User Intent Discovery Using Search Logs and Social Network Analysis

WAEL K. HANNA
Information Systems Dept, Computers & Information Faculty
Mansoura University
Mansoura
EGYPT
wael_karam1@yahoo.com

AZIZA S. ASEM
Information Systems Dept, Computers & Information Faculty
Mansoura University
Mansoura
EGYPT
dr_aziza2@hotmail.com

M. B. SENOUY
Computer and Information Systems Dept
Sadat Academy for Management Sciences
Cairo
EGYPT
badr_senousy_arcoit@yahoo.com

Abstract: - With the continuous growing of applications of internet and Web 2.0, users have the opportunity to publish data over the Web. Search engines face many difficulties to return search results whose rankings based on users' intents. All search engines provide search log of the user by tracking their online searches through recording their queries and click information besides browsing history has been stored at the client side. Also, social networks provide a powerful tool for extracting the users' interests from profile and activities of user's different social networks. This paper presents a new proposed method of enabling personalized Web search for users based on their extracted interests and intents from search logs and composite social networks. This paper explores various extracted features and intents from previous resources. Then clustering the users' extracted intents and use it to re-rank the web search results. The implementation and the evaluation of the proposed method were presented by improving the performance of the Web search engines

Key-Words: - Personalization, Search Engine, Search Logs and Social Network.

1 Introduction

Web search engines (e.g., Google, Yahoo, etc.) return many of search results for a specific search query. Often, they act as ''one size fits all'' by returning same search results for the same query by various users. However, in reality, different users may have different interests. Therefore, there is a need for the personalization of the search results returned by Web search engines for different users based on their preferences. [1]

Automated user interest modeling that involves extracting and inferring user interests without any user input present a difficulty for generating such profiles. As a simplest formulation, user interest profiles can be modeled as a Utility Matrix where users were represented by rows, and items were represented by columns and the values represent the users' levels of interest in those items on a chosen scale.[2]

User interest profiles can also be developed by utilizing domain knowledge about the users or about the items they have expressed interest in. These approaches are referred to as Content-based approaches which model user preferences by representing the item, that the user has expressed interest in, in terms of its attributes and building the user's interest profile based on that. [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].

In today's world, social networks and media such as Facebook or Twitter enable users to express their interest across several domains. They have become a popular medium for users to connect, explore, share content and express themselves. They can share URLs and videos, or post status updates and comments about topics that interest them. Every user has a group of activities within the social network, e.g., public profile information, likes, etc. [2]

Today, billions of users are now joined in multiple online social networks. In many cases, a user is concurrently a member of different networks. This is a form of composite social network. [4].

2 Related Work

"Personalization is the action of presenting the right information to the right user at the right time." To create the user profile, 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 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 [5].

Anna M. et al. In [3] proposed a technique for automatic segmentation of users' daily browsing activity into intent-related segments.

Aditi S. and Rakesh K. in [6] built a framework of an Enhanced User Profile by combined the user's browsing history and the domain knowledge to improve personalized web search.

John et al. in [1] described their approach of enabling Web search personalization for users based on their interest: (1) Activities of users in their social networks, and (2) suitable information from user's social networks. In this paper, the user' interests from social network will be extracted but in multiple social networks where a given user is concurrently a member of different networks.

Michael et al. in [2] proposed unsupervised system that figures a large range of an individual user's explicit and implicit interests from social network profile and activities without any user input. In this paper, the user' interests from social network will be extracted but in multiple social networks where a given user is concurrently a member of different networks.

Erheng et al. In [4] determined the problem of modeling multiple networks as the composite network. In this paper, the user' interests from social network will be extracted but in multiple social networks where a given user is concurrently a member of different networks.

3 Problem Definition

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

Definition. (Social Network SN). A social network (SN) is a set of entities that may be connected based on specific kind of relationship. [1].

Second, we formally define the concept of a composite social network.

Let G = {Gi = (Ui, Ei)}li=1 denote a composite social network, where Gi is the i-th component network, Ui is the user set of Gi , Ei is the user relationship of Ui, and L is the number of component networks. [4].

Each user profile composes of different sections such as basic profile information. For example, The Timeline is a picture of the user's Facebook activity such as status updates and posts along with comments and likes etc. Most of the profile sections stay static or evolve as the user adds more content or likes more pages. [2]

The two sections that can analyze are liked Pages and Timeline. A page can represent any topic and contains detailed information about it. In addition, each page also has a category such as 'Book', 'Movie' etc. assigned to it from among the Facebook Page categories. [2]

In the proposed approach, the users' interests were extracted from Search log and Social Network. Fig. 1 presents the proposed method in a layered figure that acts as a middleware between the user and the search engine. [1].

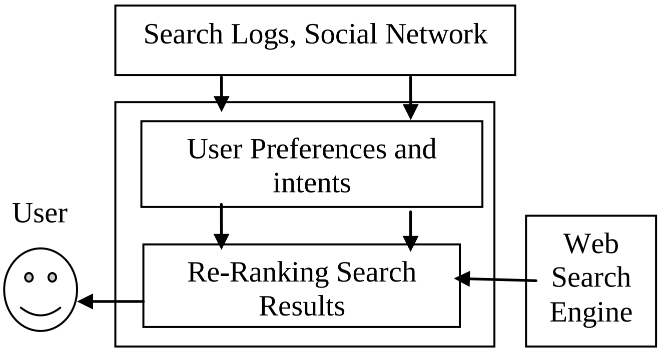
Search Logs, Social Network
User Preferences and intents
User
Re-Ranking Search Results
Web Search Engine

Fig. 1: The layered model for personalized Web search based on extracted preferences from SN.

The first layer extracts the information (intent category and intent group) from search log extract the keywords and the social network of users: extract the activities of the users in their social network, e.g. posts, likes, groups in the form of keywords. Then the second layer presented the re-ranking of the search results based on extracted preferences.

4 Method

4.1 User Interest Modeling

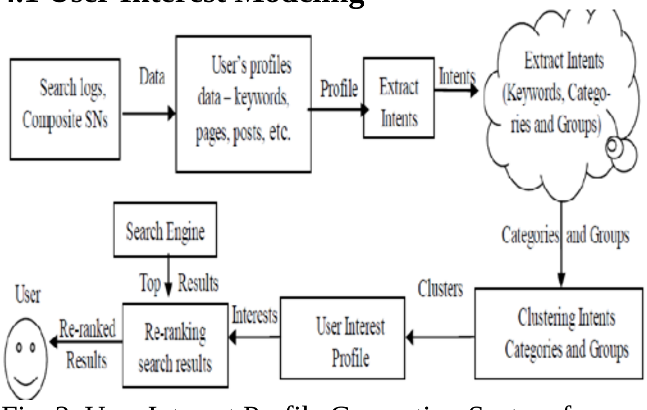
Search logs, Composite SNs
Data
User's profiles data – keywords, pages, posts, etc.
Profile
Extract Intents
Intents
Extract Intents (Keywords, Categories and Groups)
Categories and Groups
Search Engine
Top Results
User
Re-ranked Results
Re-ranking search results
Interests
User Interest Profile
Clusters
Clustering Intents Categories and Groups

Fig. 2. User Interest Profile Generation System from the SN

Fig. 2 shows an overview of the User Interest Profile Generation system. It generates a User Interest Profile (UIP) for each user, from search logs and the user's composite social network data. Then extract the intents (Keywords, Categories, and groups) to clustering the intents to decide the cluster priorities. Then the search engine reordering of the Web search results according to the preferences of the users extracted from their social network.

5 Experiment and Evaluation

5.1 Experimental setup

Standard datasets for this research problem are not existent so, the dataset had been designed. To examine the effectiveness of the proposed method, we conducted this experiment on designed data set. Search histories and social networks profiles were collected to compose the data set similar to [7].

The search logs were collected directly from the users using Google history of a group of researchers in information system filed, mostly Ph.D. students. Then collect the user's data (profiles, groups, likes and comments) of the specific group (research group) from different social networks by using netvizz and Win Automation tools that extract and analyze the interests of chosen group members.

5.2 User Interests Profile Generation

In search log analysis, we extract the keywords of issued queries and browsed URLs during one month.

In any social network, we used a simple way to model the interests of a user as key-pair values. All the activities of a user were summarized in the form of a set of keywords. For example, if a user is posting on the topic of football in his/her social network about the FIFA World Cup, in such case a simple set of keywords noting interests of the user may be recorded as {FIFA, World Cup, football}. Similar keywords are collected for users who reflect their activities on the social network [1].

Definition. (Interests List). The list of interests of a member in the social network, which may be the same user searching for information over the Web. It can be defined as a tuple [1].

Preference List:= (index, keywords (m)). Where the index is a number that represents the user ID in the social network, and keywords is a set of keywords which present the preferences of the user [1].

5.3 Feature Extraction

From each item, we extract features such as intents category and intents groups. For instance, consider the social network item: Facebook page for 'Pride and Prejudice'7. Some of the features extracted from it [2]: Intent Category: Education and Group Intent: Book. These features present a valuable insight into a user's interests.

5.3.1 The Intent Topic Categories

The topic categories used to classify the extracted intents are based on the most general categories of the Open Directory Project and Alchemy taxonomies. We use the DMOZ Search engine (the largest, most comprehensive human-edited directory of the Web) to classify the extracted intents. Also, Alchemy API has been used for classifying web pages into particular category after mapping Alchemy API taxonomies to DMOZ Categories.

Table 1: Mapping Alchemy API taxonomies to DMOZ Categories

DMOZ Categories	Alchemy Categories
Arts	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
Science	Science & Technology
Sports	Sports
Reference	References
Computers	Technology &Computing
Health	Health &Fitness
Recreation	Automotive &Vehicles, Food &Drink, Travel
Shopping	Shopping
Kids and Teens	Hobbies And Interests

5.3.2 The Intent Groups

For each previous category, we classify the intents into intent groups manually with the assistance Alchemy API.

5.4 Ranking Mechanism for Web Search

Google's uses page ranking algorithm. PageRank uses the citation graph of the Web along with the link analysis. Search results can be improved by personalization. A simplistic way to construct user interest profile is to explicitly collect the topic of interest from a user. The search results were Filtered using content similarity between the Web search results and the user interest profile. Construction of a user interest profile usually handles the user browsing behavior. [7].

The proposed method takes into account the interest list which is retrieved from the user's search logs and SNs to uses them to re-rank the Web search results. This help in displaying most relevant Web search results on top [1].

Moreover, we have narrowed the re-ordering of only the top 10 results rather than re-ordering all of the search results. This is because; there are many of the Web search engines providing millions of the search results. However, users mostly visit the top search results. Therefore, re-ordering of the top 10 search results presents a great value for the users. [1].

5.5 Clustering

In clustering, the input will be a set of extracted keywords from the user profile. The keywords are initially present in a text file as the training data. The clustering algorithm is applied to it in order to cluster the input. Using Weka (collection of machine learning algorithms for data mining tasks), SimpleKMeans algorithm [8] was used for clustering the user intents. A database is created with field's keyword, item type (search logs or social network items), its category field, and its group filed. The URLs from top ten search results is saved in a separate file and these acts as test data which is to be tested against predefined clusters of our clustering algorithm in Weka. Then the re-ranking of the top search results of the search engine. We select the top ten clusters .This was acceptable because of the average number of keywords was180, which could possibly result in 10 clusters similar to [7].

Example of cluster structure; User1: Cluster 0: Shopping Store 4%, Cluster1: Computers Programming 10%, Cluster 2 Computers Web Search 13%, Cluster4: Arts TV 6%, Cluster 5: Science Research 9%, Cluster12:Computers Social Media 5%, Cluster13: Education University 7%, Cluster15: Computers Microsoft 4%, Cluster16: Sports Football 11% and Cluster17: Recreation Food 6%.

5.5.1 Cluster Evaluation

To evaluate the clustering analysis using Weka, we record the recall and precision measures. Precision "is the ratio of the number of documents retrieved that "should" have been retrieved" [9]

$$\text{precision} = \frac{|\{\text{relevantdocuments}\} \cap \{\text{retrieveddocuments}\}|}{|\{\text{retrieveddocuements}\}|} \quad (1)\ [9]$$

Recall "is the ratio of the number of relevant documents retrieved to the number of relevant documents" [9]

$$\text{precision} = \frac{|\{\text{relevantdocuments}\} \cap \{\text{retrieveddocuments}\}|}{|\{\text{retrieveddocuements}\}|} \quad (2)\ [9]$$

Table 2. Clustering Evaluation

Cluster	TP Rate	FP Rate	Preci -sion	Re call	F-Measure	ROC Area
C0	1	0	1	1	1	1
C1	1	0.011	0.90	1	0.952	1
C2	1	0.011	0.92	1	0.963	0.995
C4	1	0.005	0.92	1	0.96	1
C5	1	0	1	1	1	1
C12	1	0	1	1	1	1
C13	1	0	1	1	1	1
C15	1	0.01	0.8	1	0.889	1
C16	1	0	1	1	1	1
C17	1	0	1	1	1	1

5.5.2 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), we have to compare pairs as shown in table 3.

Table 3. 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= [n2] 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.

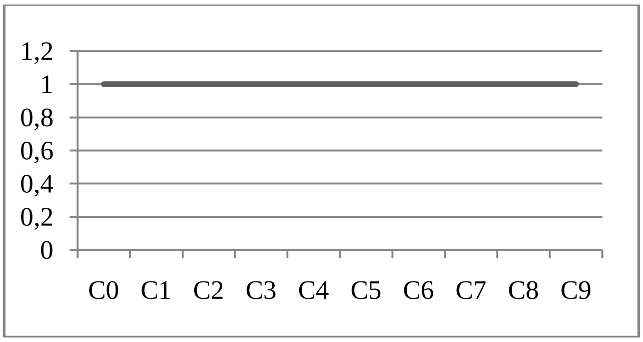

Fig. 3: Rand Index for Clusters

Fig. 3 presents Rand Index for the clusters of our method. We noticed that the Rand Index for all clusters is one. Because of clustering is depending on intent categories and intent groups.

5.5.3 Rand Index Comparison

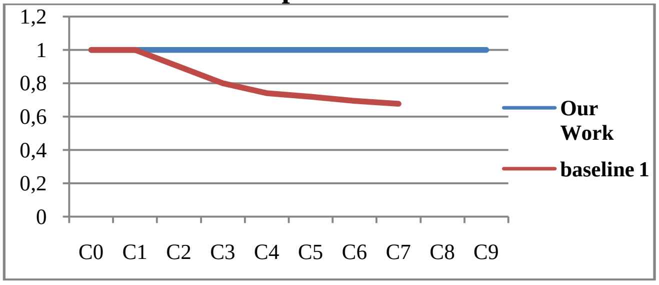

Fig. 4: Rand Index Comparison between Baseline 1 and the Proposed Method

Fig. 4 presents the Rand Index comparison between baseline 1 and the proposed method. We used [4] Work as baseline 1. We noticed that the Rand Index for all clusters is one. Because of clustering is depending on intent categories and intent groups.

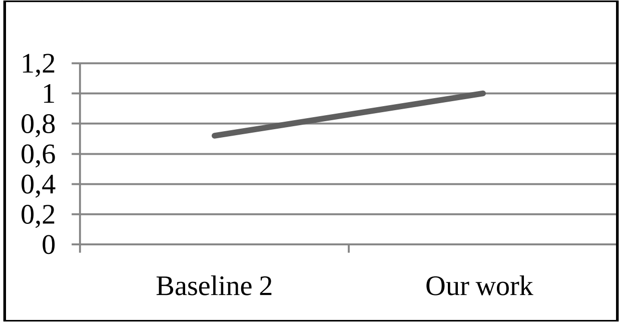

Fig. 5: Highest Rand Index Comparison between Baseline 2 and the Proposed Method

Figure 5 presents the highest Rand Index comparisons between baseline 2 and the proposed method. We used (3) Work as baseline 2. Its highest value is 0.72. But in the proposed method the Rand Index is one.

6 Results Analysis

The analysis of the result is done by discovering the top results for each query belong to each cluster. For each query, the top ten relevant search results provided by Google were collected (two experiments with different accessed times range from February 2016 to November 2016). Then classify as discussed before. Consider them as test data for our clustering algorithm to decide whether theses top ten 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.

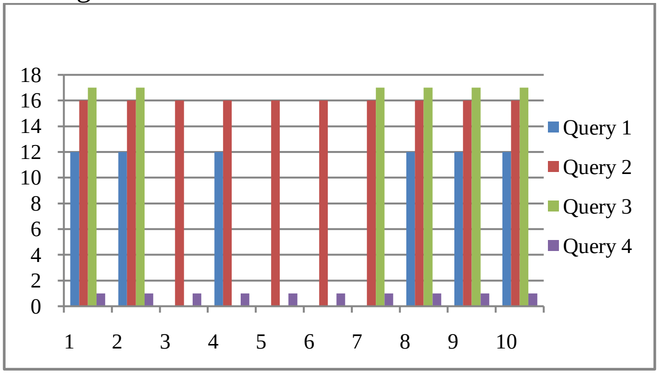

Fig. 6: Queries top ten results versus clusters

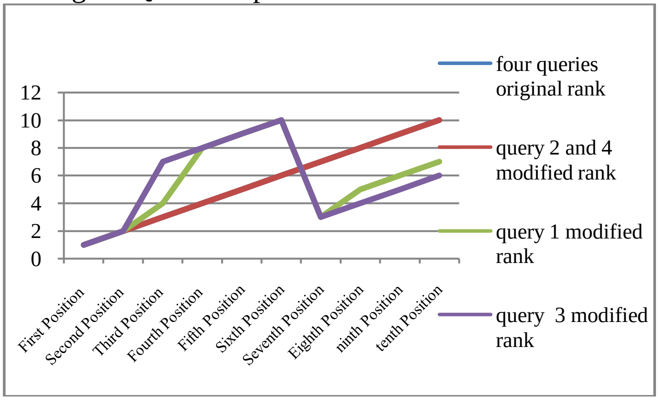

Fig.7: Queries top results original rank and modified rank

Fig. 6 presents Queries top ten results versus clusters for the four queries. And Fig. 7 presents Queries top results original rank and modified rank for the four queries. For the second 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 first query and the third query, the search engine should modify the ranking of the top results as showed in fig. 7.

6.1 Results Evaluation

6.1.1 Evaluation based on discounted cumulative gain

NDCG is an efficient measure primarily 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 we have 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). [10]

We can compute NDCG the Normalized Discounted Cumulative Gain of each rank p using the following formula:

$$\mathrm{N\,DCGp} = \frac{\mathrm{DCGp}}{\mathrm{I\,DCG}} \qquad (4)\ [10]$$

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

And DCG is Discounted Cumulative Gain =

$$\mathrm{DCGp} = \sum_{i=1}^{p} \frac{2\mathrm{reli}\text{- }1}{\log 2\ (i+1)} \qquad (5)\ [10]$$

Where p is PageRank serial number and reli is the graded relevance of the result at position i. For simplicity, suppose that on a four-point scale, the irrelevant result was given a 0 score, 1 for a partially relevant, 2 for more relevant, 3 for highly relevant according to the percentage of the traffic by Google results positions study1.

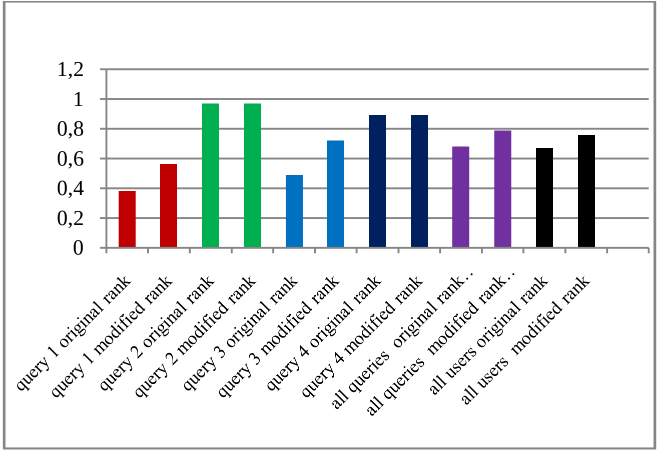

Fig. 8: NDGC for the queries for user 1 and all users

Fig. 8 presents the NDGC for the four Queries for original results and after modifying the rank. It was noticed that NDGC increased for the first and the third queries after modified the ranked. It stills the same for the second and the fourth queries. Then calculate the overall NDGC for all the four queries; this improves the search relevance from 0.68 to 0.79. Then calculate the overall NDGC for all users; this improves the search relevance from 0.67 to 0.76. This proposed method helps the search engine to discover the users' intents during the web search.

6.1.2 Mean Average Precision (MAP)

Mean Average Precision (MAP) is" the mean of the average precision scores for each query". We also calculated the average precision for each query using in our experiment. [11].

$$\mathrm{MAP} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q\,j}\sum_{j=1}^{Q_j} P(\mathrm{doc}_i) \qquad (6)\ [11].$$

N number of queries, Qj number of relevant documents for query j and P(doci) precision at i th relevant document.

Table 4 provides an overview of Mean Average Precision calculated for selected queries executed in two different cases: for original results, re-ranked results based on user interests. An improvement was noticed, with lower Mean Average Precision for the original search results, higher Mean Average Precision for re-ranked search results based on user interests, The results for Mean Average Prevision have been found suitable i.e., improvement in the Mean Average Precision in re-ranked search results using the proposed solution where more relevant search results were re-ranked to top results.

Table 4. Comparison based on mean average precision.

Result Set	MAP Experiment 1	MAP Experiment 2
Original search results	35.88	38.18
Re-ordered search results based on user preferences	41.22	43.48

7 Conclusions and Future Work

The main objective of this work is to present a new proposed method of enabling personalized Web search for users based on their extracted interests and intents from search logs and composite social networks. This paper explores various extracted features and intents from previous resources. Then clustering the users' extracted intents and use it to re-rank the web search results. The implementation and the evaluation of the proposed method were presented by improving the performance of the Web search engines.

From the experiment results, ten clusters were approved 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 have high Rand Index values.

From the results analysis, we presented top search results returned by Google as test data to

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

match them with our clusters from clustering method for four queries. 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 figures 5 and 6.

From the results re-ranking and Search Relevance, we showed how the proposed method assists in discovering the user intents that enable the search engine to help users to find what they search for in the top results by calculating the NDCG metric for the four Queries for original results and after modified rank. It was noticed that NDGC increase for the first and the third queries after modified the rank. Then, the overall NDGC was calculated for all the four queries for the first user; this improves the search relevance from 0.68 to 0.79. And finally, the overall NDGC was calculated for all queries of all users; this improves the search relevance from 0.67 to 0.76 (In the different experiment with different accessed time to top Google results for experiment's queries, the search relevance improved from 0.63 to 0.72).

Future work will include more research to evaluate the proposed method that improved the search engine ranking and its complexity on search engines performance. Expanding the experiment with a larger data set is needed. It can deploy the proposed method in a dynamic and real-time social network profile activity. Finally, a collaborative filtering could be used to improve the performance.

References:

[1] John G. R., Omair S. and Reda A., On personalizing Web search using social network analysis, *ELSEVIER*, vol. 314, 2015, pp. 55–76.

[2] Michael R., Preeti B., and Oliver B., Michael Roberts, Unsupervised Modeling of Users Interests from their Facebook Profiles and Activities, *ACM*, 2015, pp. 191-201.

[3] Anna M., Pavel S. and Yury U., Intent-Based Browse Activity Segmentation, *In Proceedings of 35th European Conference on IR Research*, ECIR, Russia, 2013, pp. 242-253.

[4] Shanmugalakshmi and venington,. *Personalized Grouping of User Search Histories for Efficient Web Search*, Applied Computational Science, 2014.

[5] Erheng Z., Qiang Y. and Wei F., User Behavior Learning and Transfer in Composite Social Networks, *ACM Transactions on Knowledge Discovery from Data*, vol. 8, 2014, pp. 1-32.

[6] Aditi S. and Rakesh K., Personalized Web Search Using Browsing History And Domain Knowledge, *In Proceedings of International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), IEEE, Ghaziabad*, 2014, pp. 493 – 497.

[7] Harshit K, Sungin L., Hong-Gee K.., Exploiting social bookmarking services to build clustered user interest profile for personalized search, *Information Sciences: an International Journal , Elsevier Science*, vol. 281,2014, pp. 399-41.

[8] Jian P. , Jiawei H. and Micheline K., *Data Mining Concepts and Techniques*, Elsevier, 2013.

[9] Senousy M.B. and Wael K. *A Comparative Study for Internet Search Engines and Web Crawlers*, M.S. thesis, SAMS, Egypt, 2011.

[10] Ruofan W., Shan J. and Yan Z., Re-ranking Search Results Using Semantic Similarity, *In Proceedings of Eighth International Conference on Fuzzy Systems and Knowledge Discovery, Shanghai*, 2011, pp. 1047 – 1051.

[11] Simone T., Chapter *An Overview of evaluation methods in TREC Ad-hoc Information Retrieval and TREC Question Answering*. In: L. Dybkjaer, H. Hemsen, W. Minker (Eds.) *Evaluation of Text and Speech Systems*. Springer, 2006.