IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Improved Accuracy and User Satisfaction by Inferring User Search Goals Based on Feedback Sessions

Rasika A. Sarambale (Rasika M. Shintre) [1], Prof. Kanchan Doke[2]

Lecturer, Computer Technology, BVIT, Navi Mumbai, India[1]

Professor, Computer Engineering, BVCOE, Navi Mumbai, India[2]

Abstract: User search goals can be defined as information on various aspects of query that user want to obtain and it can be considered as the collection of information needs for a query. Different users may have different search goals in their mind when they pass ambiguous query to a search engine. Thus, it is important to infer and analyze user search goals to improve the performance of a search engine and user experience. By clustering the proposed feedback sessions, we infer different user search goals for a query. The feedback session is combination of both clicked and unclicked URLs and this feedback session is mapped to the pseudo documents to better represent the information needs of user. These pseudo-documents are clustered using K-means clustering algorithm which produces better results than K-means clustering algorithm and reduces computation time. Finally, Classified Average Precision (CAP) evaluation criterion is used to evaluate the performance of system. In this way, the system can infer user search goals efficiently and satisfy information needs of user.

Keywords: User search goals, feedback sessions, pseudo-documents, restructuring search results, and classified average precision.

I. INTRODUCTION

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as Search Engine Result Pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain Real Time Computing information by running an algorithm on a web crawler.

Fig 1: Various Web Search Engines

Now day's web search is more booming area of research. There are so many efficient methods already presented by different authors, every method claiming their efficiency in their own ways. This area is basically defined by the uses search goals.

However, sometimes queries entered by the user may not exactly represent information needs since many ambiguous queries cover large topics and various users may want to get information on different subject when they submit the same query. For example, when the user submit "DNA" as query to Google search engine, some users may want to get information related to DNA newspaper, while other users want to get natural knowledge of DNA, as shown in Fig. 2. So, it is important and necessary to find out different search goals in information retrieval. User search goals can be defined as information on various aspects of query that user want to obtain. User search goals can be considered as the collection of information needs for a query. Finding appropriate user search goals and performing its analysis have many of advantages in enhancing performance of search engine relevance and user experience.

Some advantages are summarized as follows:

1) We can restructure web search results according to user search goals. In this, search results are grouped together with the same search goal. Thus, users with different search goals can find what information they want.

2) User search goals which are represented by the keywords can be used in query recommendation; thus, the users can take help of the suggested queries to form their queries more precisely.

3) The distributions of user search goals are useful in applications such as re-ranking web search results which contain different user search goals.

Copyright to IJARCCE

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Fig 2: Example of user search goal for the query "DNA" and its distribution

In this system, we give solution to discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. First propose to infer user search goals for query by clustering our proposed feedback sessions. The feedback session is defined as the series of both clicked and unclicked URLs and ends with last URL that was clicked in a session from user click-through logs. Then, propose an optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Based on the feedback session construct the pseudo document for analysing the accurate result. This pseudo document consist of keywords of URL's in the feedback session. This is called as enriched URL's. The enriched URL's are clustered and form a pseudo document. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have a high similarity in comparison to one another but are very dissimilar to object in other clusters. After constructing the pseudo document Web search results are restructured based on documents collection detail. Evaluation Criterion Classified Average precision (CAP) to evaluate performance of restructured web search results.

II. LITERATURE SURVEY

Effective organization of search results is critical for improving the utility of any search engine. The utility of a search engine is affected by multiple factors. While the primary factor is the soundness of the underlying retrieval model and ranking function, how to organize and present search results is also a very important factor that can affect the utility of a search engine significantly. In Table I, Compare the different methods those are presented to solve the problem of user search goals.

In [2], Lee et al. study whether and how can automate goal-identification process. First present results from a human subject study that strongly indicates the feasibility of automatic query-goal identification. They stated that majority of queries have a predictable goal. Taxonomy of query goals based on two types: Navigational queries and Informational queries. Two features are used for the prediction of user goal:

A. Past user-click behavior:
If a query is navigational, users will primarily click on the result that the user has in mind.
Therefore, by Observing the past user-click behavior on the query, we can identify the goal.

B. Anchor-link distribution:
If users associate particular query with a particular website then most of the links that contain the anchor will point to that particular website. Hence by observing the destinations of the links with the query keyword as the anchor, we can identify the potential goal of the query. Their experimental evaluation shows that by combining these features they an correctly identify the goals for 90% of the queries.

R. Jones and K.L. Klinkner [3], defined session boundaries and automatic hierarchical segmentation of search topics. In this approach, analysis of typical timeouts used to divide query streams into sessions and the hierarchical analysis of user search tasks into short-term goal and long-term missions is done. This method only identifies whether a pair of queries belong to the same goal or mission but does not care about what the goal is in detail.

In [4], Wang and Zhai learn interesting aspects of queries by analyzing the clicked URLs directly from user click-through logs to organize search results. However, this method has limitation since the number of different clicked URLs of a query may be small.

In this approach, two deficiencies by (1) learning "interesting aspects" of a topic from Web search logs and organizing search results accordingly; and (2) generating more meaningful cluster labels using past query words entered by users. Evaluate method on a commercial search engine log data. Compared with the traditional methods of clustering search results, this method can give better result organization and more meaningful labels.

In [5], [6] J.Zheg, H.Chen analyze the search results and returned by the search engine when a query is submitted. Since user feedback is not considered, several noisy search results that are not clicked by any users may be analyzed as well. In this approach, they reformalize the clustering problem as a salient phrase ranking problem. Given a query and the ranked list of documents (typically a list of titles and snippets) returned by a certain Web search engine, method first extracts and ranks salient phrases as candidate cluster names, based on a regression model learned from human labeled training data. The documents are assigned to relevant salient phrases to form candidate clusters, and the final clusters are generated by merging these candidate clusters

In [7], Li et al. define query intents as "Product intent" and "Job intent" and they try to classify queries according to the defined intents. Other works focus on tagging queries with some pre-defined concept to improve feature representation of queries [8].

However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical. In the second class, people try to reorganize search results.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

TABLE I Comparison of Methods, Techniques and Approach

Sr.No	Titles	Methods/Technique	Advantage	Disadvantage
1.	Automatic Identification of User Goals in Web Search	User click behavior and Anchor link distribution	Using goal identification task to achieve 90% of accurate results	Potentially-biased dataset
2.	Query-Sets: Using Implicit Feedback & Query Patterns to Organize Web Document	Non supervised tasks	Improve the quality as 90%	A broader comparison with online directory
3.	Learn from Web Search Logs to Organize Search Results	Commercial search engine log data and clustering	Better result organization and meaningful labels	Informative feedback information from user
4.	Generating Query Substitutions	Query pair algorithm	Increase coverage and effectiveness	Machine translation techniques
5.	Learning Query Intent from Regularized Click Graphs	Semi-supervised click graph	Improve classification Performance	Impact of seed queries and faceted query classification
6.	Varying Approaches to Topical web Query Classification	Pre vs. post retrieval classification	QC is outperforms bridging documet taxonomy as 48%	multiple approaches to improve performance
7.	Context-Aware QS by Mining Click-Through and Session Data	Offline model learning and online QS step, concept sequence suffix	Coverage and quality of Suggestions	Larger coverage area

III.PROBLEM STATEMENT

The evaluation of user search goal inference is a big problem, since user search goals are not predefined and there is no ground truth. Previous work has not proposed a suitable approach for this problem. Effective way to reorganize search results is clustering of web search result. Here in this approach reorganizing search results truly based on user search goals. These search goals represents user's interested aspect. Discover the number of user search goal for a query based upon these keywords and using k-mean clustering algorithm, forms the cluster which contain one label which is one of the aspect of query and that cluster contain links related to each other and label .And rearrange in such way that top most visited links should occur at topmost.

IV. PROPOSED SYSTEM

A. Framework of Approach

Fig. 3 shows the framework of approach. Framework consists of two parts. In the first part, all the feedback sessions of a query are extracted from user click-through logs and converted to the pseudo-documents. Then, user search goals are inferred by performing the clustering on these pseudo-documents. Each goal is depicted with some keywords. As the exact number of user search goals are not known in advance, several values are tried and the optimal value will be calculated. In the second part, the original search results are rearranged based on the user search goals inferred from the first part. Then, the performance of restructured search result is evaluated by evaluation criterion CAP and final evaluation result will be used as the feedback to get the optimal number of user search goals in the first part.

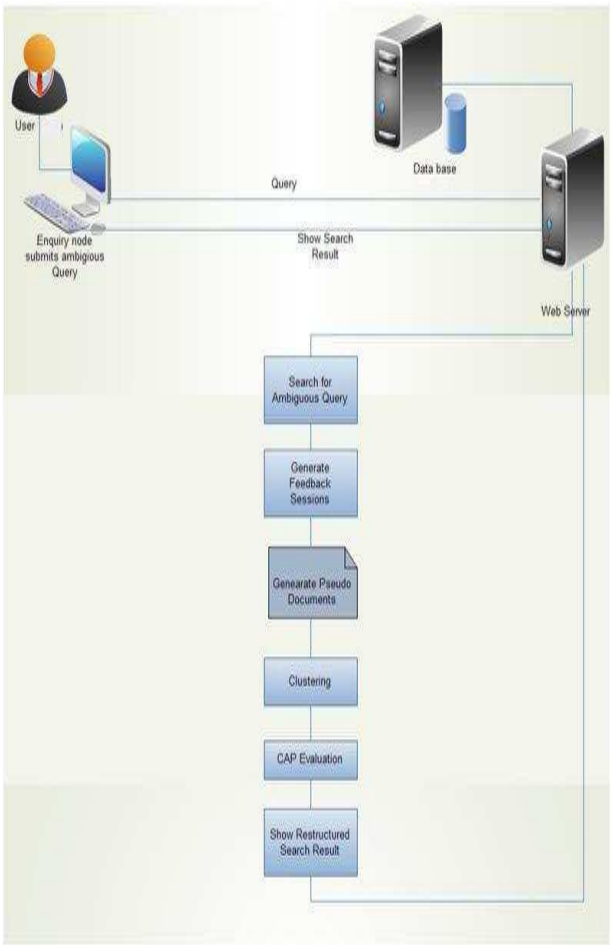

Fig 3: Framework of approach

DOI 10.17148/IJARCCE.2016.53177

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

B. System Architecture

Overall system architecture is as shown in Fig. 4. Initially user login into system and search for ambiguous query. Then system shows many search results. From this results user clicked some desired URLs, by using this clicked data system makes a feedback session. After that feedback session is mapped to pseudo document and clustering is performed. At last performance of system is calculated by using CAP evaluation criteria.

System consists of following modules to execute the designed application:-

1) Feedback Session
2) Map Feedback session to Pseudo Documents
3) Clustering User Search goals
4) CAP Evaluation

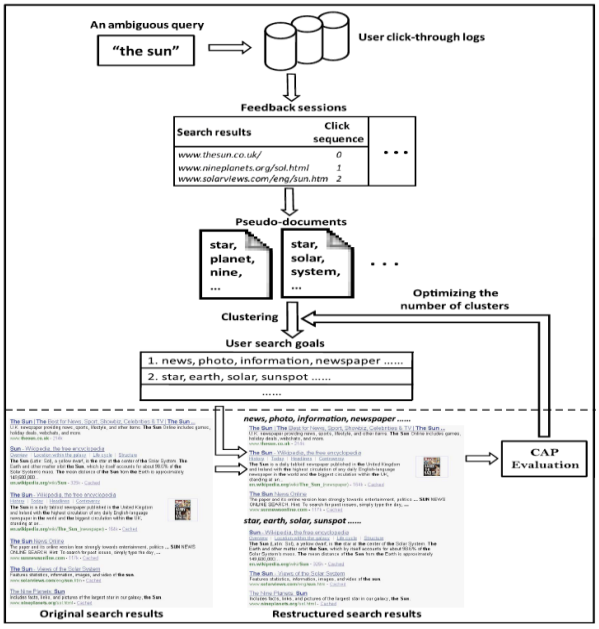

Fig 4: System architecture

1) Feedback session:

Generally, a session is used in reference to web applications. It is sequence of interaction between server and the user. The feedback session is combination of both clicked and unclicked URLs and this session stops with the last URL clicked by user. It is important that before the last click, all the URLs are scanned and analyzed by users. Thus, both the clicked and unclicked URLs before the last click are considered as a part of the user feedback. Fig. 5 shows a feedback session and a single session. In Fig. 5, the left part shows 10 search results for the query and the right part shows sequence for user clicks .Here "0" shows unclicked URLs. The single session is composed of all 10 URLs in Fig. 5, but the feedback session is consisting of seven URLs in the rectangular box.

These seven URLs again composed of three clicked URLs and four unclicked URLs. Inside the feedback session, the clicked URLs reflect what user wants and the unclicked URL tells what users do not care. It is important that the unclicked URLs after the last clicked URL should not be considered as the part of feedback sessions.

Search results	Click sequence
www.thesun.co.uk/	0
www.nineplanets.org/sol.html	1
www.solarviews.com/eng/sun.htm	2
en.wikipedia.org/wiki/Sun	0
www.thesunmagazine.org/	0
www.space.com/sun/	0
en.wikipedia.org/wiki/The_Sun_(newspaper)	3
imagine.gsfc.nasa.gov/docs/science/know_l1/sun.html	0
www.nasa.gov/worldbook/sun_worldbook.html	0
www.enchantedlearning.com/subjects/astronomy/sun/	0

Fig. 5: Feedback session in single session for the query

2) Mapping of feedback sessions to Pseudo-documents:

Feedback sessions vary a lot for different click-throughs and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. There can be many kinds of feature representations of feedback sessions. For example, Fig. 6 shows a popular binary vector method to represent a feedback session. The binary vector [0110001] can be used to represent the feedback session, where "1" represents "clicked" and "0" represents "unclicked." However, since different feedback sessions have different numbers of URLs, the binary vectors of different feedback sessions may have different dimensions. Moreover, binary vector representation is not informative enough to tell the contents of user search goals. Therefore, it is improper to use methods such as the binary vectors and new methods are needed to represent feedback sessions. For a query, users will usually have some vague keywords representing their interests in their minds. They use these keywords to determine whether a document can satisfy their needs. However, although goal texts can reflect user information needs, they are latent and not expressed explicitly. Thus, pseudo-documents can be used to infer user search goals.

Search results	Click sequence	Binary vector
www.thesun.co.uk/	0	0
www.nineplanets.org/sol.html	1	1
www.solarviews.com/eng/sun.htm	2	1
en.wikipedia.org/wiki/Sun	0	0
www.thesunmagazine.org/	0	0
www.space.com/sun/	0	0
en.wikipedia.org/wiki/The_Sun_(newspaper)	3	1

Fig. 6: The binary vector representation of a feedback session.

Mapping of feedback session to Pseudo-document includes two steps.

i) Representing the URLs in the feedback session

In the first step, we first enrich the URLs with additional textual contents by extracting the titles and snippets of the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

returned URLs appearing in the feedback session. In this way, each URL in a feedback session is represented by a small text paragraph that consists of its title and snippet. Then, some textual processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Finally, each URL's title and snippet are represented by a Term Frequency-Inverse Document Frequency (TF-IDF) vector, respectively, as in

$$T_{ui} = [t_{w1}, t_{w2,} \ldots\ldots\ldots, t_{wn}]^T$$
$$S_{ui} = [S_{w1}, S_{w2,} \ldots\ldots\ldots, S_{wn}]^T \quad \ldots\ldots\ldots(1)$$

where T_{ui} and S_{ui} are the TF-IDF vectors of the URL's title and snippet, respectively. ui means the ith URL in the feedback session. And wj(j=1,2, . ., n) is the jth term appearing in the enriched URLs. Here, a "term" is defined as a word or a number in the dictionary of document collections. t_{wj} and s_{wj} represent the TF-IDF value of the jth term in the URL's title and snippet, respectively. Considering that URLs' titles and snippets have different significances, we represent the enriched URL by the weighted sum of T_{ui} and S_{ui} , namely

$$F_{ui} = w_t T_{ui} + w_s S_{ui} = [f_{w1}, f_{w2}, \ldots\ldots, f_{wn}]^T \quad \ldots\ldots(2)$$

where F_{ui} means the feature representation of the ith URL in the feedback session, and w_t and w_s are the weights of the titles and the snippets, respectively. We also tried to set to be 1.5, the results were similar. Based on (2), the feature representation of the URLs in the feedback session can be obtained. It is worth noting that although Tui and Sui are TF-IDF features, Fui is not a TF-IDF feature. This is because the normalized TF feature is relative to the documents and therefore it cannot be aggregated across documents. In our case, each term of Fui (i.e., fwj) indicates the importance of a term in the ith URL.

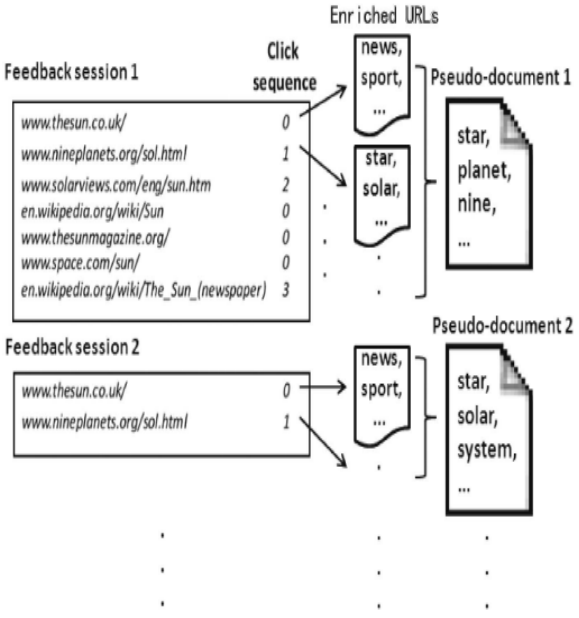

Fig. 7: Illustration for mapping feedback sessions to pseudo-documents.

ii) Forming pseudo-document based on URL
In the second step, we form pseudo-document based on URLs representation. This is done by combining the clicked and unclicked URLs. Once pseudo document is created we can infer search goals effectively.

3) Clustering the Pseudo-documents:
One of the most popular clustering methods used today is the K-means clustering algorithm. The bisecting K-means simply repeats standard K-means clustering where k is fixed. Using bisecting K-mean algorithm which will produces better clustering results.

4) Classified Average Precision evaluation:
Evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. It determines the number of user search goals for a query. Apply CAP method to evaluate the results and restructure the web results. We can obtain an implicit relevance feedback, namely "clicked" which means relevant and "unclicked" means irrelevant.

Average precision (AP) evaluates as per user implicit feedbacks. AP is calculated as:

$$AP = \frac{1}{N+} \sum_{r=1}^{N} \mathrm{rel}(r) \frac{R_r}{r} \quad \ldots\ldots(3)$$

Where, N+ is the number of clicked documents. r is the rank, N is the total number of documents that are retrieved, rel() is a binary function, the number of clicked retrieved documents of rank r or less. AP is not best solution for evaluating clustered searching results.

Thus we use new criterion "Classified AP," as

$$CAP = VAP * (1 - \mathrm{Risk})^{\gamma} \quad \ldots\ldots(4)$$

Where, "Voted AP (VAP)" is the AP of the class including more clicks. Risk is used to avoid classification of search results into too many classes. is used to adjust the influence of Risk. It is given as

$$\mathrm{Risk} = \frac{\sum_{i,j=1(i<j)}^{m} d\,ij}{C_m} \quad \ldots\ldots(5)$$

Where, m is the number of the clicked URLs. If ith, jth clicked URL are categorized into one class, then d_{ij} is set to 1 otherwise it will be 0.The term is total number of the clicked URL pairs.

V. RESULT ANALISYS

System relies on the feedback of user. Feedback are then converted into pseudo-documents which represents the keywords from the documents. After that the pseudo documents are clustered using the k-means clustering algorithm. Results are evaluated using Risk, VAP and CAP. Table II shows the keywords depiction of different queries. Those are nothing but user search goals.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

TABLE II Keywords Depiction of Different Queries

Query	Keywords used to depict user search goals
DNA	Dna, india, news, breaking
	Wikipedia, biologic
	Headlines, results, cell
SUN	Sun, wikipedia, star
	Cent, system, import
	Life, news, sport

The steps to be performed in execution of our system are as shown below:-

Step 1: Insert any query for searching.

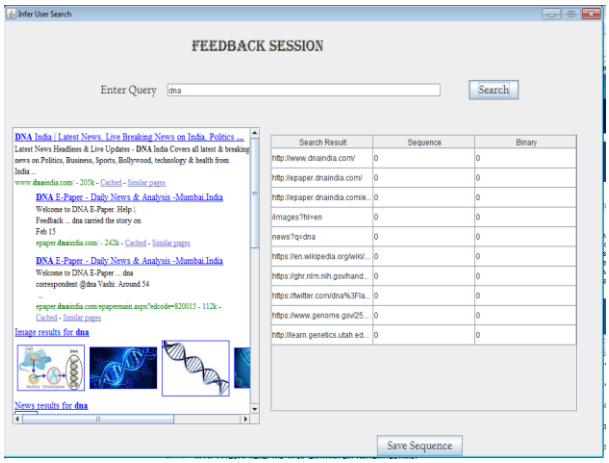

Fig 8: Insert any query for searching

Step 2: Collect the log details of user search history.

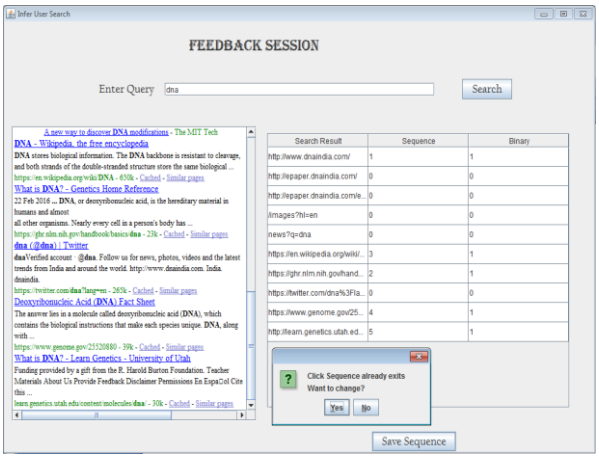

Fig 9: Collect the log details of user search history

Step 3: Depend upon the users click and unclick URL's create a feedback session document.

Step 4: From the feedback sessions create psudo documents

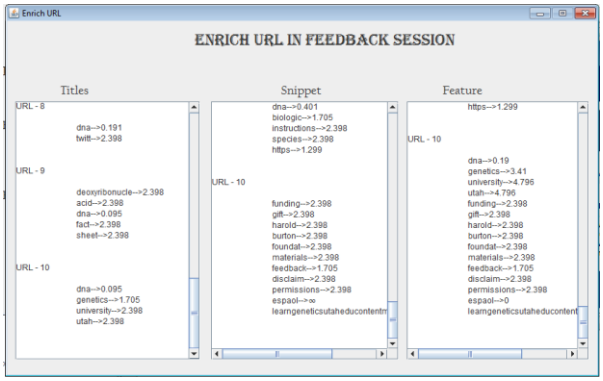

Fig 10: Users click and unclick URL's create a feedback session document.

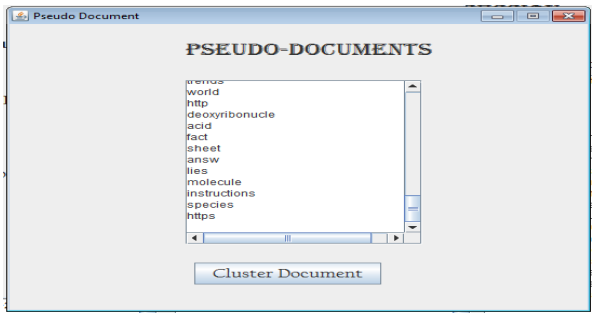

Fig 11: Psudo documents

Step 5: By using K-means Algorithm create final clusters of collected URL's according to subject

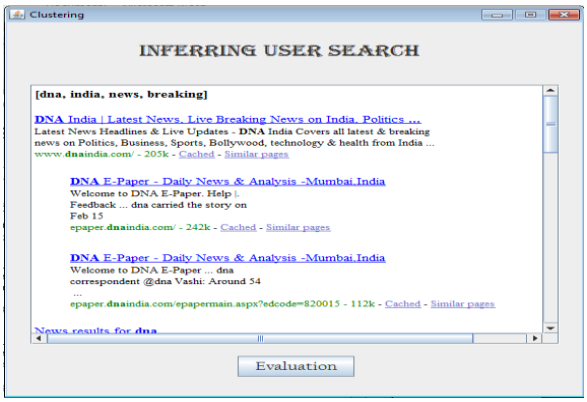

Fig 12: Cluster 1

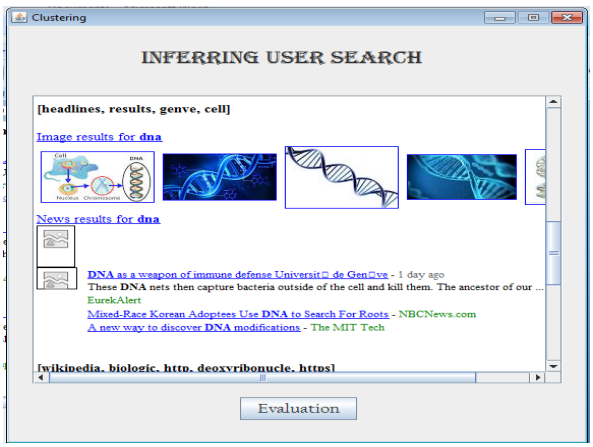

Fig 13: Cluster 2

Copyright to IJARCCE
DOI 10.17148/IJARCCE.2016.53177

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Fig 14: Cluster 3

Step 7: Forward this clusters for Classified Average Precision

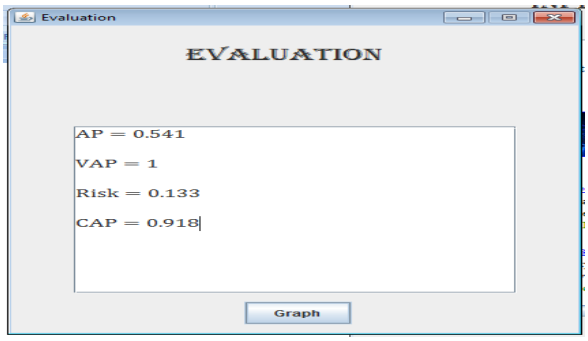

Fig 15: Values of AP, VAP, Risk and CAP

Step 8: Generate graph for AP, VAP, Risk and CAP

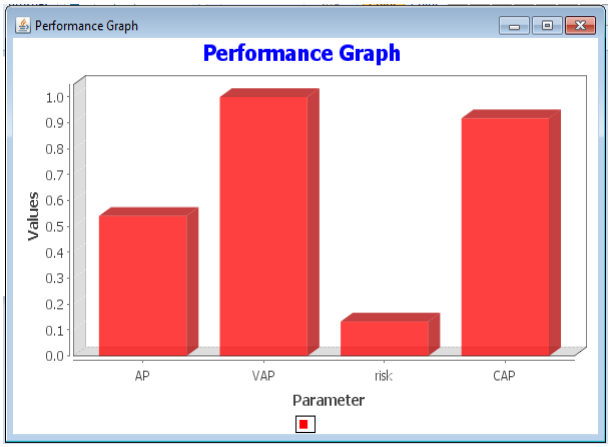

Fig 16: Performance graph

Table III shows evaluation of queries such as average precision(AP), VAP, Risk factor and CAP.

TABLE III Query Evaluations

Query	AP	VAP	Risk	CAP
Dna	0.541	1	0.133	0.918
Sun	0.592	1	0.222	0.86

VI. CONCLUSION

The user search goal using feedback sessions method focuses on inferring the user search goals by performing clustering on feedback session represented by pseudo-documents. Feedback sessions can reflect user information needs more efficiently. This system helps to the user to reduce their extra efforts while gathering information using search engine. This system can be used to improve discovery of user search goals for a similar query. This approach satisfies information needs of the user as well as saves lot of time to search ambiguous query. By using this approach we get efficient and correct search results for the query. The pseudo-documents are clustered by K-means clustering.

REFERENCES

[1] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", vol. 25, no. 3, 2013.

[2] 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.

[3] 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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for Web Query Classification," Proc. 29th Ann. Int"l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR "06), pp.131-138, 2006.

[9] 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, and 2004.

[10] 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.

[11] 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.

[12] SQL Server tutorial by Mr. Kudvenkat https://www.youtube com/ user/kudvenkat

[13] SQL Server tutorial at http://www.quackit.com/sql_server/tutorial/

[14] Rasika Sarambale, Prof. Kanchan Doke , "Inferring User Search Goals Based on Feedback Sessions" , International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4, Issue 11 November 2015

BIOGRAPHIES

Rasika A. Sarambale (Rasika M. Shintre), Received the Bachelor degree (B.E.) in Computer Engineering in 2011 from Ramrao Adik Institute of Technology (RAIT), Nerul-Navi Mumbai. She is now pursuing Master's degree in Computer Engineering at Bharati Vidyapeeth College of Engineering, Navi Mumbai. She is lecturer in BVIT, Navi Mumbai. Her current research interests include Data mining & information retrieval.

Prof. Kanchan Doke, Obtained Engineering degree (B.E.) in Computer Engineering in the year 2000 from Walchand College of Engineering, Solapur and Postgraduate degree (M.E.) in Computer Engineering from Pillai Institute of Information Technology, Panvel. She is approved Undergraduate and Postgraduate teacher of Mumbai university and having about 15 yrs. of experience. Her area of interest includes stegnography and watermarking, specially for black and white images, security, Data mining & information retrieval.

DOI 10.17148/IJARCCE.2016.53177