ISSN:0975-9646

An Effective Personalization of User Search Behavior with Semantic Framework Algorithm

Dr.N.Sivakumar[1], S.Saranya[2]
[1]*Assistant Professor,* [2]*PG Scholar*
Department of Computer Science and Engineering
Pondicherry Engineering College
Puducherry, India

***Abstract*— In web search, there are many techniques used extensively for effective retrieval of information from the web server. A user query entered into the search engine may return large number of web page results and thus, it becomes extremely important to rank these results in such a manner that it returns accurate and more relevant web results. These tasks of prioritizing the results are performed by ranking algorithms. But the search interest of every user differs with every other user uniquely. In existing system, the search engine that return results for a given user query is not uniquely based on their earlier searching behavior. Hence there may be return of web page results that are least used or accessed by the user. These results need to be filtered or less prioritized by the ranking algorithm. In our proposed system, the personalization of user search activities is taken to find search interest of each user. This personalization of user search behavior is done by storing user activities and analyzing web log by clustering frequently accessed tasks by semantic task clustering algorithm and dynamically ranks web results. Thus it uniquely ranks web results.**

Keywords— Weblog, Task Clustering algorithm, Personalization, Dynamic Ranking algorithm.

I. INTRODUCTION

There are millions of users accessing web for getting information what they need. For this the user submits the query to web server through search engine and gets web results based on given query. Thus the search engine obtains large number of results for the given query. These results must be accurate and relevant to the given query and must be immediately accessible to the user. These web results retrieved from web server are mainly based on the given keywords of query and rank it accordingly. In our proposed system, along with the retrieved web results based on keywords, we analyze the weblog and dynamically rank the web results uniquely for each user.

There are many techniques used for ranking the web pages like Page ranking, indexing, search log analysis...Etc, are used for getting the optimized web results. Hence in web searches, Web search logs are used to record the searching activities of various users in search engines as in [5]. These logs are used to analyze the most accessed web pages or web sites by users, for immediate access of web pages. Hence, it undergoes the personalization of user search profile, because the searching behavior of each user varies uniquely. Initially in our system, the user has to create an account with their details of userId and password for access the search engine. Here we uses simple authentication of userId and password to maintain unique user's searched profiles in weblog. The weblog stores user searched queries and links through search engines. Along with it stores user information's like click through rates, dwell time as in [15], bounce rate, visitation rate, ip address of users with unique user hash value that is generated for each unique user. These metrics are later used to analyze the most or frequently accessed queries by user. Likewise each link information that is visited by many users is stored in web log with hash key value uniquely. The dynamic hashing algorithm generates hash key values for maintaining unique records of the every user in web log. These hash values are mainly used to avoid disambiguation and referencing problems between different users searched tasks and to maintain unique personalized profile for each user. From each unique personalized records of the user, the search interests are extracted by applying "Task Clustering" algorithm.

The clustering algorithm involves pattern and semantic clustering method for identify user search interests [3] between the query pairs of recorded weblog information's. Once the user given the query, the search engine goes to database for retrieve of web page results. After retrieving, it compares the retrieved results with weblog records of the particular user for the given query. This referencing of weblog records are done by unique hash key values generated for particular user. Then ranking algorithms is taken dynamically based on web logged information and rank the web results accordingly. Then the search engine returns unique web results for the particular user based on their previous searched activities recorded as in web log. These web page results will display with different rank priority to different users based on their previous searched activities. These method returns the results, what user wants exactly during their web search. The main objective of the study is how to identify user search interests accurately and rank the webpage's accordingly for each user uniquely accessing the web. These goals can be achieved by applying data mining technique to the web log that records each and every activities of user searched in the web. After ranking, the mostly accessed sites by users are identified and they are accessed by user.

II. RELATED WORK

The study of the related work in [3], [4] obtains the concept of task identification from the user search history and how to analyze in an effective way. The study of personalized ranking helps to find the way to create the framework for ranking the web results based on user search behavior in [1],[6]. The task identification method in [8] obtains a way to measure the rate of tasks performed by the user inorder to identify most accessed tasks by the users during their browsing session. The concepts in [5],[12] helps to study in case of how to maintain the user search history and web directories inorder to analyze it for user satisfactory results. The dwell time metric and other web based evaluation of user interests is obtained from [15] and to mine the user tasks based on several web based metrices in this paper. The various concepts of ranking methods are studied from the papers [10], [13] and that also tells a way to rank effectively to prioritize the web results. The idea of personalization of user search is studied in the previous work [18]. Thus previous studies clearly provides an idea to explore the personalize user search behavior effectively and dynamically rank web results.

A. Motivation

The proposed work deals with analyzing the web log activities by various users and rank dynamically based on their previous search interests recorded in web log. Initially we locally create an simple search engine that searched queries based on Computer domain or computer related queries. Thus the user initially searches queries through search engine. Then the search engine retrieves the search results based on user given queries patterns from the database. Then the search results will display in the web page. The user clicks a link and access the information. After, when user clicked the particular link, the information's like username, query for the clicked link, path of the link, ip address of user, date and time, dwell time, bounce rate, click rates, visitation rates of that page information are stored in the web log. Then the pattern and semantic task clustering algorithm is applied to cluster mostly searched queries. For further accuracy the evaluation metrics of queries is considered to identify frequently accessed tasks by user. Then the weight based ranking will be undergoes in order to rank the web page results. The ranking is done dynamically for each user through dynamic mapping using hash function. Based on user searched interests the ranked results will be obtained to user.

III. PROPOSED SYSTEM

In this section, it describes how to apply the proposed task clustering algorithm and dynamically rank the web results based on analyzed user behavior from the weblog

A. *Storing Weblog Metrics*

The weblog stores the information like query, path of link clicked, ip address of user, dwell time, bounce rate, visitation rate, clicks rates, when user clicks the particular link of web results page. This information are stored with weightage values and considered for ranking the web Pages.

B. *User Segmentation*

The entire document should be in Times New Roman or The user segmentation is done by applying unique hash key values for each user in the web log to avoid referencing problem, ambiguity and duplication of records. The hash key is generated for each user actions, and stores unique information like user id, visitation rate, click rates, dwell time, bounce rate, ip address, date and time of access. Based on these information user searched profiles is segmented. The table 1 shows the information that will be storing in the weblog dynamically, when user clicks the particular links from web results page. Hence it stores the information of searched queries by the user, path of the URL (Uniform Resource Locator), dwell time or stay time of user, bounce rate, visitation rate of user to that web page and the total number of click rates made by the user to that webpage are stored in weblog.

C. *Task Clustering with Semantic Framework Algorithm*

The semantic and pattern based task clustering algorithm is applied in web logged table as shown in Table 1 of user profile in order to analyze most accessed tasks by users. The clustering is based on pattern and semantic based relevancy mechanism called pattern and semantic task clustering Algorithm. The semantic clustering is done by using WordArt tool that contains lakhs of semantic meanings and words. The semantic meanings of queries in annotated with the tool and it returns the semantically related words for the query.

Thus it undergoes four processes as follows.

1. Initially user searched query is divided into number of tokens.
2. Tagging part of speech to check syntactic roles (subject, object) and functional roles (noun, verb, pronoun, adverb) using parser.
3. Stemming is done to find root words from tokens.
4. Then check the similarity between queries and measure that similarity and relevancy between numbers of user searched queries semantically with WorldNet in order to analyze the most accessed tasks by user.

The Table 2 shows the clustered queries based on tasks similarity. The pattern clustering is done by matching patterns between the queries phrases and semantically related queries then it will be clustered and it obtains mostly accessed queries. The mostly accessed queries are analysed with their weightage value from the web log.

Table 1 Weblog Table

Queries	Url	Ip address	Dwell Time	Date and Time	Bounce Rate	Visitation Rate
What are the uses of computer	http://www.byte-notes.com/uses-computers-various-fields	101.1.1.8	243 sec	Tuesday 14/10/2014-4.58pm	No	5
What are the uses of computer	http://ecomputernotes.com/fundamental/introduction-to-computer/uses-of-computer	101.1.1.8	143 sec	Tuesday 14/10/2014-5.38pm	No	6
Mahatma Gandhi history	http://en.wikipedia.org/wiki/Mahatma_Gandhi	192.0.1.3	19 sec	Wednesday 15/10/2014-3.08pm	Yes	7
Head news of today	http://timesofindia.indiatimes.com/home/headlines	101.1.2.3	367 sec	Thursday 16/10/2014-11.09am	No	8
What are the uses of computer	http://www.byte-notes.com/uses-computers-various-fields	101.1.1.8	243 sec	Thursday 14/10/2014-4.58pm	No	9

The following algorithm clusters the most accessed tasks by user. These frequently accessed tasks are calculated with weight values and these values are considered for rank web results. The rate of total task performed by the user based on searched queries can be calculated by the task rate as follows,

Algorithm

Input: Query set Q (Session), cut-off threshold b, length N;
Output: A set of tasks S;
Initialization: S = ∅; cid: content task id
Query to task table L=∅, M=∅;
1: // Initialize queries that are same into one task
2: cid=0;
3: for i = 1: |Q|−N do
4: if M[Qi] exists then
5: add Qi into S(M[Qi]);
6: else
7: M[Qi]=cid++;
8: if |S| = 1 return S;
9: for len = 1 : n N do
10: for i = 1: |Q|−N do
11: //It undergoes clustering based on semantic and pattern based relevancy
11: // if two queries are not in the same task
12: if L[Qi]≠ L[Qi+N] then
13: // compute similarity takes T
14: T ← sim (L [Qi], L [Qi+N]);
15: if T ≥ b then
16: merge S (Qi) and S (Qi+N);
17: modify L;
18: // break if there is only one task
19: if |S| = 1 break;
20: return S;

The above algorithm finds the similarity between two queries. If two queries belongs to same task those are inserted it into the task table. If similarity measure of tasks is greater than the threshold values, it is added to queries of same tasks otherwise ignore the query. This algorithm is mainly used to accurately cluster the queries under same tasks, different and the irrelevant tasks as shown in the Table 2.

Table 2 Clustering queries based on Tasks

Same Tasks	Different Tasks	Unknown Tasks
73 queries	48 queries	10 queries

Task Rate: It can be identified by measuring similarity values between queries based on semantic and pattern based relevancy. This similarity measures between two query pairs is measured interms of Log Likelihood Ratio (LLR). It is used to measure the correlation between query phrases, which can ease the problem of generating irrelevant query pairs with high co-occurrence between the query pairs. This LLR can be determined by the following formulas.

Given two queries q1 and q2, LLR makes,

Null hypothesis
H1: Pr (q2|q1) = Pr (q2|¬q1) and

Alternative hypothesis
H2: Pr (q2|q1) ≠ Pr (q2|¬q1)

Where Pr refers to probability.

The LLR is calculated as λ, where

$$\lambda = max_{p} L(H1) / max_{p1,p2} L(H2) \qquad (1)$$

A higher LLR indicates a higher correlation between query pairs, For example, if LLR=3.84 indicates 95% confidence for rejecting H1. The following table shows overall sum of visitation rate, bounce rate, dwell time, task rate for the user searched links. These values are used to evaluate the user behavior for the particular link or URL. It helps to identify the priority of the web pages while dynamically rank the web page results. And these values are uniquely stored for each user and maintained uniquely using hask key values,

These metric values are evaluated for all the user searched queries and the links in the web as shown below. Then these values are considered for weight based ranking algorithm that taken dynamically while analyzing the weblog. The ranking or prioritizing the webpages is based on both these web page metrics and the also the task extracted after applying the task clustering algorithm in the weblog.

Dynamic Ranking

The page ranking algorithm is used to rank the web result pages based on prioritizing of the web pages. Thus here the priority of web pages is calculated by weightage values of number of visitors in the webpage, click rates in the web page, dwell time and frequently accessed queries obtained by the task clustering algorithm and the bounce rate. The hash key value is generated for each record of user query to uniquely identify the user search interest in the web. The dwell time records the staying time of user in particular web page, bounce rate represent session timeout of user in webpage, and visitation rate provided the number of visits by user and click rates represents the click made by the user on the links of the webpage or websites. Based on these values the resultant web page results are prioritized and ranked dynamically based on analyzed web log through clustering algorithm and web page metrices that helps to obtains unique web page results for each user. These unique web results are determined by the dynamic ranking with the following parameters as shown in algorithm

Algorithm

1: //PR->Page ranking of web page, p->webpage/website
2: //V(p) ->Number of visitors of webpage, W(p)->Weight value of webpage
3: //C(p)->Total Click Rates of webpage, D(p)->Dwell Time of webpage
4: //q->Queries, M (q) ->Most accessed queries, k->Hash key for unique results.
5://B(p)->Bounce Rate ,R->Ranked results.
6://Initially we calculate number of Visitors, Click Rates, Dwell Time, Frequently accessed queries and bounce Rate together to get weight of each web Page.
8://we gets this value uniquely for each user u using hask K.
9:W(p, u, k)->get([V(p).C(p).D(p).M(p)/B(p)],u,k)
10://Then we prioritize pages by compare with each other
11: Prioritize (p, u, k) ->Max ([compare (W_{F1},W_{F2},W_{F3},W_{F4}…W_{FN})], u, k)
12://Store it in ranked result of user
13: PR (p, u, k) <-Prioritize (p, u, k)
14: R<-PR (p, u, k)
15: Return R.

Table 3 Overall Evaluation of metrices for User Queries

Queries	Url	Ip Address	Visitation Rate	Bounce Rate	Dwell Time	Task Rate
What are the uses of computer	http://www.byte-notes.com/uses-computers-various-fields	101.1.1.8	99.1	22.5	92.6	87.1
Mahatma Gandhi history	http://en.wikipedia.org/wiki/Mahatma_Gandhi	192.1.2.4	89.8	34.1	86.3	72
Head news of today	http://timesofindia.indiatimes.com/home/headlines	101.1.2.3	77.3	43.2	69.2	88.9
Sports news	http://www.google.co.in/url?ved=0CBwQFjAA&url=http%3A%2F%2Ftimesofindia.indiatimes.com	192.16.3.2	76.6	55.6	62.3	93.1
Mahatma Gandhi	https://www.google.co.in/?gfe_rd=cr&ei=RZVAVKXHMaHV	192.1.2.4	72.2	69	55.2	68

Table 4 Evaluating the effectiveness of Proposed System

Methods	Task Rate	Accuracy	Precision	Efficiency	User Interest Analysis	Rate of Relevancy Measure	Recall
TSDP (Task based Session Discovery Problem)	77.1	75.20	64.23	70.2	67.89	78.13	51.36
Personalized Reranking Algorithm(PRA)	65.34	66.34	59.23	69.56	77.12	68.34	41.89
Task Trail algorithm(TTA)	78.62	76.10	66.38	55.69	79.34	73.45	32.99
Topic preference vector(TPV)	76.34	77.5	55.1	67.54	66.98	80.45	41.11
Ranking Model Adaptation Framework(RMAF)	69.40	65.98	48.34	63.34	64.75	61.09	50.90
Task Clustering algorithm with semantic Framework (TSCP-Proposed).	85.22	81.13	73.45	75.67	82.23	84.33	28.56

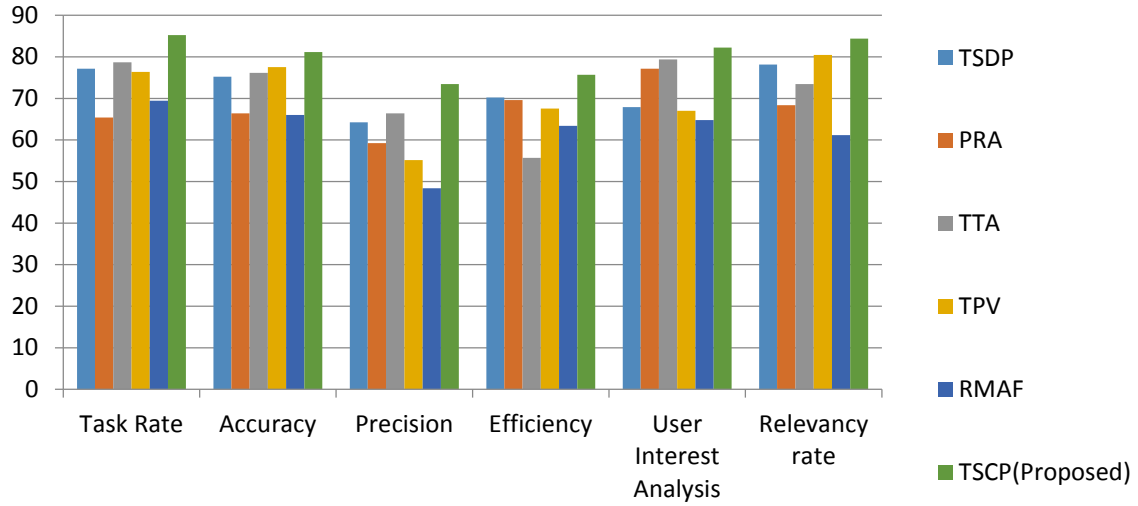

Figure 1 Performance Evaluation Chart

In this algorithm, initially the weight values of each webpage p is calculated by the metrics values of dwell time D(p), visitation rate V(p), mostly accessed queries M(p), bounce rate B(p), click rates C(p). Based on these weight values of query links, the web page links are ranked dynamically for each user uniquely. Then it returns the ranked web page results based on user searched behavior in the weblog. The dynamic ranking algorithm is used along with the hask key referencing function that is generated with each user activities in their profiles. These hash key function is mainly used to avoid disambiguation between while the dynamically rank the web pages results for each unique users that each ranking results of user varies with the ranking results of other user.

IV. Result Interpretation

The web page returning results to the web user must be absolute and needs to be ranked in an optimized way inorder to enhance the efficiency of the searching process by user in the search engine. It undergoes the process of storing the user searching activities in weblog, analyze it with task clustering algorithm and dynamically rank the web results through weight based ranking algorithm. The final results of this paper obtains the dynamically ranked web page results based on each user searched behavior as stored and analyzed from web log, and ranks the results uniquely for each user based on their searched behavior. The personalization of user search activities is taken to find search interest of each user. This personalization of user search behavior is done by storing user activities and analyzing web log by clustering frequently accessed tasks and dynamically ranks the web results. Thus it uniquely ranks web results based on user searched behavior. The following table 4 compares various parameters of proposed system with the previous work and obtains the below shown values incase of system efficiency. By comparing the values of the proposed with the existing work, it is obtained that the Task Clustering Algorithm with Semantic Framework and dynamic ranking method is foremost efficient with the previous methods of analyzing the web log data and ranking the web results accurately for the users based on their search interest or search behavior in the web during their web search. The final results of this paper obtains the dynamically ranked web page results based on each user searched behavior that is stored and analyzed from web log, and the proposed system ranks the results uniquely for each user based on their searched behavior.

V. Conclusion

It is concluded that the user search behavior of each webpage and websites are analyzed based on metrics like dwell time, visitation rate, bounce rate and click through rates along with frequently accessed tasks by the user. These metrics are necessary for analyzing the web page importance and its analyzing process. These analyzes of metrics is mainly based on user searched activities in the particular web page or website that is recorded in web log. Each user profiles are undergoes a mechanism of personalization that retrieves the web page results with unique hash key values. These hash key values helps to retrieve the contents uniquely from the web log datastore without any disambiguation. The records of each user activities are analyzed separately and uniquely by applying clustering method. The clustering of query is done by query similarity and pattern similarity from the user searched records. This information is mainly used to identify mostly searched queries and useful pages that user visited during web search. Then dynamically ranking algorithm is applied on the analysed queries from weblog and rank accordingly to the user. The final result display unique web results for unique users based on their searched behavior.

REFERENCES

[1] Wang, H, Song, Y, Chang, M.-W, He, X, White, R. and Chu W, "Enhancing Personalized Search by Mining and Modeling Task Behavior," in WWW, 2013.

[2] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", *IEEE Transactions on Knowledge & Data Engineering*, vol.25, no. 3, pp. 502-513, March 2013.

[3] Zhen Liao, Yang Song, Yalou Huang, Li-wei He, Qi He, "Task Trail: An Effective Segmentation of User Search Behavior", *IEEE Transactions on Knowledge & Data Engineering*, vol.26, no. 12, pp.3090-3192, December 2014.

[4] Liao, Z., Song, Y., He, L.-w. and Huang, Y., "Evaluating the effectiveness of search task trails," ser. WWW '12, 2012.

[5] Heasoo Hwang, Hady W. Lauw, Lise Getoor, Alexandros Ntoulas, "Organizing User Search Histories", *IEEE Transactions on Knowledge & Data Engineering*, vol.24, no. 5, pp. 912-925, May 2012

[6] Hongning Wang, Xiaodong He, Ming-Wei Chang, Yang Song, Ryen White, and Wei Chu, "Personalized Ranking Model Adaptation for Web Search", in ACM, vol.25, 2013.

[7] Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G., "Identifying task-based sessions in search engine query logs," *in WebSearch and Data Mining (WSDM)*, 2011.

[8] Ahmed Hassan, Yang Song, Li-wei He, "A Task Level Metric for Measuring Web Search Satisfaction and its Application on Improving Relevance Estimation", in ACM, 2011.

[9] Rekha Jain, Sulochana Nathawat, Dr. G.N. Purohit, "Enhanced Retrieval of Web Pages using Improved Page Rank Algorithm", in *International Journal on Natural Language Computing (IJNLC)* Vol. 2, No.2, April 2013.

[10] Neelam Tyagi, Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", in *International Journal of Soft Computing and Engineering (IJSCE)* vol-25, pp: 2231-2307, no-3, July 2012.

[11] Athanasios Papagelis and Christos Zaroliagis, "A Collaborative Decentralized Approach to Web Search," *IEEE Transactions on Knowledge & Data Engineering,* vol.42, no.5, September 2012.

[12] Dimitrios Pierrakos, Member, IEEE, and Georgios Paliouras, "Personalizing Web Directories with the Aid of Web Usage Data", *IEEE Transactions on Knowledge & Data Engineering,* vol. 22, no. 9, pp: 1331-1344, September 2010.

[13] Christina Brandt, Thorsten Joachims, Yisong Yue, "Dynamic Ranked Retrieval", in *ACM Conference on Web Search and Data Mining (WSDM)*, February 2011.

[14] Mirco Speretta, Susan Gauch, "Personalizing Search Based on User Search Histories," in *ACM Conference on Electrical engineering and Computer,* vol.25, 2004.

[15] Songhua Xu, Hao Jiang, Francis C.M. Lau, "Mining User Dwell Time for Personalized Web Search Re-Ranking," *in Twenty-Second International Joint Conference on Artificial Intelligence* (IJCAI), 2012.

[16] 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, August 2009.

[17] Tianyi Jiang and Alexander Tuzhilin, "Improving Personalization Solutions through Optimal Segmentation of Customer Bases," *IEEE Transactions On Knowledge And Data Engineering*, vol. 21, No. 3, March 2009.

[18] Fang Liu, Clement Yu, Senior Member, IEEE, and Weiyi Meng, Member, IEEE, "Personalized Web Search for Improving Retrieval Effectiveness", *IEEE Transactions On Knowledge And Data Engineering*, vol. 16, No. 1, January 2004.