Improved Search Goals with Feedback Sessions by using Precision Values

Nancy S, Sathiya Devi S

***Abstract*—Users can submit ambiguous queries to the search engine. The search engine provides the relevant information, but it does not match with the user query. To overcome this problem, User search goals can be used. It defines the information needed for a query. To discovers different user search goals for a query by clustering the feedback Sessions. It consists of both clicked and unclicked URLs. Each URL consists of title and snippets. In the proposed system, genetic algorithm can be used to find the optimal solution for the weightage of title and snippet. Second feedback session is mapped to the pseudo document. Clustering the pseudo documents and each cluster can be considered as one search goal and depict them with some keywords or goal text. Evaluation of clustering is an important problem. So, Classified Average Precision can be used to evaluate the performance of user search goals inference and restructuring results.**

***Index Terms*—User search goals, feedback session, Genetic Algorithm, pseudo-document, classified average precision, restructuring search results**

I INTRODUCTION

User can submit ambiguous queries to the search engine. They provide the relevant information, but it does not match the user query. Ambiguous query refers to a query that has more than one meaning. Another problem needs of Specific information because many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. Use lot for different queries, finding suitable predefined search goal classes is very difficult and impractical. To overcome these problems, User search goals can be used. It defines the information needed for a query. In Query Recommendation ,Query is submitted to the search engine they suggests the list of related queries The Related queries consist of some issues they are user searching the same information may phrase their queries differently user tried different queries until they are satisfied with the result [1].Query Classification Web query topic classification is a complexity in information science. Task is to allocate a Web search query to one or more predefined categories, based on the topics. The importance of query classification is underscored by a lot of services provided by Web search. Application is to provide better search result pages for users with interests of different categories.

Topical web query classification [3] can be used to improve the search services efficiency and effectiveness. This classification is accessible for use in the retrieval process.

Query can be used to retrieve the document before or after the classification. For example, the users submit a query "apple" may expect to see Web pages related to the apple fruit, or they may prefer to see products or news related to the computer company. Search result pages can be grouped according to the categories predicted by a query classification algorithm. Yet, the computation of query classification is non-trivial. Different from the document classification tasks, queries submitted by Web search users are usually short and ambiguous; also the meanings of the queries are growing over time. Therefore, query topic classification is much more difficult than conventional document classification tasks.

Next Session Boundary Detection, It is a heuristic-based technique and works on the basis of a geometric interpretation of both the time gap between queries and the similarity between them in order to flag a topic shift. Search engine returns millions of search results. If user submits the query to the search engine they provide the relevance information but it does not matches the user query. So restructuring the web search results can improve the search engine performance. It is one of the application of the user search goals .In restructuring search results the original search results can be reconstructed. The goals of this method are to evaluate whether User Search Goals are inferred properly or not.

Restructuring Web Search Result is one of the Application of User Search Results. The goals of the method is to evaluate whether User Search Goals are inferred properly or not and Restructuring search results. This method is based on Feedback Session and generating Pseudo-documents.

Section 2 explains the works related to the concepts, section 3 explains the framework. Section 4 discusses the results and discussion .Section 5 concludes the concept.

II RELATED WORKS

In recent years, many works have been carried out towards the user search and their Section. Many Authors are still working to produce a method which can give efficient search results for user.

R.B. Yates et al. [2] suggest that improving the search engine Results using Query Recommendation algorithm but using uncertain keywords is the problem.

U. Lee et al. [7] describes about Automatic Identification of User Goals in Web Search. It is based on past user-click behavior and anchor-link distribution technique. It can be used to identify the user behavior.

***Manuscript received April, 2014*.**

***S. Nancy**, Department of Computer Science and Engineering, University College of Engineering (BIT Campus), Tiruchirappalli, India,*
***S. Sathiya Devi**, Department of Computer Science and Engineering, University College of Engineering (BIT Campus), Tiruchirappalli, India,*

S. Beitzel et al. [3] discuss how to improve the search services efficiently. It is based on topical query classification.

H. Cao et al. [8] suggest that improving the usability of Search engine. It is implemented in context aware query suggestion technique. Sequence suffix and query suggestion algorithm can be used but the drawback is computational cost is high.

H. Hamada et al. [4] suggest that related queries using Ranking technique and T. Joachims et al. [6] describes about Optimizing Search Engines using click through Data. Support vector machine approach can be used for ranking. but the drawback is training data is only generated from relevance document by expert. It is difficult and expensive.

B. Cao et al. [5] describes the collaborative ranking and Optimization Algorithm. It improves the search performance and understanding user behaviors. Bradley-Terry model and Matrix Factorization Model can be used in this paper. But the drawback is bias and sparseness problem.

III FRAMEWORK

The framework can be divided in to two parts, in the upper part is to generate the feedback session and pseudo-document and the bottom part is reconstructed search results, it is obtained from original search results and user search goals can be generated from the upper part.

A. Feedback Session

Feedback session can be generated. It consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a session. Clicked and unclicked URLs can be represented by 0 and 1. A single session contains only one query. In the feedback session user click sequence can be calculated and stored in the user click through logs. The Feedback session can be clustered in this process since to generate different user search goals for a query.

To combine the enriched URLs in the feedback session using user click through logs to forms a pseudo-document.it reflect the information needs of a users. Next to map the feedback session to pseudo-documents can be generated.

TABLE I. FEEDBACK SESSION

Search Results	***Click Sequence***
www.thesun.co.uk	0
www.nineplanets.org/sol.html	1
www.solarviews.com/sun.html	0
www.thesunmagazine.org	2

B. Map feedback session to pseudo document

To Map Feedback session to Pseudo Document, The Pseudo document can be generated by the following steps;

- *Using URLs in the Feedback session*

First step to extract the titles and snippets from URLs appearing in the Feedback Session. Some textual process can be implemented to those titles and snippets. Such as stemming, removing stop words. Finally to find the term-frequency and inverse document frequency can be calculated along with the weightage of titles and snippets.

$$tf\ idf(w) = tf\ .\ log\frac{N}{df(w)} \quad (1)$$

Where tf(w) are the term frequency(no of word occurrence in title and snippet) and df(w) is the document frequency .Here N is the number of all documents.

Each URL's title and snippet are represented a Term frequency-Inverse Document Frequency vector.

$$T_{ui} = [t_{w1}, t_{w2}, t_{w3} \dots \dots t_{wn}]T$$
$$S_{ui} = [s_{w1}, s_{w2}, s_{w3} \dots \dots s_{wn}]\ T \quad (2)$$

T_{ui} is the URL's title and S_{ui} is the URL's Snippet.

C. Genetic Algorithm.

Genetic Algorithm is adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics.

- Genetic Algorithm Process

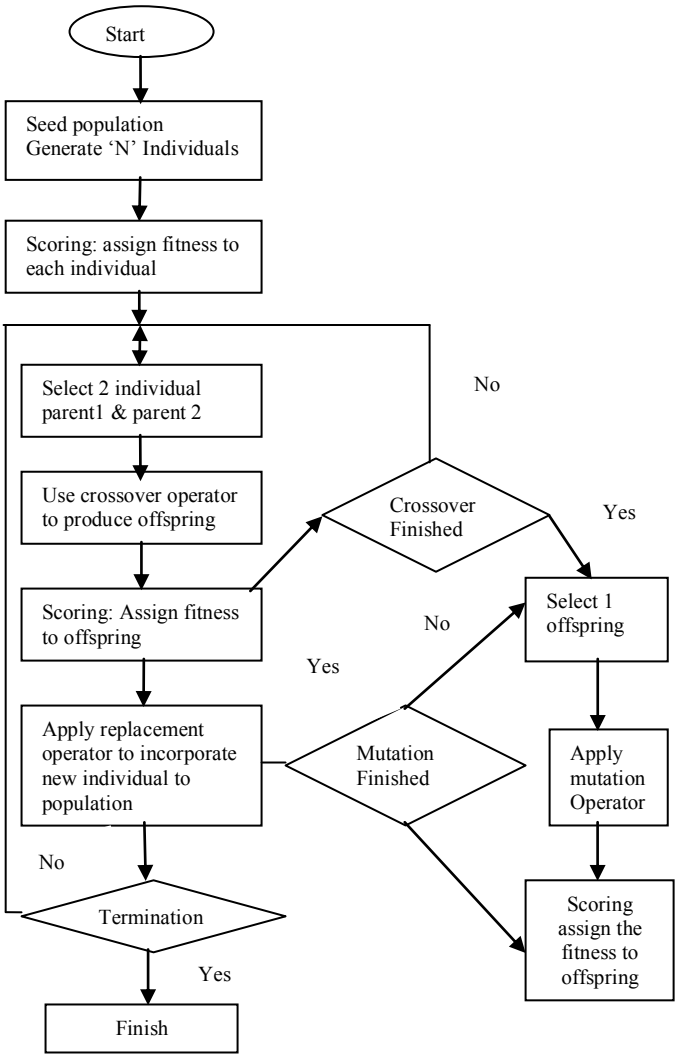

Fig 1. *Genetic Algorithm process*

- Optimization

Optimization is a process of formulating a mathematical model that is expressed in terms of functions and the find a solution. It consist of the following parameters,

- Objective Function – The function is to be either minimized or maximized.

- Set of Unknown variables – it affects the Objective function.
- Set of constraints – It allows the unknown (variables) to take as certain values, but exclude Others.

i. **Fitness** : $f(x) = x^2$
ii. **Cross over** : 1 –point cross over or cross over Probability
iii. **Mutation** : yes or no
iv. **Selection/Replacement**: Rowlet -wheel selection
v. **Termination** :Maximum no of iteration or when Saturation

➢ Probability

$$p_i = \frac{F_i}{\sum_{j=1}^{n} F_j} \quad (3)$$

Finally the optimal value can be used to find the weightage of title and snippets using genetic algorithm.

➢ ***Weightage Calculation***

$$F_{ui} = \omega_t T_{ui} + \omega_s S_{ui} \quad (4)$$

ω_t is the weightage of the title and ω_s is the weightage of the Snippet.

Equation (2) is the feature representation of the URL in the feedback session. Above processes can partitions the search results as titles, description or snippets, pictures and click sequence separately. Finally the Feedback session can be clustered in this process since to generate different user search goals for a query. Next to map the feedback session to pseudo-documents can be generated.

➢ *Forming pseudo-document*

Clicked and Unclicked URLs can be used in the Feedback session can be used. The feature representation of a feedback session can be described below.

Clicked and Unclicked URLs can be used in the Feedback session can be used. The feature representation of a feedback session can be described below

$$F_{fs} = [f_{fs}(w1), f_{fs}(w2) \ldots\ldots f_{fs})]T \quad (5)$$

Where, F_{fs} is the feature Representation of the feedback session and f_{fs} is the value of the term w.

$$f_{fs(w)} = (w)\left\{ \begin{matrix} \sum_M [\, f_{fs}(w) - f_{ucm}\,(w)]^2 - \\ \lambda \sum_L [f_{fs}\ (w) - f_{u\bar{c}l}(w)]^2 \end{matrix} \right\}, f_{fs}\ (w) \in\ _{I_c} \quad (6)$$

I_c be the interval

$$[\mu f_{uc}\ (w) - \sigma f_{uc}(w), \mu f_{uc}\ (w) + \sigma f_{uc}(w)]$$

$I_{\bar{c}}$ be the interval

$$[\, \mu f_{u\bar{c}}(w) - \sigma f_{uc}\,(w), \mu f_{u\bar{c}}(w) + \sigma f_{u\bar{c}}(w)]$$

Where, $\mu f_{uc}\,(w)$ is the mean and $\sigma f_{uc}\,(w)$ is the mean square error.

λ Parameter can be useful for balancing the importance of clicked and unclicked URLs.

$$f_{fs}(w) = 0, I_c \in I_{\bar{c}}$$

To generate the pseudo-documents to better represent the feedback sessions for clustering. It can efficiently reflect the User information needs. Clustering the pseudo documents and each cluster can be considered as one search goal and to depict them with some keywords or goal text.

D. Generating Pseudo document

Pseudo-document can be used to understand the user search goals. it can efficiently reflect the User information needs .In the Pseudo-document the similarity between documents can be calculated using Euclidean distance and cosine score measure.

The similarity between the two documents can be computed as

$$sim_{ij} = \cos(F_{fsi}, F_{fsj}) \quad (7)$$

$$= \frac{F_{fsi}\ .\ F_{fsj}}{|\, F_{fsi}\ |\ |\ F_{fsj}\ |}$$

Distance between two feedback sessions is

$$\text{Dis} = 1 - \text{Sim}_{i,j} \quad (8)$$

For clustering k-means algorithm can be used. It is one of the partitioning methods. Clustering refers to grouping of similar objects. In K-means, centroid can be calculated for each cluster. The advantage of k-means algorithm is a large number of variables can be easily computationally faster than compared with other techniques and it produce tighter cluster, especially if the clusters are globular. These are technique implemented in Pseudo-document.

$$F_{center\ i} = \frac{\sum_{k=1}^{Ci} F_{fsk}}{C_i}, (F_{fsk}\ \in Cluster\ i) \quad (9)$$

Where F_{center} is the i $_{th}$ clusters and C_i is the number of the pseudo document in the i $_{th}$ clusters.

After, clustering can be performed. Each cluster can be considered as user search goals highest value in the center point of the document can be depict as keywords or goal text.

E. Restructuring Results

Feedback information is needed to determine the best cluster. Using the pseudo document the keywords or goal text can be generated.

Finally, the original search results are reconstructed and the performance can be evaluated using classified average precision. The categorization can be performed by choosing the smallest distance between the user search goals vector and URL vectors. Goals of this method are to evaluate whether User Search Goals are inferred properly or not.

➢ Performance Measure

Average precision can be calculated at the point of each relevant document.

(10)

$$AP = \frac{1}{N} + \sum_{r=1}^{N} rel(r)\frac{R_r}{r}$$

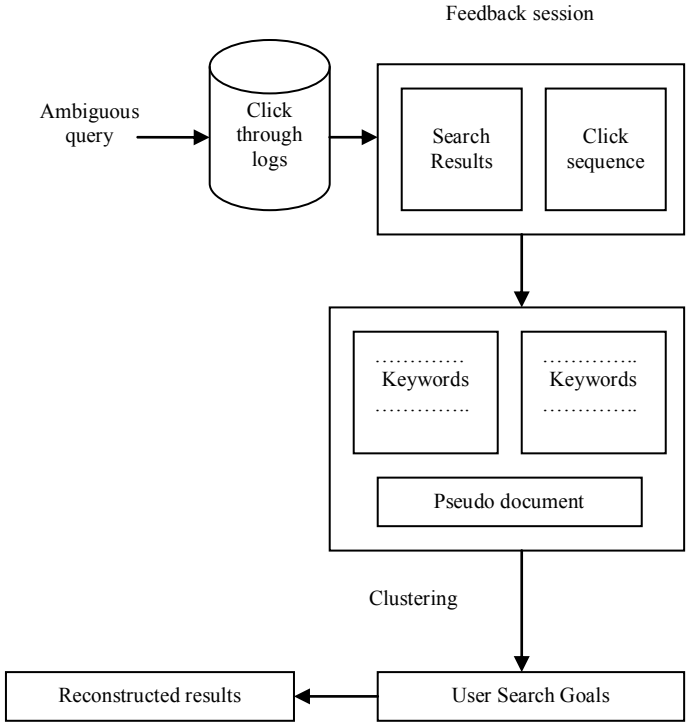

Fig 2. *Architecture Diagram*

Where N is the number of relevant document,
R is the rank and
R_r is the number of relevant document of rank r or less.

- *Calculate the Risk*

$$Risk = \sum_{i,j=1(i<j)}^{m} \frac{d(i,j)}{C_m^2} \quad (11)$$

Where i is the i $_{th}$ clicked URL's
j is the j $_{th}$ clicked URL's.

B. Classified Average Precision (CAP)

$$CAP = VAP \times (1 - Risk)^{\gamma} \quad (12)$$

Where VAP is the voted average Precision and γ is adjust the influence of Risk.

Classified average precision can evaluate the performance of restructuring results. The running time depends on the number of feedback session. Finally the performance can be measured using average precision.

IV CONCLUSION

In the existing system, weightage of title and snippet is the major problem. To overcome this problem in the proposed system, Genetic Algorithm can be used to find the optimal solution for the weightage of title and snippets. It can enhance the performance of user search goals and restructured results.

The complexity of our Approach is low and it can be used in reality easily. Through this approach the search results can be easily categorized or restructured.

REFERENCES

[1] Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Session," IEEE Transactions on knowledge and data Engineering, vol.25, No. 3, march 2013.

[2] R. B. Yates, C. H. Ado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. International Conference of Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[3] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th International ACM SIGIR Conference of Research and Development (SIGIR '07), pp. 783-784, 2007.

[4] H. M. Zahera, G.F. El Hady, W.F Abd El-Wahed," Query Recommendation for Improving Search Engine Results" *Proceedings of the World Congress on Engineering and Computer Science 2010*, Vol I WCECS 2010, October 20-22, 2010, San Francisco, USA.

[5] B. Cao, D. Shen, K. Wang, Q. Yang, "Click through Log Analysis by Collaborative Ranking" *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* 2010 (AAAI-10),

[6] T. Joachims, "Optimizing Search Engines Using Click through Data," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02)*, pp. 133-142, 2002.

[7] 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.

[8] 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.

S. Nancy, she has completed her B. Tech in Raja college of Engineering and Technology, Madurai affiliated to Anna University - Chennai, India and currently a final year M.E student at the Department of Computer Science and Engineering, University college of Engineering (BIT Campus), Tiruchirappalli, India. Her areas of interest are Web Mining and Data mining.

Dr. S. Sathiya Devi, Assistant Professor, Department of Computer Science and Engineering, University college of Engineering (BIT Campus), Tiruchirappalli. Her areas of interest are Web Mining and Image Processing.

