International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Inferring User Search Goals Using Feedback Session

Harshada P. Bhambure[1], Mandar Mokashi[2]

[1, 2]Department of Computer Engineering, KJCOEMR,Pune, India

Abstract: ***The aim of topic is to discover the number of different user search goals for a query and representing each goal with some keywords. We first infer user search goals for a query by clustering feedback sessions. For that, we use a concept of pseudo document, which is the revised version of feedback session. At the end, we cluster these pseudo-documents to infer user search goals and represent them with some keywords. Since the evaluation of clustering is also an important problem, we used evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. The clustering is done by bisecting k means where in the existing system it is done by k means clustering. The new algorithm increases the efficiency of result. After the segmented result formation, the result in the every segment is reorganized as per number of clicks of URLs. The link which is clicked more number of times will appear at first location in the segment. This reduces the time requirement for searching.***

Keywords: Classified Average Precision (CAP); Clustering; Feedback session; Pseudo-document; Segmented Result; User Goals

1. Introduction

Web mining is also one of the applications of data mining techniques to extract data from web. Web mining is basically divided into three types, web usage mining, web content mining and web structure mining. Web usage mining is used to find the requirements of user on the internet. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data. In the web structure mining graph theory is used to represent the hyperlink structure of internet. Web content mining is the mining, extraction and integration of useful data from web page content.

In the existing system, the user enters the desired query and result get appears in the list format. In which there is no bifurcation as per different goals of the query. For every user there may be several goals for several users. So that time required to find the exact result increases.

Inferring and analysis are two important aspects to improve the user search results. Every time when user enters a query he has different goals in mind. To identify that goal the inferring technique is used and to check its relevance it performs the analysis of result. When the user enters the query “paper” the search engine will give different results. The results may be based on the links which gives the details of papers or links related to newspapers. In this, the search engine doesn’t know about the user goal therefore it gives the different links of different domains. So this method does not satisfy the user requirements. Therefore there is necessity to find out the user interest and distribute the results as per goals. To categorize the goals the inferring technique is used. In addition to this, the organization of segmented result is also necessary. This organization will keep the previously clicked queries at first location.

The need of the proposed method is to find the exact goal of the query. This will improve the result and help the user to find the exact document they want. For this proposed method concepts of feedback session and pseudo document is used. Bisecting k mean clustering is used to divide the result into the different categories as per there domains. To organise the categorized result, number of clicks of links are stored in the feedback session. As per these number of clicks the results are reorganised in the segment. So that the link which is clicked more number of times will appear at first location in the categorized list.

2. Literature Review

R. Baeza-Yates, C. Hurtado, and M. Mendoza[1] suggests that, the search engine gives the list of related results. These results are based on the previously searched queries or such technique can be used to tune or redirect the user. In this method the clustering algorithm is used. The clustering is done on the basis of previously fired queries. It clusters the semantically similar queries. It does not only gives the clustered data but it also ranks the suggested list of result. The ranking is done on the basis of two conditions,

1. Similarity of a queries to the input query
2. Observation that measures the attention of the user attracted towards the result of the query.

The combination of both these conditions measures the user interests. In the given algorithm, query session is considered for giving the result. The query session also considers the rank of clicked URL. The relevance ranking is measured by using two components similarity of query and support of query. Query clustering can be done in two steps,

1. Calculation of Query similarity

The query similarity between two queries can be calculated by creating term weight vector for each query. Term is weighted by considering the number of occurrences and number of clicks of the documents in which the term is appeared. Different techniques can be used to check the vector similarity. In this paper cosine function is used for it.

2. Computing the clusters.

The k means clustering algorithm is used in the paper for computation. In the k means algorithm, different clusters are considered as single node. It results into suggesting related

queries using clustering process on the query log. But this method has following disadvantages.

Disadvantages:
1) The method is not useful for large amount of log data.
2) Query expansion (that is completion of user query by considering the logs) is not used.

Advantages:
1) Considers the query logs.
2) Ranking is done on result

Doug Beeferman, Adam Berger[2] proposed algorithm in which the method of agglomerative clustering is used on the bipartite graphs of unique queries and unique URLs. Bipartite graph is a graph in which vertices are divided into two disjoint sets A and B, such that every vertex in A is connected with one vertex in B. This bipartite graph is made up of two parts, on one side white vertices which represents the different queries and on the other side black vertices which represents the related URLs. These black and white vertices are mapped with each other as per their relevance. The agglomerative clustering is done on these black and white vertices.

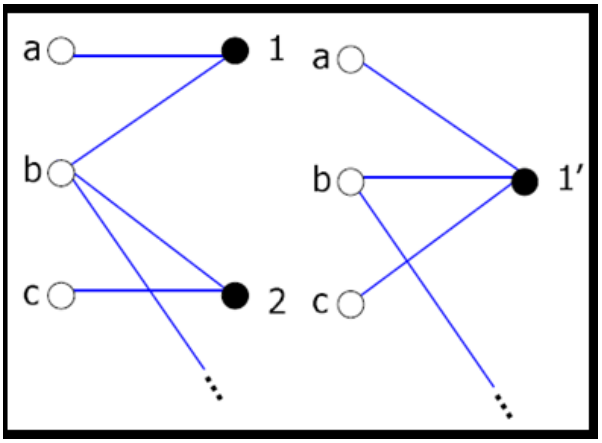

Figure 1: Agglomerative Clustering [2]

This paper results into different clusters of URLs, based on different queries. But one issue raised in this paper is not solved in it. The theory is not proposed to prove the combination of content ignorant and content aware clustering. The result is experimented with three methods for creating list. First method is baseline where standard Lycos query suggestion algorithm is used. The second method is full replacement where all list get replaced with frequently accessed search requests. In the third method Hybrid approach is used.

Disadvantage:
1) It is content ignorant.
2) Only click through logs are considered

Advantages:
1) Increases performance due to connectivity between queries

Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, Hang Li[3] proposed algorithm in which overcomes major disadvantage of content ignorance. This can be overcome in two steps. The first step is called as the offline model learning step. This step is used for data distribution where clustering is done through bipartite graph on the queries. Queries are represented in the form of concepts. Basically the sequence of queries is considered and if queries are fired in particular sequence by many users then they are considered as of the same concept.

The next step is online query suggestion model. In this context of user search queries is taken into consideration. The context is captured along with the query sequence submitted by user. Sequence suffix tree is used to give the result to the user. The figure given below represents the "Context Aware Approach". In the online model learning step, click through bipartite graphs are constructed from log session. It mines the concept from it and builds a concept sequence suffix tree from session in the data. The online query suggestion step matches the concept of current users query with concepts of sequence suffix tree and suggests the most matched query. The major advantage of this is that it is tested for large amount logs and queries.

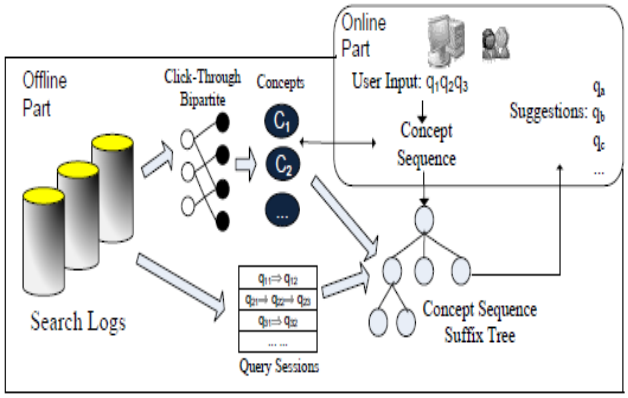

Figure 2: Framework Diagram [3]

Advantages:
1) The major advantage of this is, it is tested for large amount logs and queries.

Disadvantages:
1) Sequence of queries is considered for query suggestion

The searching process would give good results when the user reaches to his desired site in short amount of time. In the existing system, the list of all related URLs get displayed on the screen when user enters any particular query. This method takes time to find the exact desired site or source of information.

Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, Hang Li[4] proposes the method of categorization to reduce the searching time. Categorization also helps in focusing the category of interest rather than browsing through all the results sequentially. Suppose the user fires a query "Jaguar". The system will categorize links under automotive, animals, computer and internet etc. So that user can easily get the desired set of links in one category. The tricky part of this proposed work is that, some results do not fit into any category. So there is need to make one extra group called as not categorized group.

Clustering is used to categorize the results. But clustering the search result is not always best solution for categorization. The deficiencies in this approach are: 1. The cluster derived is not always related with the user interest. 2. The generated cluster labels are not informative enough for user to select the desired cluster.

Advantages:

1) Rearranged as per its relevance to the entire query session rather than considering the conventional approach of most recent single query.

Disadvantages:

1) Automation is not used for query log formation.

3. Existing System

Many works about user search goals analysis have been investigated. They can be summarized into three classes: query classification, search result reorganization, and session boundary detection. In the first class, people try to infer user goals and intents by predefining some specific classes and performing query classification accordingly. However, since user needs changes for different queries,so that finding suitable search goal is very difficult. In the second class, people try to reorganize search results. But this may involve many noisy search results that are not clicked by any users. In the third class, aim of people is to detect the session boundry. However, this only identifies whether pair of queries belongs to the same goal or not. In the existing system k means algorithm is used for clustering, in which the result depends on the k value. If the value of k is large then it will take exponential time to find the final cluster.

Disadvantages of Existing System:

1. User's goal is not identified.
2. Many Noise search result will be shown in which user is not interested

4. Proposed System

Our system contains four different phases. First is feedback session which is combination of clicked and un-clicked URL's. Second is a pseudo document that represents the feedback sessions in more meaningful manner. Third is clustering, that clusters these pseudo documents in appropriate user search goal's. For the clustering bisecting k means is used. This algorithm gives the better results than k means algorithm. Fourth phase is organization of clustered data. And finally CAP method to evaluate the performance of our clustering.

Advantages

1. User's goal get identified
2. Data will be shown according to user's interest.

Process Summary:

1] User Enter Query
2] Search Cluster Data is present or not for query if present show cluster data with Google data otherwise show only Google data.
3] User click on interested URL after that Generate Feedback Session based upon clicked and un-clicked URL.
4] Get Title, Snippets, URLs, Click URL Count, Unclick URL Count in Feedback Session
5] Separate Title and Snippet From Feedback Session. Remove Duplicate title and snippet
6] Generate Pseudo Doc.
7] Apply K-Means clustering algorithm to these pseudo documents.
8] Organize the segmented result by considering the number of clicks.
8] Implement AP, VAP, Risk, CAP
9] Show different search goals to user in a segmented format

Architecture Diagram

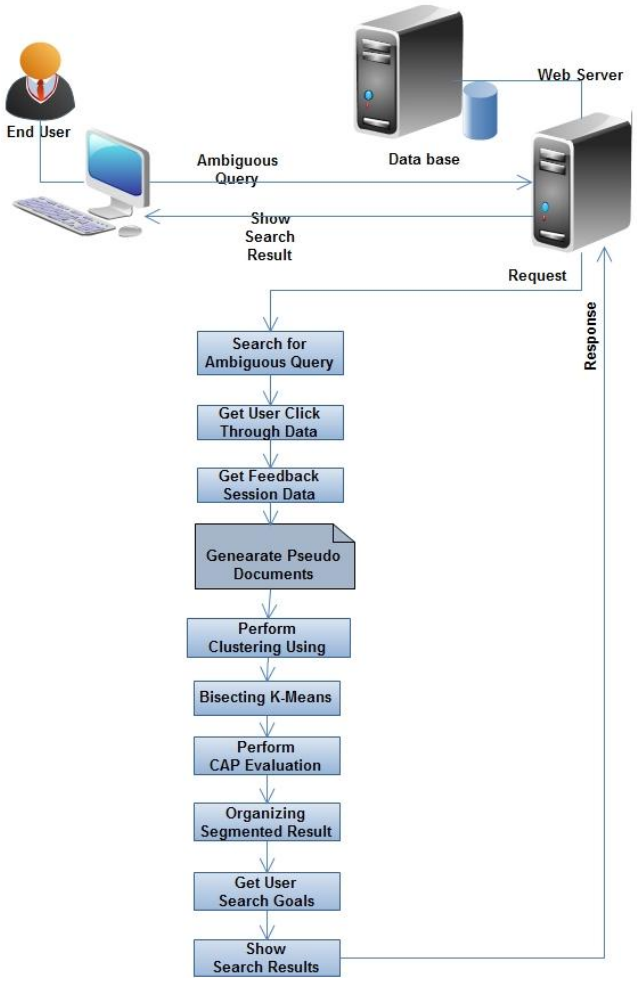

Figure 3: System Architecture

The entire system is divided into 4 parts.

Phase 1:

The first part is feedback session. Feedback session collects the data from googles database. The feedback session consists of title and snippet. Every URLs title and snippet are represented by Term Frequency-Inverse Document Frequency(TF-IDF). It saves the both clicked and un-clicked URLs up to last clicked URL.

Phase 2:

In the step 2 the pseudo document is formed by using the feedback session. In the pseudo document both the clicked and un-clicked URLs are considered. Some textual processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Sum of term frequency and inverse

document frequency is stored as a feature representation of document in F.

Phase 3:
The next step is to find out the user goal by applying clustering algorithm on pseudo document. The similarity between documents is checked by using cosine function. Distance is calculated from that cosine function for clustering algorithm. After clustering, every cluster is considered as a different user goal.

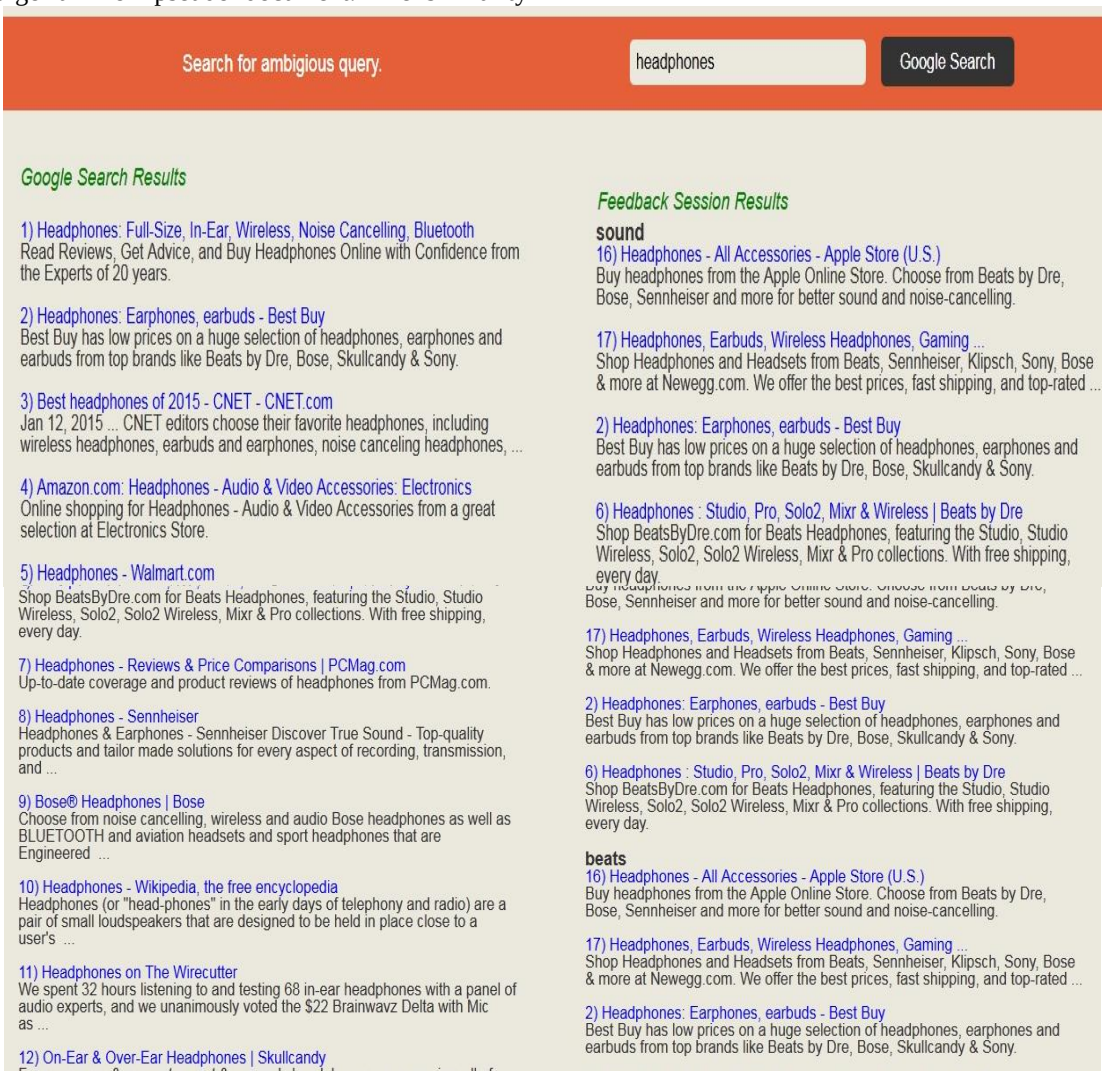

Figure 4: Comparison between list result and Segmented result as per user search goals.

Phase 4:
In this step, the evaluation of clustered is done. For the evaluation method of Classified Average Precision (CAP) is used. To calculate this CAP, the values of Average Precision and Risk is required.

AP is the average of precisions computed at the point of each relevant document in the ranked sequence.
$AP=1/P+\sum rel(r)Rd/r$
Where, P: Total Number Of Retrieved Documents
r: Ranking Of Document
Rd: Number Of Relevant Retrieved Document of rank r

"Voted AP (VAP)" this is the AP of the class including more clicks. There should be a risk to avoid classifying search results into too many classes by error. So we propose the Risk. We propose a new criterion "Classified AP," as:
CAP=VAP*$(1\text{-Risk})^{\gamma}$

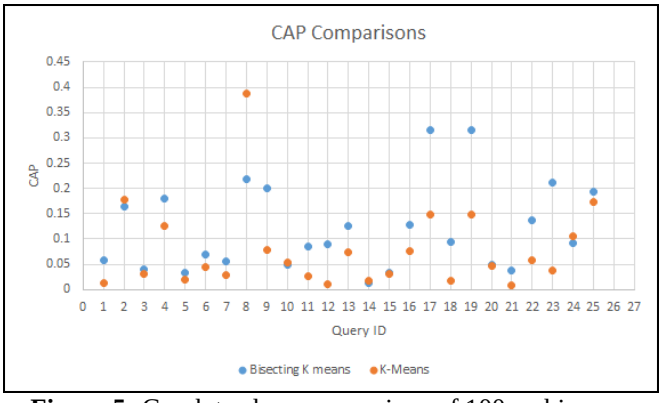

Figure 5: Graph to show comparison of 100 ambiguous results using k means and Bisecting k means

The above given graph represents the comparison between CAP results of k means and expected results from the bisecting k means. K-mean depends on the value of k, if k is large then it takes exponential time to find out the final clusters; because It has random in nature. In each iteration it

finds the random clusters for k clusters and then find out the best k clusters from those random clusters. If the value of k is less then it finds the solution in less time. But as we maximize the k, time for execution is exponential increases. If we use bisecting k-mean it takes less time than k-mean for clustering.

5. Conclusion

In this topic, a new approach is proposed in this of inferring user search goals by using the feedback session and pseudo document. In the feedback session both the clicked and the un clicked URLs ones before last click are stored. Pseudo document is made from mapping of feedback session. By performing clustering operation on this pseudo document will result into finding the user search goals which are depicted by keywords. To find out the user search goals the bisecting k means algorithm is used over the k means clustering. In the proposed work it will rearrange every segment as per the number of clicks of URLs in previous usage. So that the link which has the highest number of clicks will get appear at first position in the segment. Finally, criterion of CAP is formulated to evaluate the performance of user search goal inference. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods.

6. Acknowledgment

We thank the referees for their valuable suggestions to improvise the content and quality of this paper. The author is grateful to our principal for availing necessary facilities which helped for successful completion of work. We acknowledge the diligent efforts of our Head of the Department to guide us towards implementation of this paper.

References

[1] 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. 588-596,2004.
[2] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416,2000.
[3] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, Hang Li," Context-aware query suggestion by mining click-through and session data" ISBN: 978-1-60558-193-4,Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. (SIGKDD '08), pp. 875-883, 2008.
[4] 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.
[5] 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
[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, 2005.
[7] 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.
[8] 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,
[9] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01),pp. 162-168, 2001..
[10] T. Joachims, "Evaluating Retrieval Performance Using Click through Data," Text Mining, J. Franke, G. Nakhaeizadeh, andI. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003