A novel method for indexing and restructuring search results with Feedback Sessions

Veeresh Ramappa[1], Dr. G. F. Ali Ahammed[2]

***Abstract*— Data Mining refers to extracting or "mining" knowledge from large amounts of data. It is also called as knowledge mining from data. Search engine is one of the most important applications in today's internet. Users collect required information through the search engine in the internet. Generally a query given to the search engine can have different meaning for each user. i.e., each user has different search goals for the given query. The understanding of these different search goals helps in providing better results to the user. Firstly, different user search goals are identified with the help of the feedback session. Feedback session is one in which the last URL is the clicked one. Secondly Feedback sessions are converted to the binary vector which can efficiently reflect user information needs.**

***Index Terms*—Binary vector, Data mining, Indexing, Restructuring search results, User search goals.**

I. INTRODUCTION

Search engine is one of the most important applications in today's internet. Users collect required information through the search engine in the internet. When query is given to the search engine it can have different meaning for each user. For example, when the query "jaguar" is submitted to a search engine, some users are interested in cars; some might be looking for animals, and some might be looking for jaguar fragrances. as shown in Fig. 1.

As in the market basket analysis where the market analyst identifies the items that are frequently purchased, similarly here we analyze the search engine log reports to infer user search goals to improve search engine relevance, enhance the quality and delivery of internet information services to the end user, and improve the performance. User search goals are defined as the information on different aspects of a query that user groups want to obtain.

The understanding and analysis of user goals has many advantages like,

Firstly, Reorganizing web search results [3],[5],[4] according to user search goals by grouping search results with same information need as shown in Fig. 2.

This can be useful to other users with different search goals to find easily what they want.Secondly, user search goals can be represented by some keywords thus helping in query recommendation [6], [8], [7]; This can be helpful to other users to form their query more effective. Third, Reranking web search results according to different user search goals.

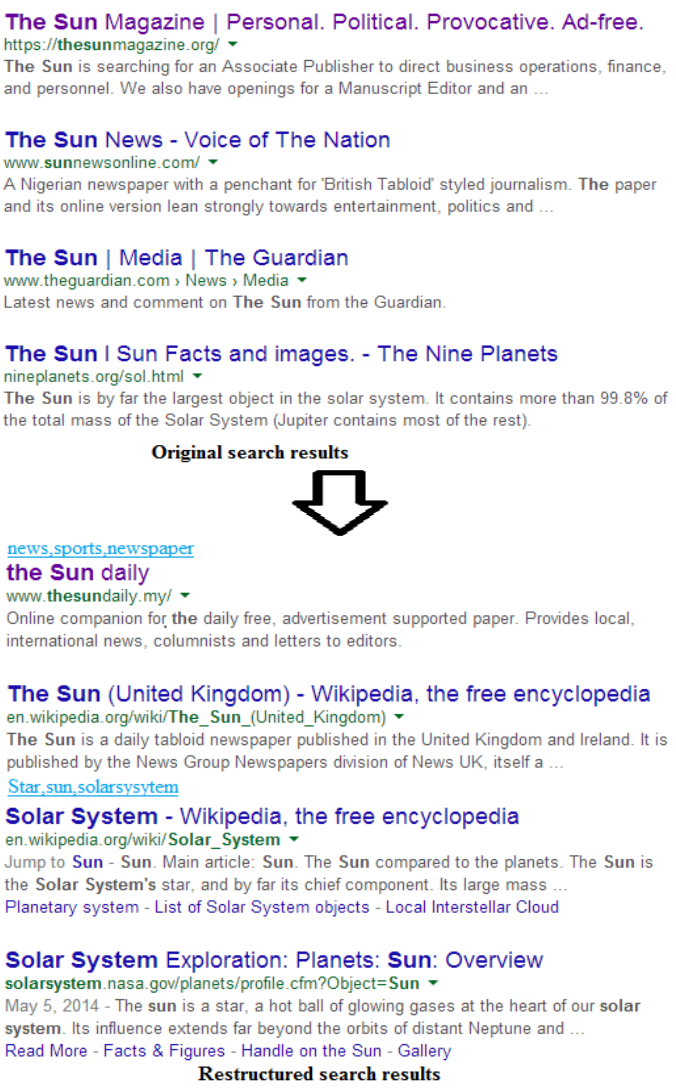

Fig. 2. Restructuring of web results.

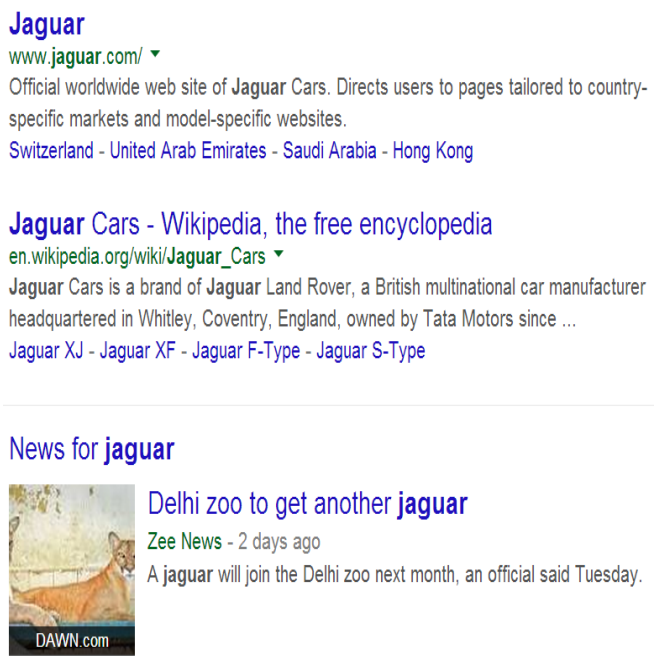

Fig. 1.The examples of the different user search goals for the query "jaguar".

Here, the number of ambiguous user search goals for a query are discovered and each goal is identified some keywords. Firstly, different user search goals are identified with the help of the feedback session. Feedback session is one in which the last URL is the clicked one. Secondly Feedback sessions are converted to the pseudo-documents which can efficiently reflect user information needs.

To sum up,

- Firstly, different user search goals are identified with the help of the feedback session. The feedback session contains both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs.
- Secondly, as the feedback sessions vary lot for different queries therefore feedback sessions are converted to binary vector which can effectively reflect the information need of a user.
- Finally to understand the user search goals efficiently using Average precision (AP) algorithm.

The rest of the paper is organized as follows: The framework of our approach is presented in Section 3. The proposed feedback sessions and their representation namely binary vector are described in Section 4. Section 5 describes the restructuring of search results. Section 6 describes evaluation and section 7 concludes the paper.

II. RELATED WORK

Some of the works done on inferring the search engine goals are discussed here.

A. Automatic identification of user goals in web search.

In this paper the goal of a user is classified into two predefined [10], [9] categories:

a. Navigational.
b. Informational.

- **Navigational queries:** A query is considered navigational when a user already has a particular website in his mind and the goal is simply to reach that particular site. Note that for such a query, the user may either have visited that site before, or just assume such a site exists. For a navigational query, typically users will only visit the correct website they have in mind.
- **Informational queries:** If the user don't have particular page in the mind and visits multiple pages then these are known as informational. For an informational query, typically users do not pre-assume a particular Website to be the single correct answer, and they are willing to click on multiple results.

There 2 types of user goal identification task:

- **Past user-click behaviour:** This method is based on the intuition that the user's goal for a given query may be learned from how users in the past have interacted with the returned results for this query. If

in the past users had clicked on a single website corresponding to the one they have in mind then the query is navigational. On the other hand, if the goal is informational, in the past users should have clicked on many results related to the query. Thus by observing how the results for a particular query have been clicked so far, authors tell whether the current user who issues that query has a navigational or an informational goal.

- **Anchor-link distribution:** Another feature that we may use is the destinations of the links with the same anchor text as the query. As a result, if we extract all the HTML links with the anchor text PubMed, we expect to find that a dominating portion of these links point to that single Website; On the other hand, for a informational query hidden markov model, because of lack of a single authoritative site, we expect that the links with the anchor text hidden markov model point to a number of different destinations.

Authors showed that by combining these features (Past user-click behaviour & Anchor-link distribution) it was possible to correctly identify 90% of user goal accurately.

B. Learning query intent from click graphs.

In this paper the authors present the use of click graphs in improving the query classifiers. Previous works on query classification have primarily focused on improving feature representation of queries this is done by augmenting queries with search engine results. In this work, the authors investigate a completely orthogonal approach — instead of enriching feature representation. Specifically, the authors infer class memberships of unlabeled queries from those of labeled ones according to their proximities in a click graph. Moreover, the authors regularize the learning with click graphs by content-based classification to avoid propagating erroneous labels. The authors demonstrate the effectiveness of the algorithms in two different applications, job intent and product intent classification. In both cases, the authors expand the training data with automatically labeled queries by over two orders of magnitude which lead to significant improvement in classification performance.

C. Building bridges for query classification

User queries are usually short and they are classified by Web query classification (QC). The applications of QC are page ranking, advertisement and sometimes query personalization. Here a bridging classifier is built by the authors this would be generally in the offline mode. Later via the intermediate taxonomy the user queries are mapped to the target categories in the online mode. Bridging the classifier needs to be done only once as there is no need of retaining a novel classifier for each category this is mainly because of leveraging the similarity distribution over the intermediate taxonomy. For narrowing down the scope of the intermediate taxonomy, here the authors introduce a new approach based on which the authors classify the queries. Both effectiveness and efficiency of the online classification can be improved by category selection.

D. Bringing order to the web automatically categorizing search results.

Here in this paper authors developed a user interface that organizes Web search results into hierarchical categories. Various algorithms such as text classification algorithms were used to automatically classify arbitrary search results into an existing category structure on the fly. Here the typically ranked list of search results is compared with the new category interface. Based on this comparison it was seen that category interface is superior both in objective and subjective measures. Most of users liked the category interface much better than the list interface, and category interface were 50% faster at finding information than that of ranked list. Showing the results in categories or organizing search results allows users to focus on items in which they are interested rather than having to browse through all the results one by one.

E. Learning to Cluster Web Search Results.

Generally the query when given to the search engine returns result in the form of long list which is arranged based on their ranking and the user need to go through the list to search the result which they need. The user has to examine all the titles carefully to identify their results. When multiple sub-topics of the given query are mixed together then this takes lot of time. For example, when a user submits query "jaguar" to any of the search engine and is interested in finding results related to "animals" then the web user need to search all the pages and the result may be present in 7th, 8th or 87th page.

In order to solve this problem authors proposed a solution which says- cluster the search results into various groups. This grouping of result allows web users to find the relevant results very easily. Clustering methods don't require Pre-defined categories are not required for the clustering methods as in classification methods. Thus, clustering methods can be adapted for more different queries.

Given a query and the ranked list of search results, our method here the procedure is as follows, firstly the whole list of titles and snippets are phrased, and then all possible phrases (n-grams) from the contents, followed by calculating several properties for each phrase like phrase length, phrase frequencies, document frequencies, etc. These properties are then combined into a single salience score by using a regression model learned from previous training data is then applied to combine these properties. Based on the salience score all the phrases are ranked accordingly, and the top-ranked phrases are taken as salient phrases. Once the salient phrases are identified they are named as candidate clusters, these are further merged according to their corresponding documents.

Here the user feedback session is not considered and the search results returned by the search engine when the query is submitted are analysed directly. Therefore many noisy results that are not clicked by the users are also analysed.

III. FRAMEWORK OF OUR APPROACH

The framework of our approach is shown in Fig. 3. Here first the feedback session is created from Click through logs.

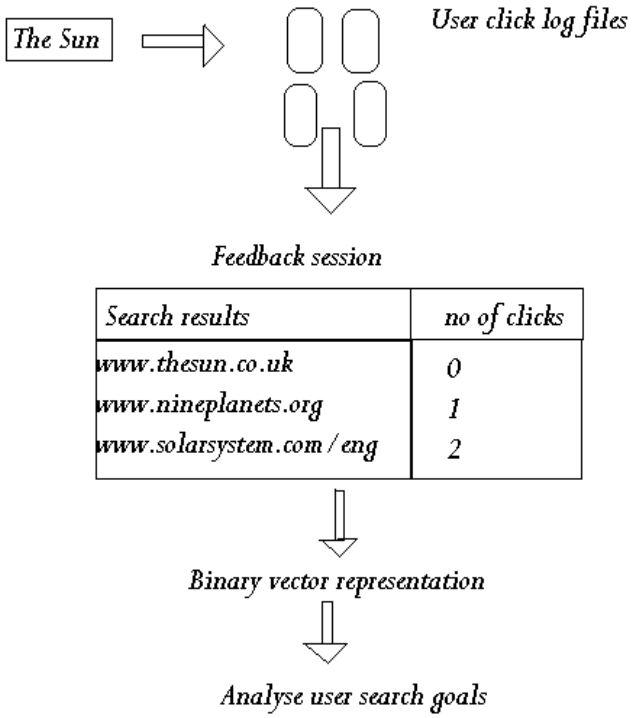

Fig.3. Proposed system architecture.

Feedback session is one which is used to analyse the user search goals and contains both clicked and un clicked URLs. As the feedback session varies for different queries so they are converted to binary vectors. With the help of this binary vector the user search goals are analysed and understood.

IV. FEEDBACK SESSIONS, BINARY VECTOR

A. Representation of feedback session

In search engine, session means number or series of queries to meet the user requirements [11]. In this paper, we make use of a session having only one query, which is different from the conventional session. Search engine logs contain the activities of the users and also reflect what each user needs and what user doesn't need. Feedback sessions are extracted from the log files, these feedback sessions are used to analyse user search goals and consists of both clicked and unclicked URLs.

Generally it is seen that user scan the search result from top to bottom so the URLs in the feedback session before the last clicked URL would be seen, scanned and evaluated by the web user. Fig 5. shows the feedback session.

Search Results	No. of times URLs clicked
URLS IN THE BOX FORM FEEDBACK SESSION	
www.bangaloresunday.com	0
http://timesofindia.indiatimes.com/	2
http://ieeexplore.ieee.org/	3
www.thehindu.com/	4
www.lonelyplanet.com/	0
www.deccanchronicle.com/	0
en.wikipedia.org	0

Fig.4. Representation of Feedback sessions.

Here, as seen in the Fig 5 there are 7 URLs for the query "sun". Here he numbers indicate the number of times the URLs clicked and "0" indicate uncliked URLs. The URLs within the box indicates feedback sessions and it includes 4 URLs out of which 3 are clicked and remaining 1 is unclicked. The feedback session tells what user care about and what he/she is interested in.

B. Mapping feedback sessions to binary vectors

Generally feedback sessions vary a lot for different click-throughs and queries so it is unsuitable to directly use feedback sessions for inferring user search goals. Therefore, some representation method is needed to describe feedback sessions in a more efficient and coherent way. There are various ways of representing feedback sessions. For example, Fig. 6 shows a popular binary vector method to represent a feedback session. Same as Fig. 5, search results are the URLs returned by the search engine when the query "the sun" is submitted, and "0" represents "unclicked" in the click sequence. The binary vector [0111000] can be used to represent the feedback session, where "1" represents "clicked" and "0" represents "unclicked." Binary vector representation is an easy and effective method for representing feedback sessions without requiring much effort.

Search Results	No. of times URLs clicked	Binary vector
www.bangaloresunday.com	0	0
http://timesofindia.indiatimes.com/	2	1
http://ieeexplore.ieee.org/	3	1
www.thehindu.com/	4	1
www.lonelyplanet.com/	0	0
www.deccanchronicle.com/	0	0
en.wikipedia.org	0	0

Fig. 5. The binary vector representation of a feedback session.

V. RESTRUCTURING WEB SEARCH RESULTS

When the query "the sun" is submitted to the search engine, it can have different meaning like some users might be looking for newspaper and for some users it might be looking for solar system. Fig 7 shows the URLs before restructuring, as seen there are 4 URLs 1st URL is unclicked, 2nd URL is clicked twice, 3rd URL is clicked thrice and 4th UR is clicked four times as seen in fig 5. Therefore, the URLs are restructured according to the clicked sequence. Fig 8 shows the URLs after restructuring so these list of URLs are displayed to the user when next time the user login to search for the same query.

http://www.bangaloresunday.com

http://timesofindia.indiatimes.com

http://ieeexplore.ieee.org

http://www.thehindu.com/

Fig.6. URLs before restructuring.

Fig. 7. URLs after restructuring.

VI. EVALUATION BASED ON RESTRUCTURING WEB SEARCH RESULTS

The evaluation of result is a big problem, since goal of each users are not predefined and there is no ground proof. Previous work has not proposed a suitable approach on this task. Considering that if user search goals are analyzed properly, then restructuring of search results can also be done correctly, since restructuring web search results is one application of analyzing the user search goals. Therefore, an evaluation method based on restructuring web search results is proposed to evaluate whether user search goals are inferred properly or not.

A. Restructuring Web Search Results

Since search engines always return millions of search results, so it becomes necessary to organize them to make it easier for users to find out what each user wants. Restructuring web search results is an application of analyzing user search goals. Then, the evaluation based on restructuring web search results will be described.

B. Evaluation Criterion

In order to apply the evaluation method the single sessions are used. Because using user click-through logs it is possible to get implicit relevance feedbacks, namely "clicked" means relevant and "unclicked" refers to irrelevant. A possible evaluation criterion is the average precision(AP) which evaluates according to user implicit feedbacks. AP refers to the average of precisions computed at the point of each relevant document in the ranked sequence.

$$AP = \frac{1}{N^+} \sum_{r=1}^{N} rel(r) \frac{R_r}{r},$$

Where, N^+ refers to number of clicked documents.

r refers to the rank,

N refers to the total number of retrieved documents,

rel(r) refers to a binary function and R_r refers to the number of relevant retrieved documents.

VII. CONCLUSIONS

In this paper, a novel approach has been proposed to infer user search goals for a query by analysing feedback sessions represented by binary vector. First, we introduce feedback sessions to be analysed to infer user search goals rather than search results or clicked URLs directly from the search results. Therefore, feedback sessions can reflect user information needs more efficiently. Feedback session consists of both clicked and unclicked URLs.

Second, as feedback sessions vary a lot for different queries so we map feedback sessions to binary vector to better represent feedback sessions. Finally, a new criterion AP is formulated to evaluate the performance of user search goal inference.

Finally to conclude this paper presents a novel approach for analysing user search goal also feedback session is proposed. This feedback session is mapped to binary vector to better represent feedback sessions, but binary vector representation is not informative enough to tell the contents of user search goals. In future some new methodology would be used to represent feedback session so that the contents are more informative to tell the contents.

REFERENCES

[1] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[2] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[3] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.

[4] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588 596, 2004.

[5] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[6] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. For Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[7] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005.

[8] X. Li, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346, 2008.

[9] R. Jones and K.L. Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 699-708, 2008.

[10] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

[11] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[12] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," Proc. 15th Int'l Conf. World Wide Web (WWW '06),pp. 387-396, 2006.

[13] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs, Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[14] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 41-50, 2008.

[15] A New Algorithm for Inferring User Search Goals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng, 2013.

Veeresh Ramappa received his B.E degree in computer science and Engineering from VTU, Belgaum. He is currently pursuing his M.Tech degree in computer science and Engineering at Centre for PG Studies, VTU Bengaluru. His Research interest includes Data Mining and Web Mining.

Dr. G. F. Ali Ahammed received his B.E degree from Bangalore University, Bangalore. The M.Tech degree from VTU, Belgaum and the Ph.D. degree from Sri Krishnadevaraya University, Anantapur (A.P). His research interest includes Computer networks, cloud computing and data mining. Presently he is working as an associate professor, Department of Computer Science & Engineering, VTU Centre for Post Graduate Studies, Bengaluru. He has published 6 papers in international Journals and conferences.

