Scientific Journal Impact Factor (SJIF): 1.711
IJMTER
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

USER SEARCH FEEDBACK BEHAVIOR USING FUZZY C MEAN

Darshana S. Parikh[1], Prof. S.M. Patil[2]

[1]Student, M.E. computer department, B.V.C.O.E. Navi Mumbai,
[2]Head of IT Department, B.V.C.O.E. Navi Mumbai

Abstract- When a query is submitted to search engine, user have in mind a fixed goal. Search engine gives thousands of results for such a query. Most of them are not useful for user so time and energy is wasted. For increasing retrieval precision, some new method provides manually verified answers to Frequently Asked Queries (FAQs). In this paper we are clustering the feedback session by using Fuzzy c-means algorithm. Also we use method to map feedback sessions to pseudo-documents which can efficiently reflect required data. Then, we evaluate the "Classified Average Precision (CAP)" of restructured web search results.

Keywords: Search Goal, Feedback Session, Fuzzy C Means (FCM) Algorithm, Classified Average Precision (CAP).

I. INTRODUCTION

In the area of web mining, more importance is given to fast and accurate extraction of information. Query suggestions provided by the search engine will help to find the user needs. But it may cover broad topics, so this may not be solution for achieving a better search result. Also same queries have different goals for different users. The analysis of user search goal improves the relevance and user satisfaction of the search engine. This method analyzes the user query and restructure the search results.

In our approach, priority is given to the current user's history; also page ranking algorithm is used to arrange the search results inside a cluster. Our method uses the user feedback to reduce the number of clusters. When the number of links inside a cluster is too large, it can be solved by clustering inside the cluster. Hence it will result more accurate search results. CAP evaluation is used to evaluate the performance of restructured web search results.

II. RELATED WORK

In this section we compare the different methods those are present in this section very clearly.

- Query recommendation using query logs given query submitted into a search engine. All previous submitted queries store in one location. Select one query identifies related similar queries information. Next to perform the cluster of similar queries or group of similar queries information. Consider the preferences or relevance of each and every query and assign the ranking. In this implementation user selection is simple and submits the query in search engine very easy. Using this approach there is no improvement in display of the results content in webpage. Users are not satisfying with current approach results [13].

- Rosie Jones, Kristina Lisa Klinkner in 2008 studied real sessions manually labeled into hierarchical tasks, and show that timeouts, whatever their length, are of limited utility in identifying task boundaries, achieving a maximum precision of only 70%. They proposed and evaluate a method for the automated segmentation of users' query streams into hierarchical units [2].

- H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li in 2008 proposed a novel context aware query suggestion approach which is in two steps. A concept sequence suffix treeis constructed as the query suggestion model. By looking up the context in the concept sequence suffix tree, our approach suggests queries tothe user in a contextaware manner [3].

- U. Lee, Z. Liu, and J. Cho in 2005 studied whether and how we can automate the goal-identification process. They proposed two types of features for the goal identification task: user-click behavior and anchor-link distribution. The experimental evaluation shows that by combining these features we can currently identify the goals for 90% of the queries studied. [6].

III. FUZZY C MEAN SYSTEM

In this our approach is to infer user search goals for a query by clustering feedback sessions. Then, we use a optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords Framework of the approach:

The framework consists of two levels.

A. The upper level
Feedback sessions are first extracted from user click-through logs. These are mapped to pseudo documents. User search goals are inferred by clustering the pseudo documents, with some keywords

B. The lower level
The original search results are restructured based on the user search goals inferred. Then evaluate the performance of restructuring search results by CAP criterion. This evaluation result is used as the feedback to select optimal number of search goals in the upper part.

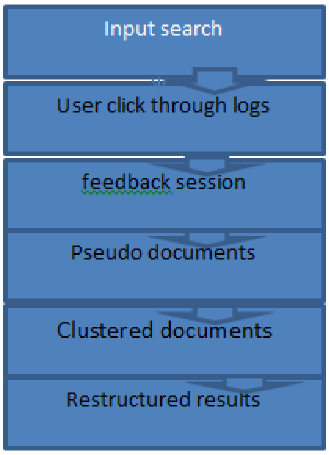


Fig 1 Data flow of proposed system

1. FEEDBACK SESSION

Session is used in reference to web application. It is a sequence of connection between user and server. The feedback session is combined of clicked and unclicked URLs. All URLs before the last click are scanned and analyzed by users and all these URLs are considered as feedback. Left part of the figure shows search results for the query and the right part of the figure shows sequence of user's clicks. Here '0' shows unclicked URLs. And number shows clicked URL's.

The clicked URL reflects that user wants and unclicked URL tells that user does not care. The feedback session includes both clicked and unclicked URLs in a single session. The clicked URLs tell what users require. Unclicked URLs reflect what users do not care about. The unclicked URLs after the last clicked URL are not includes in the feedback session.

2. FORMING PSEUDO DOCUMENT

Mapping of feedback session to pseudo-document have two steps:

A. Map feedback sessions to pseudo-documents

It is unsuitable to use the feedback sessions directly for inferring user search goals. Thus some representation is needed to describe feedback sessions in a coherent way. Binary Vector Method can be used to represent the feedback session where Clicked URL=1 Unclicked URL=0.

Representing URLS in the feedback session

First enrich the URLS with additional textual contents by extracting the titles and snippets. In this way each URL in a feedback session is represented by a small text paragraph that contains titles and snippets. Then some textual processes are implemented to those paragraphs such as: Transforming all letters, stemming, Removing stop words. Finally, each URL's title and snippet are represented by TF-IDF(Term Frequency-Inverse Document Frequency) vector.

T ui = [t w1 , t w2 , … t wn]^ T
S ui = [s w1 , s w2 , … s wn]^ T

Where Tui and sui are the TF-IDF vectors of the URL's title and snippet. Ui is the i-th URL in the feedback session Wj is the j- th term appearing in the enriched URL .Since title and snippet have different significances, we represent enriched URLs as weighted sum of Tui and Sui.

Fui = wtTui + wsSui = [fw1, fw2, .. fwn]^T

Fui indicates the importance of a term in the ith URL. Title is given a weight 2 throughout this paper. Similarly each URLs of a feedback session is represented and finally pseudo-documents are formed.

F fs =[f fs (w1), f f2(w2),… f fs (wn)]^T
F fs (w)= arg min{∑[f fs (w)- f ucm (w)]^2
-ʎ ∑[f fs (w)-~ f uc (w)]^2}

3. FUZZY C MEANS ALGORITHM

In Fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The Fuzzy C-Means algorithm (FCM) is used in the areas like computational geometry, data compression and vector quantization, pattern recognition and pattern classification. Fuzzy C-Mean (FCM) is an unsupervised clustering algorithm that has been applied to wide range of problems involving feature analysis, clustering and classifier design. The main features of that algorithm were the (i) use of a fuzzy local similarity measure, (ii) shielding of the algorithm from noise-related hypersensitivities. FCM clustering techniques are based on fuzzy behavior and they provide a technique which is natural for producing a clustering where membership weights have a natural interpretation but not probabilistic at all In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely too just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. FCM clustering which constitute the oldest component of software computing are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target recognition .More the data is near to the cluster center more is its membership towards the particular cluster center. The basic idea of fuzzy c-means is to find a fuzzy pseudo-partition to minimize the cost function .Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise .Fuzzy c-means algorithm uses the reciprocal of distances to decide the cluster centers. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After iteration of membership and cluster centers are updated according to the formula.

The FCM algorithm converges to a local minimum of the c-means functional. Hence, different initializations may lead to different results .The minimization of the c-means functional represents a nonlinear optimization problem that can be solved by using a variety of methods, including iterative minimization, simulated annealing or genetic algorithms

4. CLUSTERING PSEUDO-DOCUMENTS USING FUZZY C MEANS ALGORITHM

Fuzzy c-means (FCM) is a technique of clustering which allows one piece of pseudo documents data to to best semantics similarity of the pseudo terms in the documents. In this algorithm the same given data or pseudo documents does not go completely to a well definite cluster, based on the fuzzy membership function only the pseudo documents the cluster groups are formed in efficient manner with possible number of the groups at user feedback sessions. In the FCM approach, instead, the same given datum does not belong exclusively to a well-defined cluster, but it can be placed in a focal point way. In this case, the membership function follows a flatter line to designate that each datum may go to frequent clusters with different standards of the membership constant. In fuzzy clustering, each position has a degree of belong to clusters, as in belong to two or more clusters. It is the way to solve how the data with similar pseudo documents are clustered according fuzzy logic, rather than belong totally to just one cluster .Thus, points on the edge of a cluster might be in the cluster to a smaller degree than points in the midpoint of cluster. It is based on selection of the degree membership function.

5. EVALUATIONS BASED ON RESTRUCTURING WEB SEARCH RESULTS

User search goals are not predefined; hence evaluation of its inference is a big problem. If user search goals are inferred properly, the search results can be restructured properly. Thus an evaluation method 'Classified Average precision' is proposed. It helps to select the best cluster number it is an application of inferring user search goals. The inferred ones are represented by the feature representation of each URL in the search result. Then categorize them into a cluster cantered by the inferred search goals. This is done by choosing smallest distance between URL vector and user-search goal vectors. Thus restructuring is complete. From user click through logs, we get the relevant and irrelevant feedbacks.

Parameters for evaluation method is as follows

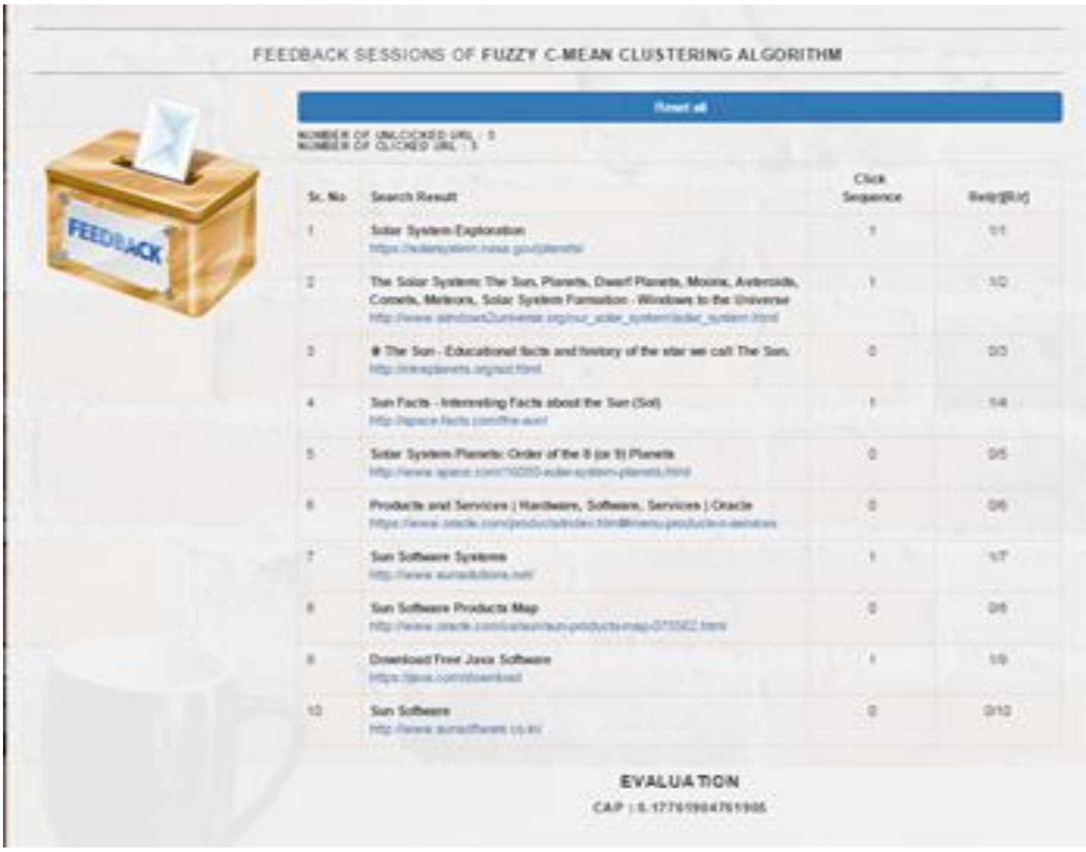
- Average Precision(AP)
- Voted Average Precision(VAP)
- Classified Average Precision(CAP)

VI. EXPERIMENTAL RESULTS

Before conclusion of the results and remarks of the paper the major part is the evaluation of the results from the experiments with classification results from each user search goal inference us a major problem , since user search goals are not predetermined and there is no ground truth. It is necessary to develop a metric to evaluate the performance of user search goal inference objectively. In this section finally measure the performance of the semantic similarity with FCM and accessible pseudo documents based clustering Measure the performance of the system with parameters like Classified Average Precision (CAP), Voted AP (VAP) which is the AP of the class including more clicks namely, risk to avoid classifying search results and average precision (AP).The corresponding AP, VAP, CAP and Risk values are measured Between user search Goal with cosine similarity and Use search Goal with semantic similarity values are shown.

Method →	K MEAN	FUZZY C MEAN
CAP (User A)	0.10	0.17
CAP (User z)	0.23	0.30
CAP (User B)	0.50	0.59
CAP (User C)	0.67	0.75
CAP (User D)	0.30	0.36
CAP (User E)	0.57	0.611
CAP (User F)	0.5	0.55
CAP (User G)	0.38	0.43
CAP (User H)	0.68	0.71
CAP (User I)	0.74	0.81

It shows that the User search goal With Semantic similarity(Fuzzy c mean) results are better than User search goal with cosine similarity measure(K mean).



FEEDBACK SESSIONS OF FUZZY C-MEAN CLUSTERING ALGORITHM

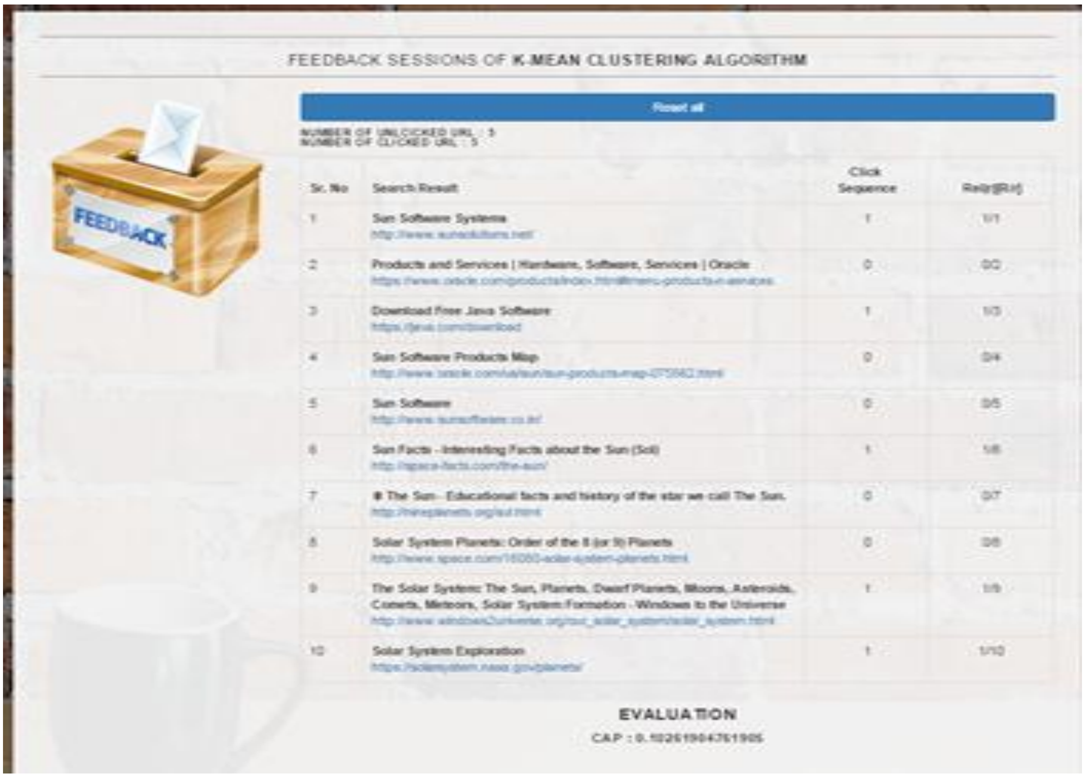
NUMBER OF UNCLICKED URL : 5
NUMBER OF CLICKED URL : 5

Sr. No	Search Result	Click Sequence	Rel@[R/i]
1	Solar System Exploration https://solarsystem.nasa.gov/planets/	1	1/1
2	The Solar System: The Sun, Planets, Dwarf Planets, Moons, Asteroids, Comets, Meteors, Solar System Formation - Windows to the Universe http://www.windows2universe.org/our_solar_system/solar_system.html	1	1/2
3	● The Sun - Educational facts and history of the star we call The Sun. http://nineplanets.org/sol.html	0	0/3
4	Sun Facts - Interesting Facts about the Sun (Sol) http://space-facts.com/the-sun/	1	1/4
5	Solar System Planets: Order of the 8 (or 9) Planets http://www.space.com/16080-solar-system-planets.html	0	0/5
6	Products and Services | Hardware, Software, Services | Oracle https://www.oracle.com/products/index.html#menu-products-n-services	0	0/6
7	Sun Software Systems http://www.sunsolutions.net/	1	1/7
8	Sun Software Products Map http://www.oracle.com/us/sun/sun-products-map-075562.html	0	0/8
9	Download Free Java Software https://java.com/download	1	1/9
10	Sun Software http://www.sunsoftware.co.in/	0	0/10

EVALUATION

CAP : 0.17761904761905

Fig.2. classified average precision (CAP) FCM method



FEEDBACK SESSIONS OF K-MEAN CLUSTERING ALGORITHM

NUMBER OF UNCLICKED URL : 5
NUMBER OF CLICKED URL : 5

Sr. No	Search Result	Click Sequence	Rel@[R/i]
1	Sun Software Systems http://www.sunsolutions.net/	1	1/1
2	Products and Services | Hardware, Software, Services | Oracle https://www.oracle.com/products/index.html#menu-products-n-services	0	0/2
3	Download Free Java Software https://java.com/download	1	1/3
4	Sun Software Products Map http://www.oracle.com/us/sun/sun-products-map-075562.html	0	0/4
5	Sun Software http://www.sunsoftware.co.in/	0	0/5
6	Sun Facts - Interesting Facts about the Sun (Sol) http://space-facts.com/the-sun/	1	1/6
7	● The Sun - Educational facts and history of the star we call The Sun. http://nineplanets.org/sol.html	0	0/7
8	Solar System Planets: Order of the 8 (or 9) Planets http://www.space.com/16080-solar-system-planets.html	0	0/8
9	The Solar System: The Sun, Planets, Dwarf Planets, Moons, Asteroids, Comets, Meteors, Solar System Formation - Windows to the Universe http://www.windows2universe.org/our_solar_system/solar_system.html	1	1/9
10	Solar System Exploration https://solarsystem.nasa.gov/planets/	1	1/10

EVALUATION

CAP : 0.10261904761905

Fig. 3 . classified average precision (CAP) K mean meth

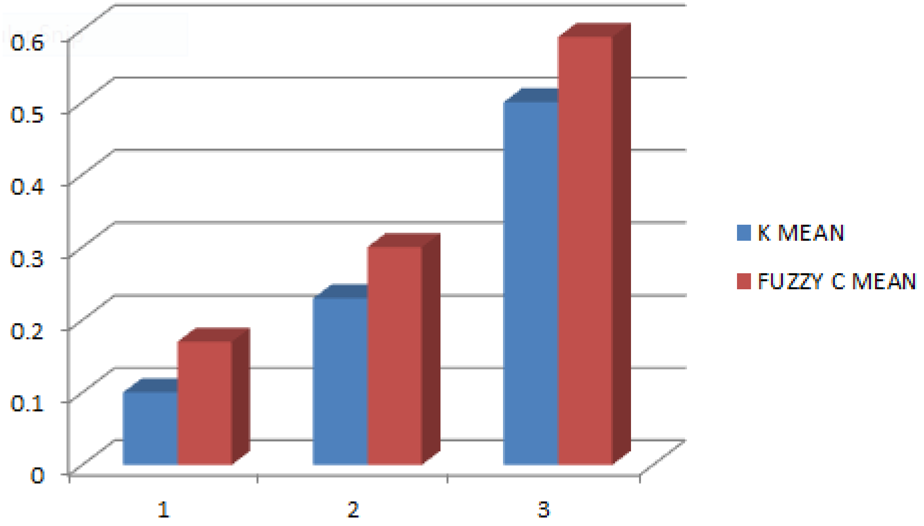


Fig 4. FCM AND K MEAN

V. CONCLUSION

In this paper Semantic similarity based FCM approach has USED to infer user search goals for a query by clustering its feedback sessions represented by pseudo documents. Primarily we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds with semantic similarity based measures and then pseudo documents can supplement the URLs with additional textual contents including the titles and snippets. Experimental results on client click-through logs from a commercial search engine reveal the efficiency of FCM method. In future work can be done in the following manner user can search the query in the feedback we automatically derive the optimal value to improve the feedback session results

REFERENCE

[1] J.I.Sheeba, Dr.K.Vivekanandan, “A Fuzzy Logic Based On Sentiment Classification”, 2014.

[2] 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.

[3] 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 ACMSIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’08), pp. 875-883, 2008..

[4] Vicenc¸ Torra, “Fuzzy c-means for fuzzy hierarchical clustering”, 2012.

[5] L.Suganya, Dr.B.Srinivasan, “Efficient Semantic Similarity Based Fcm For Inferring User Search Goals With Feedback Sessions”, 2013

[6] 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, 2005Eugene Agichtein, Eric Brill, Susan Dumais, “Improving Web Search Ranking by Incorporating User Behavior Information”, 2006

[7] Ji-Rong Wen, Jian-Yun Nie, Hong-Jiang Zhang, “Clustering User Queries of a Search Engine”, 2001.

[8] Rohini B. Mothe, V.S.Deshmukh, “A Novel Approach to Cluster Search Result based on Search Goals”, 2014

[9] S.Niveditha, T.Malathi, S.R.Sivaranjhani, “Efficient Information Retrieval using Fuzzy Self Construction Algorithm”, 2014

[10] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. 2006, “Learning user interaction models for predicting web search result preferences”. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’06. ACM, New York, NY, USA, 3–10.

[11] T. Joachims, “Evaluating Retrieval Performance Using Click through Data,” Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physics/Springer Verlag, 2003.

[12] T. Joachims, “Optimizing Search Engines Using Click through Data,” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’02), pp. 133-142, 2002.

[13] L..li, S. Otsuka, and M. Kitsuregawa “Query Recommendation Using Large-Scale Web Access Logs and Web Page Archive,” LNCS 5181, pp. 134–141, (2008), Springer-Verlag Berlin Heidelberg 2008