Clustering user Queries of Search Engine for Restructuring Web Search Results: A Survey

Salve Bhagyashri Girdhar
PG Student, Department of Computer Engineering
North Maharashtra University
SES's R. C. Patel Institute of Technology, Shirpur
Shirpur, India

R. B. Wagh
Assistant Professor, Department of Computer Engineering
North Maharashtra University
SES's R. C. Patel Institute of Technology, Shirpur
Shirpur, India

***Abstract*— Data Mining refers to mine information from large amounts of data. It also called as knowledge mining from data. Web mining is the application of data mining to mine knowledge from web data including hyperlinks between documents, web document. In today's internet the search engine is one of the most important applications. For an ambiguous query different users may have different information requirements but the search engine does not satisfy user information requirements properly on the various aspects upon submission of same query. The conclusion and evolution of user search goals can be very useful in improving search engine appropriately. It also improves user knowledge. Therefore a unique approach is used to analyzing user queries from various search engine records. The Feedback sessions are clustered to find out different user search goals for a query. Feedback sessions are constructed from user click-through logs and can efficiently reset the information needs of users. Pseudo-documents are generated through feedback sessions for clustering. Finally Classified Average precision (CAP) algorithm is used to understand the user search goals efficiently.**

Keywords- user queries; search engine results, classified average precision; Resturucturing web search results; k-means clustering.

I. INTRODUCTION

Web mining is the application of data mining to mine knowledge from web data including hyperlinks between documents, web document. It is also called as information mining from data. World Wide Web has become an important source of information and services. Mining of interesting information from web data has become more popular .The web is huge, dissimilar and dynamic and as a result of that web mining has involved lot of attention in recent time. Web mining generally divided into three main types, i.e. structure mining, content mining, and usage mining. Each one of these areas are associated mostly, but not exclusively to these .Three major types of data found in the Web mining.

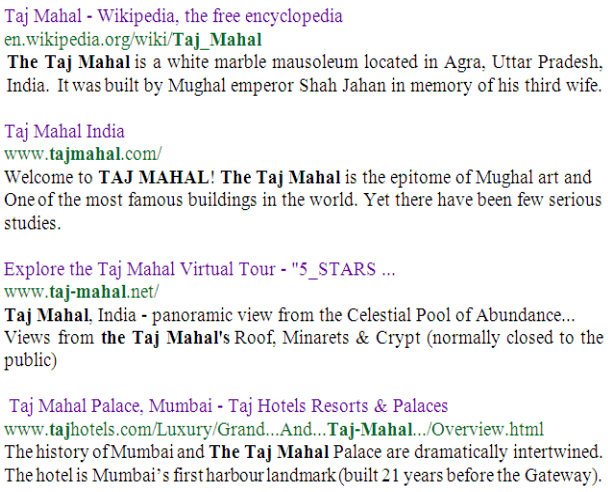
a). **Content Mining:** The real data that the document was designed to give to its users. In general this data consists mainly of text and multimedia.
b). **Structure Mining**: This data describes the organization of the content within the Web. This includes the organization inside web page, internal and external links and the website hierarchy.
c). **Usage Mining:** This data describes the use of a website or search engine, reacted in the web server's access logs as well as in logs for specific applications.

A. Information Retrival

IR is training and recovery of specific information from stored data. Information retrieval aims at defining systems able to provide a fast and effective content based access to a large amount of stored information. The IR is used for searching documents. The aim of an IR system is to estimate the importance of document to user information needs expressed by means of query. It allows fast access to large amount of data .The information is any kind of multimedia, web pages, textual. Therefore, Information retrieval is important for data mining, text mining [1].

B. User Search Goal

User search goal cab be considered as the cluster of information needs for a query. User goal is different information about query that user want one of the most important application in internet is 'web search engine '.When user submitted keywords in the search engine to obtain the information or web pages they want but sometime user don't get information they want accurately. The meaning of query is wide or different user may give same query. For ex. The keyword "The Tajmahal" is submitted to the search engine. The search engine gives different results related to the keywords.


Taj Mahal - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Taj_Mahal**
The Taj Mahal is a white marble mausoleum located in Agra, Uttar Pradesh, India. It was built by Mughal emperor Shah Jahan in memory of his third wife.

Taj Mahal India
www.**tajmahal**.com/
Welcome to **TAJ MAHAL**! **The Taj Mahal** is the epitome of Mughal art and One of the most famous buildings in the world. Yet there have been few serious studies.

Explore the Taj Mahal Virtual Tour - "5_STARS ...
www.**taj-mahal**.net/
Taj Mahal, India - panoramic view from the Celestial Pool of Abundance... Views from **the Taj Mahal's** Roof, Minarets & Crypt (normally closed to the public)

Taj Mahal Palace, Mumbai - Taj Hotels Resorts & Palaces
www.**taj**hotels.com/Luxury/Grand...And...**Taj-Mahal**.../Overview.html
The history of Mumbai and **The Taj Mahal** Palace are dramatically intertwined. The hotel is Mumbai's first harbour landmark (built 21 years before the Gateway).

Figure1. Example of Different User Search Goal

The keywords "The Tajmahal" gives information about the white marble mausoleum in the city of Agra and the information about Taj hotel in Mumbai .The diagram shows the result given by query "The Tajmahal", but sometime the necessary page is not available on the starting page the user has to search multiple pages .It takes lots of time and user has to spend more time in searching a particular page.

II. RELATED WORK

Lee et al describes automatic identification of user search goals in search engine in which they define query classification consider user goal as informational and navigational. In navigation the user have particular goal or website in their mind and their aim is to reach that particular site. In informational queries user don't have particular website in their mind they visit multiple pages .Since what user care about varies a lot for different queries finding suitable predefined search goal difficult and impractical [3].

Li et al describes learning query intent from regularized click graphs, in which they define query as intent. The query is classified in to two type job intent and product intent .The job intent and product intent expand the training data to improve the performance of classification. The aim of query intent is to enriching the amounts of data by using semi-supervised learning. In query intent classifiers uses only query words or phrases for well work [4].

Wang et al. proposed web search log to organize search result which allows a users to navigate into relevant documents quickly. In which interesting aspects of query learn from click through log. Clustering the search result is best way to organize the search result. In clustering of search results user find documents quickly. The drawback of this approach is that the member of the cluster label does not provide right information, so it's difficult to identify right cluster and the cluster do not depends on the attractive aspects of the users and the cluster exposed do not necessarily correspond to the interesting aspects of a topic from the user point of view [5].

Jones et al. defined session boundaries and automatic hierarchical segmentation of search topics .Most analysis of user search relevance and performance take single query as unit of search engine relevance. In which queries are group together by task or session and they used to identify boundaries .This method only identifies whether pair of queries belongs to the same goal or not and they doesn't care about what the goal is [6]?

Chen et al. describes bringing order to the web automatically categorizing search results. In which search results are categorize in to hierarchical category. Organizing search results allows user to focus on things in categories of importance rather than having to browse through all the results serially [7].

S. Beitzel et al. Varying Approaches to Topical Web Query Classification. Web queries are classified based on the behaviors or some similarities. This classification of query improving retrieval effectiveness and efficiently. The query is used to retrieving a document before or after a query classification. Bridging classifier maps the document taxonomy onto query classification problem and it provide sufficient training data. We find that training classifier explicitly from manually classified queries to the bridged classifier by in score. The preretrieval classifier is worse than bridged classifier. It requires snippets from retrieved documents [8].

Shen et. al. describes building bridges of web classification in which short and ambiguous user queries classified into target queries. The bridges describe build in offline mode on an intermediate taxonomy. In which there is no need of new categories for new target categories so bridging classifier taxonomy once. It improves both efficiency and effectiveness of online classification [9].

Cao et al. describes context aware query suggestion by mining Click through and session data. In which query selection plays important role in search engines .It improves the search engine usability. Query suggestion approach consist two way offline model learning step and concept sequence suffix tree. This approach consist not only the current queries but also the recent queries in same session to provide more meaningful suggestion [10].

III. METHODOLOGY

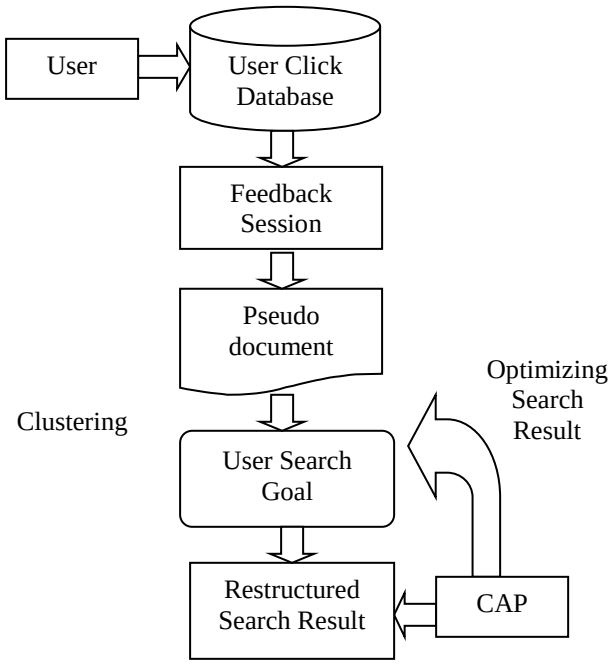


Figure 2.System Overview

A. *Feedback Session*

In feedback session the session is defined as sequence of successive queries to satisfy single information need. In single session containing only one query which is introduced. Feedback session is comprehensive of whole session. The feedback session consists of both the clicked and un clicked URL. Figure 2. shows a rectangular box shows the feedback session in which it consist both clicked and un clicked URL. The unclicked URL are mark by 0 and the clicked URL are mark by the order in which URL are clicked. In rectangular box the left part shows all the links related to query and right part shows clicked and unclicked link and the session is end

with last URL that was clicked in a single session. Before last clicked all URLs are scanned and it consists both clicked and un clicked links. In fig., the 3 links are clicked with their particular sequence and 4 links are unclicked that denoted by 0.The URL clicked after the last URL are not considered in the feedback session. The clicked URL shows what user want and the unclicked shows what user don't want The feedback session gives detail knowledge about user's need[11].

URL's	Clicked Sequence
www.en.wikipedia.org/wiki/**Taj_Mahal**	1
www.**tajmahal**.com/	0
www.**taj-mahal**.net/	2
www.**taj**hotels.com/	3
www.lonelyplanet.com/	0
https://www.sscnet.ucla.edu	0

Figure 3. Feedback Session

B. Mapping Feedback Session to Pseudo Document

Feedback session is different for different click through log database. The feedback session is changes for different database. These are not suitable to use feedback session directly for inferring user search goals. There are many types of feature representation for feedback session .binary vector presentation consists of 0 and 1. Here, 1 represent clicked link and 0 represent unclicked link but binary vector is not give enough information because one linked is clicked multiple time but its show 1 so user can't get the information that how many times the link is clicked. Therefore new method is used to represent feedback session. In clicked sequence method shows sequence of clicked URL .It shows how many time the link is visit. So we map feedback session to pseudo document. Two steps to building feedback session to pseudo document [12].

1) Representing the URL's in the feedback session

In this step we enrich the URL with additional textual content by extracting title and snippets .Then some textual process are implemented those paragraphs sum ,transferring all the letters to lowercase ,removing stop words and stemming. Finally, the title and snippet generated from URL are performed by term frequency –Inverse document frequency (TF-IDF) vector.

$Tu_i=\{tw_1.,tw_2 \ldots tw_n\}$

$Su_i=\{sw_1,sw_2 \ldots sw_n\}$

Where,

Tu_i:-TF-IDF vectors of the URL titles

Su_i:- TF-IDF vectors of the URL snippets

μ_i:-ith URL in the feedback

w_i:-j term in the enrich URL. Where,j={1,2,..n}

Tw_j:-TF-IDF value of the jth term in URL's title

Sw_j:- TF-IDF value of the jth term in URL's snippet

enriched URL represented by weighted sum is

Fu_i:-feature representation of ith URL

Wt:-weight of the title

Ws:-weight of the snippet

2) Forming Pseudo document based on URL representation

This method combine both clicked and unclicked URL in the feedback session .Based on the assumption that term in the vectors are independent .We can perform optimization on each dimension independently.

$$F_{fs}= \{F_{fs}(w_1), F_{fs}(w_2) \ldots . F_{fs}(w_n)\}^T$$

Where,

F_{fs}=feature representation

C. Inferring User Search Goals by Clustering Pseudo-Documents

With the proposed pseudo-documents, we can infer use Search goals. In this section, we will describe how to infer user search goals and depict them with some meaningful keywords.

1) K-means Clustering Algorithm

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). A set of cluster resulting from a cluster analysis can be referred To as a clustering .Cluster analysis has been widely used in many application such as business intelligence, image pattern reorganization, web search ,biology and security. In Web search application a keyword search may often return a very large number of hits (i.e. page relevant to the search) due to the extremely large no. of pages

Algorithm 1: k-means Clustering Algorithm

Input:

k:the number of clusters

D:a data set containing n objects.

Output:

A set of k clusters

Methods:

1. choose k objects from Das the initial clusters centers;

2. Repeat

3. (Re) assign each objects to the clusters to which the object is the most similar, based on the mean value of the objects in the cluster;

4. update the clusters means ,that is calculate the mean value of the objects for each cluster;

5. until no change;

end for

2) Cosine Similarity

Cosine similarity is a measure of similarity that can be used to compare documents or give a ranking of documents with respect to a given vector of query words. The similarity between two pseudo-documents is computed as the cosine score of and the distance between two feedback sessions and cluster pseudo-documents by K-means clustering which is simple and effective. The similarity between two pseudo-documents is computed as the cosine score of Ffsi and Ffsj

As follows:

$$Sim_{ij}=\text{COS}\,(F_{fs_i}, F_{fs_j}) \quad (1)$$

$$\frac{F_{fs_i}, F_{fs_j}}{|F_{fs_i}|,|F_{fs_j}|} \quad (2)$$

And the distance between two feedback sessions is

$$Dis_{ij=1-Sim_{IJ}} \quad (3)$$

3) Performance Evolution Metrics

Average Precision:-precision is the fraction of the document retrieved that is relevant to the user information need average precision compute the average value.

$$AP=1/N^{+}\sum_{r=1}^{n} \text{rel(r)Rr/r} \quad (4)$$

Where,

N^{+}:-number of relevant or clicked documents

r =rank

N=total no. of retrieval documents

Rel():binary function relevance of given rule

R_r=no. of relevance documents of rank r or less than r.

VAP:-The VAP is class of AP including more clicks namely votes .if two classes are same then the class which have larger value is considered as VAP

$$\text{Risk}=\sum_{i,j}^{m} = \frac{1(i,j)d_{ij}}{c_m^2} \quad (5)$$

Where,

m = N0. of clicked URL

$c_m^2=\frac{m(m-1)}{2}$

It calculates the no. of clicked URL's pairs they are not in the same class.

Classified Average Precision (CAP)

A new criterion is used to calculate the performance of restructuring web search result. The criteria are Classified Average precision (CAP)

$$CAP=VAP\ X\ (1\text{-}Risk)r \quad (6)$$

λ=Adjust the function in Risk of CAP

Risk=VAP=voted average precision

CAP is depends on both risk and VAP

IV. CONCLUSIONS

This system is used to improve the discovery of user search queries by clustering user feedback session represented by pseudo document. The new criteria classified average precision is used to analyze the performance of user search goals. The user search goals is used for restructuring web search result ,so user can find the exact information very efficiently and appropriately and user can find what they want very easily, quickly and appropriately.

ACKNOWLEDGMENTS

I express my sincere thanks to my guide Prof. R. B. Wagh. The faith & confidence shown by him in me, boosted me and motivated me to perform better. He is a disciplinarian yet always so approachable and warm at heart.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval," ACM Press, 1999.
[2] 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.
[3] 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.
[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] 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.
[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] 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] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf.
[9] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for WebQuery Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf Research and Development in Information Retrieval (SIGIR '06), pp. 131-138, 2006.
[10] 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.