ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2015.0135

VALIDATING THE PERFORMANCE OF PERSONALIZATION TECHNIQUES IN SEARCH ENGINE

A. Suruliandi[1], T. Dhiliphan Rajkumar[2] and P. Selvaperumal[3]
Department of Computer Science and Engineering, Manonmaniam Sundaranar University, India
E-mail: [1]suruliandi@yahoo.com, [2]dhilipanrajkumar@gmail.com, [3]selvaperumal.p@gmail.com

Abstract

User profiling is an important and basic component in personalized search engine. Search engines respond to a user's query by using the bag-of-words model, which matches keyword between the query and web documents but ignore contexts and users' preferences. Personalized search greatly improves the search results as of the results provided by the search engine without personalization. In this paper, the performance of personalized search based on content analysis and personalized search based on user group have been evaluated. In personalized search based on content analysis the contents are traced by finding the user's browsed documents and search history, which reduce the users search time. In user profile only user preference alone is taken into consideration. The experimental results show that the personalized search based on user group method having higher precision and recall rate than the content analysis method.

Keywords:

Search Engine, Personalization, User Profile, Content Analysis

1. INTRODUCTION

Every day, millions of searches are conducted on search engines such as Google, Yahoo!, and Bing, etc. As of today, the indexed web contains at least 3.53 billion pages. In fact, the overall web may consist of over 1 trillion unique URLs, more and more of which is being indexed by search engines every day. Out of this, users typically search for the relevant information that they want from search engines. Lot of users is submitting the queries in short and confusing manner. From the study of M. Jansen [8], it is found that the average query size on a well-used search engine was only 2.35 terms. These small queries are not likely to be able to accurately convey what the user really wants. Because of that, lots of results retrieved may be irrelevant to the user due to ambiguous queries. The problem that the search engines face is that the queries are different and often quite fuzzy and/or ambiguous in terms of user requirements. The reason for the problem is the keyword-based query interface, is very difficult for users to express what they need. In addition, search engines do not utilize user information. Hence personalization of search query is important.

Personalization is the process of presenting right information to the right user at the right moment. To do this, information about the user is collected by the system, analyzed and the results of the analysis are stored in the user profile. To do personalization two personalized search strategies are used they are content analysis based personalization and personalization based on user profile [22]. Content analysis is a methodology for studying the content and to learn the links between each data. With this the meaning and relationship of each data can be understood easily. K.W.T. Leung [11] offers a definition of content analysis as "Any technique for making inferences by objectively and systematically identifying specified characteristics of messages". User profile is used to store the details of each user. In this, behavior of each user, their habits etc are monitored. The profile of users is automatically learned from users past queries and click-through in search engine logs. The query log of a search engine is to trace the user search query and keywords. Also the date and time of the query are recorded, the query terms clicked by the users are noted, and finally, the pages viewed by the users and their rank in the search result listing are recorded [14]. Hence in this paper, two personalization strategies such as content analysis and user profile performance are validated.

1.1 MOTIVATION AND JUSTIFICATION FOR THE ANALYSIS

In recent times, current search engines retrieve results based on web reputation rather than user's interests. The relevant result that comes in first few pages has lower chance of meeting user interest. So there is a need for employing personalization technique in search engine. Some researchers have noticed that personalization varies in effectiveness for different queries. Kenneth Wai-Ting [11] proposed that personalized content based clustering of search engine provides query suggestion for individual users. In this, they extract concepts from the web-snippets of the search result returned from a query.

K.W. Ting [12] proposed personalization based on user profile. In this, user profiles are employed to group similar queries according to user needs. In the personalization search based on user group various user profiling strategies such as P_{Click} and $P_{Joshaims}$ are considered and then the click-through is collected to forecast the user's conceptual needs [10][16]&[18]. In user profile, user searched contents are traced, user behaviors like time spent for reading online document and which area user is interested etc, can be found out. The drawback in user group method is user does not provide correct information and user's interest may change over time. These are some problems of creating user profile. No one insist that which method is best whether content analysis or user group method. Motivated by this, content analysis and user group method are taken up for comparison. Personalization of search engine is used to minimize the search time of user and to provide users needed information accurately. Justified by these, the two methods content analysis based on personalization and personalization based on user group are validated. Thereby, the best personalization technique can be identified.

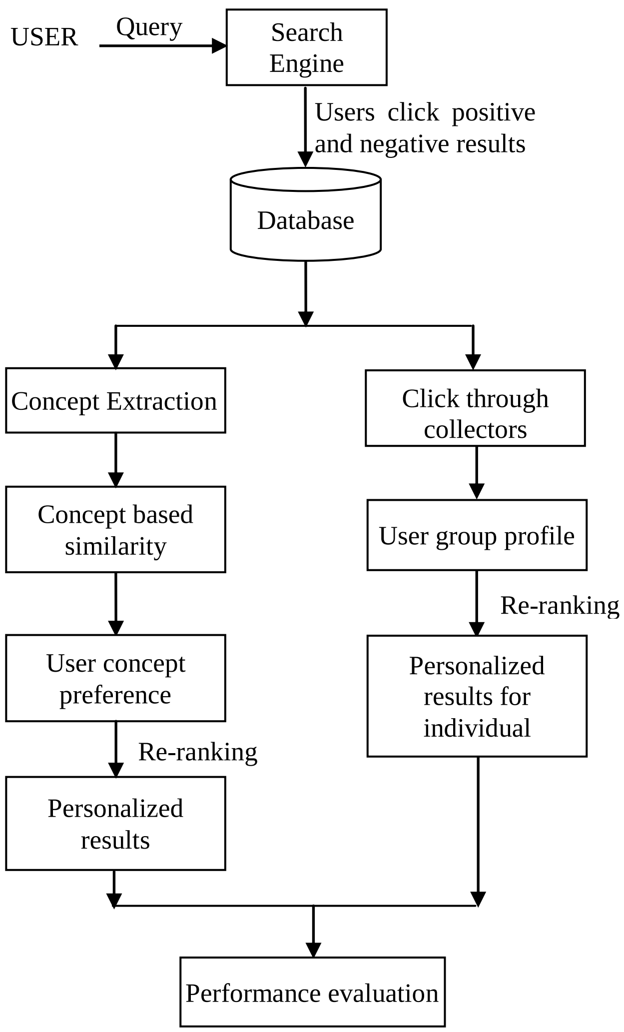

Fig.1. Block Diagram for performance evaluation of personalization based on content analysis and user group

1.2 OUTLINE OF THE PROPOSED WORK

The Fig.1 is the overall working model of personalization based on content analysis and user group. First, user's queries are got and the results are traced using search engine. The users clicked queries are positive, while others are negative and these are stored in the database. The concept and concept relationship are measured using similarity values and personalization is achieved using the user's concept preference. In the user group method, the user results are traced using some of the click methods like P_{Click}, $P_{joachims\text{-}C}$ methods and the results are maintained in click-through collectors. For each individual, their details are stored and their searched contents are maintained in user profile. With the clicks made by the individual their interest are measured and the results are re-ranked. Finally, performance is evaluated using precision, recall and f-measure.

1.3 ORGANIZATION OF THE PAPER

The remaining sections are organized as follows: In section 2, the personalization techniques have been discussed. In section 3, a detailed statistics of the data set used in the experiment and the comparative study of content analysis and user group are described. The section 4 focuses on conclusion and future work.

2. PERSONALIZATION TECHNIQUES

In this section, two personalized search strategies content analysis and personalized search based on user groups are discussed.

2.1 CONTENT ANALYSIS BASED PERSONALIZATION

Content analysis based personalization automatically traces user's interest by finding the user's browsed documents and search history. Content analysis method is used to achieve personalization effect thereby it reduces the search time of user. User's queries and their selected snippets are categorized into concept hierarchies, which are gathered to generate a user concept preference [3]. When the user selects a query, each of the returned snippets is also classified. Some other personalized approach use lists of key words to represent user interests. To do this, first concept is extracted from web snippets, finding concept relationship between queries and finally user's concept preference is developed by combining extracted concept, concept relation and user's click-through.

2.1.1 Extracting Concepts from Web-Snippets:

The concept is extracted by employing the well-known problem of finding frequent item sets in data mining. When user submits a query, a set of results with web snippets are returned. In this, if a keyword or phrase exists frequently in the web-snippets of a particular query, it represents an important concept related to the query because it exists with the query in the top documents. To measure the interest of a particular keyword or phrase k_i extracted from the web-snippets:

$$\text{support}(k_i) = \frac{sf(k_i)}{n} \cdot |k_i| \tag{1}$$

Table.1. Example of Concept Extracted for the Query "Data Mining"

Concept	**support**
Analysis	0.1111
Databases	0.1111
Computer	0.1333
Predictive	0.1556
information	0.2667
Previously	0.1333
Business	0.1111

where, $sf(k_i)$ is the snippet frequency of the keyword or phrase k_i (i.e., the number of web-snippets containing k_i), n is the number of web-snippets returned, and $|k_i|$ is the number of terms in the keyword or phrase k_i. If the support of a keyword or phrase k_i is greater than the threshold s, and k_i as a concept for the query q. Table.1 shows an example set of concepts extracted for the query "data mining." Before concepts are extracted, stop words, such as "the," "on," "and," etc., are first removed from the snippets. The maximum length of a concept is limited. This process not only reduces the processing time, but also avoids extracting meaningless concepts.

2.1.2 Extracting Concept Relationship:

In extracting concept relation, signal-to-noise formula from data mining to establish the similarity between keywords k_1 and k_2. The two keywords from a query *q* are similar if they coexist frequently in the web-snippets arising from the query *q*.

$$sim(k_1, k_2) = \frac{n * df(k_1 \cup k_2)}{df(k_1) * df(k_2)} \Big/ \log n \quad (2)$$

where, *n* is the number of documents in mass, *df*(*k*) is the document frequency of the keyword *k*, and $df(k_1 \cup k_2)$ is the joint document frequency of k_1 and k_2. The similarity $sim(k_1,k_2)$ obtained using the above formula always lies between [0, 1]. In the search engine context, two concepts k_i and k_j can coexist in the web snippets,

$$sim_{R,snippet}(k_i, k_j) = \log \frac{n * sf_{snippet}(k_i \cup k_j)}{sf_{snippet}(k_i) * sf_{snippet}(k_j)} \Big/ \log n \quad (3)$$

Table.2. Example Concepts Extracted Relationships for the Query "Data Mining"

Concept1	Concept2	Relations
Data	Mining	0.891304347826
Data	Analysis	0.173913043478
Data	knowledge	0.152173913043
Data	Discovery	0.108695652174
mining	Data	0.891304347826
mining	Analysis	0.173913043478
Mining	Knowledge	0.152173913043
Mining	Discovery	0.108695652174

where, $sf_{snippet}(k_i \cup k_j) / sf_{snippet}(k_i \cup k_j)$ the joint snippet frequencies of the concept k_i and k_j in web snippets. $sf_{snippet}(k_i) . sf_{snippet}(k_j)$ is the snippet frequency of the concept k_i and k_j respectively.

2.1.3 Developing User Concept Preference:

The concept relationship is processed by considering user's click-through. The user clicked queries are called user positive preference and others are user's negative preference. When user clicks on the query, the weight of the extracted concept is incremented by 1 to show the user interest. For other concept that are related to the user's query are also incremented to the similarity score. If the concept is closely related to the user's positive result, then it is incremented to the higher value. Otherwise, it is incremented to the small fraction close to zero. By this, user concept preference is created.

2.1.4 Re-ranking the Results:

When the user submits a query to the search engine the desired results for that query is displayed to the user with title snippet and URL (Uniform Resource Locator). The results are re-ranked using the original list and the conceptual similarity to the user's profile. With the search result title and snippets a user profile re-ranking is created by calculating the conceptual similarity between each snippets and the user interested one. The similarity between them is calculated using cosine similarity function.

$$sim(user_i, snippet_j) = \sum_{k=1}^{N} Wt_{ik} + Wt_{jk} \quad (4)$$

where, Wt_{ik} is weight of concept in user profile and Wt_{jk} weight of concept in each snippets. The snippets are re-ranked by the conceptual similarity using the conceptual rank. The final rank is calculated using the weighted scheme.

$$\textit{Final rank} = \alpha * \textit{conceptual rank} + (1-\alpha) * \textit{Google rank} \quad (5)$$

where, α has value between 0 and 1. When α is 0 conceptual ranks is not given any weight and it is equal to original rank assigned by Google. If it is 1 then Google rank is ignored and new conceptual rank is assigned and provides user needed results.

2.2 PERSONALIZATION BASED ON USER GROUP

In most of the personalized search techniques, the information provided by the user is considered to create user profile. There are some strategies that include the preference of a group of users to accomplish personalized search. In this paper, three user profiling methods that use concept based and use user's preferences. They are P_{Click} $P_{Joachims}$ and $P_{Click+Joachims}$, which deals with the click histories of a group of users with similar interest [10][12][16]. In P_{Click} concept-based user's positive preference are considered, whereas $P_{joachims-c}$ was based on users' document preference.

2.2.1 Click-Based Method (P_{click}):

The concepts extracted for a query q using the concept extraction method discussed in section 2.1.1 provides the possible concept space for the query *q*. The concept may cover more than what the user actually needs. For example, when the user searches for the query "apple," the concept derived from the concept extraction method contains the concepts "macintosh," "ipod," "apple shops," and "fruit." If the user is interested in "apple" as a fruit and clicks on pages containing the concept "fruit," the user profile represented as a weighted concept vector should record the user interest on the concept "apple" and its related contents whereas, concepts such as "macintosh," "ipod," etc are downgraded. For this the count value is incremented for the fruit and their neighbor content to one.

Table.3. An Example for Click through Data

Doc	Clicked	Search Result	Extracted Concept
d1	Clicked	Apple Computer	iPhone
d2		Apple Fruit	Tree, Farm
d3	Clicked	Apple Store	Macintosh
d4		Apple corps	Fruit
d5	Clicked	iPad	iPad, Apple store

Therefore the following formulas are used to capture the user's degree of interest w_{ci} on the extracted concepts e_i, when a web-snippet W_j is clicked by the user (denoted by click(c_j)),

$$click(W_j) \rightarrow \forall e_i \in c_j, w_{ci} = w_{ci} + 1 \quad (6)$$

$$\begin{aligned} &If\ sim(e_i, e_j) > 0 \\ &click(W_j) \rightarrow \forall e_i \in c_j, w_{ci} = w_{ci} + sim_R(e_i, e_j) \end{aligned} \quad (7)$$

where, W_j is a web-snippet, w_{ci} represents the user's degree of interest on the extracted concept e_i, and e_j is the neighborhood concept of e_i.

When a web-snippet W_j has been clicked by a user, the weight w_{ci} of concepts e_i appearing in W_j is incremented by 1. For other concepts e_j that are related to e_i on the concept relationship graph, the weights w_{ci} for concepts "fruit," "apple farm" "juice" and "apple grower" are incremented according to the similarity score, because they are related to the concept "Apple fruit". The weights w_{ti} of the concepts "mac os", "software", "applestore", "iPod", "iPhone" and "hardware" remain zero showing that the user is not interested in information about "apple store".

2.2.2 Joachims-C Method ($P_{joachims\text{-}C}$):

Joachims [10] introduced a technique entirely based on click through data to learn ranking function. Joachims et al. presented an empirical evaluation of understanding click through evidence. It is believed that every user would search results from top to bottom. If a user skipped a document d_i before clicking d_j (where rank of d_j > rank of d_i), one must have searched d_i and determined not to click on it. According to Joachims' the original proposal would extract the user preference as d_j <*r*' d_i. Joachims method was based on users' document preferences. If a user has skipped a document d_i at rank *i* before clicking on d_j at rank *j*, one must have scanned the d_i and decided to skip it. Thus the user preference for document d_j is more than document d_i. In this, document based method is converted to concept based method. For all the concept c_1, c_2............c_i extracted for the query *q*, the user selected contents are stored in the corresponding weight values W_{c1}, W_{c2}...............................W_{cn}, creating concept profile for the query *q*.

$$P_{joachims\text{-}c} = (W_{c1}, W_{c2} \ldots\ldots W_{cn}) \quad (8)$$

2.2.3 Combined P_{click} and $P_{joachims\text{-}C}$ Method (P_{click}+$P_{joachims\text{-}C}$):

In this work, it is observed that P_{Click} method is used to capture user's positive results. Joachims method is used to capture only negative preference. By combining both the results, good precision and recall value can be achieved. The user profiles P_{Click} and $P_{Joachims\text{-}C}$ can be combined using the formula:

$$W(P_c+P_J) = W(P_c)+W(P_J), \text{ if } W(P_J)<0, \quad (9)$$

$$W(P_c+P_J) = W(P_c), \text{ otherwise}, \quad (10)$$

where, $W(P_C+P_J) \in P_{Click+Joachims\text{-}C}$, $W(P_C) \in P_{Click}$ and $W(P_J) \in P_{Joachims\text{-}C}$.

The combined user profile method is applied over Google search results and re-ranked them based on the user interested results. For this group level re-ranking is used. For calculating the similarity for group of users following formula is used.

$$sim(u_1, u_2) = \frac{c_1(u_1).c_1(u_2)}{\|c_1(u_1)\|\|c_1(u_2)\|} \quad (11)$$

where, u_1 and u_2 are *user*$_1$ and *user*$_2$ and c_1 category vector of web page. The historical clicks made by similar users to re-rank the search result is calculated by,

$$sim^{G-czlick}(q, p, u) = \frac{\sum_{u_s \in s_u(u)} sim(u_s, u)|clicks(q, p, u_s)|}{\alpha + \sum_{u_s \in s_u(u)} |clicks(q.u_s)|} \quad (12)$$

where, *q*, *p* and *u* are query, web page, and user respectively. G-Click for group clicks, u_s is user similarity score and α has value between 0 and 1. The experimental results are given in the following section.

3. EXPERIMENTAL RESULTS

3.1 DATA SETS

For the combined approach proposed by this work, the Google API is used. For collecting the user searched data and to validate the work, Google search results for 10 days in November 2013 are taken. The default snippet counts are set to 100. The entire log is too large. In this, the user clicked contents, their positive and negative preference are collected. Table.4 is the statistics for the tested queries.

Table.4. Statistics of the Tested Queries

Statistics	
Number of users	30
Number of queries assigned to each user	5
Number of test queries	100
Maximum number of retrieved URL for a query	50
Number of extracted concept for a query	156

Some of the queries that are used for the evaluation are ambiguous queries, entitie names and general terms. Table.5 shows the queries used for evaluation of personalization of search results.

Table.5. Queries Used For Evaluation of Personalization Techniques

Types	Queries
Ambiguous	apple, tiger, sun, penguin, java
Entity names	dell, Disney, divya
General terms	maps, flower, music, network

Most users get the needed result within first 10 search results. To find this, relative click frequency at result position *p* is calculated by first computing the frequency of click at position *p* for each query. These queries are then averaged across various queries, so the relative frequency of a click is at the top [15]. This result reflects the fact that search engine does a reasonable job of ranking results. The graph shows that most of the times user gets the web content from the top 10 results. Hence in this, the first 50 results are taken up and it is stored in the database. The threshold for concept was set to 0.5 which was stable and the threshold for establishing concept relationship is set to zero.

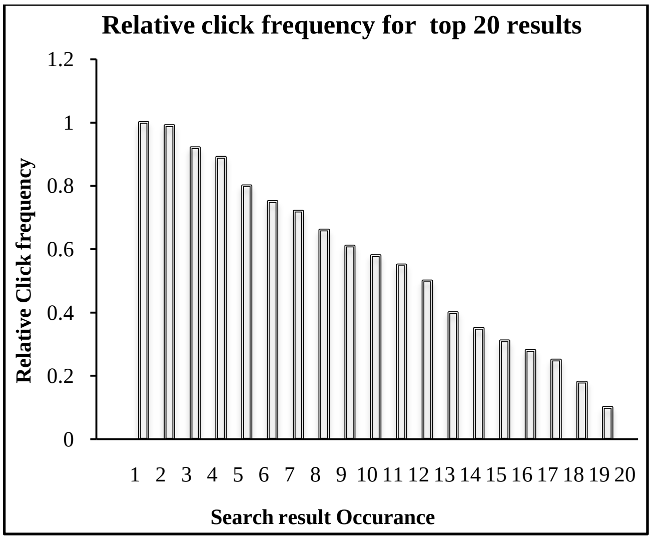

Fig.2. Relative click frequency for top 20 results

3.2 PERFORMANCE METRICS

The results of the search engine are taken into consideration and the performance of content analysis and user profile strategies are compared. For performance measure the standard precision, recall, and f-measure are computed as defined below:

$$Precision = \frac{|Q_{relevant} \cap Q_{retrieved}|}{|Q_{retrieved}|} \quad (13)$$

$$Recall = \frac{|Q_{relevant} \cap Q_{retrieved}|}{|Q_{relevant}|} \quad (14)$$

$$F-measure = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (15)$$

where, Q-relevant is the set of queries that are relevant to user and Q-retrieved is the total number of queries extracted.

3.3 RESULTS

The user profiles are engaged by the personalized search methods to group similar queries together according to user's needs. The user group method trace the user's similar searched contents and groups them. In content analysis method only the user clicked contents are taken and the concept relationship is made which finally produces the result. The two methods are employed to collect the relevant results. The results for the user query are captured and the results are processed. The best Precision, Recall and f-measure values for the User Profile techniques such as P_{click}, $P_{Joachims}$ and $P_{click+Joachims}$ methods are shown in Table.6.

Table.6. Best Precision, Recall and f-Measure Values for the User Profile Techniques

Techniques	Precision	Recall	f-measure
P_{click}	0.7626	0.6868	0.7227
$P_{Joachims}$	0.6517	0.6972	0.6736
$P_{click+Joachims}$	0.8437	0.8658	0.8546

Table.7. Comparison of Web Result without Personalization, Personalization of Content Analysis and User Group Method Using Bing Api and Google Api

	BING API			GOOGLE API		
	Precision	Recall	f-measure	Precision	Recall	f-measure
Web result without personalization	.675	.815	.702	.7150	.8231	.7652
Personalization using Content Analysis	.723	.751	.737	.7206	.7513	.7356
Personalization using User Group	.695	.657	.657	.8437	.8658	.8546

Table.8. Comparison of Personalization of Content Analysis and User Group Method Using Precision, Recall and f-Measure

Strategies	Precision	Recall	f-measure
Web result without personalization	.7150	.8231	.7652
Personalization using Content Analysis	.7206	.7513	.7356
Personalization using User Group	.8437	.8658	.8546

From Table.6, it is noted that click-based personalization performs well on repeated queries. In the user group techniques the combined use of P_{click} and $P_{joachims}$ provide good results and is clearly shown in Table.6. From this, it is noted that profile based strategies are stable because of easy implementation and it provides better result for the frequently used queries. It is clearly understood that content analysis method does not work well for personalization of search engine and Table.7 shows that the comparison of user profile methods provide better precision, recall and f-measure while using GOOGLE API. It is found that personalization based on user profile out performs web result without personalization and personalization using content analysis and it is clearly shown in Table.8.

4. CONCLUSION

In this paper, personalized search based on content analysis and personalized search based on user group method are compared. User Group method improves search engine performance by finding information needs for individuals. This method makes use of click through data. In the content analysis based search engine, the concept and concept relationship are extracted from the web snippets. From the experimentation it is inferred that, user Group method works well and the results for the search queries are re-ranked, thereby improving search accuracy. Compared to click based method, user group method is stable. To get standard search results, re-rank and display them is less and fast response is desirable. In continuation of the search process grouping/clustering of search results and query suggestion on related queries can be included.

REFERENCES

[1] E. Agichtein, E. Brill and S. Dumais, "Improving Web Search Ranking by Incorporating User Behaviour Information", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 19-26, 2006.

[2] A. P. Avinash, Prashant M. Narayankar, M. Muniraj and Jeevitha, "Clustering Method: Search Result Based Web Personalization", *International Conference on Advances in Computer and Electrical Engineering*, pp. 94-97, 2012.

[3] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton and G. Hullender, "Learning to Rank Using Gradient Descent", *Proceedings of the 22nd International Conference on Machine learning*, 2005.

[4] Charanjeet Dadiyala, Pragati Patil and Girish Agrawal, "Personalized Web Search", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 6, pp. 567-570, 2013.

[5] Z. Dou, R. Song and J. R. Wen, "A Large-scale Evaluation and Analysis of Personalized Search Strategies", *Proceedings of the 16th International Conference on World Wide Web*, pp. 581-590, 2007.

[6] Fang Liu, Clement Yu and Weiyi Meng, "Personalized web Search for improving Retrieval Effectiveness", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 1, pp. 28-40, 2004.

[7] S. Gauch, J. Chaffee and A. Pretschner, "Ontology-Based Personalized Search and Browsing", *Journal on Web Intelligence and Agent System*, Vol. 1, No. 3-4, pp. 219-234, 2003.

[8] M. Jansen, A. Spink, J. Bateman and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web", *Proceedings of the ACM SIGIR Forum*, Vol. 32, No. 1, pp. 5-17, 1998.

[9] J. Jayanthi and S. Rathi, "A Personalized search Framework for industrial Safety and health information Retrieval", *Journal of Scientific and Industrial Research*, Vol. 73, No. 6, pp. 407-414, 2014.

[10] T. Joachims, "Optimizing Search Engines Using Click through Data", *Proceedings of the ACM SIGKDD*, 2002.

[11] Kenneth Wai-Ting Leung and Dik Lun Lee, "Deriving Concept-Based User Profiles from Search Engine Logs", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 7, pp. 969-982, 2010.

[12] Kenneth Wai-Ting Leung, W. Ng and Dik Lun Lee, "Personalized Concept-Based Clustering of Search Engine Queries", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 11, pp. 1505-1518, 2008.

[13] F. Liu, C. Yu and W. Meng, "Personalized Web Search by Mapping User Queries to Categories", *Proceedings of the 11th International Conference on Information and Knowledge Management*, pp. 558-565, 2002.

[14] Mark Levene, "*An Introduction to Search Engines and Web Navigation*", John Wiley & Sons, 2011.

[15] X. Shen, B. Tan and C. Zhai, "Implicit User modelling for Personalized Search", *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 824-831, 2005.

[16] M. Speretta and S. Gauch, "Personalized Search Based on User Search Histories", *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 622-628, 2005.

[17] Susan Dumais, Eugene Agichtein, Robert Ragno and Eric Brill, "Learning User Interaction Models for Predicting Web Search Result Preferences", *ACM, SIGIR Conference on Research and Development on Information Retrieval*, pp. 3-10, 2006.

[18] Q. Tan, X. Chai, W. Ng and D. Lee, "Applying Co-training to Click-through Data for Search Engine Adaptation", *Proceedings of the Database Systems for Advanced Applications Conference*, 2004.

[19] J. Teevan, S. T. Dumais and E. Horvitz, "Beyond the Commons: Investigating the Value of Personalizing Web Search", *Proceedings of the Workshop New Technologies for Personalized Information Access*, 2005.

[20] J. R. Wen, J. Y. Nie and H. J. Zhang, "Query Clustering Using User Logs", *ACM Transactions on Information Systems*, Vol. 20, No. 1, pp. 59-81, 2002

[21] Y. Xu, K. Wang, B. Zhang and Z. Chen, "Privacy-Enhancing Personalized Web Search", *Proceedings of the World Wide Web Conference*, pp. 591-600, 2007.

[22] Zhicheng Dou, Ruihua Song, Ji-Rong Wen and Xiaojie Yuan "Evaluating the Effectiveness of Personalized Web Search", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 8, pp. 1178-1190, 2009.

[23] www.wikipedia.org/wiki/Content_analysis

[24] Open Directory Project, http://www.dmoz.org/, 2009.