International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

Vol.4., Issue.2., 2016
(Mar-Apr)

RESEARCH ARTICLE

ISSN: 2321-7758

THE CLASSIFIED AVERAGE PRECISION TECHNIQUE FOR EFFECTIVENESS OF DATA EXTRACTION GOALS

KAVITHA.D[1], DHIVYA.S[2]

[1,2]Valliammai Engineering College, SRM Nagar, Kattankulathur
Affiliated to Anna University.

ABSTRACT

The inference and analysis of data extraction goals can be very useful in improving search engine relevance and user experience. In this paper, we propose a novel approach to infer search goals by analysing search engine query logs. A framework to discover different 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. This approach is to generate pseudo-documents to better represent the feedback sessions for clustering. A new criterion "Classified Average Precision (CAP)" is to evaluate the performance of inferring user extraction goals. Results are presented using user click-through logs from a commercial search engine to validate the effectiveness of proposed methods.

Key words—classified average precision, clustering, feedback sessions, pseudo document

INTRODUCTION

In Web search applications, queries are submitted to search engines to represent the information of relevant data [5]. However, sometimes queries may not exactly represent the specific information needs since many ambiguous queries may cover a broad topic and different users need to get information on different aspects when they submit the same query [1]. Therefore, it is necessary and potential to capture different data extraction goals in information retrieval[4].

First, we can restructure web search results according to search goals by grouping the extracted results with the same search goal [3].Second, user search goals represented by some keywords can be utilized in query recommendation thus the suggested queries can help users to form their queries more precisely[8]. Third, the distributions of user search goals can also be useful in applications such as re ranking web search results that contain different data extraction goals.

II. COMPARISON BETWEEN EXISTING AND PROPOSED SYSTEM

Three methods are compared. They are described as follows:

1. It Clusters the top 20 search results to get user search goals. First, we program to automatically submit the queries to the search engine again and crawl the top 20 search results including their titles and snippets for each query[6]. Finally, we cluster these 20 search results of a query to infer user search goals by clustering.

Table 2.1 Shows the calculation of the document searched by means of given query.

	Our Method	Method 2
Average Risk	Average VAP	Average VAP
10	2	1.5

20	2.2	1.547
30	2.2	1.8
40	2.4	1.84
50	2.51	1.7
60	2.54	1.6
70	2.6	1.5
80	2.54	1.57
90	2.659	1.88

2. It clusters the different clicked URLs directly. In user click-through logs, a query has a lot of different single sessions[7]; however, the different clicked URLs may be few. First, we select these different clicked URLs for a query from user click- through logs and enrich them with their titles and snippets as we do in our method[12].Finally, we cluster these different clicked URLs directly to infer user search goals.

Table 2.2 Shows the calculation of clicked URL'S.

Average Risk	Our Method Average VAP	Method 1 Average VAP
10	2	1.68
20	2.2	1.68
30	2.2	1.39
40	2.4	1.37
50	2.51	1.56
60	2.54	1.29
70	2.6	1.54
80	2.659	1.54
90	2.4	1.47

To show that when inferring user search goals, clustering our proposed feedback sessions are more efficient than clustering search results and clicked URLs directly, we use the different framework and clustering method. Comparison of three methods for 20 queries[2]. Each point represents the average Risk and VAP of a query when evaluating the performance of restructuring the search results[9]. Some data reorganization is performed to the data set. The performance evaluation and comparison are based on the restructuring web search results[10].

We compare three methods for all the 20 queries. Compares our method with Method I and compares ours with Method II. Risk and VAP are used to evaluate the performance of restructuring search results together[11]. Each point represents the average Risk and VAP of a query. If the search results of a query are restructured properly, Risk should be small and VAP should be high and the point should tend to be at the top left corner[14]. We can see that the points of our method are closer to the top left corner comparatively[13]. Analyze the Advantages of Clustering Feedback Sessions .In this section, we will give some intuitive explanation showing why clustering feedback sessions namely pseudo- documents is better than the other methods when inferring user search goals[15].

III.OVERALL SYSTEM ARCHITECTURE

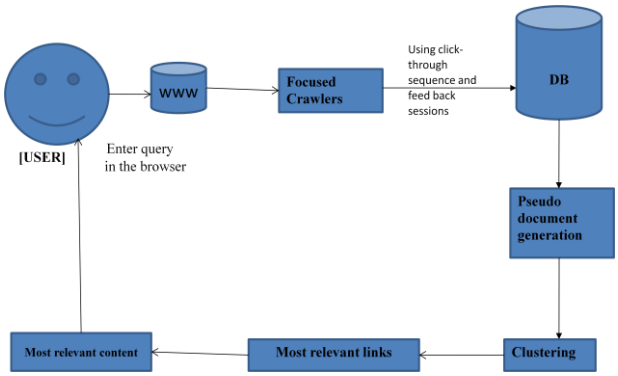

Fig 3.1 Architecture for Extracting Most Relevant Content.

Description: First the user enters the query in the browser. After that using click-through sequence and feedback sessions the data is stored in the database. After that the pseudo document is generated. Finally the most relevant link and content is got by means of clustering.

IV. PROPOSED ALGORITHM

Mean shift clustering: In statistics and data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Voronoi cell. The problem is computationally difficult (NP-Hard), however there are efficient heuristic algorithms that are commonly employed that converge fast to a 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 cluster centers to model the data, however k-means clustering tends to find Clusters of comparable spatial extend, while the expectation-maximization mechanism allows clusters to have different shapes.

Overview

A clustering algorithm An approximation to an NP-hard combinatorial optimization problem .It is unsupervised "*K*" stands for number of clusters, it is a user input to the algorithm. From a set of data points or Observations (all numerical), *K*-means attempts to classify them into *K* clusters. The algorithm is iterative in nature.

ALGORITHM 1

Step1 :begin initialize DOM tree T, Node *textnode *node
Step 2 :textnode = BeginTextNode(T);
Step 3 :do
Step 4 :{
Step 5 :node = textnode;
Step 6 :text = node
Step 7 :text;
Step 8 :}
Step 9 :do
Step 10 :{
Step 11 :if (node)
Step 12 :parent isn't linebreak node
Step 13 :node = node parent;
Step 14 :continue;
Step 15 :}
Step 16 :if node is left-most child node
Step 17 :text = enter + text;
Step 18 :if node is right-most child node
Step 19 :text = text + enter;
Step 20 :add text to text list;
Step 21 :break;
Step 22 :until node parent == null;
Step 23 :textnode = NextTextNode(textnode);
Step 24 :until textnode = null;
Step 25 :end

ALGORITHM 2

Text-To-Tag Ratio pseudocode
Step1 :input
Step2 :h ←HTML source code begin
Step 3 :Remove all script, remark tags and empty lines for each line k to numLines(h) do
Step 4 :x←number of non-tag ASCII characters in h[k]
Step 5 :y←number of tags in h[k]
Step 6 :if y = 0 then
Step 7 :TTRArray[i] ← x else
Step 8 :TTRArray[i] ← x / y end if
Step 9 :end for
Step 10 :return TTRArray end

V.CLASSIFICATION OF DATA EXTRACTION GOALS.

Due to its usefulness, many works about data extraction goals analysis have been investigated. They can be summarized into three classes: query classification, search result reorganization, and session boundary detection.

In the first class, people attempt to infer relevant data and intents by predefining some specific classes and performing query classification accordingly .consider data extraction goals as "Navigational" and "Informational" and categorize queries into these two classes. Define query intents as "Product intent" and "Job intent" and they try to classify queries according to the defined intents. Other works focus on tagging queries with some predefined concepts to improve feature representation of queries. However, since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical. In the second class, people try to reorganize search results. However, this method has limitations since the number of different clicked URLs of a query May be small. Other works analyze the search results returned by the search engine when a query is submitted. Since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well.

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 information on different aspects when they submit the same query. For example, when the query "the sun" is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun.

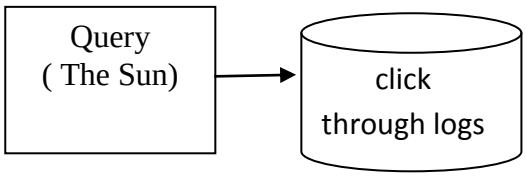

Fig 5.1 Query is Passed to click through Logs.

B. Retrieval Of Link Based On Click Sequences And Feed Back Sessions

We need to restructure web search results according to user search goals 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.

The feedback session consists of both clicked and unlocked 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 unlocked 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.

Search result	click sequence
www.thesun.co.uk	0
www.solarviews.com/sun.html	1

Fig 5.2 Click Sequences At the Backend.

C.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 unlocked URLs in the feedback session.

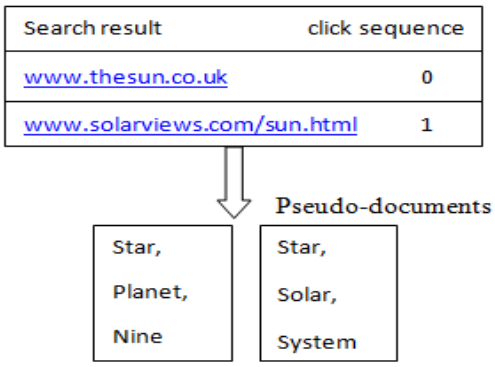

Fig 5.3 Generation of Pseudo documents.

D. Data Extraction

We cluster pseudo-documents by K-means clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we set K 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.

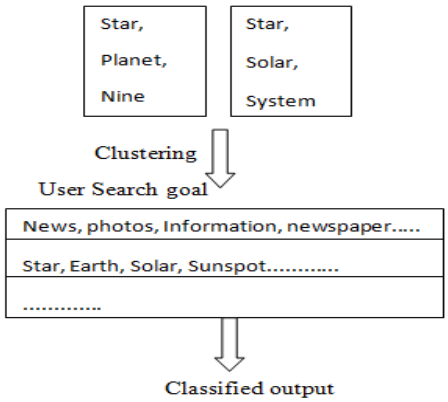

Fig 5.4 Clustering And Extracting The Most Relevant content.

VI.IMPLEMANTATION OF DATA EXTRACTION GOALS

A. Ambiguous Query

Fig 6.1 Here the query to be searched is typed.

B.Retrieval Of Links Based On Click Sequences And Feedback Sessions

Fig 6.2 The given query is searched by means of web.

Fig 6.3 The Set Of Links Are Displayed here.

Fig 6.4 The clicked link is submitted for generating click sequence.

Fig 6.5 The count of clicks are stored at the backend.

C.Pseudo Document

Fig 6.6 To Generate Pseudo Document (i.e)To Mine Pseudo Words From Click Through.

Fig 6.7 Again click search from web then submit. Here The Final Results Must Be clustered Based On Pseudo Document.

D.Data Extraction

Fig 6.8 The result is displayed by means of clustering.(i.e) After mining the pseudo words from click-through.

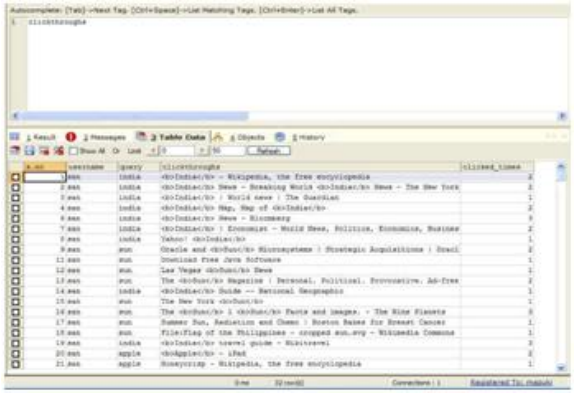

Fig 6.9 Backend

Click Sequence initially at the backend.

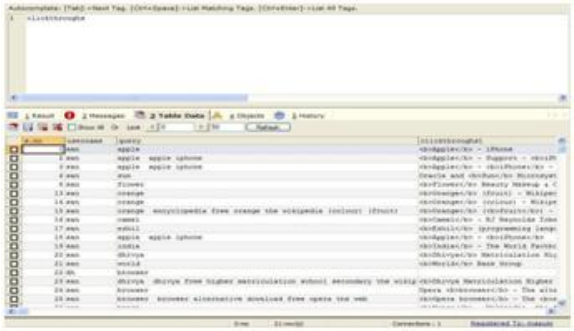

Fig 6.10 Click Sequence after the link is clicked Shows the number of times a particular link is clicked.

VII.APPLICATION OF ALGORITHM

Mean shift clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision, geostatistics and astronomy to agriculture[10]. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration.

Basic mean shift clustering algorithms maintain a set of data points the same size as the input data set. Initially, this set is copied from the input set[3]. Then this set is iteratively replaced by the mean of those points in the set that are within a given distance of that point. By contrast, *k*-means restricts this updated set to *k* points usually much less than the number of points in the input data set, and replaces each point in this set by the mean of all points in the *input set* that are closer to that point than any other (e.g. within the Voronoi partition of each updating point)[8]. A mean shift algorithm that is similar then to *k*-means, called *likelihood mean shift*, replaces the set of points undergoing replacement by the mean of all points in the input set that are within a given distance of the changing set[6]. One of the advantages of mean shift over *k*-means is that there is no need to choose the number of clusters, because mean shift is likely to find only a few clusters if indeed only a small number exist. However, mean shift can be much slower than *k*-means. Mean shift has soft variants much as *k*-means does.

VIII.CONCLUSION AND FUTURE WORK

In this paper, a novel approach has been 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 unlocked 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 CAP 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.The complexity of our approach is low and our approach can be used in reality easily. For each query, the running time depends on the number of feedback sessions. Therefore, the running time is usually short. In reality, our approach can discover user search goals for some popular queries offline at first. Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently.Our solutions can be extended in several directions Here we focused only on Text Mining (i.e.) extracting the relevant content based on the given query in the form of text. This relevant content extraction can be done for images and for videos using different algorithm thereby reducing the complexity and also it can be used in reality.Feedback sessions can be considered as a process of re sampling. If we view the original URLs in the search results as original samples, then feedback sessions can be viewed as the "processed" or "resample" samples which differ from the original samples and reflect user information needs. Without resampling, there could be many noisy URLs in the search results, which are seldom clicked by users[12]. If we cluster the search results with these noisy ones, the performance of clustering will degrade greatly.

Comparison of our method with Method 1

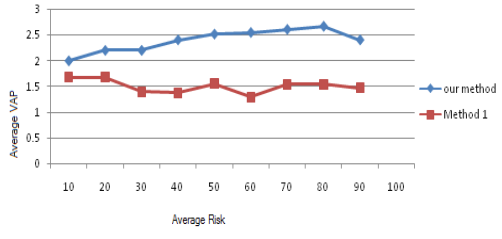

Fig 7.1 Graph For Average Risk of query.

Feedback session is also a meaningful combination of several URLs. Therefore, it can reflect user information need more precisely and there are plenty of feedback sessions to be analyzed[2]. The solid points represent the clicked URLs mapped into a 2D space and we suppose that users have two search goals: the star points belong to one goal and the circle points belong to the other goal[10]. They represent a feedback session which is the combination of several clicked URLs. (In order to clarify the problem, we consider that feedback sessions only consist of click URLs here[14].

Comparison of our method with Method 2

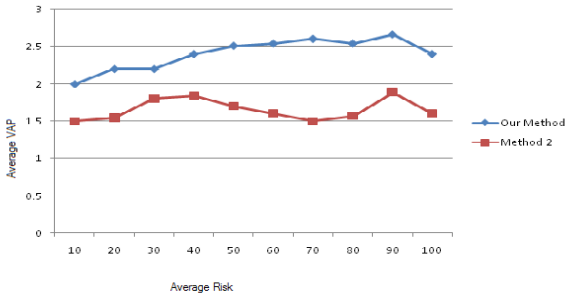

Fig 7.2 Graph After Restructuring Web Results

IX.REFERENCES

[1]. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.represent class boundaries. Supposing that the clicked URLs have two classes, the points in the left figure are hard to be segmented directly, while

[2]. R. Baeza-Yates, C. Hurt ado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf.

[3]. D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf.

[4]. S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "VaryingApproaches to Topical Web Query Classification," Proc. 30th Ann.

[5]. H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li,"Context-Aware Query Suggestion by Mining Click-Through,"Proc.

[6]. Daxin Jiang,Jian Pei,Qi Hei,"Context Aware Query Suggestion by Mining Clickthroughs and Session Data".

[7]. K.M.Sam,C.R.Chatwin,"Ontology Based Text Mining For Social Network Analysis".

[8]. Yuvan Hong,Jaideep Vaidhya and Haibing Lu,"Search Engine Query Clustering Using Top-K Search Results".

[9]. Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. Lee Giles,"Automatic Identification Of Informative Sections Of Web Pages".

[10]. Jingqi Wang, Qingcai Chen, Member, IEEE SMC, Xiaolong Wang, Member, IEEE SMC, Hongzhi Guo," Basic Semantic Units Based Web Page Content Extraction".

[11]. Fernandes–Villamor,M.Garijio,(2013)" A Frame work For Goal Oriented Discovery Of Resources In The Restful Architecture," IEEE Early AccessArticles ,Pages 1.

[12]. Hao Chan,(2012)"Bringing Order To The Web :Automatically Categorizing search Results,"School of Information Managemant And Systems Universityof california.

[13]. Huanhuan Cao,Daxin Jiang Jian Pei(2008),"Context Aware Query Suggestion By Mining Click Through And Session Data,"Future Information Technology And Management Engineering.

[14]. Jingqi Wang,Qingcai chen(2008),"Basic Semantic Units Based Web PageContent Extraction,"Intl Conf, Pages 1489-1494.

[15]. Steven M.Beitzel,E.Jensen(2011),"Varying Approahes To Topical Web Query Classification,"Proc.30th Ann.