ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

A Survey on User Search Goal Inferring System

Pramod S. Gurav, Ankush G. Gavane, Swapnil A. Shinde, Saurabh S. Mali, Anil A. Lokhande,

Prof. M.D. Jitkar

Student, Dept. of Computer Engineering, Dr. D. Y. Patil COET., Kasaba Bawada, Kolhapur, Maharashtra, India

Student, Dept. of Computer Engineering, Dr. D. Y. Patil COET., Kasaba Bawada, Kolhapur, Maharashtra, India

Student, Dept. of Computer Engineering, Dr. D. Y. Patil COET., Kasaba Bawada, Kolhapur, Maharashtra, India

Student, Dept. of Computer Engineering, Dr. D. Y. Patil COET., Kasaba Bawada, Kolhapur, Maharashtra, India

Student, Dept. of Computer Engineering, Dr. D. Y. Patil COET., Kasaba Bawada, Kolhapur, Maharashtra, India

Student, Dept. of Computer Engineering, Dr. D. Y. Patil COET., Kasaba Bawada, Kolhapur, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, Dr. D. Y. Patil COET., KasabaBawada, Kolhapur, Maharashtra, India

ABSTRACT: Internet has become a non-detachable part of human beings throughout the world. But Internet is an ocean of information that provides you enormous details on whatever topic you search on the web. Many researchers have made an excellent effort to infer the user search goals through user profiles, user searching history or user searching knowledge and pattern but most of the techniques failed as it's not that the user will always try to search the same contents or documents over the internet. Another technique to guess the user goals made use of user location to find the location specific queries and answer them. Thus we are going to analyze all the algorithms implemented so far for the user goal search.

KEYWORDS: Hidden Web Crawler, Query Optimization, Search engines, Metadata, document frequency, term weights.

I. INTRODUCTION

In many websites the search engine are widely used for finding the user need. As the queries are short in size i.e. normally two or three words. But these queries gives an ambiguous results. These result does not exactly matches to the user's expectations. Manytimes different search engine produces different search result. So that non useful results arises and those are fail to satisfy the user's expectations. Therefore we have proposed a user search goal inferring system to match the relevant search result with user's needs. In this we are treating the user's need as a cluster. This will be very useful to improve the performance of search engine. We can able to redesign the result by grouping the needs of the user at different time. The user need can assigned by a word on which the clustering will be done. Depending upon the clustering the result are ranked. For better searching, many methods were invented to make searching more effective like classification of query, recognition of search results, and session limit detection. However, this method has limitations since the number of different clicked URLs of a query may be small. Other works analyse the search results returned by the search engine when a query is submitted.

Therefore, there is no standard or optimal way to issue queries to search engines, andit is well recognized that query formulation is a bottleneck issuein the usability of search engines. Most text classification research focuses on classifying documents, which contain enough terms to adequately train machine learning approaches. The task of classifying web queries is different in that web queries are short, providing very few inherent features. Therefore, most approaches use the documents retrieved by a query as features to classify it.

For example, the user has entered a query 'phoenix' in Google search engine. Basically it should produce the results

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

for phoenix as a bird. But it is displaying the result of a shopping mall in pune. The expected result is found to user but it is not ranked as a first result. Many times user have to search for many pages of search results to find his need. Every time user wanted to submit query 'phoenix' it will firstly shows the result of mall instead of bird.

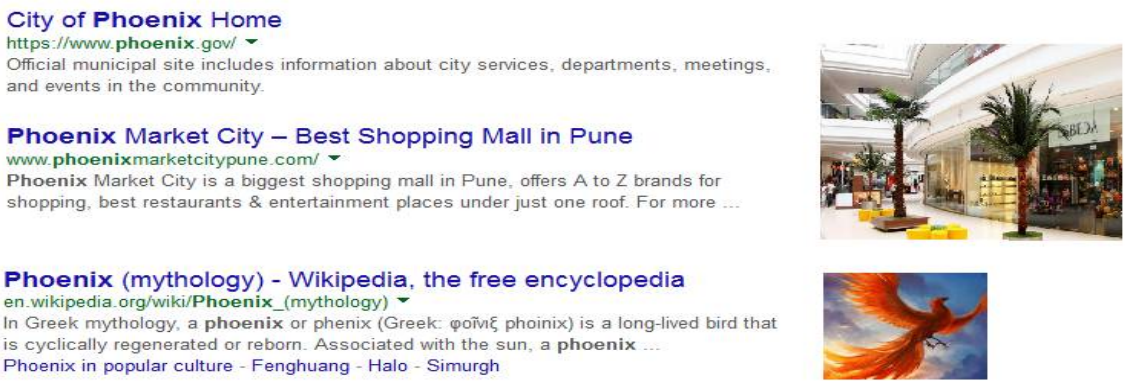

Figure: 1. Variation In output of query 'phoenix' submitted by user.

II. LITERATURE SURVEY

Many scientists and developer had done work on optimization of search query. They have represented that by writing design and implementation of their research. These research are mainly focuses to the retrieval of user specific and relevant result. We have studied those research paper which includes:

Collaborative Filtering of query logs:

In this paper the query log and its semantic relations are implicitly captured in the sequence of users submitting queries and clicking results. It is a method to represent a query in vector space. It generate a graph from the query-click bipartite graph and graph produced by query log. Measures of these graph shows the shading of colour on the user search goal. It provide an experimental analysis on the quality of the relations, showing that most of them are relevant.It uses a query suggestion algorithm for effective search result. This algorithm is stated as follow:

A. *Query Suggestion Algorithm*

1: A converted bipartite graph G ¼ ðV þ [V; EÞ consistsof query set V þ and URL set V. The two directededges are weighted using the method introducedin previous section.

2: Given a query q in V þ, a sub graph is constructed byusing depth-first search in G. The search stops whenthe number of queries is larger than a predefinednumber.

3: As analyzed above, set ¼ 1, and without loss ofgenerality, set the initial heat value of queryq fq ð0Þ ¼ 1 (the choice of initial heat value will notaffect the suggestion results). Start the diffusionprocess using

f ð1Þ ¼ e R f ð0Þ.

4: Output the Top-K queries with the largest values invector f ð1Þ as the suggestions.

The limitation is, it only identifies whether a pair of queries belong to thesame goal or mission but does not care about what the goalis in detail. It only identifies whether a pair of queries belong to thesame goal or mission but does not care about what the goalis in detail.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

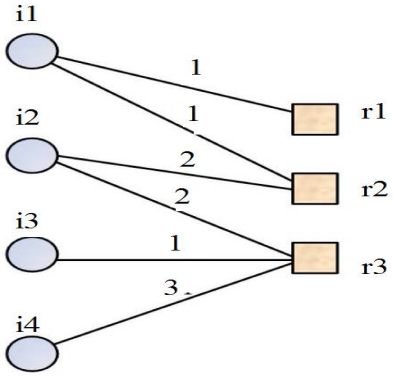

Figure: 2 Graph for query suggestion.

B. *Zealous algorithm:*

This provide the privacy to user search log. It creates a histogram input search queries then it removes the result which is having frequency below the threshold. It eliminates the items whose noisy frequencies are smaller than another threshold. In web search applications queries are submitted to search engine. Search history is formed from the user submit a query and click the URL's.

A query may contain well-formed natural-languages, or keywords or phrases. Once a user query is input to the search engine the list of documents is presented to the user with a document title. Then it generate a histogram on the basis of threshold values.

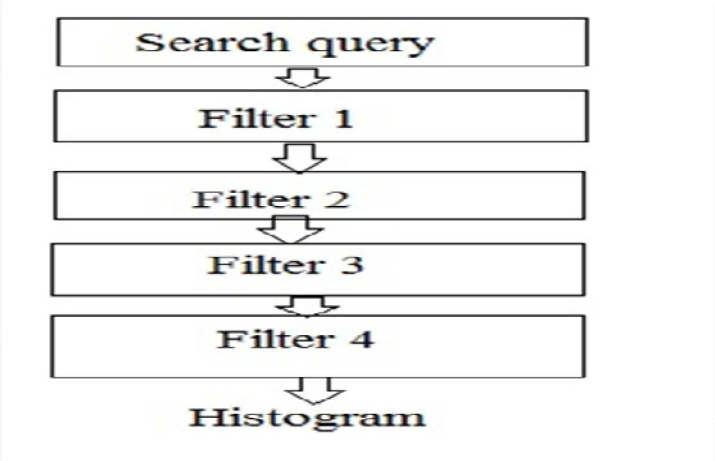

Figure: 3 Flow of Zealous algorithm

Disadvantages of this method is, this method does not maintain the feedback sessions. So it creates more noisy results.

C. *Web Query Classification:*

It examines two issues, pre and post retrieval of classification on the basis of search results. It compare and combine query classifiers which are applied before gathering the retrieved documents, a document classifier trained from pages in the ODP , and explicit query trained on the retrieved documents of classified queries. This provides enough training data to effectively test our explicit classifiers, as compared to only the 111 training queries. For the post-retrieval classifiers, all support vector machines, this training data set was used to build the model and tuning data to select the threshold at which we report F1 on testing.

In this classification is done on the basis of 3 steps:

1. Exact matching classification.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

2. SVM Classification.
3. Bridge Classification.

The limitation is that it produces more complex and irrelevant result. To reduce this we have to perform category selection.

III. PROPOSED SYSTEM

In our system the user submits the query into the browser. The search engine searches the relevant information according to the user query. The user actions are stored in the user click through logs. From the user click through logs each and every session is analysed and generates the feedback session. The user search goals are obtained according to the feedback sessions.The restructure result is produced for the user query based on the user search goal. Every user search the same query with different intensions. For example if user A and B both typed same query in a search engine. Suppose their query is 'apple'. The user wants the information about apple fruit and user B wants the information about Apple Company. Then according to their click through logs and their searching behaviour the clustering is done. This clustering make effect in searching when both users A and B wants to find same query with different intensions.

This click through log is nothing but the feedback survey of all the result of search queries. This survey will help user to find the relevant result. Depending upon this feedback the pseudo documents are created. After that depending upon the users interest the click through report is generated. Using this report clustering of the user search result is done. Then applying Cap evolution technique the classified output is displayed. This classified output is nothing but the expected result which user wants to search.

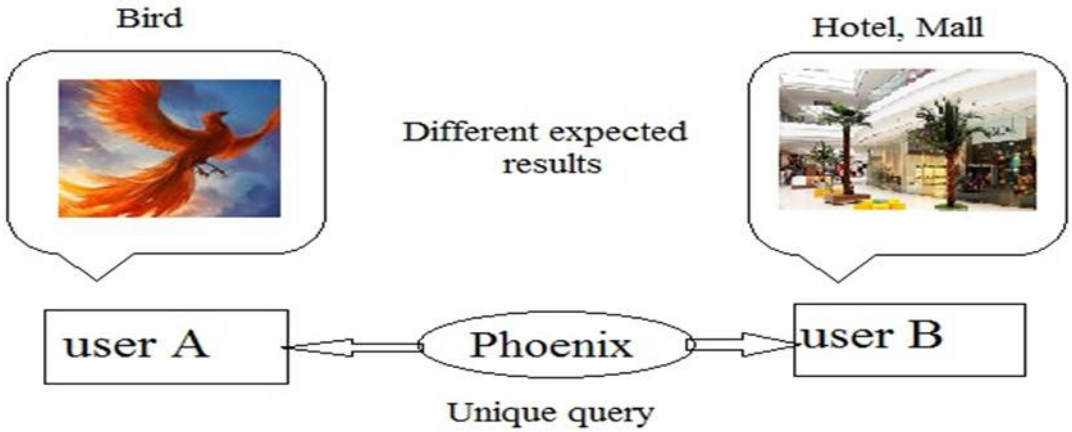

Figure: 4 Example of user goals.

A. *Feedback session:*

The feedback sessions are nothing but the clicked and unclicked URL's by the user in the result set. The clicked URLs representswhat users need and the unclicked URLs represents whatusers do not need about. The unclickedURLs after the last clicked URL should not be included intothe feedback sessions since it is not certain whether theywere scanned or not.The feedback session can tells what user need is and what kind of result he expect. The feedback sessions are numbered on the basis of user click data. The click sequence is get stored into session. When next time user searches same query then the session will execute the same click through sequence to find an accurate result which user wants.A feedback session is represented by a small paragraph of text that consists of its title and some short data. Then, sometextual processes such as transforming all the letters to lowercases, stemmingand removing stop wordsare implemented to those text paragraphs. Then each URL is represented by some term frequency. Then the weight of each URL is obtained by some mathematical operations. Then these query frequency and URL weight is use to produce pseudo documents.

B. *Pseudo Documents:*

The efficient feedback session represented by pseudo documents. User may clicked on so many links, so that there may be the creation of many feedback sessions. In that all feedback sessions the documents which are having more efficiency than others are known as pseudo documents.In this the click sequence is re-ranked depending upon the user

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

clicks. For different search results different feedback sessions are maintained. For this we have used one vector known as binary vector. The binary vectors represents the methods require for feedback sessions. With the help of pseudo documents we can easily make guess about user's goals.

For the generation of pseudo documents we combines both clicked URL and unclicked URL. Then after the calculation of document frequency and URL weight the exact match of user's expected result is evaluated. This result is then stored in pseudo document for further future guessing of user need. Whenever in future user enters same or relevant query in search engine then these pseudo document will produce the result which user wants.

C. *Evaluation of re-designed web search results:*

Since the user search goal is not fixed, the evaluation of redesigned search result becomes more difficult. There is no approach invented yet to evaluate search goals. Therefore, we propose an evaluation method based on redesigning web search resultsto evaluate whether user search goals are guessed properlyor not. User search goals are represented by thevectors and the feature representation of each URL inthe search results can be computed.Then, weare going to categorize each URL into a cluster cantered by theinferred search goals. In this we are doing categorizationby selecting the smallest value between the URLvector and user-search-goal vectors.

IV. CONCLUSION AND FUTURE WORK

Here we have introduce feedback sessions to analysed to infer usersearch needs rather than using search results or clicked URLs.Both the clicked URLs and the unclicked ones before the lastclick are considered as user implicit feedbacks and takeninto account to construct feedback sessions. Here we have maintain the sequence of most relevant search results to represent need of user. We have used the concept of pseudo documents to design the feedback sessions. This concept will make the searching easy to user. And it is producing most relevant results.

REFERENCES

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query RecommendationUsing 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 aSearch Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416,2000.

[3] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "VaryingApproaches to Topical Web Query lassification," Proc. 30th Ann.Int'l ACM SIGIR Conf. Research and Development (SIGIR '07),pp. 783-784, 2007.

[4] 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 DataMining (SIGKDD '08), pp. 875-883, 2008.

[5] H. Chen and S. Dumais, "Bringing Order to the Web: AutomaticallyCategorizing Search Results," Proc. SIGCHI Conf. HumanFactors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[6] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant TermSuggestion in Interactive Web Search Based on ContextualInformation in Query Session Logs," J. Am. Soc. for InformationScience and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[7] T. Joachims, "Evaluating Retrieval Performance Using ClickthroughData," Text Mining, J. Franke, G. Nakhaeizadeh, andI. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.