, Susan Dumais, "Improving web search Ranking by incorporating user behavior information" ACM 1-59593-369-7\06\0008.

[6] Semantic web personalization by Palakron Achananuparp, College of Information Science and Technology, Direxel University.

[7] Xiaojian Ding, Yuancheng Li, Yinliang Zhao, "A framework of user model based on Semi-supervised techniques", IEEE International Conference on e-Business Engineering, 2008.

[8] Pradipta Biswas and Peter Robinson, " A brief survey on user modeling in HCI

[9] Perugini, S. Goncalves, M.A and Fox E.A, "Recommander System Research. A connection centric survey , Journal of Intelligent Information System, Vol 23,no.2,PP.107-143,2004.

[10] X.Shen,B.Tan, and C.Zhai, "Context-Sensitive information retrieval using implicit feedback", proceedings of SIGIR 2005, pages 43-50,2005.

[11] Xuehua Shen, Bin Tan , Cheng Xiang Zhai, " Exploiting personal search history to improve search accuracy-personal Information management", A SIGIR 2006 workshop.

[12] F-Gasparetli and A . Micarelli ,"User profile generation based on a memeory retrieval theory" in proceedings of the first international workshop on web personalization recommender systems and intelligent user Interfaces WPRSIUI 2005, reading UK Oct 2005.

[13] Ahu sieg , Bamshad mobasher, Robin burke, "Learning ontology-Based user profiles : A sermantic approach to personalized web search", IEEE intelligent information Bulletion, Vol 8 no.1\ Nov 2007.

[14] T.Joachims, L.Granka,B.Pang, H.Hembrooke and G.Gay , "Accurately Interprating click through Data as conference on research. Proceedings of the ACM", conference on research and development on information Retrieval (SIGIR),2005

[15] G.R .Xue, H.J. Zeng, Z.Chen ,Y. Yu, ,W.Y.Ma, W.S.Xi and W.G Fan, "Optimizing web search using web click through data", in proceeding of the conference on information and knowledge management (CIKM),2004.

[16] Ganesan Velaythan, Seiji Yamada "Behavior-Based web page evaluation",

[17] M.Claypool, D.Brown, P.Lee and M.Waseda, "Inferring user interest in IEEE Internet compiling", 2001.

[18] S. Fox, K. Karnawat, M.mydland, S.T. Dumais and T.white, "Evaluating implicit measures to improve the search experience", iIn ACM Transactions on information Systems, 2005.

[19]]X. Shen, B. Tan, and C. Zhai, "UCAIR toolbar: A personalized search toolbar (Demo)", in Proceedings of SIGIR 2005, page 681, 2005.

AUTHORS BIOGRAPHY

Rekha Chandramohan is a research scholar from Madurai Kamaraj University. She has completed her post graduation in computer science. Her area of interest is networks, data mining and especially in semantic. She can be contacted through

Usha Rani Janakiraman is a research scholar from Madurai Kamaraj University. She has completed her post graduation in computer application. Her area of interest is OOPS, data mining and especially in Web crawling. She can be contacted through

Dr. Iyakutti Kombiah is a CSIR Emeritus Scientist, School of Physics, Madurai Kamaraj University, Madurai, Tamil Nadu, India and Visiting Professor, Institute for Materials Research, Prof. Kawazoe Lab. Tohoku University, Sendai, Japan. His research interests are Computational Physics and Software Engineering.