Journal of Theoretical and Applied Information Technology
30th April 2014. Vol. 62 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PERSONALIZED WEB SEARCH METHODS – A COMPLETE REVIEW

[1]J.JAYANTHI, [2]DR.S.RATHI
[1]Associate Professor, Department of Computer Science and Engineering, SCT, Salem
[2] Assistant Professor(SG), Department of Computer Science and Engineering, GCT,Coimbatore
E-mail: [1]jayamithu2002@gmail.com, [2]rathi@gct.ac.in

ABSTRACT

Internet's foremost information retrieval service is the World Wide Web. It serves as a platform for retrieving variety of information that are associated with research, education , marketing, sports, games, politics, Finance, etc. The utter volume of information growth leads to the information overload on the Internet. Search engines are the collection of programs that facilitates information retrieval from the Internet. Even though the search engines do a good job of retrieving content from the Internet, users often feel disoriented about the result retrieved. Hence, no matter who the user of the search engine is, if the same query is provided as input to the search engine, the results returned will be exactly the same. The need to provide users with information tailored to their needs led to the development of various information personalization techniques. Personalization aims to provide users with what they need either by asking explicitly or implicitly. Web Personalization is conventionally defined as the process of tailoring web pages to satisfy the individual user needs by adapting different approaches. Several personalized web search models were developed based on web link structure, web contents, user queries, user profiles, browsing history etc. A Personalized Web Search has various levels of effectiveness for different users, queries, contexts etc. Personalized search has been a most important research area and many techniques have been developed and tested, still many issues and challenges are yet to be explored. This paper concentrates on the analysis, comparison and application of many personalized web search approaches that are being widely used today .Hence the motivation of this survey is directed towards to understand the web personalization processes, benefits, limitations and future trends.

Keywords: *Data Preprocessing, User Modeling, Page Ranking Strategies, Learning Methods,Personalization.*

1. INTRODUCTION

Personalization is an effort to uncover most relevant documents using information about user's target, domain of interest, browsing history, query context etc. that gives higher value to the user from the large set of results. As the web content is growing exponentially, more and more sophisticated methods are required to deliver the relevant content to the individual user. The most common difficulties encountered when searching the Web are [4]: i) Problems with the data itself ii) Problems faced by the users trying to retrieve the data they want iii) Problems in understanding the context of search requests and iv) Problems with identifying the changes in user's information need. The fundamental reason behind all the problems is scale of the web that limits its utility. Scaling leads to information over load, web users spent more time on filtering out the relevant results, Search engines may not able to provide different results for people with different intensions and context for the same query. Hence the significance of Personalization is [1][2][3], to customize the Web for individual users by filtering out the irrelevant results and identify relevant results.

The key steps of Web Personalization Process include i) Web Data Preprocessing ii) User Modeling in Personalization iii) Recommending Personalized Page Ranking Strategies. Every step of a Personalization process requires adaptability because of the change in the user's interest and instant information growth. The fundamental motivation is to learn and understand the steps that generate useful and actionable knowledge about users that in turn used for personalizing an application. The structure of the survey is as follows. The section II presents web data preprocessing methods that discover useful patterns

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

from the Web data. Section III presents user modeling techniques that models the user information needs with the help of user's transactional data such as search queries, dwell time, navigation patterns, domain of interest etc. Section presents page ranking algorithms used to tailor the Web pages to individual user's characteristics or preferences, Section V presents the open issues and Section VI concludes the essence of this survey.

2. WEB DATA PREPROCESSING

Data preprocessing is the process to convert the raw data into the data concepts necessary for the further applying it in building user profiles. It identifies unique users and their session data. A Session data are the different information source utilized in the personalized web search process. It could be in any one of the following forms.(i) **Web Page:** A document on the World Wide Web and each page is identified by a unique URL .The content of the page can be a simple text, images or structured data such as information retrieved from the databases.(ii) **Web Structure:** Hyper link structure of the web pages thereby becomes a directed graph. The nodes are the web pages and the directed edges connect different pages.(iii) **Web Usage Data:** It is a web site usage representation in terms of visitors IP address, date and time of Access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.(iv) **User profile data provide information about the users of a Web site**. The user profile contains demographic information (such as name, age, country, marital status, education, interests etc) for each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

2.1 Preprocessing Methods

Data preprocessing in personalization consists of data cleaning, user identification, session identification, feature identification of visited pages and path identification. Hence the input for the preprocessing step is a user session file that gives an exact account of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is the set of the page accesses that occur during a single visit to a Web site. **Matthijs and Radlinski[25]**captured web usage data such as Page URL ,visit duration, session date and time ,length of the source HTML using Firefox add on called Alter Go. Term extraction algorithm is used by **Leung et al. [31]** to summarize the web pages text into a set of important keywords. The Algorithm uses the C/NC method that uses combination of linguistic patterns and statistical information to score each term.C-value is defined as the relation of the cumulative frequency of occurrence of a word sequence in the text, with the frequency of occurrence of this sequence as part of larger proposed terms in the same text. The NC-value component corresponds to the final step of the ATR(Automatic Term Recognition) process, aiming at the refinement of C-value estimations based on candidate term phrase context. Term re extraction is done using Viterbi Algorithm. Open NLP(Natural Language Processing) tools are used to extract noun phrases. Term list filtering is done by removing infrequent words or words that are not in WordNet Dictionary.

User's input query is forwarded to general purpose search engines .**Oyama et al.[43]** used Domain Specific keywords called as Keyword spices are effectively discovered from the web document using the machine learning techniques. It enhances the web search results relevancy. Automatic text filtering is used to classify documents into relevant and non-relevant ones. Algorithm to extract keyword spices first classifies the collected web pages into two classes T (Relevant to the domain) and F(irrelevant to the domain).HTML tags from the initially collected web pages are removed and nouns are extracted as keywords. Two disjoint subsets are created $D_{training}$ and $D_{validation}$to form an initial keyword spices and simplify the same. Decision Tree algorithm discovers the keyword spices from the web documents and convert them into Boolean expressions. Hence Web Document classification is done using decision trees made up of Keywords. Classification process is started at the root of the tree and it will proceed with the relevant branch of the tree and end up on the target leaf node.

Peng et al. [26] built user profile by tracking clicked search results with reference to the Google directory. It is known as user topic tree where topics are linked in a tree structure. Each topic in the user topic tree is one of the topics in Google directory. It stores the value of the node visited count. It represents the degree of interest. **Sugiyama et al. [9]** gathered user profile data using the browsing history. Preferences of the user are treated as ephemeral and persistent nature. Ephemeral profile is constructed using the data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

gathered during current session. Persistent profiles are constructed exploiting the user's behavior of web searching N days ago. For each web page hp(r), number of distinct terms t_k is computed. Time spent on the web page is also deciding about the relevance of the web page.

Liu et al. [27] modeled user's search history using the following information items: Queries, relevant documents and related categories. A document retrieved by a search engine with respect to the query and category. User behaviors such as user page clicks, duration before the next click, user's save and print action are observed. User's search history is represented by Document-Term (DT) and Document-Category (DC) matrix. Category-Term (CT) matrix is used to represent the user profile. DT is constructed from the queries and their relevant documents. The value of DT(i,j) is identified by normalized TF*IDF weight scheme. Stop words are removed and if a term appears only once in the relevant document, then it will be removed from the search history. If the occurrence of a term is more than 5 words away from each query term then the occurrence of term is from the search history. For each row in the DT, there is a corresponding row in DC. Columns of DC are the set of related categories. Each row in the DC indicates set of categories related to the query/document. CT represents the user profile, where each row represents the category of interest to the user, is a vector of weighted terms.

Kim et al. [28] built the User Interest Hierarchy from a set of interesting web pages using a Divisive Hierarchical Clustering (DHC) algorithm. User's interest are organized from general to specific.DHC algorithm determines strong and weak correlation values between pair of words that appear in the same document. Hence it measures the co-occurrence of words in a document and then builds a weighted undirected tree where the vertex represents a word and weight denotes the correlation value. The algorithm recursively partitions the graph into sub graphs called clusters. Edges with weak weights are removed. **Kim et al. [29]** generated a session interest concept (SIC) based on the user's query.SIC is defined as a pair of intent and extent where extent covers set of keyword features extracted from the selected document. Hence the information need is modeled as a "concept network" which is a network structure of session interest concepts. Keywords are extracted from the selected document by computing the TF-IDF weights of each term. Terms with higher TF-IDF values will be selected from each document.TF-IDF weights for each term occurring in the entire document will be added and top scoring terms with accumulated weights are isolated. A new concept that is generated will be combined into the current concept network, by computing concept similarity measures.

Preprocessing in web search personalization takes variety of inputs such as search queries, relevant concepts from the web dictionaries such as WordNet, Wikipedia etc, visited URLs, search history etc. The outcome of the preprocessing activity is used to identify and build User profiles. User interested concepts and features are weighted and utilized in computing personalized page rank scores. **[25],[9]** methods are suitable for Implicit behavior based relevance method and **[31],[43],[26],[27][28][29]**are suitable for Implicit content based relevance methods. It is evident that if we combine both methods, it will yield better results.

3. USER MODELING IN PERSONALIZATION

User modeling is an essential part of a Personalized Web Search. It is the process of developing personal preferences of the users in terms of user's browsing history, knowledge about the world, likes and dislikes etc. So the current research challenge of personalization is directed towards user modeling and representation methods. Several approaches were proposed that accurately identifies the user context and organizes the information in such a way that it matches the particular context. Models are built as ontological profiles which contain the derived interest scores with reference to the concepts in the domain ontology. **Sieg et al. [5]** used a spreading activation algorithm to maintain the interest scores of the concepts based on the user's ongoing behavior. Initially each ontological user profile is the instance of the reference ontology for the given query and it is assigned with a value of one as user interest. User context is maintained and updated incrementally based on user's ongoing behavior. The main idea is to activate other concepts following a set of weighted relations during propagation and at the end obtain a set of concepts and their respective activations.

Kim et al. [7] defined a Probabilistic profile that is used to describe users, queries or web sites .It is called as a RLT (Reading Level and Topic) profile which describes about the distribution of reading level and topic. For example, the user profile might be associated with the URLs of previously clicked search results, or a website

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

profile might be associated with the URLs making up the website content. They have used automatic text classifiers to compute the RLT profiles (distributions over reading level and topic) for each URL in the set. Finally, they aggregate the distributions of the individual URL profiles to obtain the combined RLT profile of the entity. Profiles can also be constructed from other profiles: a user's profile could be computed not only based on the web pages visited by the user, but alternatively using the profiles of websites visited by the user, or the profiles of queries issued by the user.

Cordon et al. [38] proposed a Multi objective Genetic Algorithm to automatically learn persistent fuzzy linguistic queries for text retrieval applications. These queries are able to speak for user's long-term standing information needs in a more understandable profile structure. Genetic fuzzy system will be able to build different queries for the same information need in a single run, with a different trade-off between precision and recall. **Sugiyama et al. [9]** proposed a system which monitors the user's browsing history and updates his/her profile whenever his/her browsing page changes. When the user submits a query the next time, the search results adapt based on his/her user profile. They have constructed each user profile based on the following two methods: (i) Pure browsing history, and (ii) Modified collaborative filtering. User Profile construction based on pure browsing history considers both short term(ephemeral) and long term(persistent) preferences of the user. In Persistent preferences, the profile is constructed exploiting the user's browsing history of Web page from today and N days ago. User Profile Construction Based on Modified Collaborative Filtering Algorithms proposed by **Dasdan et al.[15]** is neighborhood-based method, where a subset of users is first chosen based on their similarity to the active user, and a weighted combination of their rating is then used to produce predictions for the active user. They have proposed the following two methods: (i) user profile construction based on the static number of users in the neighborhood, and (ii) user profile construction based on dynamic number of users in the neighborhood.

In the method proposed by **Li et al. [10]** user profiles are composed as the independent models for long term and short term user preferences. Long term interest is represented as a taxonomic hierarchy and short term interest is represented as visited page-history buffer. Dynamic adaptation strategies are devised to capture the accumulation and degradation changes of user preferences, and adjust the content and the structure of the user profile to these changes. Long term model is a part of the Google directory. It means that the topics associated with the clicked search results were only used to construct the model. Hence the interested topics are linked as a tree structure called as user topic tree. Each node in the user topic tree has a value of the number of times the node has been visited. This value is called the "Topic Count", and represents the degree of preferences. Page-History Buffer (PHB) is framed for the short terms model. Based on the ability of the search engine the most recently clicked pages with a fixed size are stored in the PHB cache. Cache management is done continuously by keeping track of the most recent accesses of search results. As a result, the Least Frequent Used Page Replacement (LFUPR) reflects the changes of the short term model.

Sun et al. [12] focused on utilizing click through data to improve Web search. The click through data is represented by a 3-order tensor, on which they perform 3-mode analysis using the higher order singular value decomposition technique to automatically capture the latent factors that govern the relations among these multi-type objects: users, queries and Web pages. A tensor reconstructed based on the CubeSVD (Singular Value Decomposition) analysis reflects both the observed interactions among these objects and the implicit associations among them. From the click through data, they can construct a 3-order tensor **A** $\in R^{U\times Q\times P}$, where U,Q,P are sets of users, queries and pages respectively. Each element of tensor **A** measures the preference of [u, q] pair on page p. In the simplest case, the co-occurrence frequency of u, q and p can be used. After tensor **A** is constructed, the CubeSVD algorithm can be applied on it. CubeSVD approach is to apply HOSVD on the 3-order tensor constructed from the click through data. The input is the click through data and the output is the reconstructed tensor **A**. **A** measures the associations among the users, queries and Web pages. The elements of **A** can be represented by a quadruplet [u,q,p,w], where 'w' measures the likeliness that user 'u' will visit page 'p' when 'u' submits query 'q'. Therefore, Web pages can be recommended to 'u' according to their weights associated with [u, q] pair.

Sontag et al. [8] proposed a generative model of relevance which can be used to infer the relevance of a document to a specific user for a search query. The user-specific parameters of this generative model constitute a compact user profile.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

They showed how to learn these profiles from the user's long-term search history. They presented two probabilistic models and inference algorithms for computing the probability that a document d is relevant to user 'u' for the query 'q'. A document about topic T_d is assumed relevant to a user looking for topic T_u if both:1. Topic T_d satisfies a user with information need T_u, and2. Given that the document's topic matches that of the search intent, the document is relevant to the query. The first criterion is measured by the variable $cov_u(d,q) \in \{0; 1\}$, which represents the extent to which T_d "covers" the information need T_u. The second criterion is measured by the variable $\Psi(d,q) \in [0, 1]$, which they called as the non-topical relevance score, corresponding to the user-independent probability that the document is relevant to this query. This score is assumed to be comprised for large number of user independent signals such as the match of the query to document text or anchor text, aggregate user behavior for this query, etc.

Daoud et al. [22] presented a method for modeling and inferring user's intention via Data Mining technique. They have defined two levels of intention (action intention and semantic intention) and differentiated user's intentions from user's preferences. Two linguistic features (keyword and concept features) are extracted for intention modeling. They have used association rule to mine proper concepts of corresponding keywords. Naïve Bayes classifier is a learned intention model. They also modified the algorithm to support incremental learning. Feature extraction algorithm based on Apriori algorithm and WordNet concept hierarchy is used. Association rules are generated based on which the extraction is done. A user intention model can be learned from the user logs. Each user action record contains a text part and a tag of action as well as other important information that may reflect the user's intention. The XML format is adopted when user's log data is recorded, "Action Type" is one of the five intentions, (browse, click, query, save and close).

Tang et al. [24] formalizes the profiling problem as several subtasks: profile extraction, profile integration, and user interest discovery. They proposed a combined approach to deal with the profiling tasks. They developed a classification model to identify the relevant documents for the user from Web. A Tree-structured Conditional Random Fields (TCRF) to extract the profile information from the identified documents was proposed. The 'name ambiguity problem' (several users with the same name), while integrating the profile information extracted from different sources, was also dealt with the probabilistic model. Finally they have used probabilistic topic model to model the extracted user profiles, and construct the user interest model. **Michlmayr et al. [21]** created user profiles from tagging data using Add-A-Tag algorithm. Aggregated data for a user's bookmark collection is created by counting the occurrence of tags. It accounts the structural and temporal nature of tagging data. A graph with labeled nodes and undirected weighted edges in which nodes correspond to tags and edges correspond to the relationship between tags. Each time a new tag is used, a new node for this tag is added to the graph. Each time a new combination of tags is used, a new edge with weight 1 between the corresponding nodes is created in the graph. If two tags co-occur again, the weight for the corresponding edge is increased by 1. In the second step, a user profile is derived from the resulting graph by selecting the top k edges with the highest weights and their incident nodes. Table I shows the most popular user models.

The search process is personalized by considering the User Models that will hold the searchers personal attributes and preferences. The model is built with the help of user's static and dynamic interests. It is done by tracking and aggregating the user interactions with the system. The set of attributes include User's search query, behavioral attributes such as Click-through data, dwell time, reading time, navigation path etc. We started living in an active environment where the environment is sensed automatically and based on which it responds. This technique is recently developed, often called as pervasive or ubiquitous computing. It assures the novel ways to deliver information to the people within their current environment. Thus the scope of the user modeling is directed towards the development of "a ubiquitous user modeling framework" for search engines which is capable of delivering relevant information about the user in the active environment and make relevant parts of a user model available when and where it is needed.

4. RECOMMENDING PERSONALIZED PAGE RANKING STRATEGIES

There are several ways to retrieve the documents relevant to the query. The research efforts on re-ranking web search results are categorized into the following classes of strategies.

(i) Explicit relevance judgments
(ii) Implicit relevance judgments
(a) Content-based implicit measures

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(b) Behavior-based implicit measures

4.1 Explicit Relevance Judgments

The trouble-free way to verify whether a result retrieved for a query is relevant to the user is to explicitly ask that user. Explicit judgment allows us to scrutinize the uniformity in relevance assessments across judges in a controlled setting. Advantage of this method allows us to examine the consistency in relevance assessments across judges in a controlled setting. Following are the limitations (i) It is cumbersome for people to give explicit judgment because it consumes additional time and effort from the users.(ii)It is difficult to gather sufficient data to generalize across a broad variety of people, tasks and queries.(iii)It is captured outside an end-to-end search session.

4.2 Implicit Relevance Judgments

Implicit data can be generated by users' interaction with their service. Implicit measures are easier to collect and allow us to explore many queries from vast variety of searchers. The two most common implicit measures used for personalization are (a) Content-based Implicit Relevance Judgments, (b) Behavior-based Implicit Relevance Judgments.

4.2.1 Content-based implicit relevance judgments

This type of measure uses a textual representation of users' interest to deduce the results which are relevant to their current need. Content-based profile captures all of the information created, copied or viewed by an individual. It also includes web pages viewed, e-mail messages sent or received, calendar items and documents stored on the client machine. Benefit of using this method is information about millions of users and millions of queries can be obtained and shows better performance than pure text-based algorithm and content-based algorithm. But, it is having following disadvantages (i) Activities of users are influenced by presentation of results (ii) Performance is lower than regular Web Ranking methods (iii) Currently updated information in the web repositories will not be reflected in the web search results dynamically.

Topical Interest based Ranking covers the spatial factors such as Queries used, Query usage count, Relevancy between the query and the document, query and the user profile, context of the query with reference to the ontologies or web dictionaries. The above factors normally support to develop the knowledge based user models. **Sieg et al. [5]** utilized the user context to personalize search results by re-ranking the results returned from a search engine for a given query. Re-ranking the search results based on the interest scores and the semantic evidence in the user profile is done. A term-vector r is computed for each document r Є R, where R is the set of search results for a given query. The term-weights are obtained using the *tf-idf* formula.

To calculate the rank score for each document, first the similarity of the document and the query is computed using a cosine similarity measure. Then, the similarity of the document with each concept in the user profile to identify the best matching concept is computed. Once the best matching concept is identified, a rank score is assigned to the document by multiplying the interest score for the concept, the similarity of the document to the query, and the similarity of the specific concept to the query. If the interest score for the best matching concept is greater than one, it is further boosted by a tuning parameter .Once all documents have been processed, the search results are sorted in descending order with respect to this new rank score.

Li et al. [10] proposed a method in which, when the user submits a query to the search engine, the search results are re-ranked by semantic, defined as the degree by which the search result is similar to the user profile. Weighing (WT) the degree of preferences of a node in the user topic tree is done. The larger the WT is, the more interested the user is in one topic. For one search result, the number of the CSim value is size (User Topics). One user topic tree represents one user. They define the semantic similarity between one search result and the user topic tree as the maximum value among all the values (i = 1, 2, · · · , size(User Topics)).

$$\text{FinalRank}(\text{User}, \text{j}) = (1-\gamma)\text{CSim} * (\text{User}, \text{j}) + \gamma * \text{PageRank(j)} \quad (1)$$

$CSim * (User, j)$ represent the semantic similarity between one search result and the user topic tree. γ is a smoothing factor ranges between [0-1].

Result diversity and Query reformulation techniques are followed by **Radlinski and Dumais [11]**to improve the personalized result. They have used 100 results and to personalize search using query reformulations by introducing diversity into those results. Given a query q, they generate a set of k related queries R(q) and take $(100/_{k+1})$ results from each query in R(q) and from q. They have developed Most Frequent (MF) method, Maximum Result Variety (MRV) method, Most

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Satisfied (MS) method for Query reformulation. Result diversity is achieved using the relevance-feedback approach for reranking developed by **Teevan et al. [5]**. They proposed a modification of the standard BM25 weighting scheme, in which relevance information is obtained from a local representation of a user's interests.

$$w_t = \log \frac{(r_t + 0.5)(N - n_t + 0.5)}{(n_t + 0.5)(R - r_t + 0.5)} \quad (2)$$

n_t,r_t-number of documents in N and R that contain term t; N-number of documents in the corpus; R-number of documents with relevance feedback;w_t-user's interest.

$$\text{match}_{\text{unigram}}(d_i, u) = \sum_{t \in d_t} w_t \quad (3)$$

$$\text{match}_{\text{bigram}}(d_i, u) = \sum_{t_i, t_j \in d_t} w_{t_i, t_j} \quad (4)$$

Unigram-individual words; bigram-pair of adjacent words; d_i-document, u-user; match(d_i, u)-match between document and user interest; w_t-user's interest.

In the method proposed by **Kim et al. [28]** personal page score is computed based on the number of interesting terms and how interesting these terms are in a web page. A web page that contains more interesting terms will be treated as the more relevant page. Four different characteristics are used to score a term .They are Length of the term, Frequency of the term, Emphasis of a term and Depth of a node. The Weighted scoring function and Uniform Scoring Function is adopted to incorporate the re-rank in the search engine results. For a given web page, p_j, the *personal* and *public* page score (PPS) is calculated using the formula,

$$PPS_{p_j} = c \times R\left(S_{p_j}\right) + (1 - c) \times R\left(GOOGLE_{p_j}\right) \quad (5)$$

Where the function $R(GOOGLE_{p_j})$ returns the rank of a web page, p_j with the public page score of $GOOGLE_{p_j}$; $R\left(S_{p_j}\right)$ is the rank of a web page p_j with the personal page score S_{p_j}.

The formula for uniform term scoring (US) function is,

$$S_{t_i} = -\log_2 P(D_{t_i}) - \log_2 P(L_{t_i}) - \log_2 P(F_{t_i}) - \log_2 P(E_{t_i}) \quad (6)$$

The formula for weighted term scoring (WS) function is,

$$S_{t_i} = -w_1 \log_2 P(D_{t_i}) - w_2 \log_2 P(L_{t_i}) - w_3 \log_2 P(F_{t_i}) - w_4 \log_2 P(E_{t_i}) \quad (7)$$

Where, $P(D_{t_i})$ - level/depth of a UIH node, $P(L_{t_i})$ – length of a term, $P(F_{t_i})$ - frequency of a term, $P(E_{t_i})$ - emphasis of a term, w_1-0.4, w_2=w_3=w_4=0.2.

The system proposed by **Liu et al. [27]** first maps each user query to a set of categories and returns the top three categories. Then they provide methods to improve retrieval effectiveness using categories as a context of the user query. There are three modes of personalization. In the baseline mode, the user query is submitted without specifying any category. In the semi automatic mode, system determines the top three categories which are matched with the user's interest by consulting the user. In the automatic mode, the system automatically picks the top category or the top two or three categories without consulting the user. DOC-WO-C (Documents Retrieved without Specifying Categories) is retrieved in the baseline mode. DOC-W-C_i (Documents Retrieved by Specifying the top ' i ' category) are retrieved in the semi automatic or automatic mode. Then the retrieved lists of documents DOC-WO-C and DOC-W-C_i are merged in such a way that resulting set has exactly the same cardinality as DOC-WO-C which is done by modified voting based merging scheme [26]. The weight of a list is given by,

Where,

rank-C = 1, if rank of category C with respect to the query is$W_j = rank - C * square - root(sim - C) * num - C$ (8) 1; 0.5 if rank is 2; 0.25 if rank is 3.

num-C is the number of retrieved documents in the list; rank-C and sim-C is 1 in semi-automatic mode else rank-C is 0.5 and sim-C is 0.1.

Sontang et al. [8] propose an end-to-end system for personalization and focused on three main problems: representation, learning and ranking. They developed a probabilistic model for predicting the relevance of a document to a particular user with respect to the query. In this general formalization they assume that there are only user specific latent variables and document specific latent variables. To find whether a document's content satisfies the user's information need these two variables are combined. A trained global ranking function was used to find how a specific user differs from the population as a whole and then the relevance probability is de-convolved into the probability that the page is relevant to particular query intent. Then by considering the user profile they recomputed the probability of relevance.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\Pr\left(rel_u(d,q)=1\middle|\theta_u,q,d,\varphi(q,d)\right) = \varphi(d,q)\sum_{T_d}\Pr(T_d|d)\,\alpha(T_d) \quad (9)$$

$\propto(T_d)=\sum_{T_u}\Pr(T_u|\theta_u,d)\Pr(cov_u(d,q)=1|T_u,T_d)$;$\theta_u - userprofile$; q – query; $\Pr(T_u|\theta_u,d)$ – distribution over topics; d- document; $T_d - documentvariable$; $T_u - uservariable$;$(d,q) - nontopicalrelevancescore$.

Daoud et al. [41, 22] personalized the search using a graph-based user profile issued from a predefined ontology. By combining graph based query profiles, user profile is constructed over a search session. Personalization is achieved by re-ranking the search results of the queries related to same session. The user profile G_u^0 is initialized by the profile of the first query submitted by the user. It is then updated by combining it with the query profile $G_q^{s+1}$of a new related query submitted at time s+1. When the correlation value is greater than the predefined threshold value, the search results are re-ranked by combining their original score with their contextual score using the following formula,

$$S_f(d_k) = \gamma * S_i(q,d_k) + (1-\gamma) * S_c(d_k,G_u^s) \quad (10)$$

The contextual score is computed using the formula:

$$S_c(d_k,G_u^s) = \frac{1}{h}.\sum_{j=1\ldots h}\text{score}(c_j) * \cos(\overrightarrow{d_k},\overrightarrow{c_j}) \quad (11)$$

Where,

$S_c(d_k,G_u^s)$- Contextual score; G_u^s- user profile at time s; $S_i(q,d_k)$- original score; c_j- concept; q – query; d_k- document.

4.2.2 Behavior-based implicit relevance judgments

This type of measure uses people's behavior such as their past interactions with search result lists, click-through data from the logs etc. The Performance is better than pure text-based algorithms. Some of the disadvantages are(i) Intent for a query may vary widely among each individual. (ii)Performance is lower than behavior-based and other web ranking methods. The goal of **Collins et al. [6]** was to show how modeling reading proficiency of users and the reading difficulty of documents can be used to improve the relevance of Web search results. Web users differ widely in their reading proficiency and ability to understand vocabulary, depending on factors such as age, educational background, and topic interest or expertise. Hence it is clear that there is a need for improvement in ranking search results at an appropriate level of reading difficulty. To address this problem, they described a tripartite approach based on user profiles, document difficulty, and reranking.

First, the snippets and Web pages can be labeled with reading level and combined with Open Directory Project (ODP, www.dmoz.org) category predictions. Second, they described how a user's reading proficiency profile may be estimated automatically from their current and past search behavior. Third, they use this profile to train a re-ranking algorithm that combines both relevance and difficulty in a principled way and which generalizes easily to broader tasks such as expertise-based re-ranking. In this view, the overall relevance of a document is a combination of two factors: a general relevance factor, provided by an existing ranking algorithm, and a user-specific reading difficulty model, based on the gap between a user's proficiency level and a document's difficulty level. While users may self-identify their desired level of result difficulty, such information may not always be provided. They investigate methods for estimating a reading proficiency profile for users based on their online search interaction patterns. The reading level of user can be defined by,

$$p(\text{u likes level of d}\,|r_u,r_d) = \exp(-(r_d-r_u)^2) \quad (12)$$

Where, u - user; d - document; r_u - reading level of user u; r_d - reading level of document d.

Xu et al. [32] proposed a personalized re-ranking algorithm through mining user dwell times derived from a user's previously online reading or browsing activities. They acquire document level user dwell times via a customized web browser, from which they infer concept word level user dwell times in order to understand a user's personal interest. According to the estimated concept word level user dwell times, proposed algorithm can estimate a user's potential dwell time over a new document, based on which personalized webpage re-ranking can be carried out. They represent each document D_i previously read by a user u_k, as a collection of concept words. Modeling Document Level User Dwell Time and Concept Word Level User Dwell Time both are done for reduced case and full case. They have considered an extremely simplified case where the whole document consists of only one concept word, which may occur multiple times known as reduced case. A document consists of multiple concept words of which some may occur multiple times is treated as a full case. User's potential dwell time for each document is predicted based on the estimated concept word level user dwell time. The document D_x as a collection of concept words via the document

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

representation procedure is first represented. Then the predicted user dwell time $\varphi(u_k,D_x)$ on the new document D_x is derived. Thus the search result set is reranked based on the respective predicted dwell times for user $u_k$in a search session.

$$\vartheta(C_i, u_k, 1) \triangleq \sum_{D_j \in D_k} \Delta_{u_k}(D_j, C_i) \quad (13)$$

Where, $\Delta_{u_k} - worddensity; C_i - concept; u_k - user; D_j - document; D_k - documentset$

Dou et al. [39, 17] proposed a large scale evaluation framework based on query logs and five different search algorithms (two click-based ones and three topical based ones) for evaluating their performance. They have followed two strategies person-level and group level. In the person level re-ranking, whenever a user submits query, the web pages clicked by the user in the past are more relevant than those seldom clicked by the user. This approach is named as P-Click. The formula for P-Click is,

$$S^{P-Click}(q, p, u) = \frac{|Clicks(q, p, u)|}{|Clicks(q, \blacksquare, u)| + \beta} \quad (14)$$

|Clicks (q,p,u)| - number of clicks on web page *p;* |Clicks(q,■,u)| - total number of clicks for query *q* by *u*; β is a smoothing factor.

They also implemented a long-term user topical interest known as L-Topic. When a user(u) submits a query(q),each returned web page(p) is first mapped to a category vector and the page category vector is computed. The similarity between user interests and a web page is used to re-rank search results. The formula for L-Topic is,

$$S^{L-Topic}(q, p, u) = \begin{cases} sim(c_l(u), c(p)) & \text{if } sim(c_l(u), c(p)) \geq t \\ 0 \text{ if } sim(c_l(u), c(p)) < t & \end{cases}, \quad t \in [0,1] \quad (15)$$

Finally they have implemented K-Nearest Neighbor collaborative algorithm as a group-based personalization .The historical clicks made by all similar users in a group to rerank the search results are used. It is denoted as G-Click.

$$S^{G-Click}(q, p, u) = \frac{\sum_{u_s \in S_u(u)} sim(u_s, u)|Clicks(q, p, u_s)}{\beta + \sum_{u_s \in S_u(u)} |Clicks(q, \blacksquare, u_s)|} \quad (16)$$

User similarity is computed based on long-term user profiles:

$$sim(u1, u2) = \frac{c_l(u1).c_l(u2)}{||c_l(u1)||||c_l(u2)||} \quad (17)$$

The K-Nearest neighbors are obtained based on the user similarity:

$$S_u(u_{a)}) = \{u_s | rank(sim(u_a, u_s)) \leq K\} \quad (18)$$

Kajaba et al. [40] proposed a combination of usual search engine and a dedicated search engine. They augment the search query using additional keywords from IP's profile, which is maintained in the form of keyword cloud. The augmented query is submitted to the usual search engine and it will return the results decorated with keywords describing a particular document. Search process will take place in three phases. In first phase, regular search is performed by Yahoo and it returns list of results decorated with extra keywords. For each hit, the system will determine whether the keyword is included in IP's profile. In second phase, for each $word_k$, ρ_k index is calculated using the formula,

$$\frac{p_k}{P}.\lambda + \frac{i_k}{avg(i)}.(1-\lambda) = \rho_k \quad (19)$$

Where, $\frac{p_k}{P}.\lambda$ is the natural context of the search engine and $\frac{i_k}{avg(i)}.(1-\lambda)$is the context of IP's profile. In the last phase, the search engine will be re-run using rewritten query and the results will be displayed to the IP.

5.OPEN RESEARCH ISSUES

This survey would help us to identify the open issues to be solved in the Personalized Web Search Domain. There are scopes in refining(i) Profile building Approaches (ii) Storage Pattern (iii) Page Ranking measure (iv) Sentiment based factor weighting.

6. CONCLUSION

Personalized web search is an active ongoing research field that focuses on the retrieval of the relevant web page results based on the user preferences. This paper covers the personalization process in various stages. In each stage various techniques and algorithms have been discussed. From the preprocessing stage, it is evident that the Implicit Behavior Based Personalization extracts the user session attributes such as url's visited, navigation patterns, dwell time, reading time etc as a result of preprocessing activity. Implicit content based Personalization extracts the related concepts for the query given by the user from the retrieved set of documents. If we combine above two methods of preprocessing, it will be useful to build complete user profile. Different methods of building user models are also discussed. There are several types of classification of user interest are discussed like short term interest, long term interest, static interest, domain interest etc. The adaptiveness quality of the profile should be concentrated in the future to describe the convergence level of the profile. The suitability of implicit and explicit relevance method is discussed.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

It is evident that deriving the set of rules and designing an intelligent framework that will predict the appropriate method to be used so that it will yield relevant results. Thus the motivation behind the personalization is to enhance quality of rankings. The future of personalization research is directed in the following aspects like building complex profiles with minimal user interaction, better accuracy of interest prediction ,building profiles of ubiquitous nature which senses the environmental contexts automatically and profiles with social inputs such as common interest group, social networks etc. This survey will direct the researchers to develop a Personalized Web Search framework that will produce customized results to the user.

REFERENCES

[1] Shahabi C. and Yi-Shin Chen Y.S. 2003, Web information personalization: challenges and approaches, Third International Workshop on Databases in Networked Information Systems, pp. 5-15.

[2] Allen C., Kania D., Yaeckel B. 2001, One-to-One Web Marketing: Build a Relationship Marketing Strategy One Customer at a Time, 2nd edition. John Wiley and Sons, New York.

[3] EetuMakela 2005, 'Survey of Semantic Search Research', Proceedings of the Seminar on Knowledge Management on the Semantic Web, Department of Computer Science, University of Helsinki.

[4] Lawrence Steve and Giles Lee 2000, Accessibility of information on the Web, Intelligence, vol. 11, pp.32-39.

[5] Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. InCIKM'07: *Proceedings of the ACM Conference on information and knowledge management*, pages525–534, New York, NY, USA, 2007. ACM.

[6] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, "Personalizing Web search results by reading level," in Proceedings of the ACM International Conference on Information and Knowledge Management(CIKM '11), New York, NY, USA, pp. 403–412, 2011. ISBN 978-1-4503-0717-8.doi: 10.1145/2063576.2063639. URL http://doi.acm.org/10.1145/2063576.2063639.

[7] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, "Characterizing Web content, user interests, and search behavior by reading level and topic," in Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '12), New York, NY, USA, pp. 213–222, 2012. ISBN 978-1-4503-0747-5.doi: 10.1145/2124295.2124323. URL http://doi.acm.org/10.1145/2124295.2124323.

[8] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais,and B. Billerbeck, "Probabilistic models for personalizing Web search," in Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM '12), New York, NY, USA, pp. 433–442, 2012. ISBN 978-1-4503-0747-5.doi: 10.1145/2124295.2124348. URL http://doi.acm.org/10.1145/2124295.2124348.

[9] Sugiyama, K., Hatano, K. and Yoshikawa, M. (2004). Adaptive Web search based on user profile constructed without any effort from user. In *Proceedings of WWW '04,* 675-684.

[10] L. Li, Z. Yang, B. Wang, and M. Kitsuregawa. Dynamicadaptation strategies for long-term and short-term user profile to personalize search. In AP Web/WAIM, pages 228–240, 2007.

[11] FilipRadlinski and Susan T. Dumais. Improving personalized web search using result diversification. In SIGIR, pages 691-692, 2006.

[12] J.-T. Sun, H.-J.Zeng, H. Liu, Y. Lu, and Z. Chen, CubeSVD: A novel approach to personalized Web search, in WWW 2005: Proceedings of the 14th International Conference on World Wide Web, ACM Press, 2005, pp. 382-390.

[13] Kent, M.M. Berry F.U. Luehrs Jr., and J.W. Perry (1955), Machine literature searching VIII. Operational criteria for designing information retrieval systems. American Documentation 6(2):93-101

[14] B. Carterette and R. Jones (2007), Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks, NIPS'07.

[15] Dasdan, A., Tsioutsiouliklis, K., Velipasaoglu, E. "Web search engine metrics for measuring user satisfaction", 2009. On-line, retrieved from http://dasdan.net/ali/www2009/web-search-metrics-tutorial-www09-part6a.pdf.

[16] Y. Zhu, L. Xiong, C. Verdery. Anonymizing user profiles for personalized web search.In WWW, 2010.

[17] N. Matthijs and F. Radlinski.(2011). Personalizing Web search using tong term browsing history.In Proceedings of the ACM

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

WSDM Conference on Web Search and Data Mining, pp. 25–34.

[18] J. Castellà-Roca, A. Viejo, J. Herrera-Joancomartì, Preserving users' privacy in web search engines, Computer Communications 32 (2009) 1541–1551.

[19] Hu J, Chan P (2008) Personalized web search by using learned user profiles in re-ranking. In: Workshopon knowledge discovery on the web, KDD conference. pp 84–97.

[20] Sieg, B. Mobasher, and R. Burke. Ontological user profiles for representing context in web search. Web Intelligence and Intelligent Agent Technology Workshops,2007 IEEE/WIC/ACM International Conferences on, pages 91–94, Nov. 2007.

[21]Michlmayr, E., Cayzer, S., &Shabajee, P. (2007). Learning User Profiles from Tagging Data and Leveraging them for Personal(ized) Information Access. Tagging and Metadata for Social Information Organization; a workshop at WWW2007, Banff, Alberta, Canada Retrieved December 5, 2007 from http://www.ibiblio.org/www_tagging/2007/paper_29.pdf.

[22] Mariam Daoud, Lynda Tamine-Lechani, and MohandBoughanem.Using a graphbased ontological user profile for personalizing search. In James G. Shanahan, SihemAmer-Yahia, IoanaManolescu, Yi Zhang, David A. Evans, AleksanderKolcz, KeySun Choi, and AbdurChowdhury, editors, CIKM'08: Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, pages 1495–1496. ACM, 2008.ISBN 978-1-59593-991-3.

[23] Sugiyama, K., Hatano, K. and Yoshikawa, M. (2004). Adaptive Web search based on user profile constructed without any effort from user. In Proceedings of WWW '04, 675-684.

[24] J. Tang, L. Yao, D. Zhang, and J. Zhang.A combination approach to web user profiling. ACMTKDD, 5(1):1–44, 2010.

[25] N. Matthijs and F. Radlinski.(2011). Personalizing Web search using tong term browsing history.In Proceedings of the ACM WSDM Conference on Web Search and Data Mining, pp. 25–34.

[26] Peng, Xueping, ZhendongNiu, Sheng Huang, and YuminZhao."Personalized Web Search Using Clickthrough Data and Web Page Rating."*Journal of Computers* 7, no. 10 (2012): 2578-2584

[27] Liu, Fang, Clement Yu, and WeiyiMeng. "Personalized web search for improving retrieval effectiveness."*Knowledge and Data Engineering, IEEE Transactions on* 16.1 (2004): 28-40.

[28] Kim, H. R., and Philip K. Chan. "Personalized ranking of search results with learned user interest hierarchies from bookmarks." In *WEBKDD*, vol. 5, pp. 32-43. 2005.

[29] Kim, Han-joon, Sungjick Lee, Byungjeong Lee, and Sooyong Kang. "Building concept network-based user profile for personalized web search." In *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, pp. 567-572. IEEE, 2010.

[30] Giannopoulos, Giorgos, Theodore Dalamagas, and TimosSellis. "Collaborative Ranking Function Training for Web Search Personalization." In Proceedings of the 3rd International Workshop PersDB. 2009.

[31] Leung, KW-T., DikLun Lee, and Wang-Chien Lee. "Personalized web search with location preferences."In Data Engineering (ICDE), 2010 IEEE 26th International Conference on, pp. 701-712.IEEE, 2010.

[32] Xu, Songhua, Hao Jiang, and Francis Lau. "Mining user dwell time for personalized web search re-ranking." In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, pp. 2367-2372. AAAI Press,2011.

[33] Noll, Michael, and ChristophMeinel. "Web search personalization via social bookmarking and tagging." The Semantic Web (2007): 367-380.

[34] Nakagawa, Miki, and BamshadMobasher. "A hybrid web personalization model based on site connectivity." In Proceedings of WebKDD, pp. 59-70. 2003.

[35] Lee, Hyun Chul, and Allan Borodin. "Criteria for Cluster-Based Personalized Search." Internet Mathematics 6, no. 3 (2009): 399-435.

[36] Liu, Fang, Clement Yu, and WeiyiMeng. "Personalized web search by mapping user queries to categories."In Proceedings of the eleventh international conferenceon Information and knowledge management, pp. 558-565.ACM, 2002.

[37] Mobasher, Bamshad, Honghua Dai, Tao Luo, and Miki Nakagawa. "Discovery and evaluation of aggregate usage profiles for web personalization." Datamining and knowledge discovery 6, no. 1 (2002): 61-82.

[38] Cordón, Oscar, Enrique Herrera-Viedma, and Marıa Luque. "Fuzzy Linguistic Query-based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

User Profile Learning by Multiobjective Genetic Algorithms." InEvolving Fuzzy Systems, 2006 International Symposium on, pp. 261-266. IEEE, 2006.

[39] Dou, Zhicheng, Ruihua Song, Ji-Rong Wen, and Xiaojie Yuan. "Evaluating the effectiveness of personalized web search." Knowledge and Data Engineering, IEEE Transactions on 21, no. 8 (2009): 1178-1190.

[40] Kajaba, Michal, PavolNavrat, and Daniela Chuda. "A simple personalization layer improving relevancy of web search." Computing and Information Systems Journal 13, no. 3 (2009): 29-35.

[41] Daoud, Mariam, Lynda Tamine-Lechani, and MohandBoughanem. "A contextual evaluation protocol for a session-based personalized search."In Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (In conjunction with European Conference on Information retrieval-ECIR), Toulouse, France, Springer. 2009.

[42] J. Teevan, S. T. Dumais, and E. Horvitz.Personalizing search via automated analysis of interests and activities. In SIGIR 2005, pages 449–456, 2005.

[43]Oyama, Satoshi, Takashi Kokubo, and Toru Ishida. "Domain-specific web search with keyword spices." Knowledge. And Data Engineering, IEEE Transactions on 16, no. 1 (2004): 17-27

Table 1
User Models and Learning Methods

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

User Model	Learnirg Method	Taxonomies Used	Input	Output
User Context Model ontology[5]	Spreading Activation Algorithm	Open Directory Project	User interested Web Documents, User interest concepts with updated activation values.	Ontological user profile with interest scores.
RLT profile[7]	Probability distribution of reading level and Topic level	Open Directory Project	Web search log data, Click entropy details	Expert Web site profile and Expertise profiles
multi-granular linguistic query profiles	A Multi objective Genetic Algorithm To Learn Linguistic Persistent Queries	-	Communicationsof the ACM document set, queries, relevant documents.	multi-granular linguistic query profiles
User profile as User Term weight matrix[9]	User Profile Construction Algorithm Based on Pure Browsing History and Modified Collaborative filtering algorithm	-	User interested Web pages	User Term weight matrix as user profile
User Topic Tree[10]	Least Frequent Used Page Replacement (LFUPR)	Google Directory	User clicked web pages. Google directory	User topic Tree as a Profile
SVD matrices[12]	Cube SVD Algorithm	ODP	Click through Data	latent associations among the multi-type data objects: user, query and Web page that improves the web search.
Probabilistic Model[8]	User specific and independent inference algorithms	ODP	BING logs	Weighted Topics and preferred results.
User Intension Model[22]	Naïve Bayes classifier	Word Net	Web logs, User actions	User Intension Model
Graphical Model[24]	Tree-structured Conditional Random Fields method, Hidden Markov Random Fields Method(HMRFs)	ARNET MINER, DBLP bibliography	Web Logs of researchers	User Profiles for expert finding
Concurrence Network[21]	Add-Tag-Algorithm	-	User book marks	Adaptive user Profiles