International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Enhanced Approach for Construing & Reorganizing User Search Result Using Feedback Session

Sultana N. Sayyad[1], Deepak S. Tamhane[2]

[1], [2] M. Tech Student, Department of Computer & Science Engineering, MLRIT, Andhra Pradesh, India

Abstract: ***As different users may have different search goals or different search queries for same search results. There may be broad topic or ambiguous query which user wants to search through search engine. The reconstruction and analysis of user search results can be very useful in improving search engine relevance & user experience. In this paper we propose a enhanced approach to construe user search goals by analyzing search engine query logs. Primarily we propose a framework to find out different search results for a query by grouping the proposed feedback session. Feedback session is generated from no of clicks per query logs (click through logs) and can efficiently reflect need of user. Second we propose a enhanced approach to generate duplicated document or pseudo document to represent the result of grouping, n last we propose a new criterion "Classified Average Precision (CAP)" to conclude the performance of construing user search results. Concluded Results are presented using click-through logs from a commercial Search Engine to verify and validate the effectiveness of our proposed methods.***

Keywords: Web mining, pseudo documents, information retrieval, Web text analysis, searching, grouping.

1. Introduction

Now a day's internet plays very important role in our life. The WWW is a huge resource for people for web search application. For searching their queries people always prefer to use search engine. The search engine is always responsible to search different kinds of information like text, audio, video, images and many more... for this purpose user needs to submit a query to search engine. Queries which are to search engine represents the user needs. When query gets submitted search engine represents all results regarding that query. All results need not to be useful to user. Sometimes query may not represent the exact information need of user. N also there are chances of ambiguity in user queries & different user wants different results on same query [2]. So there is many search results but not useful for user.

In this paper our main purpose is at searching the no. of variant search result for a ambiguous query and representing each result with some keyword automatically. First we propose a new aspect to assume user search results for ambiguous query by grouping defined feedback session. Feedback session is defined as the series of all clicked & unclicked URLs by user and ends with the Last URL that was clicked in same session from user Click-Through Logs. Then we propose one method for perfection to map feedback session to pseudo documents. Then we group these pseudo documents to restructure user search results and represent them with some keywords. Since we also propose a concluding criterion classified average precision (CAP) to evaluate the performance of the restructured web search results[1].

2. Motivation

Different User may have different information goals according to his queries. This project is basically used for broad topic and ambiguous queries because ambiguous queries mean that it has different search result for same query. Hence when user want to get information about particular query it displays the different results. It means that user query is ambiguous. The construing and analyzing user search results can be very useful in improving search engine pertinence and user experience.

3. Literature Survey

3.1 Automatic identification of user goals:

U. Lee, Z. Liu, and J. Cho [2] , proposed automatic identification of user search goals. They stated that majority of queries have a predictable goal. Taxonomy of query goals based on two types:

3.1.1 Navigational queries

In this type, user has a particular result for his query in mind and is primarily interested in visiting that result. User may either have visited that resulted site before, or just assumes such a site exists. Here, users will only visit the correct sites.

3.2.2 Informational queries

These are the queries where user does not have a particular search result in his/her mind or intends to visit multiple resulted pages to learn about the topic. User is seeking WebPages that provide background information about a particular query. Users click on multiple results because they do not have any knowledge about a particular website to be single correct result.

Here, two features are used for the prediction of user goal:

a) **Maintain user Click through logs:**
If a query is navigational, users will first click on the result that the user has in already in his mind. Therefore, by observing the all previous 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 information regarding the query 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.

3.2 Classification of query according to information need of the user

D. Shen, J. Sun, Q. Yang, and Z. Chen [3], published a work on classifying web queries into a set of target categories where the queries are very short and there are no training data. Here, intermediate axonomy is used to train classifiers bridging and target categories so that there is no need to collect training data. Classifier bridging is used to map user queries to target categories.

3.3 Reorganizing Search Results

X. Wang and C.-X Zhai [4], proposed grouping of search results which restructure it and allows a user to navigate into significant documents quickly. These aspects restructure search results learned from search engine logs. Steps of this aspect are as follows:

Given a query,
1. Get its related information from search engine query logs. Working set is generated by using this information
2. Learn the appearance of search result from information in the working set. These appearances correspond to users interests.
3. Each appearance of search result is labeled with representative query.
4. Categorize and Restructure the search results of the input query according to the aspects.

3.4 Grouping of Web Search Results

H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma [5], researched on renormalizing the clustering problem. This approach consists of four steps:

1. Search result fetching
2. Document parsing and phrase property calculation
3. Salient phrase ranking
4. Post-processing.

a query and ranked list of search results. Firstly, the whole list of titles and snippets is parsed, extracts all possible phrases from the contents and calculates several properties for each phrase such as document frequencies, phrase frequencies. Then the regression model is applied to combine these properties into a single salience score. Phrases are ranked according to salience score and the top ranked phrases are taken as salient phrases. In post processing, filter out the pure stop words

Disadvantages:
Feedbacks are not considered. So, noisy results that are not clicked by user may be analyzed.

3.5 Session Boundaries

R. Jones and K.L. Klinkner [6], defined session boundaries and automatic layered segmentation of search topics. In this approach, Query streams are divided into session and the layered analysis of user search tasks into short term goal and long-term missions is done by analyzing typical timeouts. Timeout is an elapsed time of 30 minutes between queries which signifies that the user has discontinued searching. Here, combination of different set of related, time limited, query log and web search features can state mission limits and goals. Hence, best approach to grouping queries within the same results may build on first identifying the limits then matching subsequent queries to existing segments.

Disadvantages:
It 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.

Table1. Previous Work

Previous Research Papers	*Result/Conclusion*
Z. Chen	Worked on Query classification Limitations- Experiment was conducted on a potentially-biased dataset
H. Chen	Organizes search results into a hierarchical category structure. Limitations- Query aspects without user feedback have limitations to improve search engine relevance
Wang , Zhai	clustered queries and learned aspects of similar queries Limitations- This method does not work if we try to discover user search goals of any one single query in the query cluster rather than a cluster of similar queries
R. Jones and K.L.Klinkner	Introduce search goals and missions to detect session boundary hierarchically Limitations- Their method only identifies whether a pair of queries belong to the same goal or not and does not care what the goal is in detail.

4. Proposed Work

Overall system architecture is as shown in fig 1. There are mainly four modules in this system.

1. Capturing feedback Session.
2. generate pseudo documents
3. grouping pseudo document
4. restructuring based on User Search Goals

4.1 Feedback Session

In this module we define feedback session. Feedback session consist of both clicked and unclicked URLs using query click-through-log. This session ends with previous URL which was clicked last in a single session. It is motivated that before the

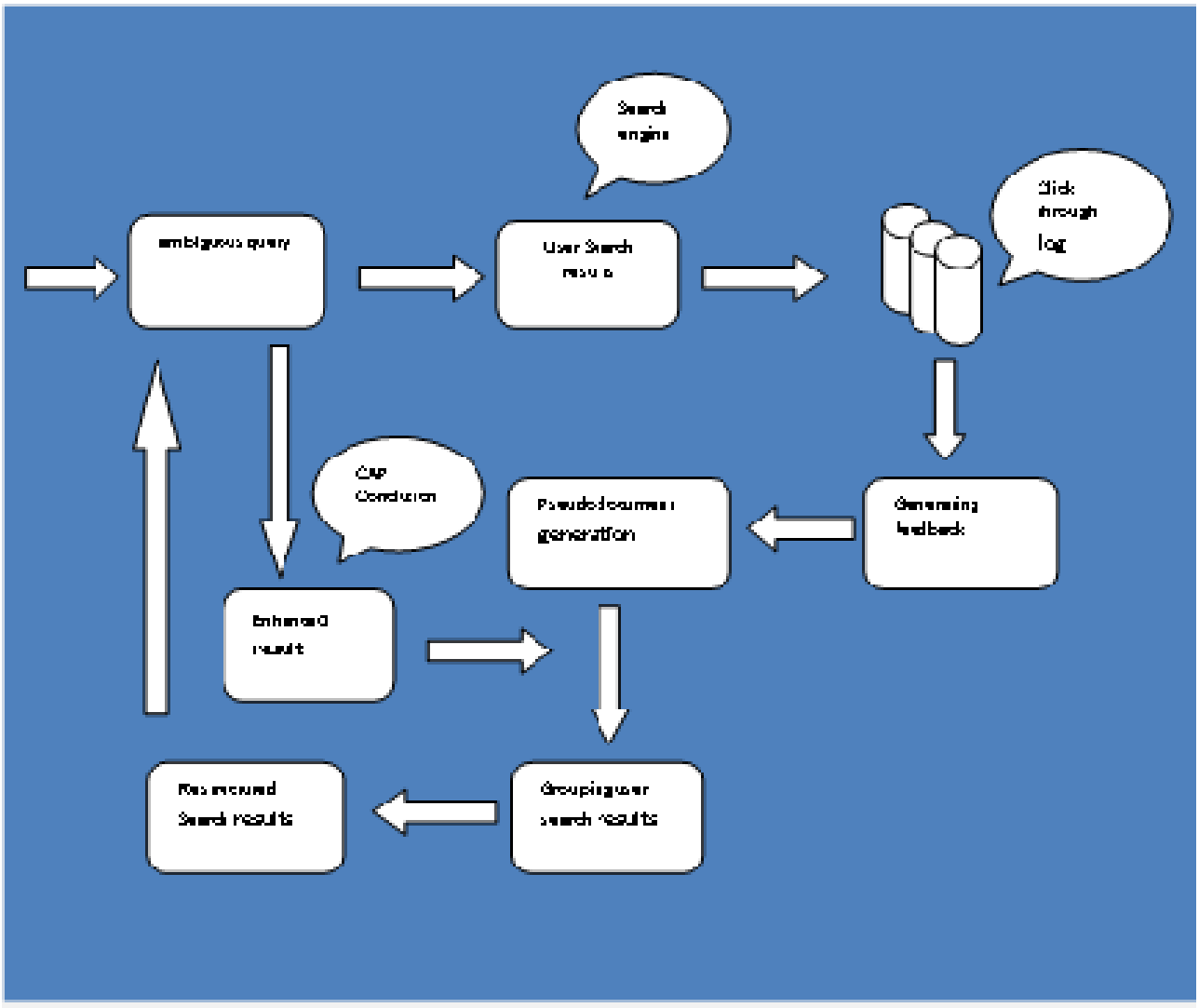

Figure 1: System Architecture

last click of URL, all the URLs have been seen and concluded by the users. Hence, apart from the clicked URLs, the unclicked URLs before the last click should be a part of the user feedbacks. Fig.2. shows an example of a feedback session and a single session.

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

Figure 2: Feedback session in single session

Fig. 2 shows the 10 serach result for the query "the Sun". in fig 2. Left part shows the search result and left part shows the click log for that search result. In click-through log '0' indicates the unclicked URL whereas number other than '0' indicates the no of click sequence for particular search result. As shown in fig2 last clicked URL is 7th URL in search result as shown in rectangular box. All unclicked URL are skipped after last clicked URL. And all unclicked URL before last click is included in feedback session.

4.2 Generating Pseudo Documents Using Feedback Session

There is need of some representation method to describe feedback session in more efficient way. There can be many kinds of representation of feedback session. One of those is Binary Vector Method in which '0' represents unclicked URL & '1' represents clicked URL. But binary Vector Method is Not so informative. So to build the pseudo documents there are two step included. They are as follows

4.2.1. 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 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 (1), respectively

$$T_i=[t_{w1}+ t_{w2+} t_{w3}..... t_{wn}]^T$$

$$S_i=[s_{w1}+ s_{w2+} s_{w3}.....s_{wn}]^T \quad (1)$$

where Ti and Si are the TF-IDF vectors of the URL's title and snippet, respectively. And w_j(j=1,2,….n) is the j^{th} term appearing in the enriched URLs. Considering that URLs' titles and snippets have different significances, we represent the enriched URL by the weighted sum of Ti and Si , namely

$$F_i = w_tT_i + w_sS_i = [f_{w1,} f_{w2,}....., f_{wn}]^T \quad (2)$$

where Fi means the feature representation of the i^{th} URL in the feedback session, and w_t and w_s are the weights of the titles and the snippets, respectively.

4.2.2 Formation of Pseudo-Document

We propose an optimization method to combine clicked and unclicked URLs in the feedback session to obtain a feature representation. Let R_{fs} be the feature representation of a feedback session, and $r_{fs}(w)$ be the value for the term w. Let R_c (c=1,2,….,m) & R_{uc}(uc=1,2,….n) be the feature representations of the clicked and unclicked URLs in this feedback session, respectively. Let $r_c(w)$ & $r_{uc}(w)$ be the values for the term w in the vectors

$$R_{fs} = \left[r_{fs}(w1), r_{fs}(w2), \dots\dots, r_{fs}(w3)\right]^T, \quad (3)$$

$$r_{fs}(w) = \arg \min_{r_{fs}(w)} \left\{ \sum_m \lambda \sum_n \left[r_{fs}(w) - R_c(w)\right]^2 - \left[r_{fs}(w) - R_{uc}(w)\right]^2 \right\},$$

$$r_{fs}(w) \in I_c \quad (4)$$

λ is a parameter balancing the importance of clicked and unclicked URLs. When λ in (3) is 0, unclicked URLs are not taken into account. On the other hand, if λ is too big unclicked URLs will dominate the value of $rfs(w)$. In this paper, we set λ to be 0.5.

4.3 Grouping of Pseudo Document

As in equation (3) and (4), each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document is R_{fs}. The similarity between two pseudo-documents is computed as the cosine score of R_{fsi} and R_{fsj}, as follows

$$\text{Sim}_{i,j} = \cos(Rfsi, Rfsj)$$

$$= \frac{R_{fs_i} . R_{fs_j}}{\left|R_{fs_i}\right|\left|R_{fs_j}\right|}$$

And the distance between two feedback sessions is

$$\text{Dis}_{i,j} = 1 - \text{Sim}_{i,j}.$$

To cluster pseudo documents K-means clustering is used which is very simple and effective. To check the optimal values of clustering we have a evaluation criterion.

4.3 Restructuring based on User Search Goals

since the exact number of groups are still not determined when construing user search results, a feedback information is needed to finally determine the best cluster number hence, it is necessary to develop a metric to conclude the performance of user search result induction objectively. Considering that if user search goals are induced properly, the search results can also is restructured properly, since restructuring web search results is one application of construing user search goals. Therefore, we propose an evaluation method based on restructuring web search results to evaluate whether user search goals are inferred properly or not.

4.4.1 Restructuring Web Search Results

Since search engines always cause millions of search results, it is necessary to arrange them to make it easier for users to find out what they want. Restructuring web search results is an application of construing user search goals. We will introduce how to restructure web search results by inferred user search goals at first. Then, the evaluation based on restructuring web search results will be described. Then, we can group each URL into a cluster centered by the construed search goals. In this paper, we perform Grouping by choosing the smallest distance between the URL vector and user-search-goal vectors. By this way, the search results can be restructured according to the inferred user search goals.

4.4.2 Evaluation Criterion

In order to apply the evaluation method to large-scale data, the single sessions in user click-through logs are used to minimize manual work. Because from user click-through logs, we can get inherent significant feedbacks, namely "clicked" means significant and "unclicked" means not significant. A possible evaluation criterion is the average precision (AP) (1) which evaluates according to user implicit feedbacks. AP is the average of precisions computed at the point of each relevant document in the ranked sequence, as shown in

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

N^+ is number of relevant documents, r is rank
N is total number of retrieved documents
rel() binary function on the relevance of given rank
Rr is number of relevant retrieved documents

VAP (voted AP) is the AP of the class with more clicks as votes. Here URL's in the single session are restructured into two classes, bold-faced and unbold faced. VAP still does not satisfactorily work. So, there should be a risk to avoid classifying search results into too many classes (1).

$$\text{Risk}= \frac{\sum_{i,j=1}^{m}(i<j)^{d_{i,j}}}{c_m^2} \quad (2)$$

This calculates normalized number of clicked URL pairs that are not in same class. Here, m is number of clicked URL's[1].

$$C_m^2 = \frac{m(m-1)}{2} \qquad (3)$$

C is the total number of clicked URL pairs (1). CAP is extension of VAP as,

$$CAP = VAP \times (1 - risk)^r \qquad (4)$$

CAP selects the AP of the class that user is interested in and takes the risk of wrong classification into account. r is used to adjust the influence of risk on CAP.

5. Result Analysis

System relies on the feedback of user. Feedbacks 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 1.1 shows the keywords depiction of different queries. Those are nothing but user search goals.

6. Snapshots

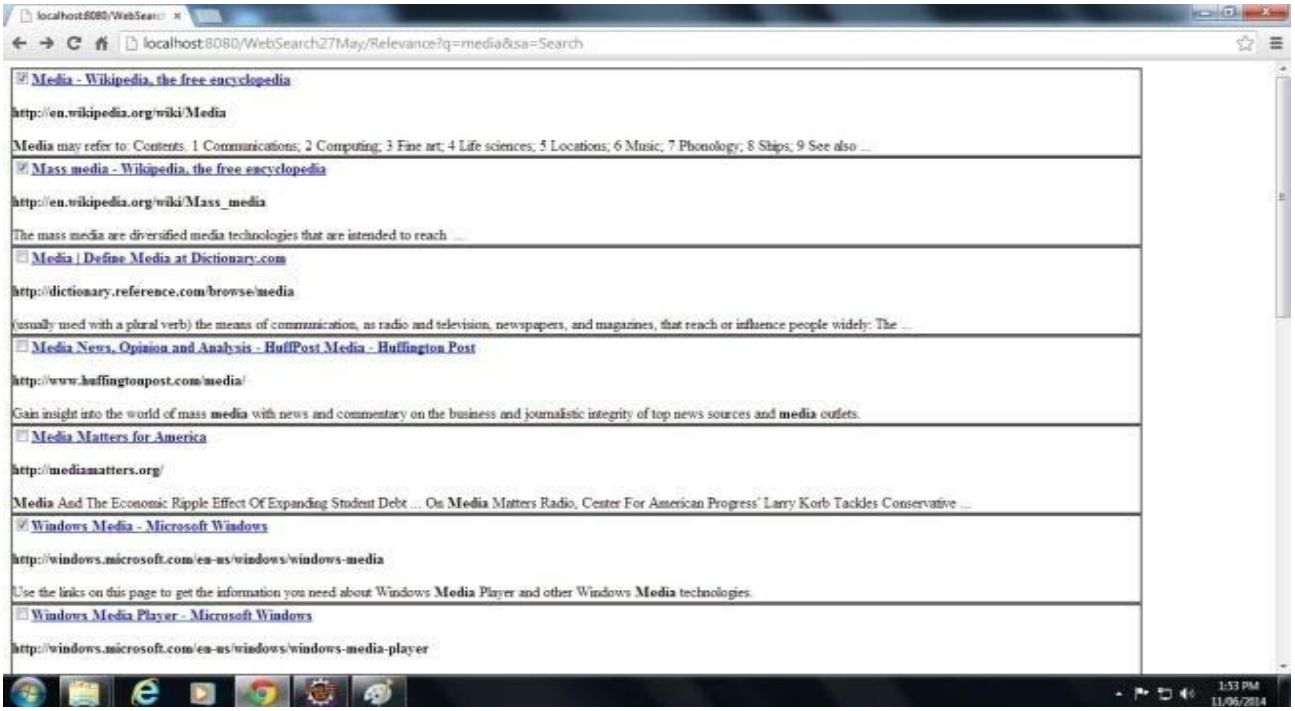

Figure 3: Snapshot of original results

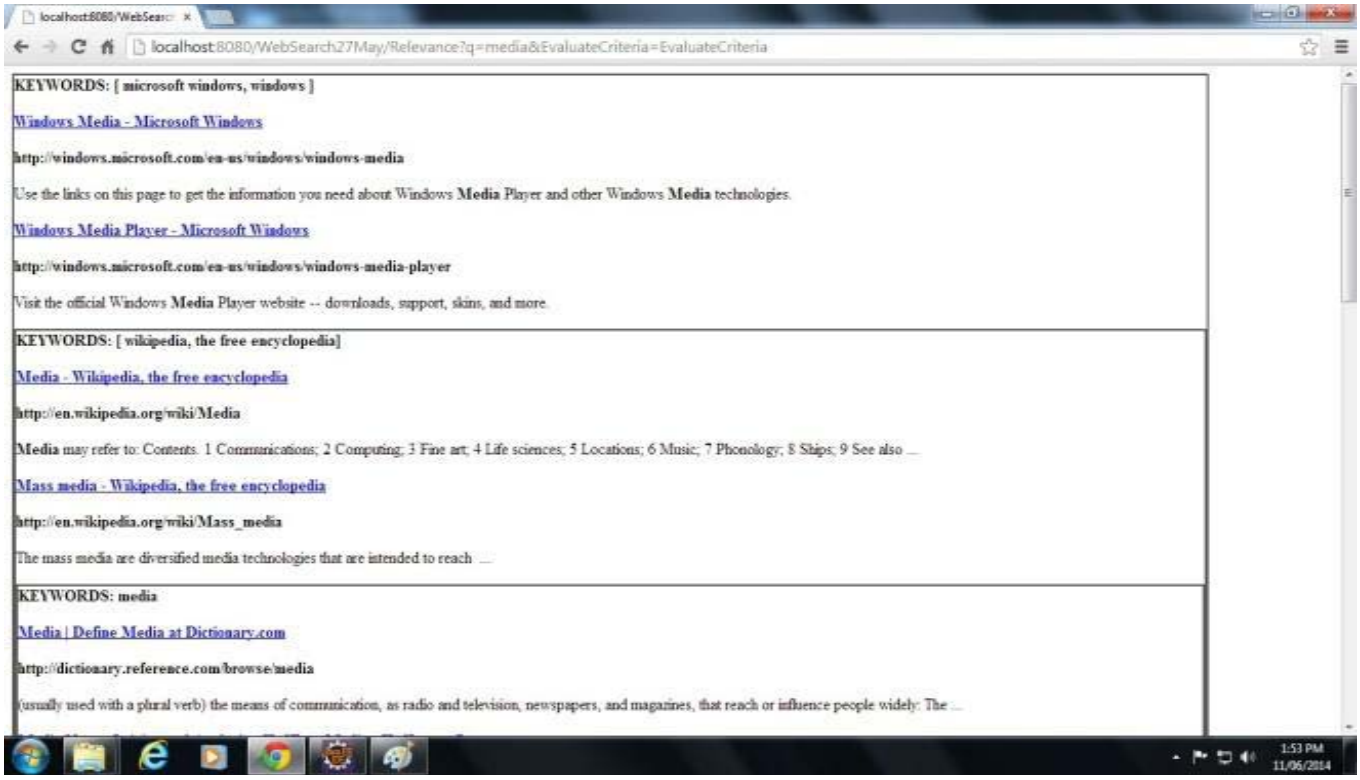

Figure 4: Snapshot of ReOrganised results

Table 2: Keyword Representing Different Queries

Query	*Keywords used to represent user search goals*
Taj	India
	Mahal
	Hotel
Apple	Apple, Wikipedia
	Operating System
	Official
Numerology	Android apps Google play
	Numerology, Wikipedia
	Astrology
Ginger	hotel
	Software
	Health Benefits

Table 3 shows evaluation of queries such as mean average VAP, risk factor and CAP

Table 3: Query Evaluations

Query	*Mean average VAP*	*Risk*	*CAP*
Apple	0.715	0.6	.612
Numerology	0.333	0.3	0.602
Taj	0.444	0.67	0.549

7. Conclusion

In this paper we studied some problems associated with feedback session record. Feedback session can record limited number of URLs. So that user can analyze few URLs. In this case we have increased the size of feedback session. So that user can analyze more number of URLs. In pseudo documents keywords are present which are clustered according to hierarchical clustering. We used hierarchical clustering for searching topic-subtopic wise. From this method user can easily find out his/her information need within small time. We studied and implemented feedback session and mapping of these feedback session to the pseudo documents. Finally we also implemented performance method to evaluate search results. This approach is used to improve searching. The proposed system framework is useful and feasible to be used with real world search systems. It will help users to search information more precisely.

References

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, “A New Algorithm for Inferring User Search Goals with Feedback Sessions”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 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] 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.

[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] 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.

[7] 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.
[8] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.
[9] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.
[10] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Clickthrough Data as Implicit Feedback, Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

Author Profile

Sultana N. Sayyad received the Bachelor degree (B.E.) in Information Technology in 2009 from SVPM COE, Malegaon (Bk). She is now pursuing Master's degree in Computer Science Engineering at MLR, Institute of Technology, Hyderabad. She is currently working as a Lecturer in Al-Ameen College of Engineering, Koregaon Bhima, Pune

Deepak S. Tamhane, received the Bachelor degree (B.E.) in Information Technology in 2009 from SVPM COE, Malegaon (Bk). He is now pursuing Master's degree in Computer Science Engineering at MLR, Institute of Technology, Hyderabad. He is currently working as a Lecturer in PGM College of Engineering, Wagholi, Pune