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

Learning to Cluster Feedback Session for Identification of User Search Objective

Manjiri M. Kokate[1], Poonam D. Lambhate[2]

[1]Computer Engineering Department, JSCOE, Pune University, India

[2] Head of Information Technology Department, JSCOE, Pune University, India

Abstract: ***In web search applications, the WWW is a huge resource for people. This resource uses search engines to search the information. For this purpose the queries are submitted to search engine to represent the needs of the users. Sometimes queries may not exactly represent the actual objective of user. There may be ambiguous query for different users. There are mainly two aspects for improving search engine relevance 1) identification of user search objectives and 2) analysis of user search objectives. In this paper, we propose a new method to identify user search objective by analysis process. This process is operated on search engine query logs. First, we propose a skeleton for discovering different user search objective for a query by clustering the proposed method of feedback sessions. Feedback sessions are calculated by considering user click-through logs. By using these users click-through logs the information needs of users can be efficiently identified. Second, we propose a new approach for generating pseudo-documents for best representation of the feedback sessions for hierarchical clustering. Hierarchical clustering gives the relationship between all the keywords in the corresponding cluster. Finally, we propose a new technique for actual evaluation of the performance of identified user search objectives. For the better effectiveness of our proposed methods results are presented using user click-through logs.***

Keywords: Web mining, pseudo documents, information retrieval, Web text analysis, Searching, Hierarchical clustering

1. Introduction

In the recent years searching information from the web within time has more importance. The different users have different search objectives when they submits query to the search engine. Sometimes queries may not exactly represent the actual objective of user. As there are so many ambiguous queries and different users may want to get information on different aspects when they submit the same query. In fig 1 it shows that user has entered query" the dawn" which has ambiguous meaning. One can want information of Spacecraft and another may want to locate home page of Pakistan English newspaper. User search objective is the information on different aspects of a query that user groups want to obtain. There are mainly two aspects for improving search engine relevance 1) inference of user search objective and 2) analysis of user search objective.

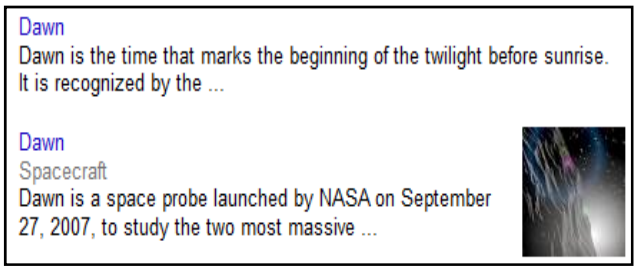

Figure 1: Example of different user objectives for" the dawn" query

Both aspects have some advantages summarized as follows.

First, we can restructure web search results obtained [1], [11] according to user search objectives. Restructuring can be done by grouping the search results with the same search objective. By using this, we can easily find out users with different search objectives. And also we can find what exactly user want. Second, We are representing user search objectives by some keywords. These keywords can be utilized in query recommendation [6], [12], [13].By using recommendation, we find out the suggestions about queries which can help users to form their queries more detail. Third, the search objectives are distributed for the use of reranking of documents. These advantages are categorized into three classes: Query classification, restructuring of search results and to detect limit of session. In the first class, users attempt to infer user search objective and find out in which class this query is included. Lee et al.[9] consider user goals as "Navigational" and "Informational" .Then they categorize queries into these two classes. Li et al. [4] define query intents as "Product intent" and "Job intent" .Then they classify queries into these two defined intents. But finding exact class is very difficult and impractical. In the second class, users reorganize the search results. Wang and Zhai [1] learn aspects of queries by analyzing the clicked URLs which are directly taken from user click-through logs for the organization of search results. In this method number of

different clicked URLs of a query may be small. This is limitation of this method. Other methods analyze the search results which are returned by the search engine when a query is submitted [11].Clicked as well as unclicked URLs are considered. Therefore, this kind of methods cannot identify user search objective. In the third class, users are supposed to detect session boundaries. Jones and Klinkner [5] predict goal and mission boundaries to hierarchically segment query logs. However, their method only identifies whether a pair of queries belong to the same goal or mission. It does not care about what the goal is in detail. There are different kinds of technique for improving the search results and finding the exact users need. We first propose a approach to identify user search objective for a query by clustering our proposed feedback sessions. The feedback session is the series of both clicked and unclicked URLs. For keeping the feedback we have use Click through logs. In the existing feedback system it contains limited number of URLs. In this paper the numbers of URLs are increased in the Click through log. For better understanding of user search objective, we propose a optimization method. In this method, feedback sessions are mapped to pseudo-documents. By this mapping process one can efficiently reflect user information needs. At last, we group these pseudo documents to identify user search objective by using Hierarchical clustering. By using Hierarchical clustering we can find out the hierarchy among all the keywords. Finally the evaluation of clustering is also an important problem. To solve this problem we propose a evaluation criterion classified average precision (CAP).CAP is used to evaluate the performance of the restructured web search results.

2. Related Work

H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li [6],put forth a two step novel context-aware query suggestion approach . In the model-learning phase, to deal with data sparseness, summarization of queries into concepts is done by clustering a click-through compound. Then, a concept series suffix tree is created from session data as the query suggestion model. At the time of online query suggestion step, a user's search framework is imprisoned by mapping the query sequence, suggested by the user to a series of concepts. Our approach, recommends queries to the user in a context-aware mode. X. Wang and C.-X Zhai, [1] proposed that clustering of search results is a best way to organize search results. It allows to group similar type of documents. Developing search results are easy but the organization of search results is critical. They proposed learning interesting aspects of a topic from web search logs and generation of cluster labels which has specific meaning. All this implementation is performed on search engine log data. U. Lee, Z. Liu, and J. Cho [9] proposed Automatic identification of a user goal for a Web query. Through a human subject study, they showed that about 60% of the queries considered have expected goals autonomous of users. This Study more suggested that for the other 40% of the queries with anticipated goals, a search engine may be able to utilize simple procedures to identify and handle them individually. Then they proposed two categories of effective features in identifying the goal of a query: past user-click behavior and anchor-link distribution. Their evaluation showed that using a combination of the proposed features can correctly identify the goals for 90% of the queries studied. Results showed that features clearly outperformed the existing features. S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder [7], evaluated three differing approaches to topical web query classification and found that training explicitly from classified queries outperforms bridging document taxonomy for training by as much as 48% in F1. They also showed that pre-retrieval classification using only the query string can provide surprisingly effective results, enabling adjustments to the retrieval process to improve effectiveness and efficiency. Fusion of multiple approaches did not yield improved performance. Hua-Jun Zeng, Zheng Chen, QiCai He, Jinwen Ma[11], reformalized the search result clustering problem as a supervised salient phrase ranking problem. Several properties, as well as several regression models, are proposed to calculate salience score for salient phrase. Experimental results generate correct clusters with short names (thus hopefully is more readable), thus could improve users' browsing efficiency through search result. They further investigated several problems on search result clustering. First, extract syntactic features for keywords and phrases to assist the salient phrase ranking. Second, current clustering is still a flat clustering method. Hierarchical structure of search results is necessary for more efficient browsing. Third, some external taxonomies such as Web directories contains much knowledge which is familiar to Web users, thus a combination of classification and clustering might be helpful in this application. X. Li, Y.-Y Wang, and A. Acero [4], presented a semi-supervised learning approach to query intent classification with the use of click graphs. The work differs from previous works on query classification in that we aim at drastically expanding the training data in an attempt to improve classification performance. This ALLOWS USING relatively unbiased features, namely words/phrases in queries themselves, despite their sparseness. They achieved this goal by mining a large amount of click-through data, and inferring class memberships of unlabeled queries from those of labeled ones in a principled fashion. Moreover, we used content-based classification to regularize this learning process, and jointly performed graph-based learning and content-based learning in a unified framework. R. Jones and K.L. Klinkner [5], shown that a diverse set of syntactic, temporal, query log and web search features in combination can predict goal and mission boundaries well..Classifiers achieve at least 89% accuracy in all four tasks, and over 91% in all but one task, matching within the same goal. Additionally, shown that the task of matching queries within the same interleaved goal or mission is harder than identifying boundaries. This may indicate that the best approach to clustering queries within the same goal or mission may build on first identifying the boundaries, then matching subsequent queries to existing segments. It may also be effective to use multi-task machine learning to join the tasks of identifying mission and goal boundaries together. The utility of adopting a hierarchical model for the grouping of user queries will allow us to more easily model what type of task the user may be doing when querying, e.g. is the user performing a series of searches with information needs which are the same, or are the information needs only peripherally related? This may help to determine when the user is performing a more complicated task, vs. a simpler task. Including the interleaving in the model allows more accurately measure the length of time or number of queries a user needs to complete tasks. If ignored the fact that

a more involved task may be interrupted with other needs for information, loses the ability to model these more involved tasks. The work sets the stage for evaluating search engines, not on a per-query basis, but on the basis of user tasks.

3. Implementation

A. Proposed Framework

Fig. 1 shows the framework of our approach. The proposed system framework is an enhancement to techniques introduced in [1].The main motive of proposed system is to identify search objective and return search results within few time. In feedback session it keeps more number of URLs as compared with previous method. New framework makes use Feedback session and Hierarchical clustering.

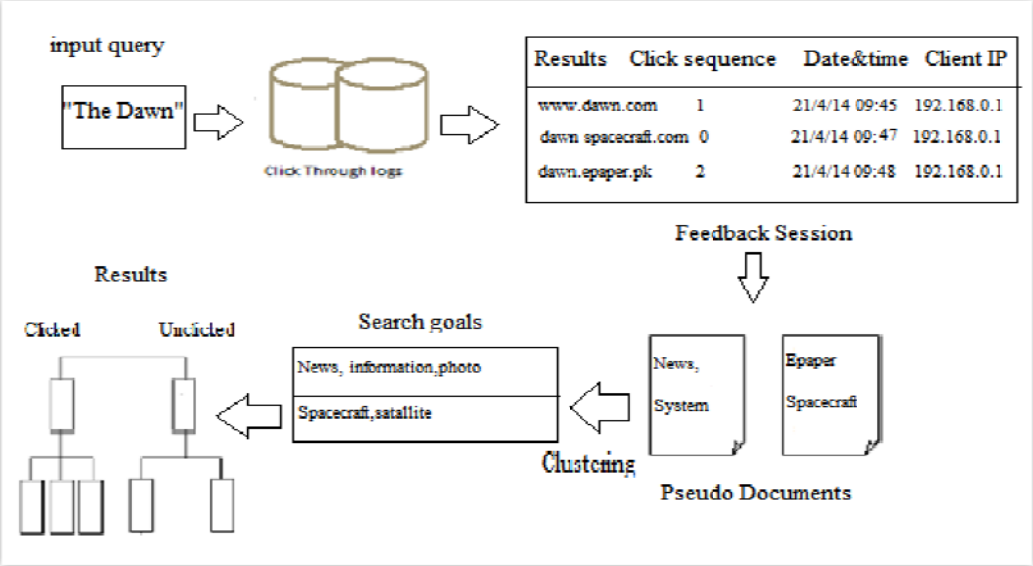

Figure 2: Proposed System Model (Framework)

B. Detail of System Model

Proposed framework consists of different parts as follows part 1: In the first step queries are submitted to the search engine and all these queries are stored into Click through logs. Different user's has different aspects for searching information. Feedback sessions of a query are first extracted from user click-through logs. In the feedback session different types of information is stored like "Click sequence, Date/Time, IP address". The feedback session is useful for inferring user search objective. In the proposed system we can consider number of feedback session for analysis. These feedback sessions are mapped to the pseudo-documents. Pseudo-documents consist of keywords which represent the user's information need. By using these pseudo-documents we can easily find out the user search objectives inferred by clustering. Initially we do not know the exact number of user search objective so that several different values are tried. From these values optimal value will be determined by the feedback from the part 2.

Part 2: In this the clustering of the feedback session takes place. One can cluster the retrieved results but in the proposed system feedback sessions are clustered. Clustering of the feedback session is more efficient than the clustering of the retrieved results. Finally Then, we evaluate the performance of retrieved results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the part 1.

In this paper focus is on the feedback sessions, pseudo documents, hierarchical clustering. For identification of user search objective following procedure we have to follow:
Procedure

1. Feedback Session:
When user submits the query to the search engine we do not know about exactly need of user. So for the identification of user search objective we are maintaining feedback session. Feedback session is the series of queries and some clicked search results. In the previous method it keeps the record of limited URLs. But in the proposed method number of URLs is increased. By using feedback session we can easily identify user search objective. In this we are considering feedback session for only single query. Therefore for the single query feedback session is also single. The proposed framework consists of feedback session with both clicked and unclicked URLs and they ends with the last URL which was clicked in single session. Before the last click all previous URLs have been scanned and they are evaluated by users. Feedback session consists of 0 which indicates that corresponding URL is unclicked. In feedback session only three URLs are shown. In that one URL is clicked and two URL's are unclicked. Clicked URL indicates what exactly user need. Unclicked URL shows that user doesn't want the information related to this URL. Therefore, for identification of user search objective it is best way to analyze the feedback session then the analyzed search results or clicked URLs.

2. Mapping of feedback session to pseudo document:
There different methods for representation of feedback session. One of the methods is known as "Binary Vector Method" to represent feedback session. When the query "the

sun" submitted to the search engine 0 represents unclicked in the click sequence. For example, binary vector [0110001] can be used to represent feedback session [14]. In that 1 represents "clicked" and 0 represents "unclicked". The binary vector method has disadvantage that doesn't give enough information to identify user search objective.

New method is proposed for representation of feedback session. In this method the feedback session is mapped to pseudo documents. Pseudo documents consist of keyword to determine whether document can satisfy their need. Hence pseudo documents can be used to identify user search Objective.

3. Algorithmic strategy for proposed framework:

1) Generate pseudo documents of URL from feedback session.
2) Pseudo document contains all the keywords from given web page's title and description.
3) Also all the stop words and stemming words are removed.
4) Apply K-means clustering algorithm to form a group of relevant keywords from pseudo documents so that each cluster represents one user search objective.
5) Next step is to organize words from one cluster into topic-subtopic hierarchy by using clustering.

4. Mathematical Model

Identify the processes as P.
S= {I, O, P....}
P= {WC, IE}
* WC is Web Crawler
* IE is Information Extraction.
Where
S-Proposed System
I= {Q | where Q is input query string}
O= {Ri | Ri is output strings that gives relevant search results}
WC= {URL, MAX, CP}
MAX is maximum no of URL to be crawled
CP is output of Web Crawler which is Crawled web pages.
IE= {CP, NLP Techniques, Info}
CP is input which is crawled web pages given to IE

5. Results and Discussion

In this paper we studied some problems associated with feedback session record. Feedback session can record limited number of URLs. So that user can analyse few URLs. In this case we have increased the size of feedback session. So that user can analyse 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.

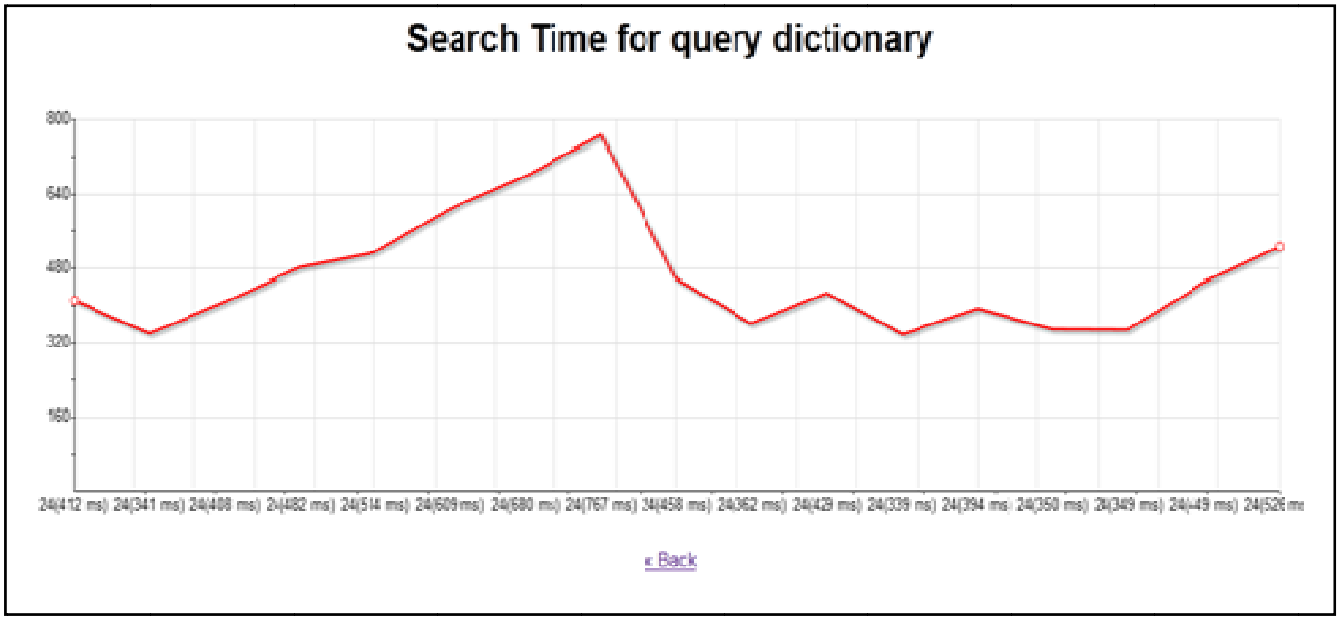

Figure 2: Graph showing Retrieval time for user query

A graph is plotted as number of results versus retrieval time (milliseconds) .The graph shows retrieval time taken for query "dictionary".

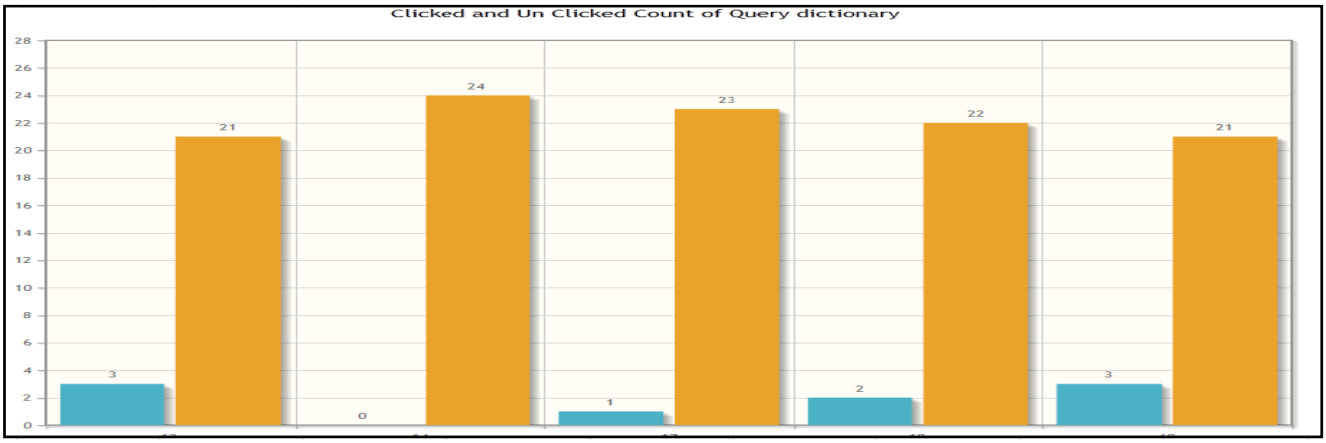

Figure 3: Graph showing Clicked and unclicked count for user query

The results obtained for user input query can be shown as a bar graph. The clicked count and unclicked count are plotted as sessions versus count for clicked and unclicked count

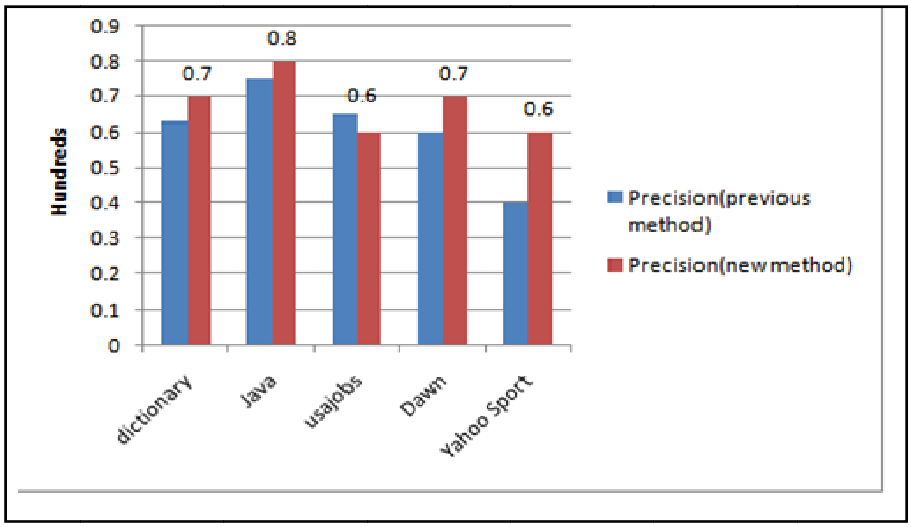

Figure 4: Precision comparison for previous method and proposed method

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.

4. Acknowledgment

I wish to express my sincere thanks and deep gratitude towards Dr. M. G Jadhav [Principal JSCOE] and my guide Prof. P. D. Lambhate for her guidance, valuable suggestions and constant encouragement in all phases. I am highly indebted to her help in solving my difficulties which came across whole Paper work. Finally I extend my sincere thanks to respected Head of the department Prof. S. M. Shinde and Prof. M .D. Ingle [P.G Co-Ordinator] and all the staff members for their kind support and encouragement for this paper. Last but not the least, I wish to thank my friends.

References

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[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] 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.
[9] 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.
[10] 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.
[11] 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.
[12] R. Baeza-Yates, C. Hurtdo, 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
[13] 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 "IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013.
[14] 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- Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

Author Profile

Manjiri M. Kokate received the B.E. degree in Information Technology from Pune University; India in 2011.She is pursuing M.E.in Computer Engg. at University of Pune. Her research interests include data mining, Information Retrieval, Web mining.

Poonam D. Lambhate received the B.E. and M.E. degrees in Computer Engineering from Solapur University and Shivaji University, India. She is currently Head of Information Technology Department, JSCOE, Pune University, India. Her research interests include Image processing, Information retrieval, swarm intelligence; data mining. She is member of ISTE.