Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 1, January 2015, pg.275 – 283

RESEARCH ARTICLE

Categorization of Search Results using Feedback Sessions

Suraj M. Pagare[1], P. M. Yawalkar[2]

[1]Department of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, Savitribai Phule Pune university, India

[2] Department of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, Savitribai Phule Pune university, India

[1] surajmpagare75@gmail.com; [2] prashant25yawalkar@gmail.com

Abstract— Different users may have different user search goals for an ambiguous query while submitting it to search engine. In information retrieval, inference and analysis of user search goals are very useful. It can help to improve search engine relevance and user experience. So, for inferring user search goals, feedback sessions are used. Feedback sessions are constructed from search engine query logs. It consists of both clicked and un-clicked URLs and it reflects information needs of users easily. It clearly shows what a user requires and what does not care about. To describe the feedback session Pseudo documents are generated. Pseudo documents are used to approximate the goal texts in user minds. By using pseudo documents user search goals are discovered and displayed with some keywords for categorizing search results.

Keywords— feedback session, pseudo documents, user search goals, classified average precision

I. INTRODUCTION

In web search applications, user submit query to search engine to get knowledge or information about query. If the query is ambiguous, it can cover several topics. So, search results returned by search engine consist of several results. Some of the results that user does not want from search engine to be returned, are also get displayed. User has to find out particular results and then he clicks on URLs of the results to redirect to that website. For example if the query submitted by user is “kitkat”, some user may want to locate the official homepage of famous chocolate company or some may want to get information of the android version of operating system of mobile shown in Fig. 1.

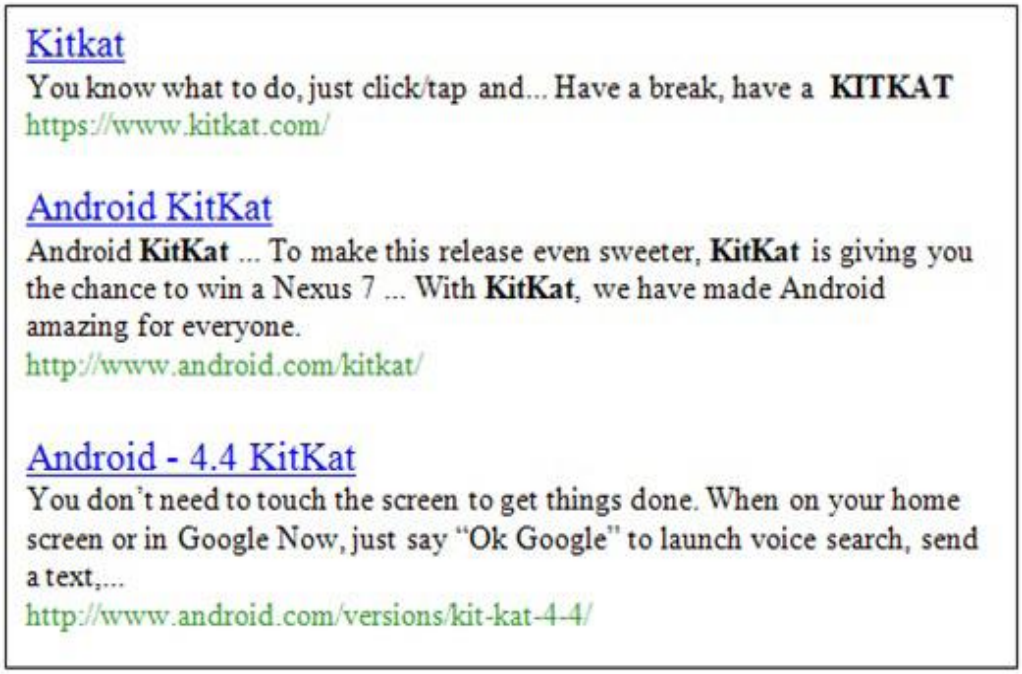


Fig. 1 Results returned by search engine

So, it is necessary to capture user search goals for a query in information retrieval. By using search goals user can find easily what he want to find. Search Goals System uses feedback sessions to infer user search goals. From these feedback sessions pseudo documents are generated. Finally, these pseudo documents are clustered by using clustering algorithms to get the user search goals. This can help user to find required results for a query easily.

II. LITERATURE SURVEY

There are various methods that are proposed for getting search goals. Some of these methods are classified in three types such as classification of query, reorganization of the search results and detection of session boundary. These methods are explained below.

A. *Classification of queries*

To infer search goals and intent of the queries some specific classes are predefined. According to these classes queries are classified. Lee et al. consider user goals as "Navigational" and "Informational" and queries are categorized into these two classes [2]. X. Li et al. defined the query intent as "Product intent" and "Job intent" and the queries are classified according to these predefined intents of the query [3]. So, these methods have limitation because finding predefined search goals and intent of the query is very difficult and improper. Search goals and intents of the query are varies for a different users and different queries.

B. *Reorganization of search results*

In this method, interesting aspects of queries are learned by considering and analyzing only the clicked URLs by user [4]. Clicked URLs are considered from click-through logs. This method has limitation because number of clicked URLs by users may be small. Some methods analyze search results returned by search engine when query is submitted by user [5] [6]. These methods do not consider the feedback of user. Many noisy search results that are not clicked by users are also get analyzed. So, search goals are not inferred precisely by using these methods.

C. *Detection of session boundary*

R. Jones et al. predicted the mission boundaries and goal boundaries initially [7]. These boundaries are used to segment the query logs hierarchically. Here, search goal is an atomic information need that results in one or more queries and search mission as a related set of information needs those results in one or more goals. This method only detect whether a pair of queries belong to same mission or same goal. So, it does not consider the goal in detail.

The methods to infer search goals for query are proposed in [1], [8]. Here, feedback sessions are used to infer user search goals. Pseudo documents are formed from feedback session and these pseudo documents are clustered by using K – means algorithm. Each cluster of pseudo documents represents one user search goal.

These user search goals are used to categorize search results. But the problem in this approach is some clicked URLs by user are categorized in different category. To overcome from the drawbacks of the previous methods Search Goals System is proposed.

III. SEARCH GOALS SYSTEM

The overall system architecture of system is shown in Fig. 2. Main components of the system are Search Engine, Creation of Feedback Session, Generation of Pseudo Documents, Concluding User Search Goals, and Categorizing Search Results. These are explained below.

A. Search Engine

The User can access the Search Goals System by login to system. User submit query to system to get information about the query. System directly forwards query to search engine i.e Google. Google apis are used to get the results for a query directly from Google. These results are stored in database for further processing. The results retuned by the system to user are the normal results. i.e. there is no categorization. Each result has URL, URL title and URL snippet, where snippet is the small description about that URL.

B. Creation of feedback session

When user enters a query in search engine, it returns various results. From these results user clicked on the URLs that user wants to visit. While user searching the results he scans and examines the URLs from top to bottom. So the URLs that are not related to the topic that user actually want to visit, are skipped by the user. So, the URLs from first URL to last URL that was clicked by the user are considered in feedback session. The feedback session for the query "kitkat" is shown in Fig. 3.

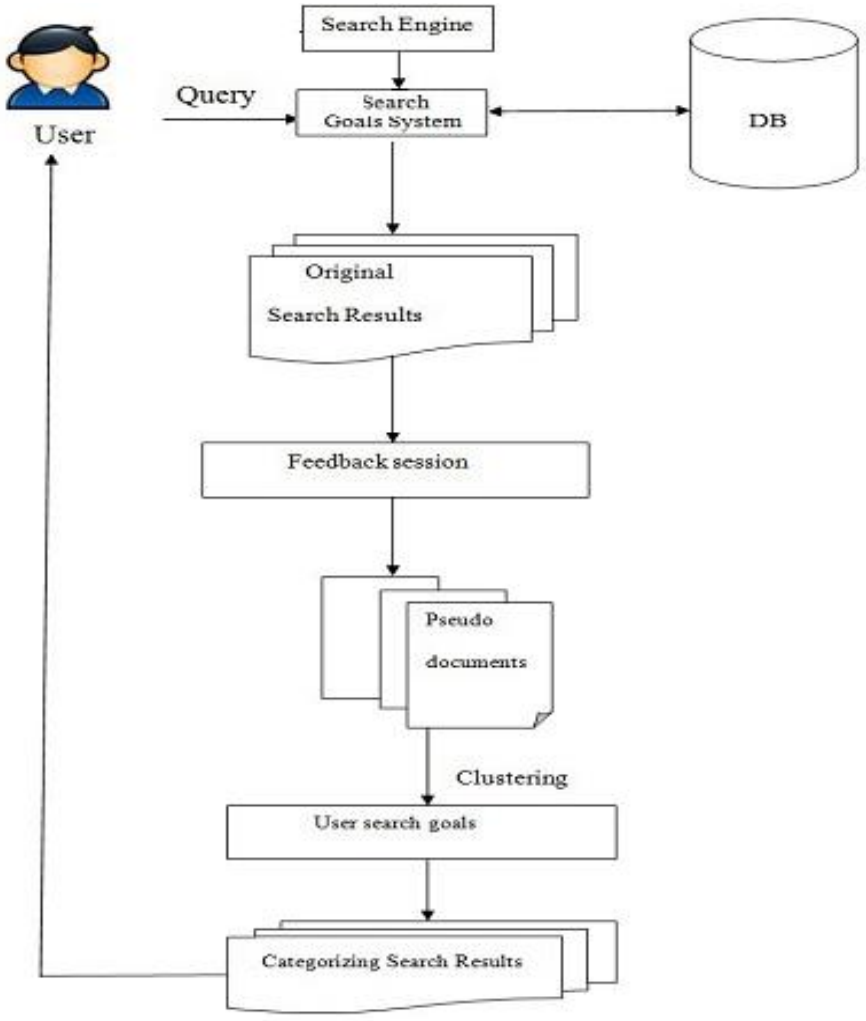
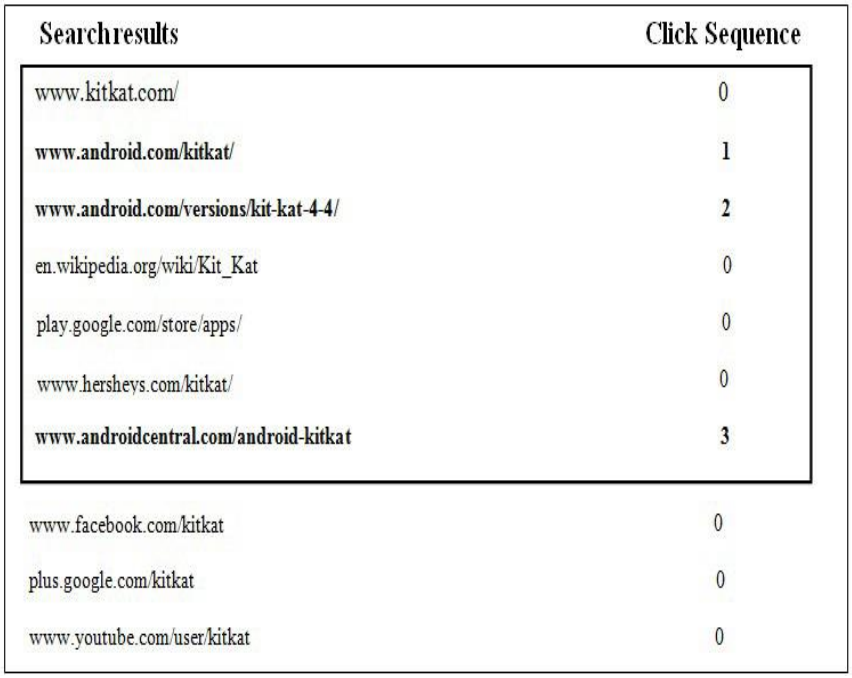


Fig.2 Architecture of the system



Search results	Click Sequence
www.kitkat.com/	0
www.android.com/kitkat/	**1**
www.android.com/versions/kit-kat-4-4/	**2**
en.wikipedia.org/wiki/Kit_Kat	0
play.google.com/store/apps/	0
www.hersheys.com/kitkat/	0
www.androidcentral.com/android-kitkat	**3**
www.facebook.com/kitkat	0
plus.google.com/kitkat	0
www.youtube.com/user/kitkat	0

Fig. 3 Feedback session for query "kitkat"

Left part of Fig. 3 shows the ten results for the query "kitkat" and the right part shows the user's click sequence where "0" means un-clicked ones. The rectangle region is the feedback session for the query "kitkat" entered by the user. It consist of seven URLs out of the 10 the URLs. These seven URLs consist of three clicked URLs and 4 un-clicked URLs. So, in proposed Search Goals system the URLs those were not actually clicked by the user are also considered. Feedback session can tell what user wants and what does not want. So, for categorizing the search results feedback session is more efficient than only considering the clicked URLs.

C. Generation of pseudo documents

It is unsuitable to directly use feedback sessions for inferring user search goals because feedback sessions varies for different click-through and queries by users. Some representation method is needed to describe feedback sessions. So, pseudo documents are used to represent it.

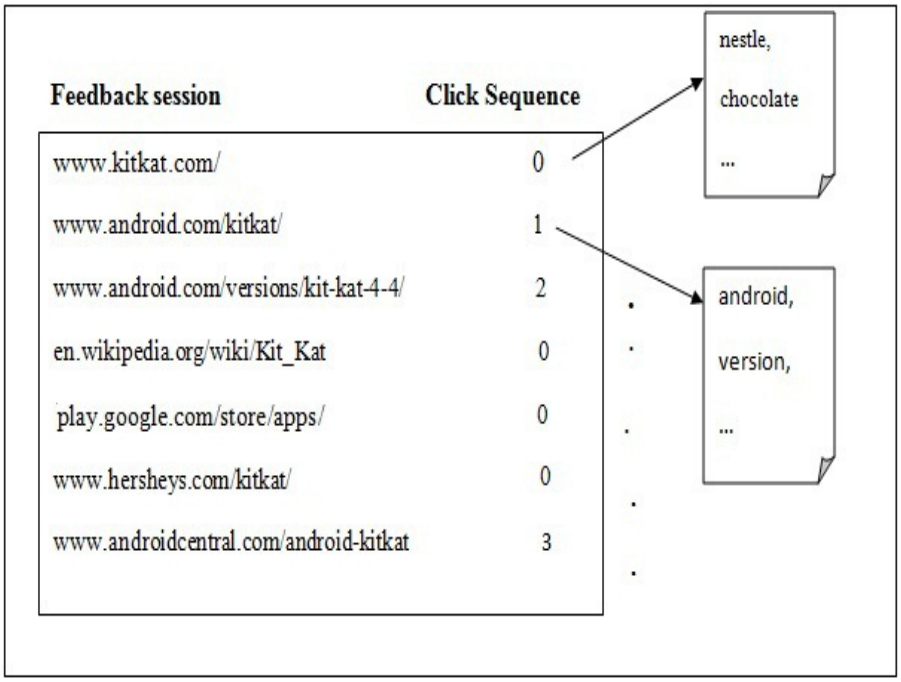


Fig. 4 Pseudo Documents for feedback session of query “kitkat”

For a query, users will usually have some keywords in their minds that representing their interests. They use these keywords to determine whether a result can satisfy their needs. These keywords are called as "goal texts". However, although goal texts can reflect user information needs, they are latent and not expressed explicitly. Therefore, pseudo documents are used to approximate goal texts. Thus, pseudo documents can be used to infer user search goals. For each feedback session pseudo documents are created. From these pseudo documents, user search goals inferred more efficiently. Each URL in feedback session has title and snippet. These titles and snippets are extracted from the results and some pre-processing such as stemming, removal of stop words is applied on these texts and then stored in documents called as pseudo documents. These documents consists only the some words of URL's title and snippet. The pseudo documents for the feedback session of the query "kitkat" are shown in Fig. 4.

D. Concluding user search goals

Pseudo documents are nothing but collection of words. Clustering techniques are applied on these pseudo documents. Clusters of pseudo documents are generated and each cluster represents one user search goal for a query. The terms with higher frequency in the cluster are called as user search goal. For a query there may be one or more user search goals.

E. Categorizing search results

Search engines always returns various search results for a query, it is necessary to organize them to make it easier for users to find out what they want. Categorizing web search results is an application of inferring user search goals. The user search goals are used for categorization of search results. Search results that are related to particular search goals are grouped in a category and the category is highlighted by the keywords of search goals. Search results are categorized according to number of search goals for a query. So, by categorization of search results user can find easily what he wants to find.

IV. ALGORITHM FOR CLUSTERING PSEUDO DOCUMENTS

The algorithm takes pseudo documents as an input. The documents that are of clicked URLs are added in one cluster and based on the similar terms in the documents of un-clicked URLs are also clustered in this cluster. So, clicked and un-clicked URLs that are of same search goals are clustered together. The remaining documents are clustered in other clusters. Algorithm is explained below.

Input: pseudo documents d1, d2,…, dn.
Output: clusters of pseudo documents

1. The documents of clicked URLs are clustered in one cluster.
2. Compare the documents of un-clicked URLs with the documents in cluster formed in step 1, if term in un-clicked URL document is found in the one of the clicked document then that document is added to this cluster.
3. The documents that are not added in step 2, compare the documents to each other, if they are of same search goal are add in separate cluster.
4. Repeat step 3 all un-added documents.
5. If the documents are not getting added in cluster in step 4, then add each individual document in separate cluster.

V. RESULT ANALYSIS

The results obtained from Search Goals System that uses algorithm based on clicked URLs are compared with the method that uses K- means algorithm for clustering pseudo documents. Fig. 5 shows the feedback session of query "kitkat" is categorized in two categories and rel() is binary function which is used to calculate relevance of rank given to clicked URL in category. Results are compared with two criterion such as VAP (Voted Average Precision) and CAP (Classified Average Precision).

Category 1	Click Sequence	rel = $\frac{\text{click seq.}}{\text{rank in category}}$
www.android.com/kitkat/	1	1/1
www.android.com/versions/kit-kat-4-4/	2	2/2
play.google.com/store/apps/	0	0
www.androidcentral.com/android-kitkat	3	3/4
Category 2		
www.hersheys.com/kitkat/	0	0
www.kitkat.com/	0	0
en.wikipedia.org/wiki/Kit_Kat	0	0

Fig. 5 rel function for feedback session of query "kitkat"

VAP of categories is calculated as,

VAP of category 1 = 1/4 (1 + 1+ 3/4)
= 0.916
VAP of category 2 = 0

And CAP is calculated by,

CAP = VAP * (1-Risk)

Where, Risk is used when clicked URLs are not categorized in same category. In Search Goals System it is always equal to 0. The computed values of VAP and CAP for some queries using K-means algorithm and Algorithm based on clicked URLs are shown in Table 1 and Table 2 respectively.

TABLE I
COMPARISON OF VAP VALUES FOR SOME QUERIES

Query ID	Query	K-means algorithm	Algorithm based on clicked URLs
1	kitkat	0.896	0.985
2	kpple	0.83	0.89
3	sun	0.793	0.823
4	apache	0.853	0.914
5	jellybean	0.775	0.814

TABLE III
COMPARISON OF CAP VALUES FOR SOME QUERIES

Query ID	Query	K-means algorithm	Algorithm based on clicked URLs
1	kitkat	0.816	0.985
2	apple	0.673	0.89
3	sun	0.692	0.823
4	apache	0.797	0.914
5	jellybean	0.675	0.814

Table 3 shows keywords that are obtained for queries from the system to display user search goals for categorizing search results.

TABLE IIIII
KEYWORDS OBTAINED FOR QUERIES FROM SYSTEM TO DISPLAY SEARCH GOALS

Query	Search goals
kitkat	android, device
	wafer, chocolate
apple	iphone, ipad
	news, aapl
sun	hot, solar
	news
apache	software, http
	TVS, bike
jellybean	android, version
	candy

Fig. 6 shows comparison of VAP values obtained by using Method-I and Method-II. Fig. 7 shows comparison of CAP values obtained by using Method-I and Method-II.

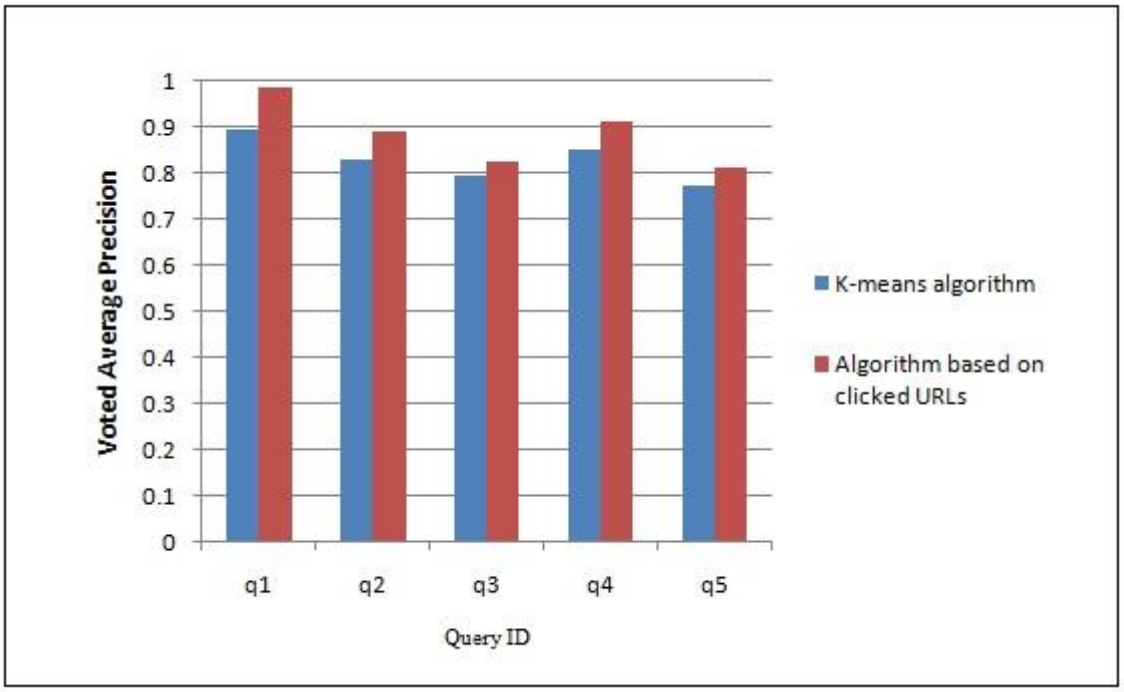


Fig. 6 Comparison of VAP of Method-I and Method-II

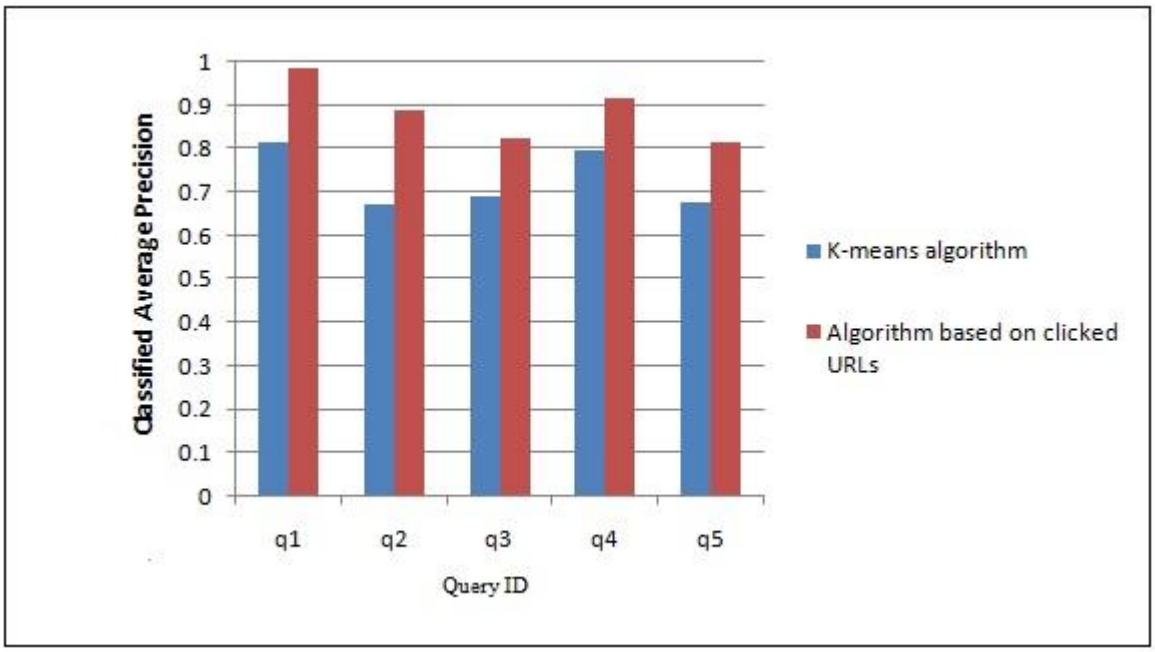


Fig. 7 Comparison of CAP of Method-I and Method-II

VI. CONCLUSION

Thus, Search Goals system is developed to get the search goals for a query. System uses feedback session for inferring search goals. The clicked URLs in feedback session shows what user wants and un-clicked URLs shows what user does not want about query. So, feedback session is efficient to get the search goals for a query. Pseudo documents are generated from feedback session and clustered to get the keywords that depict each search goal for query. If search goals of a query for a user are already inferred then the system directly shows the categorized search results to user. So, user feedback is considered by the system and by using keywords user can find easily what he wants to find.

REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, and Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions ", *IEEE Transactions on Knowledge and Data Engineering*, VOL. 25, NO.3, pp.502-513, 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] 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.

[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. 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.

[6] 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.

[7] 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.

[8] Suraj M. Pagare and P. M. Yawalkar, \Information Retrieval By mining Query Logs", Third Post Graduate Symposium for Computer Engineering, cPG-CON 2014, 28-29 march 2014.

