User Intent Discovery using Analysis of Browsing History

Wael K. Abdallah
Information Systems Dept
Computers & Information Faculty
Mansoura University
Mansoura, Egypt

Dr. / Aziza S. Asem
Information Systems Dept
Computers & Information Faculty
Mansoura University
Mansoura, Egypt

Prof. /Dr. / Mohammed Badr Senousy
Computer and Information Systems Dept
Sadat Academy for Management Sciences
Cairo, Egypt

***Abstract*—The search engine can retrieve the information from the web by using keyword queries. The responsibility of search engines is getting the relevant results that met with users' search intents. Nowadays, all search engines provide search log of the user (queries logs, click information besides browsing history). The main objective of this work is to provide features that can help users during their web search by categorizing related browsing URLs together. That will be done by identifying intent groups for each URLs category, then identifying intent-segments for each intent group. Upon clustering the query categories, groups, and intent segments search engines can improve the representation of users' search context behind the current query, this would help search engines to discover the user's intents during the web search. Through the use of the normalized discounted cumulative gain (NDCG), the experimental results show the proposed method can improve the performance of the search engine.**

Keywords—component; Information Retrieval; Search Engines; Users' Search Intents; Search Log and Browsing History

I. INTRODUCTION

With the growing of World Wide Web, web search engines have added a big value in web searching. Search engines can find what user search for on the web quickly and easily.

Users issue a query Q and a search engine returns a ranked list of URLs retrieved from the indexed collection of web pages. Developing reliable ranking techniques may be not easy because user search goals are dynamic and depend on their search intents. It is difficult to web search engines to know what the users exactly need. [1]

While searching from the web, users need results based on their interest. For the same keyword two users might require different pieces of information. For a query, a number of documents on different topics are returned by search engines. Hence, it becomes difficult for the user to get the relevant result. Moreover, it is also time consuming. Personalized web search is considered as a promising solution to handle these issues, since different search results can be provided depending upon the information needs of users. It exploits user information and search context to learn in which sense a query refer. [2].

The most important sources help to extract user preferences (i.e. query logs, search engine result page clicks, as well as browsing behavior). Many processes can be done on browsing data, so information extracted from it would become more useful [3].

The proposed method studies the user profiles based on the logs, historical from browsers and clicks. This paper presents a method for the personalization using features based on intentions. It uses resources (browsing history) to categorize URLs in order to extract groups and segments of intents.

There were many studies about the query log that interact with a search is a continuous process to provide a valuable information about the context of the query. So consideration of a given query as a part of large search process could significantly improve the performance of the search engine [3].

But the proposed method studies the browsing history of users to understand how and when information need of a user changes, and it can be very helpful in cold start problem that comes when a new user/query or both just enters the system[4]. As opposed to the previous studies, the proposed method deals not only with search engine result pages but with the whole browsing logs.

The topics categories are used to classify the browsed URLs are based on the most general categories of the Open Directory Project and Alchemy taxonomies like [2]. The second step is grouping each category into intent groups as in [5]. Then for each group, intent segmentation was used like [3]. This categorizing and grouping and intent segmentation information can improve the representation of users' search context behind the current query, this would help search engines to discover the user's intents during the web search and re rank the search results that allow users to find what are they want in the top search results.

Paper organized as the following: the first section discusses the related work. The second section defines the research problem. The third section presents the proposed method. The Fourth section tests the proposed method on data set and evaluates the results. And the last section presents the conclusions and the future work.

II. Related Work

"Personalization is the process of presenting the right information to the right user at the right time." To create the user profile (user context), it needs to collect and analyze user's personal information. User Profile information can be collected from users in two ways: explicitly, i.e. feedbacks; or implicitly, i.e. from user's browsing behavior. The user profile can be presented in the user's preferences and user's interests. Usually, there are three types of a user profile:1) Content-based profile (i.e. terms), 2) Collaborative profile (i.e. shared similar interest/preference between users' groups) and 3) Rule-based profile: first, users answering the questions about their usage of information. Second, rules are extracted from theses answers [4].

Filip and Nicolaas in [6] used complete browsing behavior to build a user interest profile, and then this model was used to re-rank search results.

Ruofan W., Shan J. and Yan Z In [7] proposed a re-ranking method by used semantic similarity to enhance the quality of search results. In the experiment, they used NDCG to evaluate the re-ranking results. The NDCG was used to evaluate our re-ranking results

Fedor et al, in [8] showed how to interact the short-term behavior and long-term behavior (in isolation or combination) to improve the relevance of search results through search personalization

Daxin J., Hang L.and Jian P., in [9] presented a survey that discussed the mining of the search log and the browsing data to improve the search engine components.

Pavel and Yury in [10] tried to solve the problem of the not existence of the search context by using the short-term browsing.

Anna M., Pavel S. and Yury U., in [3] proposed a technique for automatic segmentation of users' daily browsing activity into intent-related segments. In this paper, the proposed method will use the intent segmentation as a part of intent clustering (besides intent categories and intent groups) of browsing history to understand and discover the user intent during the web search.

Aditi Sharan and R. Kumar In [2] built a framework of an Enhanced User Profile by combining the user's browsing history and the domain knowledge to improve personalized web search. In this paper, the proposed method will use the intent categories (besides intent groups and intent segments) as a part of intent clustering of browsing history to understand and discover the user intent during the web search. Also in the proposed work, re-ranking the search engine results according to users' discovered intents instead of used the Enhanced User Profile (DMOZ Directory Domain Knowledge) for suggesting relevant pages to the user in this previous work.

Veningston and Shanmugalakshmi In [5] proposed grouping of the search query to allow to the search engine to personalize the search results according to user's interests. This paper will use the intent groups for browsed URls (besides intent categories and intent segments) as a part of our intent clustering of browsing history to understand and discover the user intent during the web search.

III. Problem Definition

First, it is important to provide some definitions: Definition. The browsing log is the recorded daily activity of a user in the browser. The browsing log composes of URLs of visited pages [3].

Definition. Browsing segments (or logical segments) is a subset of the browsing log, consisting of intent-related pages, i.e. pages visited with the same or similar search goal. [3]

Definition. Query logical session is a subset of queries, unified into one search goal (=intent). [3].

The topics categories are used to classify the browsed URLs are based on the most general categories of the Open Directory Project and Alchemy taxonomies. The second step is grouping each category into intent groups. Then for each group, intent segmentation was used as in [3] to obtain a partition of pages visited by a user into intent related goal. Next time when the user issues a query, retrieval process may take place incorporating URL category and its group and its intent segment information in addition to the current query in order to understand and discover the user's intent during the web search. This categorizing and grouping and intent segmentation information can also act to gather a user profile from browsing history which includes users search intents and interests. Thus, web search could be personalized to promote efficient web search.

IV. Method

Once all the browsed URLs are classified into predefined intent categories, groups, and segments, then the search result ranking of a user is found out as showed in fig. 1. This is done by creating a database which contains the browsed URLs and its intent classifications (categories, groups, and segments using intent classification tools such as DMOZ and Alchemy taxonomies and Page Analyzer tools). It also includes the queries and its top results from search engines (and its intent category, intent group, and intent segment). The clustering of browsed URLs is determined during the clustering phase as training data, and the clustering of top URLs is determined during the clustering phase as testing data. The URLs will be updated in the DB. Then the re-ranking of the top search results of a user.

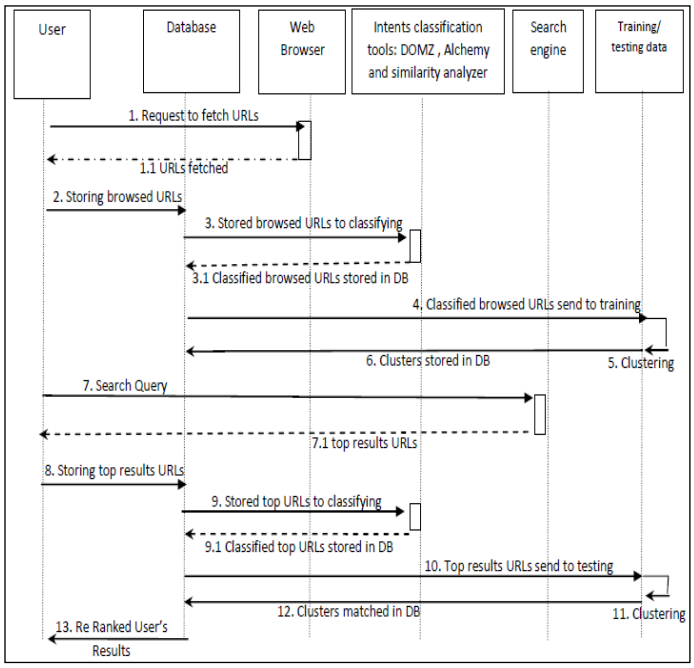

Fig. 1. Process sequence diagram

A. Intent Topic Categorizing, Intent Grouping and Intent Segmentation

a) The Intent Topic Categories : The topic categories used to classify the browsed URLs are based on the most general categories of the Open Directory Project and Alchemy taxonomies. The $DOMZ_1$ Search engine is used to classify the browsed URLs. Also, Alchemy API_2 is used for classifying web pages into particular category after mapping Alchemy API taxonomies to DMOZ Categories.

TABLE I. MAPPING ALCHEMY API TAXONOMIES TO DMOZ CATEGORIES

DMOZ Categories	Alchemy Categories
Art	Arts & Entertainment, Style & Fashion
Education	Education
Home	Family &Parenting, Home & Garden, Pets
Society	Law & Crime , Govt & Politics, Culture, Religion &Spirituality
Business	Business &Industrial, Finance, Real Estate, Careers
Games	Gaming
News	News And Weather
Sciences	Science & Technology
Sports	Sports
References	References
Computers	Technology & Computing
Health	Health & Fitness
Recreation	Automotive & Vehicles, Food & Drink, Travel
Shopping	Shopping
Kids and Teens	Hobbies And Interests

b) The Intent Groups: For each previous category, the browsed URLs were classified into intent groups manually with the assistance of DOMZ search engine and Alchemy API. For confirmation, the HTML code of the respective URL's is extracted and crawled to get the content keywords of the page and test theses keywords against the intent groups.

c) The Intent Segmentation: Split browsing logs into logical segments manually. During visited pages, search intent segment was assigned to each page. To determine the intent of each page, it is allowed to look through several pages visited after the current one. Then take a collection of pairs (d1, d2) of pages visited by one user and manually assign them segments labels S(d1, d2,Du) ∈ {0, 1}, choosing 1 if they belong to the same segment and 0 otherwise. With the assistance of similarity-analyzer $tool_3$ that basically implement the similarity features in 1) HTML code similarity URL Features measuring similarity of URLs and 2)Text similarity textual Features measuring similarity of texts. Similarity-analyzer tool measure the similarity between web pages by giving a similarity score. If the similarity value is above the specified threshold level then only these will be considered belong to one segment.

V. EXPERIMENT AND EVALUATION

A. Experimental Setup

Standard datasets for this research problem are not existent so, the dataset had been designed. In this Experiment, the Lemur toolbar and Google history are used to record the browsing history of the users (researchers in information system filed). Our Experiment is conducted for browsing history for one month for each user. The DOMZ Search engine and Alchemy API are used to cluster the browsed URLs based on the most mapping general categories of the Open Directory Project and API taxonomies. Then, clustering a set of related documents to its group in each category such as art, games, society and so on. Then, the page analyzer is used to determine the intent segment (of related documents) for each group. At last, the browsed URLs had crawled and indexed in each cluster using $dtsearch\ engine_4$. Then, keywords have been mined from the crawled web pages. Then calculate the frequency of a specific term in a specific cluster = the number of times that specific term is presented in that specific cluster.

B. Clustering

The input will be a set of URL's from user browsing history as shown in Table 2. The browsed URLs are present in a text file as the training data. The clustering algorithm is applied to it to cluster the input. By the use of $Weka_5$, the farthest first algorithm [11] is used for clustering the user's intents. A database is created with fields URL, the category field; its group filed and its segment filed. For each query, the URLs from top five search results is saved in a separate file as test data which is to be tested against predefined clusters of our clustering algorithm in Weka. The last step is the re-ranking of the top search results to improve the web search. The next section will show the clustering results of one user from the dataset as a sample.

C. Clusters

Cluster 0: Computers Web Mining S9, Cluster1: News World News S1, Cluster 2: Sports Football S2, Cluster 3: Education University S4, Cluster 4: Recreation Food S5,

1. https://www.dmoz.org/.
2. http://www.alchemyapi.com/products/demo/alchemylanguage.
3. http://tool.motoricerca.info/similarity-analyzer.phtml.
4. https://www.dtsearch.com/
5. http://www.cs.waikato.ac.nz/ml/weka/

Cluster 5: Science Academic Database S7, Cluster 6: Arts Movies S10, Cluster 7: Health Hypertension S11, Cluster 8: Computers Computer Journals S3, Cluster 9:Computers Html Programming Language, S8, Cluster 10: Health Kidney Disease S12, Cluster 11: Computers Software S13, Cluster 12: Arts Music S14, Cluster 13: Recreation Cars S15 and Cluster 14: Recreation Food S6.

D. Clusters Keywords

Now a term cluster matrix can be formed, which specifies by the frequency of the term in that cluster = the number of times the term t is present in the cluster.

TABLE IV. TERMS - CLUSTER MATRIX (TCM)

Term	Cluster0	Cluster	Cluster2	…..	Cluster
t1	C_{10}	C_{11}	C_{12}	…..	C_{1n}
t2	C_{20}	C_{21}	C_{22}	…..	C_{2n}
t3	C_{30}	C_{31}	C_{32}	…..	C_{3n}
…..	…..	…..	…..	…..	…..
t m	C_{m0}	C_{m1}	C_{m2}	…..	C_{mn}

E. Cluster Evaluation

To evaluate the clustering analysis using weka, it can record the recall and precision measures. Precision "is the ratio of the number of documents retrieved that "should" have been retrieved" [12].

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (1)$$

[12].

Recall "is the ratio of the number of relevant documents retrieved to the number of relevant documents" [12].

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

(2) [12].

TABLE II. CLUSTERING EVALUATION

CLUSTER	TP RATE	FP RATE	PRE-CISIO	RECALL	F-MEASURE	ROC AREA
C0	1	0	1	1	1	1
C1	1	0	1	1	1	1
C2	1	0	1	1	1	0.75
C3	1	0	1	1	1	0.75
C4	1	0	1	1	1	0.667
C5	1	0	1	1	1	0.75
C6	1	0	1	1	1	1
C7	1	0	1	1	1	1
C8	0	0	0	0	0	0.5
C9	0	0	0	0	0	0.5
C10	0	0	0	0	0	0.5
C11	1	0	1	1	1	1
C12	1	0	1	1	1	0.75
C13	1	0	1	1	1	0.75
C14	0	0	0	0	0	?

F. Rand Index

In order to measure the quality of clustering, Rand Index is used. Rand Index is determined as the accuracy of cluster formation. It is a measure of the similarity between two clusters. It is assumed that the two different clusters consist of the same number of data. In order to calculate the Rand Index shown in equation (3), it has to compare pairs as shown in Table 3.

TABLE III. POSSIBLE PAIRS TO COMPUTE RAND INDEX [5]

	Pairs assigned to the same cluster *(C1)*	**Pairs assigned to the different cluster *(C1)***
Pairs assigned to the same cluster *(C2)*	A	b
Pairs assigned to the different cluster *(C2)*	C	d

Count the number of pairs that fall into each of these four options a, b, c & d. C1 & C2 are the two clusters. The four options are expressed in the form of a table. In total there are possible pairs a+b+c+d= $[\binom{n}{2}]$ of n data points. Once a, b, c & d are identified, the Rand Index is computed as follows;

$$\text{RandIndex} = \frac{(a+b)}{(a+b+c+d)} \quad (3)\ [5]$$

Where a+b is assumed as the number of agreements between C1 & C2 and c+d as the number of disagreements between C1 & C2.

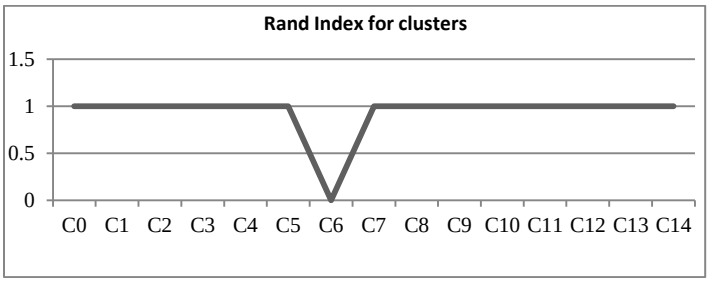

Fig. 2. Rand Index for clusters

Fig. 2 presents Rand Index for the clusters of the proposed method. It was noticed that the Rand Index for all clusters except cluster 6 is one because each cluster contains a few numbers of browsed URLs because clustering depend on many factors; intent categories, intent groups and intent segments. Rand Index of C6=0 because it contains only one URL.

G. Rand Index Comparison

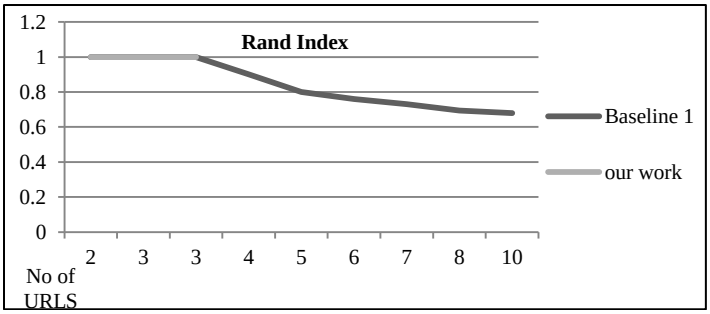

Fig. 3. Rand Index comparison between baseline 1 and the proposed method

Fig. 3 presents the Rand Index comparison between baseline 1 and the proposed method. Work [5] was used as baseline 1. It was noticed that the Rand Index of baseline 1 decrease when the number of browsed URLs increased. But in

the proposed method Rand Index is stable because each cluster contains a few numbers of browsed URLs because clustering depends on many factors; intent categories then intent groups and finally intent segments.

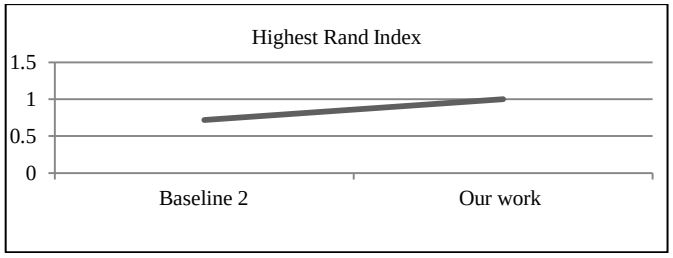

Fig. 4. Highest Rand Index comparison between baseline 2 and the proposed method

Fig. 4 presents the highest Rand Index comparisons between baseline 2 and the proposed method. Work [3] was used as baseline 2. Its highest value is 0.72. But in the proposed method the Rand Index is one because each cluster contains a few numbers of browsed URLs because clustering depends on many factors; intent categories then intent groups and finally intent segments.

VI. Results Analysis

For each query, the top 5 relevant search results provided by Google were collected (many experiments with different accessed times range from August 2015 to February 2016). Then classify them as intent categories, intent groups, and intent segments as discussed before. Consider them as test data for our clustering algorithm to decide whether theses top 5 results belong to clusters or not. If the one or more of top results belong to clusters then increase their rank positions to the top else display the original results, but if theses top results lately browsed by the user then add them to browsing history data set and re run the clustering algorithms and its followed steps. The analysis of the result is done by discovering the top results for each query belong to each cluster.

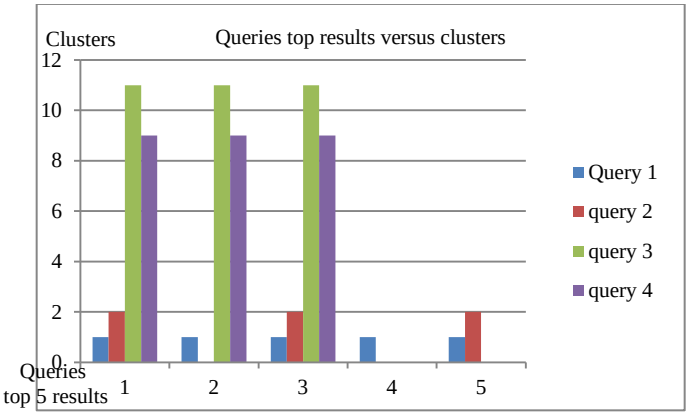

Fig. 5. Queries top 5 results versus clusters

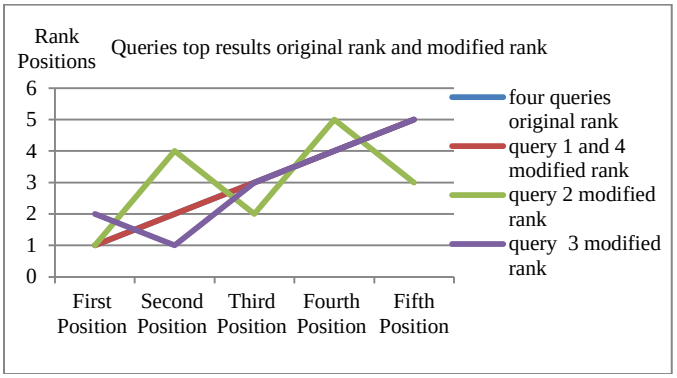

Fig. 6. Queries top results original rank and modified rank

Fig. 5 presents Queries top 5 results versus clusters for the four queries. And Figure 6 presents Queries top results original rank and modified rank for the four queries. For the first query and the fourth query, the search engine should keep the original ranking because the top results match with clusters in their same rank. For the second query and the third query, the search engine should modify the ranking of the top results as showed in the figure 6.

Search Relevance: NDCG Calculation

NDCG is an effective measure mainly used in information retrieval research to evaluate rankings of search documents according to their relevance. It measures how a ranking algorithm is in assigning the proper ranking to relevant documents. For example, if there are three web pages d1, d2, d3 whose relevance scores are (3, 2, 1) respectively (the higher score, the relevant), then the ranking of (d1, d2, d3) will achieve a higher NDCG value than the ranking of (d3, d2, d1). [7]. it can compute NDCG the Normalized Discounted Cumulative Gain of each rank p using the following formula:

$$\mathrm{N\,DCGp} = \frac{\mathrm{DCGp}}{\mathrm{I\,DCG}} \qquad (4)\ [7]$$

Where IDCG is Ideal Discounted Cumulative Gain calculated when get the search results. it has the best rank. And calculate the order of query of DCG.

And DCG is Discounted Cumulative Gain

$$\mathrm{DCGp} = \sum_{i=1}^{P} \frac{2^{reli}- 1}{\log 2\,(i + 1)} \qquad (5)\ [7]$$

Where p is PageRank serial number and reli is the graded relevance of the result at position i. For simplicity, suppose that on a five-point scale, 0 score given for an irrelevant result, 0 for a partially relevant, 1 for relevant, 1 for relevant again and 2 for perfect according to the percentage of the traffic by Google results positions study[6].

6. http://searchenginewatch.com/sew/study/2276184/no-1-position-in-google-gets-33-of-search- traffic-study

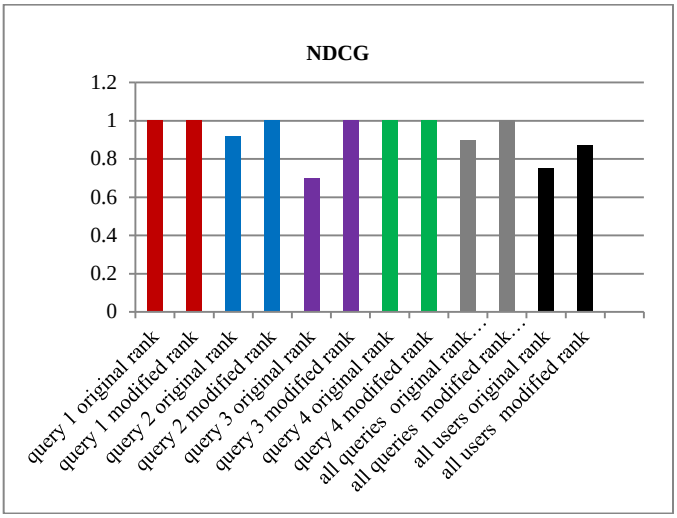

Fig. 7. NDGC for the queries for user 1 and all users

Fig. 7 presents the NDGC for the 4 Queries for original results and after modifying the rank. It was noticed that NDGC increased for queries 2 and 3 after modified the ranked. It stills the same for queries 1 and 4. Then calculate the overall NDGC for all the 4 queries; this improves the search relevance from 0.9 to 1. Then calculate the overall NDGC for all users; this improves the search relevance from 0.75 to 0.87. This proposed method helps the search engine to discover the users' intents during the web search.

VII. Conclusions

Our work's key objective is to provide features that can help users during their web search by categorizing, grouping, and segmentation of related browsing URLs together. Upon clustering the browsed URLs categories, groups, and intent segments, search engines can improve the representation of user's search context. This would help the search engine to understand better and discover the user's intent during the web search.

From the experiment results, fourteen clusters were proven by high values of recall and precision and f measure metrics. By using the Rand Index metric, it was approved that clusters of the proposed method compared to baselines had high Rand Index values because each cluster contains a few numbers of browsed URLs because clustering depend on many factors; intent categories then intent groups and finally intent segments. Also, term cluster matrix was presented, which specifies the frequency of the term in each cluster.

From the results' analysis, the top search results returned by Google were presented as test data to match them with our clusters from clustering method for four queries. It was found the first five top results of the first query had matched with the clusters 1. And it was found the three top results of the fourth query have matched with the clusters 9, so the search engine should keep the original results ranking for theses queries. It was found the first, the third and the fifth top search results of the second query matched with the clusters 2. And the second, the first, and the third top search results of the third query matched with the clusters 11, so the search engine should modify the ranking of top search results of theses queries as discussed in Fig. 5 and fig.6.

From the results re-ranking and Search Relevance, the proposed method assists in discovering the user intents that enable the search engine to help users to find what they search for by calculating the NDCG metric for the four Queries for original results and after modified rank. It was noticed that NDGC increased for queries 1 and 2 after modified the rank. Then, the overall NDGC was calculated for all the four queries for the first user; this improved the search relevance from 0.9 to 1. And finally, the overall NDGC was calculated for all queries of all users; this improved the search relevance from 0.75 to 0.87. (In the second experiment with different accessed time to top Google results for experiment's queries, the search relevance improved from 0.72 to 0.86).

Future work will include more research to evaluate the proposed method that improved the search engine ranking and its performance complexity. Expanding the experiment with a larger data set is needed. It is interesting to utilize complete knowledge about users' behavior during the web browsing. Also, it is possible to utilize complete browsing history from different resources such as social media links URLs. Also, it can develop more sophisticated similarity method between browsed web pages in segmentation level of user intents.

References

[1] Kiseleva, "Using Contextual Information to Understand Searching and Browsing Behavior", ACM , Information Retrieval , USA, pp. 1059-1059 , 2016 [the 38th International ACM SIGIR Conf. on Research and Development in Information Retrieval USA, 2016].

[2] Sharan and R. Kumar, "Personalized Web Search Using Browsing History and Domain Knowledge", IEEE, ICICT, Ghaziabad, pp. 493 – 497, 2014. [international Conf. on Issues and Challenges in Intelligent Computing Techniques (ICICT) Ghaziabad , 2014].

[3] A. Mazur, P. Serdyukov, and Y. Ustinovskiy, "Intent-Based Browse Activity Segmentation", IR, Russia, pp. 242-253, 2013. [the 35th European Conference on IR Research, Russia, 2013].

[4] D. Gupta, V. Tayal, A. Thakkar3, K.Makvana, "Cold Start Problem in Personalized Web Search", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, 2016.

[5] Veningston .k and dr. R. Shanmugalakshmi, "Personalized Grouping of User Search Histories for Efficient Web Search", Applied Computational Science, 2014.

[6] Filip Radlinski and Nicolaas Matthijs, "Personalizing Web Search using Long Term Browsing History" , ACM, Web search and data mining, USA, pp. 25-34, 2011. [the fourth ACM international conference on Web search and data mining, USA, 2011).

[7] R. Wang, S. Jiang, and Y. Zhang, "Re-ranking Search Results Using Semantic Similarity", Fuzzy Systems and Knowledge Discovery, pp. 1047 – 1051, 2011 [the Eighth International Conference on Fuzzy Systems and Knowledge Discovery, 2011].

[8] F. Borisyuk, P. Bailey, P. N. Bennett, R. W. White, S.T. Dumais, W.Chu, and X. Cui, " Modeling the Impact of Short- and Long-Term Behavior on Search Personalization", ACM, SIGIR, USA, pp. 185-194, 2012. [the 35th international ACM SIGIR conference on Research and development in information retrieval, USA , 2012].

[9] D. Jiang, H. Li, and J. Pei, "Mining Search and Browse Logs for Web Search: A Survey", ACM Transactions on Intelligent Systems and Technology (TIST), Vol, 4, 2013.

[10] P. Serdyukov and Y. Ustinovskiy, "Personalization of Web-search Using Short-term Browsing Context", ACM, information & knowledge management , USA, pp. 1979-1988, 2013. [the 22nd ACM international conference on information & knowledge management, USA, 2013].

[11] J. P. Data, J. Han, and M., Kamber, Data Mining Concepts and Techniques, Elsevier, USA:,2013.

[12] M.B.Senousy and W. Karam; "A Comparative Study for Internet Search Engines and Web Crawlers", SAMS, Cairo, Egypt, 2011.