INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

IMPROVED ALGORITHM FOR INFERRING USER SEARCH GOALS WITH FEEDBACK SESSIONS

[1]A.Sangeetha, [2]C.Nalini

[1]Phd scholor,Bharat University,Chennai, sangeethagsam@yahoo.co.in.
[2]Professor,Bharat University,Chennai, drnalinichidambaram@gmai.com.

***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. In this paper, we propose a novel approach to infer user search goals by analyzing 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 constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion to evaluate the performance of inferring user search goals using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

***Keywords*—** user query logs, Feed Back Sessions, Clustering, user Click through logs

I. INTRODUCTION

The information retrieval goal is to find the documents that are most relevant to a certain Query. The problem of information retrieval is to find the documents that are relevant to an information need from a large document. It deals with notions of Collection of documents, Query (User's information need), Notion of Relevancy. The types of information's are text, audio, video, xml structured and documents, source code, application and web services. The types of information needs are Retrospective, Prospective (Filtering). Retrospective means "searching the past". The different queries are posed against a static collection. Prospective means "Searching the future". The static queries are posted against a dynamic collection. It is time dependent. The components in information retrieval are user, process, and collection. User- What computer cares about? Process and collection tends to what we care about. The information retrieval cycle consists of five phases. Source selection, query formulation, search, selection and result. The search process consists of Index and document collection. The indexing is a Black box function; its process is not visible. The main tasks of information retrieval are indexing the documents, process the query, evaluate similarity and find ranking and display the results. The documents are searching that are most closely matching the query. The indexing consists of stop word removal and stemming and inverted index. The removal of stop word usually improves the effectiveness of information retrieval. The lists of stop words are about, afterwards, according, almost, above etc [12].

The stemming is based on suffix stripping. The reason for stemming is that the words that have similar meaning to each other. The stemming removes the some ending of words. E.g.: include, including, includes, included. A porter

algorithm is used for suffix striping. The results of indexing are based on some set of weighted keywords. The results of indexing are in the form of [10]:

D1= {(t1, w1,), (t2, w2).....}. **(1)**

Inverted file is used for retrieving the information for higher frequency.

II. PROBLEMS IN INFORMATION RETRIEVAL

- How we represent the documents with selected keywords?
- How document and query representations are compared to calculate the weight?
- Mismatching of vocabularies.
- Ambiguous query.
- Depicting of content may be incomplete and inadequate.

The effectiveness of information retrieval can be improved based on keywords. The keywords cover only the part of contents.[13] User can identify the relevant/irrelevant documents based on the weight of the words. We need to be interacting with user and getting the user feedback. The evaluation is based on recall and precision. The more information retrieval process available is open source IR tool kits.

III. WEB MINING

The World Wide Web has been dramatically increased due to the usage of internet. The web acts as a medium where large amount of information can be obtained at low cost. The information available in the web is not only useful to individual user and also helpful to all business organization, hospitals, and some research areas. The information available in the online is unstructured data because of development technologies**.** Web mining can be defined as the discovery and analysis of useful information from the World Wide Web data. [14] It is one of the data mining techniques to automatically extract the information from web documents. The three issues in the WWW are web content mining, web structured mining, web usage mining. Web structure mining involves web structure documents and links. Web content mining involves text and document and structures. Web usage mining includes data from user registration and user transaction. WWW provides a rich set of data for data mining. The web is dynamic and very high dimensionality. It is very helpful to generate a new page, lot of pages are added, removed and updated anytime. Data sets available in the web can be very large and occupy ten to hundreds of terabytes, need a large farm of servers. A web page contain three forms of data, structured, unstructured and semi structured data. A number of algorithms are available to make a structured data, one such algorithm is a fuzzy self constructing. An unstructured data can be analyzed using term frequency, document frequency, document length, text proximity.
We have to improve searching in the web by adding structured documents. Using clustering techniques we have to restructure the web information. We provide a hierarchical classification of documents using web directories Eg: Google. While increasing the annual band width in ten times its average is increasing three times, because of that the traffic management is important in web mining.

IV. LITERATURE REVIEW

A. Google keyword tool

Google's keyword tool is free online researches tool to helps the user to find the appropriate keywords**.** The tool based on ranking is shaped by two participants: Search engine company programmers, Webmasters and SEO practitioners Search engine users [2].
Google ranking algorithm is based on "click through rate" and "bounce rate".

1) Click through Rate

In a given query, how many percentage of time, the user clicks on a particular URL in a web page.

2) Bounce Rate

The bounce rate has a reverse effect; it calculates that how many searchers clicked in a particular webpage.

B. Web Cap browser side tool

Web cap is a browser side tool. It does not maintain log information. It can collect relevant information from different user based on the different user search goals. Web cap use the different inference algorithm to collect the relevant documents. It involves two steps: i) collecting the relevant information based on the user's interests ii) apply inferring algorithm and different techniques to learn about the information. Use implicit indicators associate with each user behaviour's action to compute the user's degree of interest [3]. Implicit indicators such as reading actions such as reading time, mouse clicks, mouse movements and scrollbar movements and saving action and printing action.

C. Page Ranking Algorithm

The page rank algorithm determined the rank for single page not for whole website. Based on the user clicking, the page rank can be determined. The original page rank algorithm is [4]

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + +PR(Tn)/C(Tn))

Where:

PR(A) is the Page Rank of page A,

PR(Ti) is the Page Rank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti d is a damping factor which can be set between 0 and 1.How to increase the page rank value? Sum of the page rank is same; we can add a new page to the web site. Increase the number of links with other inbound links. A one off website is considered as a small web, if we want to raise a page rank, add a new page to it. When you add a new page, be sure that link connected to the front page is correct.

D. Lexical pattern Extraction algorithm

In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. We need to identify the numerous semantic relations that exist between two given words. Efficient retrieval is based on words. The information retrieval is problematic when the words are same. [5] On the other hand, the snippets returned by a search engine for the conjunctive query of two words provide useful clues related to the semantic relations that exist between two words. A snippet contains a window of text selected from a document that includes the queried words. Snippets are useful for user while searching, user can learn the snippets in the without opening the URL. Using snippets is efficient it is helpful to download the documents from the web; it is time consuming when the document is large.

E. Construction of virtual documents

The virtual document terms are extracted from its detailed graph. [6] The virtual document of a concept consists

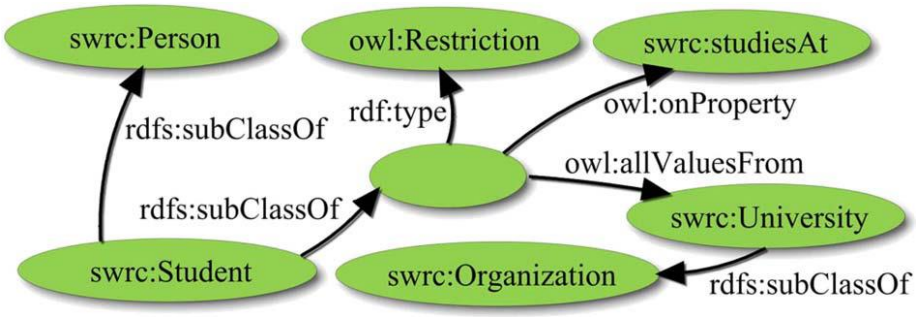


Figure1: RDF graph [6].

of terms extracted from its description graph: For each entity identified by a URI in this graph, extract its local name and label; for each literal in this graph, extract its lexical form. For example, in Fig. 1, the virtual document of swrc:Student consists of terms "Student," "subclass Of," "Person," "type," "Restriction," "on Property," "studiesAt," "allValuesFrom," and "University." It is worthy to property names are also included so that the system can support more diverse keyword queries, e.g., swrc: Student can be retrieved by the keyword query "subclass of person."

F. User and query similarity model

When the multiple users in the organization, the two users asking the same set of queries. The ranking functions have been derived on browser choices. The user similarity between the two users can be expressed as the average similarity between the ranking functions.

The goal of the similarity is to determine the ranking functions derived from the similar query for the similar user.

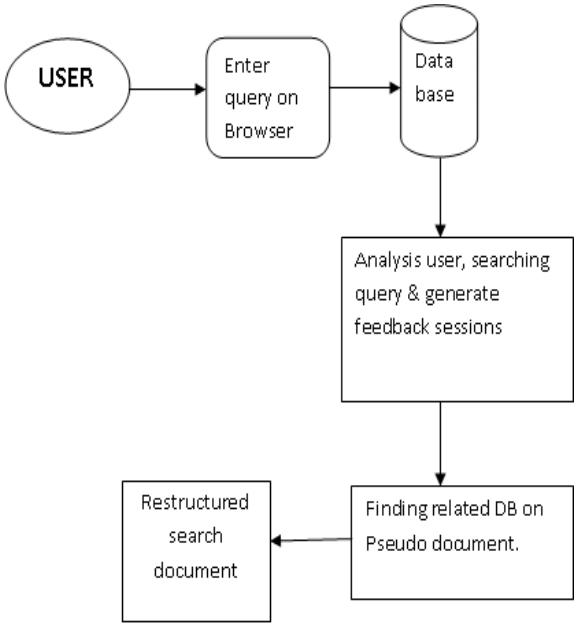
When the user in the system enters the query in the browser, the results are obtained. Let be the ranking functions derived for the each individual queries. The same query having different ranking functions. We cannot choose ranking functions randomly. In order to ascertain the correct ranking function, we use the concept called query similarity. It has two distinct approaches, query condition similarity and query result similarity [7]. distinct approaches, query condition similarity and query result similarity [7].

V. PROPOSED TECHNIQUE

We propose a framework to infer different user search goals for a query by clustering feedback sessions. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.

To find the initial data cluster center points is a challenged process. [15] This algorithm improves the efficiency of the process and speed up the calculation and automatically iterates each other in every time. The existing K means algorithm is based on some constant factors; it does not automatically each other [11]. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries. 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. We propose an propose an algorithm to estimate the semantic analysis of similarity between words or entities using web search engines. Because of numerous documents and the high and vast growth rate of the web, it is time consuming to analyze each document individually.

Architecture diagram



"Fig. 1", Shows the Architecture diagram .

Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most web search engines. Page count of a query is an estimate of the number of pages that contain the query words.

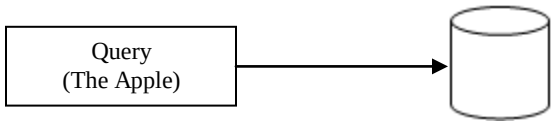
We first propose an approach to infer user search goals for a query by clustering the proposed feedback sessions. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Then, we propose a optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Finally, we cluster these pseudo documents based on user search goals and depict them with some keywords. The evaluation of clustering is also an important problem, we also propose a criterion called classified average precision (CAP) to evaluate the performance of the restructured web search results. We also demonstrate that the proposed criterion can help us to optimize the parameter in the clustering method when inferring user search goals.

VI. ILLUSTRATION OF FEEDBACK SESSIONS

A. Ambiguous Query

Queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get different types of information on different aspects when they submit the same query. We cannot guess the user behavior exactly. For example, when the query "The apple" is submitted to a search engine, some users want to learn fruit apple, while some others want to learn the apple iphone, iPod etc.

An Ambiguous Query



"Fig. 2", Query Search.

B. User Search Goals

We cluster pseudo-documents by FCM clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we set number of clusters to be five different values and perform clustering based on these five values, respectively. After clustering all the pseudo-documents, each cluster can be considered as one user search goal. The center point of a cluster is computed as the average of the vectors of all the pseudo-documents in the cluster.

C. Restructure web search results

Based on the user search goals we have to restructure the web search results by grouping the search results with the same search goal users with different search goals can easily find what they want. User search goals represented by some keywords can be utilized in query recommendation. The distributions of user search goals can also be useful in applications such as re-ranking web search results that contain different user search goals. Due to its usefulness, 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.

D. Feedback Sessions

The feedback session is formed by both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks. Feedback session can tell what a user requires and what he/she does not care about. Moreover, there are plenty of diverse feedback sessions in user click-through logs. Therefore, for inferring user search goals, it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

E. Pseudo document

In this paper, we need to map feedback session to pseudo documents User Search goals. The building of a pseudo-document includes two steps. One is representing the URLs in the feedback session. URL in a feedback session is represented by a small text paragraph that consists of its title and snippet. Then, some textual processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Another one is Forming pseudo-document based on URL representations.

In order to obtain the feature representation of a feedback session, we propose an optimization method to combine both clicked and unclicked URLs in the feedback session**.**

VII. FEATURE CLUSTERING

Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. When all the words have been fed in, a desired number of clusters are formed automatically. We then have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experimental results show that our method can run faster and obtain better extracted features than other methods.

A. Explanation of clustering

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. Depending on 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. Some 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 exactly one cluster. These indicate the strength of the association between that data element and a particular cluster.

- Choose a number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than, the given sensitivity threshold):
- Compute the centroid for each cluster, using the formula above.
- For each point, compute its coefficients of being in the clusters, using the formula above.

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights.
Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes.

VII. ASSOCIATED WORK

In recent years, many works have been done to infer the so called user goals or intents of a query. But in fact, their works belong to query classification. Some works analyze the search results returned by the search engine directly to exploit different query aspects. However, query aspects without user feedback have limitations to improve search engine relevance. Some works take user feedback into account and analyze the different clicked URLs of a query in user click-through logs directly, nevertheless the number of different clicked URLs of a query may be not big enough to get ideal results. However, their method does not work if we try to discover user search goals of one single query in the query cluster rather than a cluster of similar queries. However, their method only identifies whether a pair of queries belong to the same goal or mission and does not care what the goal is in detail. A prior utilization of user click-through logs is to obtain user implicit feedback to enlarge training data when learning ranking functions in information retrieval. In our work, we consider feedback sessions as user implicit feedback and propose a novel optimization method to combine both clicked and unclicked URLs in feedback sessions to find out what users really require and what they do not care. One application of user search goals is restructuring web search results. There are also some related works focusing on organizing the search results. In this paper, we infer user search goals from user click-through logs and restructure the search results according to the inferred user search goals.

VIII PROPOSED SYSTEM

The user enters the queries to the search engine. The queries are maintained as a log and the results will be produced based on the keywords. The search goals for a query and depicting each goal with some keywords automatically .The user's queries are saved. The feedback sessions is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs.

Combine the enriched URL's in a feedback sessions to form a pseudo document. The feedback session is based on a single session .and also it can be extended to the whole session. So besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks. For inferring user search goals it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly. Map the feedback sessions to pseudo-documents which can effectively reflect user information needs.

The highest values in the center points are used as the keywords to depict user search goals. Similar queries may not share query-terms but they do share terms in the documents selected by the users. It avoids the problems of comparing and clustering sparse collection of vectors in which similar queries are difficult to find a problem that appears in previous works on clustering.

The results are restructured based on the evaluation of web search goals. Search engines will returns millions of search results so I is necessary to organize them to make it easier for users to find what they want. The user search goals are represented as the vectors .So, perform categorization by choosing the smallest distance between the URL vector and user-search –goal vectors. By this way the results can be restructured according to the inferred user search goals.

IX. CONCLUSION

In this paper, we proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo documents. First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion 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.

REFERENCES

[1] How to Use Search Engine Optimization Techniques to Increase Website Visibility JOHN B. KILLORAN, IEEE TRANSACTIONS ON PROFESSIONAL COMMUNICATION, VOL. 56, NO. 1, MARCH 2013.

[2] WebCap: Inferring the user's Interests based on a Real-Time Implicit Feedback. Nesrine zemrili, Information system department,978-1-4673-2430-4/12-2012.

[3] A Collaborative Decentralized Approach to Web Search Athanasios Papagelis and Christos Zaroliagis, *Member, IEEE* IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 5, SEPTEMBER 2012.

[4] A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 7, JULY 2011.

[5] Correspondence Falcons Concept Search: A Practical Search Engine for Web Ontology Yuzhong Qu and Gong Cheng IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 41, NO. 4, JULY 2011.

[6] One Size Does Not Fit All: Towards User & Query Dependent Ranking For Web Databases Aditya Telang, Chengkai Li, Sharma Chakravarthy *Department of Computer Science and Engineering, University of Texas at Arlington* July 16, 2009

[7] Long-Term Cross-Session Relevance Feedback Using Virtual Features Peng-Yeng Yin, Bir Bhanu, Fellow, IEEE, Kuang-Cheng Chang, and Anlei Dong, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 3, MARCH 2008.

[8] Automated Ranking of Database Query Results, Sanjay Agrawal, Surajit Chaudhuri, Gautam Das,Microsoft Research., Aristides Gionis Computer Science Dept, Stanford University, Proceedings of the 2003 CIDR Conference.

[9] An Efficient k-M,eans Clustering Algorithm: Analysis and Implementation Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002

[10] An introduction through information retrieval Power point slides. Jian-Yun Nie ,University of Montreal,Canada.

[11] A New Algorithm for Inferring User Search Goals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng

[12] [Online]. Available: Introduction to Information Retrieval, Jian-Yun Nie University of Montreal Canada.

[13] [online]. Available: Prof. Navneet Goyal BITS, Pilani. PPt.

[14] http://www.henzinger.com/~monika.

[15] An Efficient Fuzzy c-means Clustering Algorithm. Ming-chuan and Don-lin Yang. Dept of Information Technology, FengChiaUniversity.