© 2015 IJSRSET | Volume 1 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

# Literature Survey on Web Personalization

**Sachin Pardeshi**
Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur, India

## ABSTRACT

Retrieve the most relevant information for the Web becomes difficult since the massive amount of documents existing in various formats. It is compulsory for the users to go through the long list of oddments and to choose their relevant one, which is a time overwhelming process. User satisfaction is less important in this aspect. One approach to satisfy the requirements of the user is to personalize the information available on the Web, called Web Personalization. Web Personalization is the process that adapts information or services provided by a Web to the needs of each specific or set of users, taking the facts of the knowledge gained from the users. Web Personalization can be the answer to the information overload problem, as its purpose is to provide users with what they really want or need, without having to ask or search for it unambiguously. It is a multi discipline area for putting together data and producing personalized output for individual users or groups of users. This approach helps the researchers to improve the effectiveness of Information Retrieval (IR) systems. By considering all the benefits of the Web Personalization, this paper presents elaborately the various approaches used by researchers to achieve Web Personalization in Web Mining.

**Keywords:** Information Retrieval, Semantic Web, Ontology, Web Personalization, User Profile, Personalized Search, Personalized Ontology

## I. INTRODUCTION

On the Internet, we have experienced enormous growth in systems that can personalize content transported to individual users. The science behind personalization has undergone marvelous changes in current years, yet the basic objective of personalization systems remains the same: to provide users with what they want or need without requiring them to ask for it explicitly. Personalization is the stipulation to the individual of customized products, services, information or information relating to products or service. It is a broad area, also covering recommender systems, customization, and adaptive Web sites.

Three aspects of a Web site concern its utility in providing the anticipated service to its users. These are the content provided on the Web site, the layout of the individual pages, and the structure of the entire Web site itself. The relevance of each of the objects comprising a Web page to the users' needs will clearly affect their level of satisfaction. The structure of the Web site, defined by the existence of links between the various pages, restricts the navigation performed by the user to predefined paths and therefore defines the ability of a user to access relevant pages with relative ease. However, the definition of relevance is subjective.
It is here that there is a potential mismatch between the perception of what the user needs, on the part of the Web site designer, and the true needs of users. This may have a major impact on the effectiveness of a Web site.

Personalization technology involves software that learns patterns, habits, and preferences. On the Internet, its use is primarily in systems that support ebusiness. Personalization works in this context because it helps users to find solutions, but perhaps more importantly; it also empowers e-business providers with the ability to measure the quality of that solution. In terms of the fast emerging area of Customer Relationship Management (CRM), personalization enables e-business providers to implement strategies to lock in existing customers, and to win new customers.

Daniel E. O'Leary from the University of Southern California coined the phrase "AI Renaissance" in 1997,1 to describe how artificial intelligence (AI) can make the Internet more usable. Personalization technology is part of that renaissance. In parallel with the academic progress covered in this special section, the commercial world is witness to unprecedented growth in personalization technology companies. It is sometimes difficult to find a commonality in technology foundation that spans the breadth of commercial product offerings and global academic efforts in personalization, as well as the broad cross section of emerging efforts in digital markets.

Initial attempts at achieving personalization on the Internet have been limited to check-box personalization, where portals allow the user to select the links they would like on their "personal" page. However, this has limited use since it depends on the users knowing beforehand the content of interest to them. Arguably, collaborative filtering was the first attempt at using AI for achieving personalization in a more intelligent manner.

This allows users to take advantage of other users' behavioral activities based on a measure of similarity between them. These techniques require users to divulge some personal information on their interests, likes and dislikes, information that many Web users would not necessarily wish to divulge. An alternative is observational personalization, which attempts to circumvent the need for users to divulge any personal information. The underlying assumption in this approach is that hidden within records of a user's previous navigation behavior are clues to how services, products, and information need to be personalized for enhanced Web interaction.

## WHAT IS PERSONALIZATION?

Personalization is the process of deciding - given a large set of possible choices - what has the highest value to an individual. This adds both utility and warmth to a web application, as users find what they seek faster and feel "recognized" by a site. On a more practical level Personalization is an overloaded term: There are many mechanisms and approaches (both automated and marketing rules controlled) whereby content can be focused to an audience in a one to one manner. This section delineates between the various approaches providing the student with a terminology to describe each approach in isolation. Furthermore it describes how the approaches can be combined such as using Like Minds "engines" to prioritize results from a rules based recommendation or filtering Like Minds recommendations using business rules.

## WHAT CAN BE PERSONALIZED?

Personalized content may be advertising, recommended items, screen layout, menus, news articles, or anything else accessed via a web page or software application.

- **Business benefits**

Personalization contributes to a variety of e-business goals: increasing site usability, replicating offline experience, and converting browsers to buyers, retaining current customers, re-engaging customers, and penetrating new markets.

- **Increase site usability**

By limiting navigation options, and providing direct links to desired content, personalization automatically makes a site more navigable, allowing users to find desired information, products, and services more quickly.

- **Replicate offline experience**

Replicating familiar offline experiences is a key goal/ benefit of personalization. Ideally, personalization acts as a stand-in for the Friendly Store Clerk, the person behind the counter at the corner hardware store who remembers you, suggests purchases, and helps you solve your particular problems.

- **Conversion (increased sales)**

Research shows that converting browsers to buyers has a significant impact on site revenues. Toward this end, personalization brings targeted, high-value purchase opportunities directly to the user. By positioning desired content in front of a user, personalization increases the odds that a browser will become a buyer.

- **Retention**

As the Internet matures, and success becomes measured in more than strict traffic numbers, retaining customers is crucial for any site's success. Personalization enhances site "stickiness," that is, an increased likelihood that customers will bookmark and return to your site. Customers return more frequently to sites

where they receive specific benefits, and personalization provides these benefits.

- **Re-engagement**

Often times, a customer will shop or consume information from a number of sites on the Internet. Reengagement is the process of reaching back out to a customer via email or other means to let them know you have something that they may be interested in. If such notifications are personalized, the customer will learn to trust re-engagement attempts and they will more likely be successful.

## WEB PERSONALIZATION APPROACHES

### a. Web Personalization and User Profile

Different users usually have different special information needs when they use search engines to find web information. The technologies of personalized web search can be used to solve the problem. An effective way to personalized search engines' results is to construct user profile to present an individual user's preference. Utilizing the relative machine learning techniques, three approaches are proposed to build the user profile in this paper. These approaches are called as Rocchio method, k-Nearest Neighbors method and Support Vector Machines method. Experimental results based on a constructed dataset show that k-Nearest Neighbors method is better than others for its efficiency and robustness.

### b. Techniques using User Profiles

Intelligent user profiling implies the application of intelligent techniques, coming from the areas of Machine Learning, Data Mining or Information Retrieval, for example, to build user profiles. The data these techniques use to automatically build user profiles are obtained mainly from the observation of a user's actions, as described in the previous section.

- **Bayesian Networks**

A Bayesian network (BN) is a compact, expressive representation of uncertain relationships among variables of interest in a domain. A BN is a directed acyclic graph where nodes represent random variables and arcs represent probabilistic correlations between variables (Jensen, 2001). The absence of edges in a BN denotes statements of independence. A BN also represents a particular probability distribution, the joint distribution over all the variables represented by nodes in the graph. This distribution is specified by a set of conditional probability tables (CPT). Each node has an associated CPT that specifies the probability of each possible state of the node given each possible combination of states of its parents. For nodes without parents, probabilities are not conditioned on other nodes; these are called the prior or marginal probabilities of these variables.

- **Association Rules**

Association rules are a data mining technique widely used to discover patterns from data. They have also been used to learn user profiles in different areas, mainly in those related to e-commerce (Adomavicius and Tuzhilin, 2001) and web usage (Gery and Hadad, 2003). An association rule is a rule which implies certain association relationships among a set of objects in a given domain, such as they occur together or one implies the other.

CBR is a technique that solves new problems by remembering previous similar experiences (Kolodner, 1993). A case-based reasoner represents problem-solving situations as cases. Given a new situation, it retrieves relevant cases (the ones matching the current problem) and it adapts their solutions to solve the problem. In an interpretative approach, CBR is applied to accomplish a classification task, that is, find the correct class for an unclassified case. The class of the most similar past case becomes the solution to the classification problem. CBR has been used to build user profiles in areas like information retrieval and information filtering (Lenz et al, 1998; Smyth and Cotter, 1999). For example, in (Godoy et al, 2004) CBR is used to obtain a user interest profile.

Other User Profiling Techniques Many other Machine Learning techniques have been used for user profiling, such as genetic algorithms, neural networks, kNN-algorithm, clustering, and classification techniques such as decision trees or naïve Bayes classifier. For example, Personal WebWatcher (Mladenic, 1996) and Syskill&Webert (Pazzani et al, 1996) use naive Bayes classifiers for detecting users' interests when browsing the web. Amalthaea (Moukas, 1996) uses genetic algorithms to evolve a population of vectors representing a user's interests. The user profile is used to discover and filter information according to the user's

interests. NewsDude (Billsus and Pazzani, 1999) obtains a short-term interest user profile using the k-NN algorithm and a long-term interest profile using a naïve Bayes classifier. Personal Searcher (Godoy and Amandi, 2006) uses a clustering algorithm to categorize web documents and hence determine a user's interest profile. SwiftFile uses a TF-IDF style classifier to organize emails (Segal and Kephart, 2000). CAP uses decision trees to learn users' scheduling preferences (Mitchell et al., 1994). Combinations of different techniques have also been used for building user profiles. For example, in (Martin-Bautista et al, 2000) the authors combine genetic algorithms and classification techniques (fuzzy logic) to build user profiles from a collection of documents previously retrieved by the user. In (Schiaffino and Amandi, 2000) case-based reasoning and Bayesian networks are combined to learn a user profile in a LIMS (Laboratory Information Management System). The user profile comprises routine user queries that represent a user's interests in the LIMS domain. In (Ko and Lee, 2000) the authors combine genetic algorithms and a naive Bayes classifier to recommend interesting web documents to users.

**c. Management of User Profiles**

Personalization and effective user profile management will be critical to meet the individual users' needs and for achieving eInclusion and eAccessibility. This paper outlines means to achieve the goal of the new ICT era where services and devices can be personalized by the users in order to meet their needs and preferences, in various situations. Behind every instance of personalization is a profile that stores the user preferences, context of use and other information that can be used to deliver a user experience tailored to their individual needs and preferences. Next Generation Networks (NGN) and the convergence between telephony and Internet services offer a wide range of new terminal and service definition possibilities, and a much wider range of application in society. This paper describes the personalization and profile management activities at European Telecommunications Standards Institute (ETSI) Technical Committee Human Factors, together with relevant experimentations in recent European research projects.

**d. Semantic based Personalized Search**

Personalized search utilizes the user context in a form of profile to increase the information retrieval accuracy with user's interests. Recently, semantic search has greatly attracted researchers' attention over the traditional keyword-based search because of having capabilities to figure out the meaning of search query, understanding users' information needs accurately using semantic web technology.

## WEB PERSONALIZATION AND ONTOLOGY

- **An Ontology**

ONOTOLOGY Ontology is a formal description and specification of knowledge. It provides a common understanding of topics to be communicated between users and systems [8]. As defined by Thomas R. Gruber as Ontology is "an explicit specification of a conceptualization". A conceptualization consists of a set of entities (such as objects and concepts) that may be used to express knowledge and relationships [7]. Developing a Ontology includes.

Ontologies have been proven an effective means for modeling digital collections and user context. Ontologies in the form of hierarchies of user interests have been proposed [11]. This ontology-based user modeling system integrates three ontologies: • User ontology: It includes different characteristics of users and their relationships. • Domain ontology: It captures the domain or application specific concepts and their relationships. • Log ontology: It represents the semantics of the user interaction with the system. [8]. The personalized ontology can describe different concept models for different users, although they may have the same topic. Ontology is based on two kinds of knowledge: 2.1 World Knowledge: World knowledge covering large number of topics so that the user's individual information needs can be best match 2.2 Expert Knowledge: Expert knowledge is the kind of knowledge classified by the people who hold expertise in that domain. [9]. Ontologies are ever growing, constantly ontology repositories needs to be updated with the latest click stream data.

- **The Need of Ontology Model**

Ontology is the model for knowledge description and formalization, which are widely used to represent user profile s in personalized web information gathering. When representing user profiles, many models have utilized only knowledge from either a global knowledge base or user local information. I

- **Reasons for developing an ontology**

Ontology is one of the approaches for knowledge representation. Ontology has some advantages that encourage researchers to use it. The most important advantage is the reusability and share ability (Shishehchi, Banihashem et al. 2010). Ontologies enable us to share the domain and the knowledge between applications (Yu, Nakamura et al. 2007; Shishehchi, Banihashem et al. 2010). Ontologies create machine-understandable descriptions of learning resources and provide the personalization and adaptively.

## II. WEB PERSONALIZATION AND RELATED WORK

### a. Personalization Based Web Usage Mining

Web mining is nothing but a careful systematic search and evaluation of the documents available in World Wide Web. Web-mining is related to the information and its features are divided in the procedure given below:

- Content Data: These are the documents that are available to the browser. Content mining is deriving information from the material of the web pages.[16]
- Structure Data: Web structure mining is the way of selecting information from the structure of data.[17]
- Usage data: The data which is taken from the browser is connected to the web. As cited above [18] Web Usage Mining (WUM) is the exploration and evaluation of browser access to the web information system using the data available to customize the web for user was not any new idea but was suggested way back in year 1995. [19]

### i. User-Interaction Tracking

The data about the transactions of a user with Internet is of great use for personalization. This connectivity data can be acquired in different ways: The web browser on the client side, web server logs, or representative server logs. As the importance of personalization rises, strict attention to minute details of tracking is of major importance and must be undertaken as the important feature in choosing a data source. There are many degrees of storage available in the web, especially to find out browsers access to much utilized page while browsing, user tend to refer back many a times data is directed with the help of web browser storage. Nevertheless, cache hits are not totally saved at proxy server logs, which in return effect the analyzing of user preferences and search behavior. Lin et al. (1999) [20] has invented an "access pattern collection server" to overcome the above said problem which works only when user secrecy doesn't matter. Cookey et al. (1999) [21] has used referrer and agent fields of a server log to obtain the information about the stored references that are hit back. Spiliopoulor et al. (2003) [22] analyzed the output of many such methodologies. It is found that server and proxy logs are unable to provide the temporary aspects of user communication. Time stamps stored in these logs for document demands will also have network-transmitting time. Because of the uncontrolled working of the network, the important information can't be inspected easily. Rather, if temporal characteristics are stored on the client side, hiding times of all user communications can be stored as promptly as needed. The data that is available with the user about the communication done with Internet is the most reliable and spatial. Since complete information is available with user, finding out the URL or resource of a data becomes very simple. This is a very big challenge in case of proxy or server logs. Moreover previously collecting data about the web page usage is a single person job for a proxy, but now it is rendered to all the users.

This work is known as session identification and is efficiently done at the user side. Because of the stateless connection model of the HTTP protocol, documents asked for are logged automatically in the server or proxy logs. The documents are reorganized and grouped for a better understanding and analysis and should be divided according to the key words. In Shahabi et al. (1997) [23], employed a remote agent that finds out browser communications on the user side. The information collected by every agent is saved as different semantic groups at the server so as to dismiss the user identification again. Nevertheless, collecting information at the client has a few oversights. Java scripts or Java applets are employed to run the agents, which collect data from users. For this Java program must be incorporated in the browser of a client, which may not be liked by users. Shahabi et al. (2000) [24] elaborated on this information collecting methods depending on the user-side data collecting idea.

### ii. Access Pattern Analysis

Digging in all the usage data is not possible because they are enormous in amount. The basic method is that, the value or grade of a paper is estimated according to the

number of hits that it has faced by the users. In addition, when a document is preferred that is selected first or after browsing few more documents among all the results.

Aggregate tree and hidden Markov models, which are not independent, are utilized to find out this characteristic and to imagine future references. Along with spatial features, temporal features like page view time are of much importance, especially in the surroundings of web personalization applications. Yan et al.(1996) [25] and Lovene and logic (2000)[26] believe that a paper can't be judged according to the time it is selected because sometimes some papers are not preferred due to its tough accessing process, Zipfin division and but this can be solved if the view time is combined with other characteristics , the present model which is explained is capable of combining the above said and many other qualities.

Hobasher et al have used the classical group regulation a priori algorithm to trace a frequent item sets depending on their patterns of occurrence at the browser sessions Mobasher et al [27] display that grouping methodologies give better results when compared to group regulations when used in the personalization of a web. Other set of methods, which are not independent are used to imagine future reference depending on the previous selections of a browser. These methods understand and represent important similarities among page selections. Cadez et al employ a Markov method for this Borges and Levene [28] explain a probabilistic regular grammar whose higher probability strings coincides to browsers selected access methods. Breese et al [29] carry out an experimental evaluation of expected algorithms like Bayesian division and Bayesian networks in the framework of web personalization and show that the results of these algorithms depend on the kind of application and wholeness of the usage data. Grouping to mine usage data methodology was initiated by Yan et al. [25]. With this method, browser terms are generally structured vectors. In the domestic design of the vector structure, every part of the vector shows the importance of a feature, like hit-count, for correlating to the web page. A group algorithm is used to find the browser access methods. Active user terms are divided with the help of a definite application dependent on the similarity measure like Euclidean breadth.

Presently many Algorithms were tested to access the grouping achievement in the surroundings of WUM; Perkowitz and Etzioni [33] presented a new grouping algorithm, cluster miner, which is developed to answer particular web-personalization necessities; Fu et al. [30] employ BIRCH [25], an efficient hierarchical clustering algorithm; Joshi and Krishnapuram [31] prefer a fuzzy relational clustering algorithm for WUM because they believe usage data are fuzzy in nature; Strehl and Ghosh [32] propose relationship-based clustering for high dimensional data mining in the context of WUM. Paliouras et al [34], from the machine-learning society correlate achievement of cluster miner with two other grouping procedures which are vibrant in machine-learning research, for example, auto class & self organizing maps, and display that Auto-class is better than other procedures. Mobasher et al [27] point out that a browser may exhibit features that one collected by various groups while he/she is to be divided as a single cluster. VerderMeer et al [35] examine anonymous WUM by taking dynamic profiles of browsers in association with static profiles. Dynamic clusters as a methodology to prepare the group model which can update the new developments in browsers behavior. A perfect similarity calculation, which can vary, is well estimated by the gap between partial user sessions and cluster representation is also a matter of importance.

**b. Personalization on Medical Search Engines**

To date, collaborative personalization has not been implemented on medical search engines. Among popular techniques to perform personalization are exact query matching among users of similar interest [37] and query similarity and page similarity matching [36]. These techniques have drawbacks. Users found that explicitly identifying a community of interest beforehand to be inconvenient, especially when a user wants to identify more than one area of interest Every new query entered experienced the cold start problem [38]. Query similarity measures using edit distance and user click through behavior overcame issues stated above. However, restrictions within the edit distance metric make it difficult to cater for all possible methods of similarity calculation. On the other hand, irrational searching behavior demonstrated by users [39] undermines the authenticity of user click through behavior when selecting links in the results page. Although evaluations of these techniques are valid on general search engines, the case maybe different on a specialized environment

like a medical search engine. This creates the opportunity for the application of query similarity and user click through behavior on medical search engines.

On vertical search engines, most research focus on assisting layman users in transforming a layman query into a medically focused query [41]. Techniques used in [40] and [42] transform a layman query into a medical query using Unified Medical Language System (UMLS). While [43] performs automatic conceptual query transformation, [42] manipulates the query using semantic distance with recommendations from a user's usage pattern or logs. Both these techniques attempt to provide relevant results to the user. Since the domain of search is a medical search engine, these techniques are indeed helpful to users. In another example, a controlled vocabulary called MeSH is used to perform automatic term mapping in PubMed [41]. This technique provides the opportunity for a user's query to be matched against an existing medical category. It also ensures that search results returned are matched to the users search intent. A different approach to assisting users on a medical search engine is explored in iMed [44]. iMed involves the user in the query expansion process. Initially, the user is required to select known symptoms and signs. The system then performs query expansion using an interactive questionnaire. This technique incorporates the user in the expansion process to ensure that a user's information seeking goal is preserved. However, the user now has to concentrate on the search process and search results.

**c. Algorithm for web personalization**

Various personalization schemes have been suggested in the literature. Letizia [45] is perhaps the first system which takes into account the user's navigation through a web site. This goal is achieved by using a client-side agent that records the user's behavior and gives interesting recommendations to the user herself. Yan et al. [46] propose a methodology for the automatic classification of web users according to their access patterns, using cluster analysis on the web logs. In [47], Joachims et al. describe WebWatcher, and similarly the Personal WebWatcher in [48], an intelligent agent system that provides navigation hints to the user, on the basis of a knowledge of the user's interests, the location and relevance of the items in the site, and the way in which other users interacted with the collection in the past.

In the SpeedTracer project,Wu et al. [49] use statistically dominant paths and association rules discovery, previously developed by Chen et al. [50]: each user session is mapped into a transaction and then data mining techniques are applied in order to discover the most frequent user traversal paths and the most frequently visited groups of pages. Zaiane et al. [51] propose the use of cube models to extract knowledge about the user behavior. Similarly, Buchner and Mulvenna [52] describe a knowledge discovery system which combines existing online analytical mining and marketing expertise. Very important is also the paper of Perkowitz and Etzioni [53], that first describes adaptive web sites as sites that semiautomatically improve their organization by learning from visitor access patterns. They used an algorithm (PageGather) based on a clustering methodology. In [54] Lee at al. propose an adaptive web system that analyzes user browsing patterns from their access records. The paper concentrates on the operating-efficiency of a web site that is, the efficiency with which a group of users browse a web site.

By achieving high efficiency, users spend less operating cost to accomplish a desired user goal. The paper develops an algorithm to accurately calculate the efficiency and to suggest how to increase it. Agreat number of papers also deals with time-related issues. In [55] Grandi introduces an exhaustive annotated bibliography on temporal and evolution aspects in the World Wide Web. Several time-related issues have been investigated, among which we are primarily interested in navigation time, that can be defined as the temporal dimension marking the navigation of theWeb by a user.

Differently by previous approach, Eirinaki and Vazirgiannis [56] introduce a PageRank-style algorithm which combines usage data and link analysis techniques for assigning probabilities to Web pages. Recently, there has been an increasing interest in web personalization techniques based on semantic analysis. In particular, there has been an interest in using deeper domain knowledge, often represented in the form of an ontology, as reported in [57], Anand et al. present an approach to integrate user rating vectors with an item ontology to generate recommendations.

Baraglia and Silvestri [58] introduced SUGGEST a completely onlineWeb recommender system that does

not require user intervention on the model building module, thus performing user profiling, model updating and recommendation, exploiting both logs and semantic annotation. We can finally conclude that most of the existing works try to classify a user i) while she is browsing the web site or ii) using registration information. Our main criticism stands in the fact that in some applications it is not possible to perform an "on line" classification if the number of visited pages is not sufficiently great. By the way, using the registration forms alone may result inaccurate if the interests of a user change over time. The novelty of our approach is that of proposing a classification process consisting of two phases: in the first one a pattern analysis and classification is performed by means of an unsupervised clustering algorithm, using the registration information provided by the users. In the second one a re-classification is iteratively repeated until a suitable convergence is reached.

Re-classification is used to overcome the inaccuracy of the registration information, based on the users' navigational behavior. To the best of our knowledge, our approach is the first one that uses re-classification in order to address both static and dynamic requirements.

**d. Personalized Recommendation in Social Tagging Systems**

In topic relevant partitions are created by clustering resources rather than tags. The most characteristic representatives of a cluster are recommended for users interested in a domain described by a cluster. Using clusters of resources, Flickr improves recommendation by distinguishing between alternative meanings of a query. For example, a user selecting the tag "apple" will receive several groups of resources. One group represents "fruit"; while another contains iPods, iMacs, and iPhones. A third cluster contains pictures of New York City. In [59] clusters of resources are shown to improve recommendation by categorizing the resources into topic domains. Consequently, the user may interactively disambiguate his query.

The utility of clustering extends beyond the scope of recommendation. In [60] hierarchical clustering is proposed to generate a taxonomy from a folksonomy. In [61], tag clusters are presumed to be representative of the resource content. Thus, a folksonomy of Web resources is used to move the Internet closer to the Semantic Web. Tag clustering can support tag recommendation, reducing annotation to a mouse click rather than a text entry. Well-chosen tags make the recovery process simple and offer some control over the tag-space diminishing tag redundancy and ambiguity to some degree. In [62], a group of tags are offered to the user based on several criteria (coverage, popularity, effort, uniformity) resulting in a cluster of a relevant tags.

In [63], ranking of web search was optimized using social annotations by taking in account the similarity of the query to the resources in del.icio.us. Their work is based on the assumption that folksonomies, such as del.icio.us, offer insights to the user's information needs. Our work shares this assumption as we seek to personalize the user recommendation.

In [64], a novel algorithm, FolkRank, for search and ranking in folksonomies is proposed that dependson interrelated tags, resources and users. The authors extend the commonly known PageRank algorithm to folksonomies under the assumption that users, resources and tags are important if they are connected to other important tags, resources and users in folksonomies. They use a weight passing scheme to derive the importance of an object in folksonomies. In this paper, we also adopt the idea of deriving the importance of resources to the users. Integral to our algorithm for personalization is the measurement of relevance between a user and a resource. A similar notion was previously described in [65] in which an affinity level was calculated between a user and a set of tag clusters. A collection of resources was then identified for each cluster based on tag usage. Resources were recommended to the user based on the user's affinity to the clusters and the associated resources.

**e. A Semantic Approach to Personalized Web Search**

Web search engines are essential "one size fits all" applications [66]. In order to meet the demands of extremely high query volume, search engines tend to avoid any kind of representation of user preferences, search context, or the task context [67]. Allan et al. [66] define the problem of contextual retrieval as follows: "Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs." Effective personalization of

information access involves two important challenges: accurately identifying the user context, and organizing the information in such a way that matches the particular context. Since the acquisition of user interests and preferences is an essential element in identifying the user context, most personalized search systems employ a user modeling component.

Recent studies show that users often settle for the results returned by imprecise queries, picking through them for relevant information, rather than expending the cognitive effort required to formulate more accurate queries. Since the users are reluctant to specify their underlying intent and search goals, personalization must pursue techniques that leverage implicit information about the user's interests [68], [69]. Google Personalized Search1 builds a user profile by means of implicit feedback where the system adapts the results according to the search history of the user. Many systems employ search personalization on the client-side by re-ranking documents that are suggested by an external search engine [70], [71] such as Google. Since the analysis of the pages in the result list is a time consuming process, these systems often take into account only the top ranked results. Also, only the snippets associated with each page in the search results is considered as opposed to the entire page content. Many personalization approaches are based on some type of a user profile which is a data instance of a user model that is captured based on the user's interaction. User profiles may include demographic information as well as representing the interests and preferences of a specific user. User profiles that are maintained over time can be categorized into short-term and long-term profiles. Short-term profiles can be utilized to keep track of the user's more recent, faster-changing interests. Long-term profiles represent user interests that are relatively stable over time.

Personal browsing agents such as WebMate [72] and Web- Watcher [73] perform tasks such as highlighting hyperlinks and refining search keywords to satisfy the user's short-term interests. These approaches focus on collecting information about the users as they browse or perform other activities. InfoWeb [74] builds semantic network based profiles that represents long-term user interests. The user model is utilized for filtering online digital library documents.

One increasingly popular method to mediate information access is through the use of ontologies [75]. [76], [77] in utilizing the Open Directory Project (ODP)2 taxonomy as the Web topic ontology. The ODP is the largest and most comprehensive Web directory, which is maintained by a global community of volunteer editors. The ODP taxonomy is used as the basis for various research projects in the area of Web personaliza tion [78], [79]. Chirita et al. [80] utilize the documents stored locally on a desktop PC for personalized query expansion. The query terms are selected for Web search by adapting summarization and natural language processing techniques to extract keywords from locally stored desktop documents.

Hyperlink-based approaches have also been explored as a means to personalize Web search. In Personalization the well-known Hyperlink Induced Topic Selection (HITS) algorithm [81] is enhanced with an interactive query scheme utilizing the Web taxonomy provided by the ODP to resolve the meaning of a user query. Considerable amount of Web personalization research has been aimed at enhancing the original PageRank algorithm introduced in Google. In Personalized Page Rank [82], a set of personalized hub pages with high PageRank is needed to drive the personalized rank values. In order to automate the hub selection in Personalized Page Rank, a set of user collected bookmarks is utilized in a ranking platform called PROS [83]. Instead of computing a single global PageRank value for every page, the Topic-Sensitive PageRank [84] approach tailors the PageRank values based on the 16 main topics listed in the Open Directory. Multiple Topic-Sensitive PageRank values are computed off-line. Using the similarity of the topics to the query, a linear combination of the topic-sensitive ranks are employed at run-time to determine more accurately which pages are truly the most important with respect to a particular query. This approach is effective only if the search engine can estimate the suitable topic for the query and the user. Thus, Qui and Cho [85] extend the topic-sensitive method to address the problem of automatic identification of user preferences and interests.

## III. CONCLUSION

Even though the World Wide Web is the major resource of electronic information, it lacks with efficient methods for retrieving, filtering, and displaying the information

that is exactly required by each user. With the advent of the Internet, there is a remarkable growth of data available on the World Wide Web. Hence the task of retrieving the only required information keeps becoming more and more difficult and time consuming. To reduce information overload and create customer loyalty, Web Personalization, a significant tool that provides the users with important competitive advantages is required. A Personalized Information Retrieval approach that is mainly based on the end user modeling increases user satisfaction. Also personalizing web search results has been proved as to greatly improve the search experience. This paper reviews the various research activities carried out to improve the performance of personalization process and also the Information Retrieval system performance.

## IV.REFERENCES


[1]. M. Albanese, A. Picariello, C. Sansone, L. Sansone, "A Web Personalization System based on Web Usage Mining Techniques", in Proc. of WWW2004, May 2004, New York, USA.

[2]. B. Mobasher, H. Dai, T. Luo, Y. Sung, J. Zhu, "Integrating web usage and content mining for more effective Personalization", in Proc. of the International Conference on Ecommerce and Web Technologies (ECWeb2000), Greenwich, UK, September 2000.

[3]. Jiawei Han And Micheline Kamber "Data Mining: Concepts and Techniques", 2nd ed., Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.

[4]. S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing" Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.

[5]. Y. Li and N. Zhong, "Web Mining Model and Its Applications for Information Gathering" Knowledge-Based Systems, vol. 17, pp. 207-217, 2004.

[6]. Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs" IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.

[7]. Morita M., Shinoda, Y., "Information Filtering Based on User Behaviour Analysis and Best Match Retrieval", in Proceedings of the 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 272-281.

[8]. Shardanand U., Pattie M., "Social Information Filtering: Algorithms for Automating "Word of mouth", in Proceedings of the Human Factors in Computing System, Denver, May 1995, pp. 210-217.

[9]. "CGI ocumentation", http://hoohoo.ncsa.uiuc.edu/cgi/

[10]. Xiaohui Tao, Yuefeng Li, and Ning Zhong, "Senior Member, IEEE, "A Personalized Ontology Model for Web Information Gathering", Ieee Transactions On Knowledge And Data Engineering, Vol. 23, No. 4, April 2011.

[11]. http://en.wikipedia.org/wiki/Ontology_%28 information_science%29

[12]. http://www.unicist.org/what-is-an-ontology.pdf

[13]. http://en.wikipedia.org/wiki/Ontology

[14]. Bhaganagare Ravishankar, Dharmadhikari Dipa, "Web Personalization Using Ontology: A Survey", IOSR Journal of Computer Engineering (IOSRJCE) ISSN : 2278-0661 Volume 1, Issue 3 (May-June 2012), PP 37-45 www.iosrjournals.org.

[15]. Xiaohui Tao, Yuefeng Li, and Ning Zhong, Senior Member, IEEE. "A Personalized Ontology Model for Web Information Gathering", IEEE transactions on Knowledge and Data Engineering, Vol. 23, No. 4, April 2011.

[16]. Salton, G., McGill, M.: An Introduction to modern information retrieval. Mc-Graw-Hill, New York, NY (1983)

[17]. Micarelli, A., Sciarrone, F., Marinilli, M.: Web document modeling. In Brusilovsky, P., Kobsa, A., Nejdl, W., eds.: The Adaptive Web: Methods and Strategies of Web Personalization. Volume 4321 of Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York (2007)

[18]. Olston, C., Chi, E.H.: ScentTrails: Integrating browsing and searching on the web. ACM Transactions on Computer-Human Interaction 10(3) (2003) 177-197

[19]. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Commun. ACM 30(11) (1987) 964-971

[20]. Freyne, J., Smyth, B.: An experiment in social search. In Bra, P.D., Nejdl, W., eds.: Adaptive Hypermedia and AdaptiveWeb-Based Systems, Third International Conference, AH 2004, Eindhoven, The Netherlands, August 23-26, 2004, Proceedings. Volume 3137 of Lecture Notes in Computer Science., Springer (2004) 95-103

[21]. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. Proceedings for ACM conference on Computer supported cooperative work, New York, NY, USA, ACM Press (1994) 175-186

[22]. Dieberger, A., Dourish, P., H¨o¨ok, K., Resnick, P., Wexelblat, A.: Social navigation: techniques for building more usable systems. Interactions 7(6) (2000) 36-45

[23]. Kritikopoulos, A., Sideri, M.: The compass filter: Search engine result personalization using web communities. In Mobasher, B., Anand, S.S., eds.: Intelligent Techniques forWeb Personalization, IJCAI 2003Workshop, ITWP 2003, Acapulco, Mexico, August 11, 2003, Revised Selected Papers. Volume 3169 of Lecture Notes in Computer Science., Springer (2003) 229-240

[24]. A. Broder. A taxonomy of web search. SIGIR Forum, 36(2), 2002.

[25]. B. Jansen and A. Spink. How are we searching the web? a comparison of nine search engine query logs. Information Processing and Management, 42, 2006.

[26]. O. Madani and D. DeCoste. Contextual recommender problems. In Utility Based Data Mining Workshop at KDD, 2005.

[27]. M. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 27, 1997.

[28]. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (1999) 604-632

[29]. Henzinger, M. 2000. Link analysis in web information retrieval. Bull. of the Technical Committee on Data Engrg., IEEE Computer Soc. 23 3-9

[30]. Baumgarten, M., A. G. Büchner, S. S. Anand, M. D. Mulvenna, J. G. Hughes. 2000. Navigation pattern discovery from internet data. M. Spiliopoulou, B. Masand, eds. Advances in Web Usage Analysis and User Profiling, Lecturer Notes in Computer Science 1836 70-87.

[31]. Armstrong, R., D. Freitag, T. Joachims, T. Mitchell. 1995. Web-Watcher: A learning apprentice for the world wide web. AAAI Spring Sympos. on Inform. Gathering from Heterogeneous, Distributed Environments, Stanford, CA, 6-13.

[32]. Srivastava, J., R. Cooley, M. Deshpande, P. N. Tan. 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations 1 12-23.

[33]. Lin, I. Y., X. M. Huang, M. S. Chen. 1999. Capturing user access patterns in the web for data mining. Proc. of the 11th IEEE Internat. Conf. Tools with Artificial Intelligence, Chicago, IL, 22-29.

[34]. Cooley, R., B. Mobasher, J. Srivastava. 1999. Data preparation for mining world wide web browsing patterns. Knowledge and Inform. Systems 1 5-32.

[35]. Spiliopoulou, M., B. Mobasher, B. Berendt, M. Nakagawa. 2003. Evaluating the quality of data preparation heuristics in web usage analysis. INFORMS J. on Comput. 15(2) 171-190.

[36]. Rose, D.E. & Levinson, D. (2004). Understanding user goals in web search. Proceedings of the 13th International Conference on World Wide Web, pp. 13-19

[37]. Smyth, B. (2007). A Community-Based Approach to Personalizing Web Search, Cover Feature, IEEE Computer, 40(8), pp.42-50.

[38]. Zigoris, P. & Zhang,Y.(2006). Bayesian Adaptive User Profiling with Explicit and Implicit Feedback, Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, pp. 397-40

[39]. Jansen B.J., Booth D.L., Spink A. (2008). Determining the user informational navigational and transactional intent of web queries, Information Processing Management, Vol. 44 (3), pp. 1251-1266

[40]. Can, A.B. & Baykal, N. (2007). MedicoPort: A medical search engine for all, Computer Methods and Programs in Biomedicine, Vol.86, pp.73-86

[41]. Zeng Q.T. & Tse T. (2006). Exploring and Developing Consumer Health Vocabularies, Journal of American Medical Information Association, Vol. 13 pp.24-29

[42]. Zhiyong L., Kim W. & Wilbur J.W. (2009). Evaluation of query expansion using MeSH in PubMed, Journal of Information Retrieval, Vol. 12 (1), pp. 69-80

[43]. Eysenbach, G. & Kohler C. (2002). How Do Consumers Search For And Appraise Health Information On The World Wide Web? Qualitative Study Using Focus Groups, Usability Tests, and In Depth Interviews, British Medical Journal, 2002, Vol. 24 pp.573- 577

[44]. Luo, G & Tang, C. (2008). On Iterative intelligent Medical Search, Proceedings of the 31st Annual International ACM Special Interest Group on Information Retrieval, pp. 3-10

[45]. H. Lieberman, Letizia: An agent that assists web browsing, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995, pp. 924–929.

[46]. T.W.Yan, M. Jacobsen, H. Garcia-Molina and U. Dayal, From user access patterns to dynamic hypertext linking, in: Proceeding of the Fifth International WorldWideWeb Conference, Paris, 1996.

[47]. T. Joachims, D. Freitag and T. Mitchell, Webwatcher: a tour guide for the world wide web, in: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, August 1997, pp. 770–777.

[48]. D. Mladenic, Machine learning used by personalWebWatcher, in: Proceedings of the Workshop on Machine Learning and Intelligent Agents (ACAI-99), Chania, Greece, July 1999.

[49]. K.L. Wu, P.S. Yu and A. Ballman, Speedtracer: A web usage mining and analysis tool, IBM Systems Journal 37(1) (1998).

[50]. M.S. Chen, J.S. Park and P.S. Yu, Data mining for path traversal patterns in a web environment, in: Proceedings of the 16th International Conference on Distributed Computing Systems, 1996, pp. 385–392.

[51]. O.R. Zaiane, M. Xin and J. Han, Discovering web access patterns and trends by applying olap and data mining technology on web logs, in: Proceedings of Advances in Digital Libraries Conference (ADL98), Santa Barbara, CA, April 1998.

[52]. A.G. B¨uchner and M.D. Mulvenna, Discovering internet marketing intelligence through online analytical web usage mining, ACM SIGMOD Record 27(4) (December 1998), 54–61.

[53]. M. Perkowitz and O. Etzioni, Towards adaptive web sites: Conceptual framework and case study, Computer Networks 31(11–16) (May 1999), 1245–1258.

[54]. J.-H. Lee and W.-K. Shiu, An adaptive website system to improve effciency with web mining techniques, Advanced Engineering Informatics 18(3) (July 2004), 129–142.

[55]. F. Grandi, An annotated bibliography on temporal and evolution aspects in the world wide web, TIMECENTER Technical Report TR-75, University of Bologna, Italy, September 2003.

[56]. M. Eirinaki and M. Vazirgiannis, Web site personalization based on link analysis and navigational patterns, ACM Transactions on Internet Technology 7(4) (2007), 21.

[57]. S.S. Anand, P. Kearney and M. Shapcott, Generating semantically enriched user profiles for web personalization, ACM Transactions on Internet Technology 7(4) (2007), 22.

[58]. R. Baraglia and F. Silvestri, Dynamic personalization of web sites without user intervention, Communications of the ACM 50(2) (2007), 63–67.

[59]. H. Chen and S. Dumais. Bringing order to the Web: automatically categorizing search results. Proceedings of the SIGCHI conference on Human factors in computing systems, pages 145–152, 2000.

[60]. P. Heymann and H. Garcia-Molina. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical report, Technical Report 2006-10, Computer Science Department, April 2006.

[61]. X.Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. Proceedings of the 15th international conference on World Wide Web, pages 417–426, 2006.

[62]. Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland,May, 2006.

[63]. S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. Proceedings of the 16th international conference on World Wide Web, pages 501–510, 2007.

[64]. A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. The Semantic Web: Research and Applications, 4011:411–426, 2006

[65]. S. Niwa, T. Doi, and S. Honiden.Web Page Recommender System based on Folksonomy Mining for ITNGŠ06 Submissions. Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)-Volume 00, pages 388–393, 2006.

[66]. J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harman, D. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai, "Challenges in information retrieval and language modeling," ACM SIGIR Forum, vol. 37, no. 1, pp. 31–47, 2003.

[67]. S. Lawrence, "Context in web search," IEEE Data Engineering Bulletin, vol. 23, no. 3, pp. 25–32, 2000.

[68]. X. Shen, B. Tan, and C. Zhai, "Ucair: Capturing and exploiting context for personalized search," in Proceedings of the Information Retrieval in Context Workshop, SIGIR IRiX 2005, Salvador, Brazil, August 2005.

[69]. J. Teevan, S. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities," in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, Salvador, Brazil, August 2005, pp. 449–456.

[70]. M. Speretta and S. Gauch, "Personalized search based on user search histories," in Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005, Compigne, France, September 2005, pp. 622–628.

[71]. A. Micarelli and F. Sciarrone, "Anatomy and empirical evaluation of an adaptive web-based information filtering system," User Modeling and User-Adapted Interaction, vol. 14, no. 2-3, pp. 159–200, 2004.

[72]. F. Gasparetti and A. Micarellli, "A personal agent for browsing and searching," in Proceedings of the 2nd International Conference on Autonomous Agents, St. Paul, MN, May 1998, pp. 132–139.

[73]. D. Mladenic, "Personal webwatcher: Design and implementation," Technical Report IJS-DP-7472, 1998.

[74]. G. Gentili, A. Micarelli, and F. Sciarrone, "Infoweb: An adaptive information filtering system for the cultural heritage domain," Applied Artificial Intelligence, vol. 17, no. 8-9, pp. 715–744, 2003.

[75]. H. Haav and T. Lubi, "A survey of concept-based information retrieval tools on the web," in 5th East-European Conference, ADBIS 2001, Vilnius, Lithuania, September 2001, pp. 29–41.

[76]. S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing," Web Intelligence and Agent Systems, vol. 1, no. 3-4, 2003.

[77]. D. Ravindran and S. Gauch, "Exploting hierarchical relationships in conceptual search," in Proceedings of the 13th International Conference on Information and Knowledge Management, ACM CIKM 2004, Washington DC, November 2004.

[78]. P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using odp metadata to personalize search," in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, Salvador, Brazil, August 2005, pp. 178–185.

[79]. C. Ziegler, K. Simon, and G. Lausen, "Automatic computation of semantic proximity using taxonomic knowledge," in Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, Arlington, VA, November 2006, pp. 465– 474.

[80]. P. Chirita, C. Firan, and W. Nejdl, "Summarizing local context to personalize global web search," in Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, Arlington, VA, November 2006, pp. 287–296.

[81]. H. Chang, D. Cohn, and A. McCallum, "Learning to create customized authority lists," in Proceedings of the 7th International Conference on Machine Learning, ICML 2000, San Francisco, CA, July 2000, pp. 127–134.

[82]. G. Jeh and J. Widom, "Scaling personalized web search," in Proceedings of the 12th international conference on World Wide Web, WWW 2003, Budapest, Hungary, May 2003, pp. 271–279.

[83]. P. A. Chirita, D. Olmedilla, and W. Nejdl, "Pros: A personalized ranking platform for web search," in Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2004, Eindhoven, The Netherlands, August 2004.

[84]. T. H. Haveliwala, "Topic-sensitive pagerank," in Proceedings of the 11th International World Wide Web Conference, WWW 2002, Honolulu, Hawaii, May 2002.

[85]. F. Qiu and J. Cho, "Automatic identification of user interest for personalized search," in Proceedings of the 15th InternationalWorld Wide Web Conference, WWW 2006, Edinburgh, Scotland, May 2006, pp. 72.