# User Search Goals Optimisation

Nitish Kumar, *Rajat Nagpure,Shubham Channawar,* Bhushan Naikwad , Snehal Rathi

Abstract—For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. We propose a novel approach to infer user search goals by analysing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are on structured from user click through logs and can efficiently react to the information needs of users.. Secondly, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion Classified Average Precision (CAP) to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

*Keywords—Feedback session,Cap evaluation,security and reliablity,web hosting*

## I. INTRODUCTION

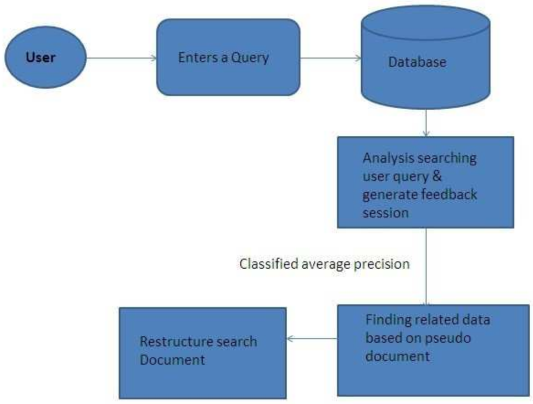
Basically discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a novel approach to infer user search goals for a query by clustering our proposed feedback sessions. Then, we propose a novel optimization method to map feedback sessions to pseudo documents which can efficiently react user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords[5]. The proposed feedback session consists of both clicked and un-clicked URLs and ends with the last URL that was clicked in a single session we propose this novel criterion Classified Average Precision to evaluate the restructure results. Based on the proposed criterion, we also describe the method to select the best cluster number.A search engine is an interesting task itself, it is beyond the scope of this work. As previously discussed, page counts are mere approximations to actual word co-occurrences in the web. However, it has been shown empirically that there exists a high correlation between word counts obtained from a web search engine. A user who searches for apple on the web, might be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining ontology's to capture these new words and senses is costly if not impossible. Generally, the session for a web search is a series of successive queries to satisfy all the single information needed and some clicked search results [1]. In this paper, we focus on inferring user search goals for the particular query. Therefore, the single session containing only 1 query is introduced, which distinguishes from all the conventional session.The feedback session in this paper is based on a single session, although it can extend to the whole session.Each feedback session can tell us what a user requires and what the user does not care about at all. Moreover, there are a plenty of diverse feedback sessions in user click-through various logs[3]. Therefore, for inferring user search goals, it is more efficient to analyze all the feedback sessions than to analyze the search results or the clicked URLs directly.

### *A. Parameters*

1. Feedback Session : Generally, a session for web search is a series of successive queries to satisfy a single information need and some clicked search results. In this paper, we focus on inferring user search goals for a particular query.
2. Pseudo Document Creation : Some representation method is needed. Popular binary vector method to represent a feedback session is used. "0" represents UN-clicked and 1 represent clicked in the click sequence.
3. Clustering : Similar documents are clustered using Fuzzy K-means clustering.
4. Ranking Search Results : Time from when a user enters a request until the first character of the response is received.

For this project we have used fuzzy c means clustering algorithm for the implementation part.Cap evaluation is also used for further implementation of this project.

Information retrieval is that activity of obtaining resources relevant to an information needed from a collection of information resources. Searches can be based on metadata or on fulltext (or other content-based) indexing. The meaning of the term information retrieval (IR) can be quite broad. Every time you look up information to get a task done could be considered as IR.User search goals can be considered as the


User
Enters a Query
Database
Analysis searching user query & generate feedback session
Classified average precision
Finding related data based on pseudo document
Restructure search Document

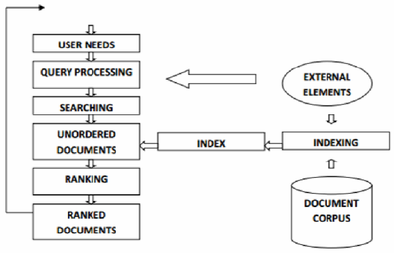
clusters of the specific information needed by the user. The inference and analysis of the user search goals can have a lot of advantages in the improving search engine relevance and various user experience. Advantages are we can restructure web search results according to user search goals by grouping the search results with the same user search goal.

*A. Fuzzy C-means Algorithm*

Basically Data clustering is the process of dividing various data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Then Depending upon the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Various examples of measures that can be used as in clustering include distance, connectivity, and intensity.

In hard clustering, data is divided into distinct clusters, where each data element belongs to the exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one clusters, and associated with each element is a set of membership levels. These usually indicates the strength of the association between that data element and the particular cluster. Fuzzy C-means clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm The algorithm minimizes the intra-cluster variance as well, but has the same problems as that of the k-means; the minimum is a local minimum, and the results depend on the initial choice of the weights.

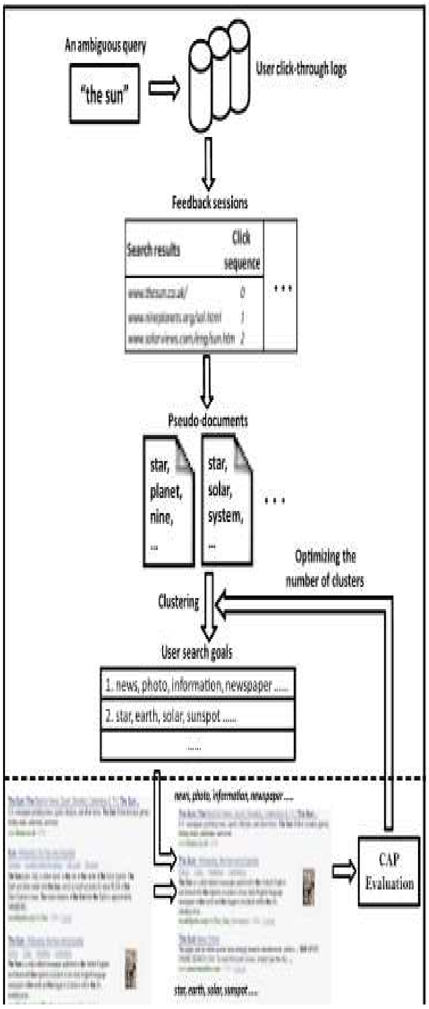


Using a mixture of Gaussian along with the expectationmaximization algorithm in a more statistically formalized method which includes some of these ideas: partial membership in various classes.

Another algorithm closely related to Fuzzy C-Means is Soft K-means.

Fuzzy c-means has been a very important tool for image processing in clustering objects in an image.Mathematicians introduced the spatial term into the FCM algorithm to improve its accuracy of clustering under noise.

Fuzzy clustering is a type of class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic. Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. The FCM algorithm attempts to partition a finite collection of n elements into a collection of c fuzzy clusters with respect to some given criterion. Like the k-means algorithm, the FCM aims to minimize an objective function. For clustering of the pseudo documents, the similarity of the documents is clustered using the fuzzy clustering. The same users in the same session can have different goals at different times. It is thus inappropriate to capture such overlapping interests of the users in crisp clusters. The fuzzy is used to discover all the different search goals. The similarity of the cluster is based on the centroid values. The search goals having least precision in one cluster may have to appear in another cluster with high precision. So discover different search goals for the users, the fuzzy clustering is used. The clusters are very informative and they are stored in the user search goals. In fuzzy c-means clustering, each purpose contains a degree of happiness to clusters, as in the formal logic, instead of happiness fully to only one cluster.Points on the sting of a cluster, is also within the cluster to a lesser degree than points within the centre of cluster. The outline and comparison of various fuzzy cluster algorithms is obtainable. The Purpose x contains a set of coefficients giving the degree of being within the kth cluster wk(x). With fuzzy c-means, the centre of mass of a cluster is that the mean of all points, weighted by their degree of happiness to the cluster. The degree of the happiness, wk(x), is said reciprocally to the gap from x to the cluster centre as calculated on the previous pass. It additionally depends on the parameter m that controls what quantity weight is given to the nearest centre. The fuzzy c-means rule is extremely like the that of the k-means algorithm. Choose variety of clusters. Assign every which way to every purpose coefficients for being within the clusters. Repeat till the rule has converged (that is, the coefficients' modification between 2 iterations isn't any quite, the given sensitivity threshold): Compute the center of mass for every cluster, mistreatment the formula on top of For every purpose, work out its coefficients of being within the clusters, mistreatment the formula on top of.


An ambiguous query
"the sun"
User click-through logs
Feedback sessions
Search results
Click sequence
www.thesun.co.uk/ 0
www.nineplanets.org/sol.html 1
www.solarviews.com/eng/sun.htm 2
...
Pseudo-documents
star, planet, nine, ...
star, solar, system, ...
...
Optimizing the number of clusters
Clustering
User search goals
1. news, photo, information, newspaper .....
2. star, earth, solar, sunspot .....
......
news, photo, information, newspaper ....
star, earth, solar, sunspot ....
CAP Evaluation

CAP evaluation

### *B. Cap Evaluation*

The evaluation of the user search goals can be done with the help of CAP (CLASSIFIED AVERAGE PRECISION).
The classified average precision basically is the calculation of precision of documents.From the user click-through logs, we can get implicit relevance feedback, namely the clicked means relevant and unclicked means irrelevant. A possible evaluation criterion is average precision (AP) which evaluates according to the user implicit feedback. AP basically is the average of precisions computed at the point of each relevant document in the ranked sequence. VAP is the voted average precision which can be used for grouping the dissimilar documents for the particular user query search. Risk is the mapping of similar and dissimilar documents for the particular user query. If there is a similarity then the mapping value is 0 and if there is no similarity between VAP and risk then the mapping value is 1[8].

### *C. Restructure Web-Searched Result*

Restructure internet search results per user search goals by grouping the search results with a similar search goal users with totally different search goals will simply notice what they require. User search goals depicted by some keywords will be used in question recommendation. The distributions of user search goals may be helpful in applications like reranking internet search results that contain totally different user search goals. As a result of its quality, several works concerning user search goals analysis are investigated. They will be summarized into 3 classes: question classification, search result reorganization, and session boundary detection.

### *D. Pseudo Document*

Map feedback session is used to create pseudo documents User Search goals. The building of the pseudo-document includes two steps. One is representing the relevant URLs within the feedback session. Uniform resource locator in a very feedback session which is depicted by a little text paragraph that consists of its title and piece. Then, some of the processes are enforced to those text paragraphs, like remodelling all the letters to lowercases, stemming and removing stop words. Another one is then Forming pseudo-document supported uniform resource locator representations. So as to get the feature illustration of a feedback session, we tend to propose an improvement methodology to mix each clicked and unclicked URLs within the feedback session itself[3].

### *E. Desired Result*

Cluster the pseudo-documents by FCM bunch that is easy and effective. Since we have a tendency to don't understand the precise variety of user search goals for each and every question, we have a tendency to set variety of clusters to be 5 totally different values and perform bunch supported these 5

values, severally. When bunch all the pseudo-documents, every cluster will be thought of mutually user search goal. Then the middle purpose of a cluster is computed because the average of the vectors of all the pseudo-documents within the cluster.

4

3

## III. EXISTING TECHNIQUES

Data clustering is the method in which we make cluster of objects that are some how similar in their characteristics. The criterion for checking the similarity is implementation dependent itself.

Clustering is often confused with the classification, but there is some difference between these two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are also to be defined.

Precisely, Data Clustering is a technique in which, the information which is logically similar is physically stored together. In order to increase the efficiency of the database systems the number of disk accesses are to be minimized[6]. In clustering the objects of similar properties are placed in one of c the lass of objects and a single access to the disk makes the entire class available.

*A. K-means Algorithm* k-means clustering is a method of the vector quantization, originally from signal processing, that is popular for cluster analysis in various data mining.K-means clustering aims in the partition of n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Thus results in a partitioning of the data space into various clusters.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and can converge quickly to the local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use the cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have various shapes.

The algorithm has a loose relationship to the k-nearest neighbors classifier, a popular machine learning technique for the classification that is often confused with k-means because of the k in the name. As One can apply the 1-nearest neighbor classifier on the cluster centers obtained by k-means to classify new data into the various existing clusters[8]. The procedure follows a simple and easy way to classify the given data set through a certain number of clusters (assume k clusters) fixed with priori. The basic idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result.The better choice is to place them as much as possible far away from each other. The next step is that to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grou page is done. At this point we need to re-calculate k new centroids as various barycenters of the clusters resulting from the previous step. After that we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid.FCM is more appropriate algorithm as compared to that of K-means.

## IV. CONCLUSION

A novel approach has been proposed to infer the user search goals for a query by clustering its feedback sessions represented by the pseudo-documents.We introduced feedback sessions to analyze to infer user search goals rather than the search results or clicked URLs. Both the clicked URLs and unclicked ones before the last click are considered as user implicit feedbacks and are taken into account to construct feedback sessions. Therefore, feedback sessions can thus reflect user information which needs more efficiently. Secondly, we map feedback sessions to pseudo-documents to approximate goal texts in user minds. The pseudo-documents can enrich the various URLs with additional textual contents including the titles and the snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, the new criterion CAP is formulated to evaluate the performance of various user search goal inference.

## ACKNOWLEDGMENT

I would like to take this opportunity to thank my internal guide Prof. SNEHAL RATHI for giving me all the help and guidance I needed. I am really grateful to them for their kind support. Their valuable suggestions were very helpful. I am also grateful to Prof. S R Sakhare, Head of Computer Engineering Department, Vishwakarma Institute of Information Technology for his indispensable support, suggestions.

## REFERENCES

[1] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, Varying Approaches to Topical Web Query Classification, Proc. 30th Ann.Intl ACM SIGIR Conf. Research and Development (SIGIR 07),pp. 783-784, 2007.

[2] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, Learning to Cluster Web Search Results, Proc. 27th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 04), pp. 210-217, 2004.

[3] U. Lee, Z. Liu, and J. Cho, Automatic Identification of User Goals in Web Search, Proc. 14th Intl Conf. World Wide Web (WWW 05), pp.
391-400,2005. M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain, Organizing and Searching the World Wide Web of Facts - Step One: The One Million Fact Extraction Challenge, Proc. Natl Conf. Artificial Intelligence (AAAI 06),2006. A New Algorithm for Inferring User Search

oals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang

Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng

[4] X. Li, Y.-Y Wang, and A. Acero, Learning Query Intent from Regularized Click Graphs, Proc. 31st Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 08), pp. 339-346, 2008.

[5] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, Accurately Interpreting Clickthrough Data as Implicit Feedback, Proc. 28th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 05), pp. 154-161, 2005

[6] M. Sulaiman1, K.Ananthajothi2, Syed Zubair Ahmed Hussainy,S.Jayakrishnan4 Department of Computer Science and Engineering Aalim Muhammed Salegh College of Engineering, Avadi,Chennai,India