International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4403 - 4408

A New Approach of Clustering Feedback Sessions for Inferring User Search Goals

Mr. Ajinkya A. Godbole
ME Student,
Department of Computer Engineering,
P.V.P.I.T., Bavdhan,
Savitribai Phule Pune University,
Pune, Maharashtra, India
ajinkyagdbl@gmail.com

Mrs. V. S. Nandedkar
Assistant Professor,
Department of Computer Engineering,
P.V.P.I.T., Bavdhan,
Savitribai Phule Pune University
Pune, Maharashtra, India
vaishu111@gmail.com

Abstract — Internet information is growing every day exponentially. In order to find out the exact required information from this web search engines has become absolutely necessary tool for the web users. It has also become more difficult to provide user the required information. When Different users provide an ambiguous query to a search engine, they might be having different search goals. Therefore, it is required to find and analyze user search goals to improve the performance of a search engine and user experience. By representing the results in cluster we find out different user search goals for a query. It has advantages in improving search engine relevance and user experience. It extends the delivery and quality of internet information services to the end user. It also improves performance of Web server system. Query classification, search result reorganization and session boundary detection are the approaches attempt to find out user search goals. But the mentioned approaches has many limitations. A new approach has been implemented that overcomes the limitations and analyze, discover user search goals using feedback sessions. This approach first takes the user search query. For each single result of the search query pseudo-documents are generated. Using K-means++ clustering algorithm, these pseudo-documents are clustered. Each cluster can be considered as one user search goal. Finally in restructured result is given to the user where each URL is categorized into a cluster centered by the inferred search goals. Then depending upon user click through, results are restructured and represented to the user in order to satisfy the information need.

Keywords- *search engines, user search goals, feedback sessions, pseudo-documents*

I. INTRODUCTION

Search engine is the most important application in today's internet. User needs some information and thus queries to internet in order to get the result. Most of the times these queries are ambiguous. Means user is expecting information in one topic is not returned by the search engine as search engine interprets the query differently. For example, when the query is "gladiator". It is hard to determine what user is expecting in result in such scenarios as query is ambiguous. It is hard for a search engine to decide if the user is interested in history of a gladiator or list of famous gladiators or the film gladiator.

Search engine is the most important application in today's internet. User needs some information and thus queries to internet in order to get the result. Most of the times these queries are ambiguous. Means user is expecting information in one topic is not returned by the search engine as search engine interprets the query differently. For example, when the query is "gladiator". It is hard to determine what user is expecting in result in such scenarios as query is ambiguous. It is hard for a search engine to decide if the user is interested in history of a gladiator or list of famous gladiators or the film gladiator.

Without looking at the context of search, search engine suggests many queries with very low accuracy. Thus it is required to capture user search goal. Information need is nothing but a user's desire to satisfy his/her need. In order to improve user search goals the inference and analysis of goals have a lot of advantages. First advantage is web search results can be restructured [9], [6], [7] according to user search goals by grouping the search results with the same search goal. Another advantage is the usage of keywords to represent user search goals in the query suggestion [10], [11], [12]. Third advantage would be reranking of web search results can also be done with the distribution of user search goals.

User search goals can be represented in following three classes: Query classification, Reorganization of Search Result and Session Boundary Detection. In Query classification, classification is done depending upon some predefined classes. User goals are either navigational or informational. In case of navigational user goal user has web page in mind. In case of informational user does not have any particular page in

mind. In case of search result reorganization user try to recognize search result. This is done either by learning aspects of queries by analyzing the clicked URLs or by analyzing search results returned by a search engine. In third method the main aim is to detect session boundaries. Feedback session ends with the last URL clicked in a session and contains both clicked and unclicked URLs.

The rest of the paper is organized as follows: Section II is about the previous and current methods in use and also comparison of different clustering methods in use. Section III talks about the implemented system working. Result analysis is discussed in Section IV. Finally paper ends with the conclusion in section V.

II. LITERATURE SURVEY

A. Automatic identification of user goals:

Uichin Lee, Zhenyu Liu, Junghoo Cho [3], proposed automatic identification of user search goals. Majority of queries have a goal which is predictable was the statement of them. Classification of query goals based on two types:

A1. Navigational queries

In case of navigational user has web page in mind. User may have visited that site before or predicts that site may exist.

A2. Informational queries

In case of informational user does not have any particular page in mind. User also may intend to visit different pages to know about the topic. In this type user keeps on exploring webpages. User does not have a guarantee which page is going to have correct required answer.

For the prediction of user goal two features are used:

1. Past user-click behavior:

In case of navigational, users has a result in the mind and will click on that result. So, user goal can be identified by Observing the past user-click behavior.

2. Anchor-link distribution:

If the user is associating query with website then links with the anchor will point to respective websites. So potential goal of the query can be identified by observing destinations of the links with the keyword of the query.

B. Web query classification

Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen[4], proposed classification of web queries into target categories where there is no training data and queries are very short. Here there is no need of collecting training data as intermediate classification is used to train target categories and classifiers bridging. Following are internal classification approaches:

B1. Classification by exact matching

It has two categories defined. First is the intermediate taxonomy and the other is target taxonomy. Given a certain category in an intermediate taxonomy, we say that it is directly mapped to a target category if and only if the following condition is satisfied: one or more terms in each node along the path in the target category appear along the path corresponding to the matched intermediate category. For example, the intermediate category "Computers\Hardware \Storage" is directly mapped to the target category "Computers\Hardware" since the words "Computers" and "Hardware" both appear along the path Computers → Hardware → Storage

B2. Classification by SVM

Query classification with SVM consists of the following steps: 1) construct the training data for the target categories based on mapping functions between categories. If an intermediate category CI is mapped to a target category CT, then the Web pages in CI are mapped into CT; 2) Train SVM classifiers for the target categories; 3) For each Web query to be classified, use search engines to get its enriched features

B3. Classifiers by bridges

It is taxonomy-bridging classifier or bridging classifier by which target taxonomy and queries are connected by taking an intermediate taxonomy as a bridge. To reduce the computation complexity category selection is performed.

C. Reorganizing search results

Xuanhui Wang and ChengXiang Zhai[5], published a work on clustering of search results. This clustering organizes it and allows a user to navigate into relevant documents quickly. Two deficiencies of this approach make it not always work well: First is the clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective; and the second one the cluster labels generated are not informative enough to allow a user to identify the right cluster. In this paper, they propose to address these two decencies by following two steps:

1. Learning "interesting aspects" of a topic from Web search logs and organizing search results accordingly
2. Generating more meaningful cluster labels using past query words entered by users.

D. Clustering web search results

Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma[6], re-formalized the search result clustering problem as a salient phrases ranking problem. Thus they convert an unsupervised clustering problem to a supervised learning problem. Although a supervised learning method requires additional training data, it makes the performance of search result grouping

 ISSN: 2321-8169
4403 - 4408

significantly improve, and enables us to evaluate it accurately. This new algorithm has following four steps:
1. Search result fetching
2. Document parsing and phrase property calculation
3. Salient phrase ranking
4. Post-processing.

First the webpage of search results is returned by some web search engine. HTML parser then analyzes these webpages and result items are extracted. Phrases are ranked according to salience score. The top ranked phrases are taken as salient phrases. Then post processing is performed which filters out the pure stop words.

E. Session boundaries

Rosie Jones and Kristina Lisa Klinkner[7], published a work on session boundaries and automatic hierarchical segmentation of search topics in Query Logs. In this work they studie real sessions manually labeled into hierarchical tasks, and showing that timeouts, whatever their length, are of limited utility in identifying task boundaries, achieving a maximum precision of only 70%. They report on properties of this search task hierarchy, as seen in a random sample of user interactions from a major web search engine's log, annotated by human editors, learning that 17% of tasks are interleaved, and 20% are hierarchically organized. No previous work has analyzed or addressed automatic identification of interleaved and hierarchically organized search tasks. They proposes and evaluated a method for the automated segmentation of users' query streams into hierarchical units.

E. Clustering algorithms

TABLE 1: Summary of existing clustering algorithms

Method	Advantages	Disadvantages
DBSCAN	Does not require you to know the number of clusters	Cannot cluster data sets well with large differences in densities
Expectation Maximization	Gives extremely useful result for the real world data set	Algorithm is highly complex in nature
Hierarchical Clustering	Ease of handling of any forms of similarity or distance	Shows good results when using small datasets
k-means Clustering	With a large number of variables, K-Means is computationally faster	Difficulty in comparing quality of the clusters produced

Clustering is the process of grouping of data. This grouping is done by finding similarities between data based on their characteristics. These groups are termed as Clusters. Clustering is a special type of classification. It is similar to database segmentation where tuples in a database are grouped together. In the implemented system K-means++ algorithm is used. This algorithm has advantages over existing clustering algorithm.

III. System Overview

A new approach overcomes the limitations of existing systems and analyze, infer user search goals using feedback sessions. This approach first takes the user search query. For each single result of the search query pseudo-documents are generated. Using K-means++ clustering algorithm, these pseudo-documents are clustered. Each cluster can be considered as one user search goal. Finally in restructured result is given to the user where each URL is categorized into a cluster centered by the inferred search goals. Then depending upon user click through, results are restructured and represented to the user in order to satisfy the information need.

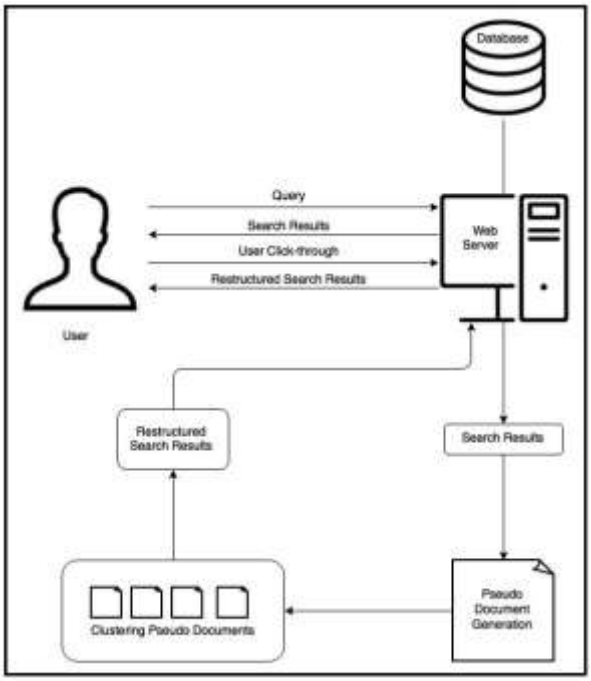

Fig 1: System Framework

Considering pros and cons of the existing approaches of inferring user search goals new method is required for finding out user's information need. Therefore, a new algorithm for inferring user search goals with the feedback sessions is effective in finding out user search goals. There are four modules of the implemented system.

A. Building pseudo-documents
B. Clustering pseudo-documents
C. Capturing feedback sessions

D. Restructuring web search results

A. Building pseudo-documents

Feedback sessions vary a lot for different click-through and queries. So, it is not recommended to directly use feedback sessions for inferring user search goals. In order to represent these feedback sessions some representation method is needed. This method should be a more efficient and coherent.

Implemented system has this new method "Pseudo-documents" to represent every single search result. These documents can be used to infer user search goals. The building of a pseudo-document is done with the help of representation of the URL. The process goes as follows:

1. Titles and snippets of the returned URLs are extracted and the URLs are enriched with this additional textual contents. In simple words, each and every URL is represented by a small text paragraph. This paragraph consists of its title and snippet.
2. It is followed by some textual processes. These processed includes stemming and removing stop words and transforming all the letters to lowercases.

B. Clustering pseudo-documents

With the pseudo-documents, system can infer user search goals. Each URL is represented by a pseudo-document and let Ffs be the feature representation of the pseudo-document. The similarity between two pseudo-documents is computed as the cosine score of Ffsi and Ffsj is

$$Sim(i,j) = cos(Ffsi, Ffsj)$$

and the distance between two feedback sessions is

$$Dis(i,j) = 1 - Sim(i,j)$$

where, i and j are two pseudo documents.

In new system clustering of pseudo-documents is done by K-means++ clustering which is simple and effective. Since the exact number of user search goals is not known for each query, K is set to the five different values (i.e., 1; 2; . . . ; 5) and clustering is done based on these five values. After clustering of all the pseudo-documents, each cluster is considered as one user search goal.

C. Capturing user clicks

A session for web search is a series of queries to fulfil a user's information need and some clicked search results. In implemented system main focus is on inferring user search goals for a particular query.

Feedback session consists of both clicked and unclicked URLs. This session ends with the last URL that was clicked in a single session. It is assumed that before the last click, all the URLs have been scanned and evaluated by users and along with the clicked URLs, the unclicked URLs before the last click are made a part of the user feedbacks. Feedback sessions are constructed with the click through logs. Clicked urls along with the count are maintained and that log is used in order to understand user preferences and likings.

D. Restructuring web search results

Search engines returns millions of results. So, it is necessary to organize them to make it easier for users to find out what they want. Restructuring web search results is an application of inferring user search goals. Vectors are used to represent inferred user search goals. Each URL's feature representation is calculated and we can categorize each URL into cluster. This is done with the help of URL vector and user search goal vector. By choosing smallest distance between URL vector and user search goal vectors URL is categorized into a cluster and the user search goals are restructured. Also based on the user clicks order of the clusters and urls in the clusters are decided. This ordered result helps user to find the frequently visited url in minimum time.

IV. Result Analysis

User search goals for a query are discovered by clustering search results. User search goals are represented by the center points of different clusters. Table II gives some examples of depicting user search goals with four keywords that have the highest values in those feature vectors. Consider the query "India" as an example. Consider that as per users click through, K-means++ has created two clusters corresponding to "India" and each cluster is represented by four keywords. From the keywords "travel, map, city, region" we can say that this part belongs to users who are interested to travel in India. From the keywords "government, elections, constituency, parliament" we can see that other users want to retrieve the information about constitution of India. Based on feedback sessions and user click through logs such clusters are made. Every cluster is one user search goal. K-means++ provides effective clustering results thus more relevant search results to users. Following table has few ambiguous queries along with respective different clusters based on user search goals.

TABLE 2: Ambiguous Queries and Keywords

Query	Keywords depicting user search goals
India	travel, map, city, region government, elections, constituency, parliament
earth	google, map, wikipedia, planet planet, solar, system, nineplanet
Lamborgini	car, history, company, overview new, auto, picture, vehicle

It can be noted k-means++ consistently outperformed k-means [2], both by achieving a lower potential value, in some cases by several orders of magnitude, and also by completing faster.

ISSN: 2321-8169
4403 - 4408

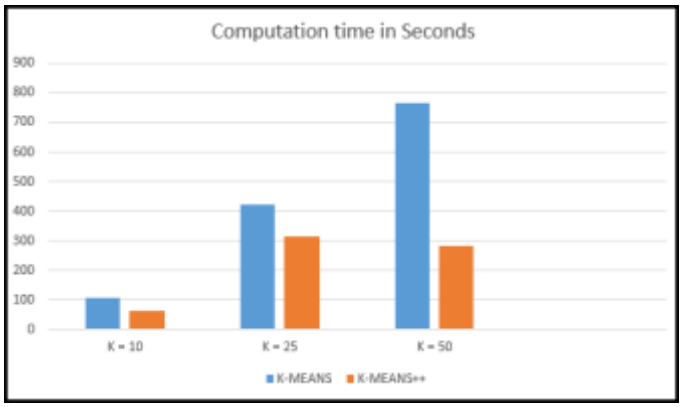

Fig 2: Computation time comparison of K-Means and K-Means++ on the Intrusion Dataset

With the synthetic examples, the k-means method does not perform well, because the random seeding will inevitably merge clusters together, and the algorithm will never be able to split them apart. The careful seeding method of k-means++ avoids this problem altogether, and it almost always attains the optimal results on the synthetic datasets.

The difference between k-means and k-means++ on the real-world datasets is also quite substantial. On the Cloud dataset, k-means++ terminates almost twice as fast while achieving potential function values about 20% better. The performance gain is even more drastic on the larger Intrusion dataset, where the potential value obtained by k-means++ is better by factors of 10 to 1000, and is also obtained up to 70% faster [2].

When performance comparison of k-means and k-means++ was done on the actual search results, it can be noted that k-means++ consistently performs better than k-means. The clustering results were delivered by k-means++ almost twice as fast as k-means. Following is the comparison graph of the same.

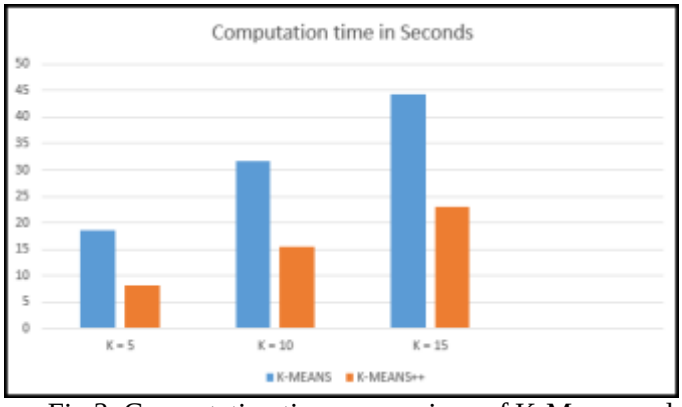

Fig 3: Computation time comparison of K-Means and K-Means++ on the Search Results

As search results are represented in clustered manner at first, user can just go through the headers of clusters with the relevant words of search query. This will help user to identify that if the respective search result cluster is going to satisfy the user's search need or not. This definitely saves user time as user is not required to go through the search results by reading title of each. This saves user's time as compared to'the traditional systems.

V. CONCLUSION

A new approach of clustering feedback sessions is used to discover user search goals. With this new approach users can efficiently find what they want and satisfy their information need. This new approach satisfies information needs of the user though user enters ambiguous query. For search results returned for a ambiguous query pseudo-document is generated. Pseudo-documents has the URLs with extra text including titles and snippets. Based on these documents user search goals are discovered and denoted with some keywords. Similar pseudo-documents are clustered together. Based on this search result in the form of clusters is returned to the user. User can click through the result returned by the system. This click log is maintained by the system and results are restructured at the same time. Finally performance of user search goals is evaluated. This framework produces efficient and correct search results for the ambiguous query. As new approach uses K-means++ algorithm, computation time is reduced along with better clustering results.

ACKNOWLEDGMENT

This is a great pleasure and immense satisfaction to express my deepest sense of gratitude and thanks to everyone who has directly or indirectly helped me in completing my Dissertation work successfully. I express my gratitude towards project guide Prof. Vaishali Nandedkar and Prof. N. D. Kale, Head of Department of Computer Engineering, Padmabhooshan Vasantdada Patil Institute of Technology, Bavdhan, Pune who guided and encouraged me in completing the seminar work in scheduled time. I would like to thank our Principal, Dr. Y. V. Chavan, for his help and cooperation.

REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013

[2] David Arthur, Sergei Vassilvitskii, "k-means++: The Advantages of Careful Seeding", IEEE 2012

[3] Uichin Lee, Zhenyu Liu, Junghoo Cho, "Automatic Identification of User Goals in Web Search" , Proc. 14th Int'l Conf. World Wide Web (WWW '05),pp. 391-400, 2005.

[4] Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen, "Building Bridges for Web Query Classification",

Proc. 29th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR '06),pp. 131-138, 2006.

[5] Xuanhui Wang and ChengXiang Zhai, "Learn from Web Search Logs to Organize Search Results" , Proc. 30th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR '07),pp. 87-94, 2007.

[6] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma, "Learning to Cluster Web Search Results" Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04),pp. 210-217, 2004.

[7] Rosie Jones and Kristina Lisa Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs" , Proc. 17th ACM Conf. Information and Knowledge Management(CIKM '08), pp. 699-708, 2008.

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

[9] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[10] 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.

[11] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

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

[13] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

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

[15] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay,"Accurately Interpreting Clickthrough Data as Implicit Feedback", Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.