clusters. Clustering algorithm takes snippets instead of whole documents as input. Unsupervised clustering problem is converted to supervised with some training data. The shorter cluster names enabled users to quickly identify topics of a specified cluster and clusters are ranked according to scores. The disadvantage was that clustering was still inefficient and a hierarchical structure of search results was necessary for efficient browsing.

2. Motivation

2.1 Query classification

Web query classification is a problem in information science. The task is to assign a Web search query to one or more predefined categories, based on its topics. The importance of query classification is underscored by many services provided by Web search. A direct application is to provide better search result pages for users with interests of different categories. For example, the users issuing a Web query "*apple*" might expect to see Web pages related to the fruit apple, or they may prefer to see products or news related to the computer company. Online advertisement services can rely on the query classification results to promote different products more accurately. Search result pages can be grouped according to the categories predicted by a query classification algorithm.

However, the computation of query classification is non-trivial. Different from the document classification tasks, queries submitted by Web search users are usually short and ambiguous; also the meanings of the queries are evolving over time. Since what users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical. Therefore, query classification is much more difficult than traditional document classification tasks.

2.2 Search result reorganization

With the exponential growth of the Internet, it has become more and more difficult to find information. *Web search* services such as AltaVista, InfoSeek, and MSN Web-Search were introduced to help people find information on the web. Most of these systems return a ranked list of web pages in response to a user's search request. Web pages on different topics or different aspects of the same topic are mixed together in the returned list. The user has to sift through a long list to locate pages of interest. Since the 19th century, librarians have used classification systems like Dewey and Library of Congress classification to organize vast amounts of information. More recently, *Web directories* such as Yahoo! and Look-Smart have been used to classify Web pages. The manual nature of the directory compiling process makes it impossible to have as broad coverage as the search engines, or to apply the same structure to intranet or local files without additional manual effort.

This method has limitations ,since the number of different clicked URLs of a query may be small. Since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well. Therefore, these kind of methods cannot infer user search.

2.3 Session Boundary Detection

Detecting session boundaries on the Web is important for several reasons. Firstly, it is important to establish a common context for various statistics relating to user sessions and frequency of user activities. More specifically, it is important to detect some boundaries in order to group related information together for other applications, such as learning techniques for adaptive search engines.

To date, however, the notion of a session on the Web has not been consistently defined, if it at all. The tendency has been to group the log data that has been made available from one user or IP address under the umbrella of one session regardless of the length of time covered by the logs. This tendency lacks a more user oriented view. Our argument is that a session on the Web can be defined as a group of user activities that are related to each other not only through an evolving information need but also through close proximity in time.

The identification of sessions themselves would be straightforward if the role behind each query/activity was known. However, automatically identifying a role is a difficult task, and requires further information about the background to the query from the session. Hence, the identification of a session has to be done through other information.

Activities in the same session are not only more likely to share the same role at a conceptual level, but also are close to each other in terms of generation time. Although there is a time gap between two adjacent activities in the same session, we think that the gap would usually be smaller than that between two activities in the different sessions. A time span called session interval could be defined in advance to be used as a threshold. Two adjacent activities are counted in two different sessions if the time between them exceeds this threshold. Hence, the identification of session boundaries or delimiters now effectively becomes a process of examining the time gap between activities and their number of occurrences and comparing with the session interval.

Ideally, a session should contain only those activities from and only from one role. In this respect, the optimal session interval should not be too large since the larger the gap, the higher the risk of grouping activities from different roles together – which results in the correctness of the role information from other queries being reduced. However, having too small a session interval also has its problems as essentially there is less information available on the role at a particular point in time. These methods only identifies whether a pair of queries belong to the same goal or mission and does not care what the goal is in detail.

3. A Novel Approach to Infer User Search Goals

A session for web search is a series of successive queries to satisfy a single information need and some clicked search results. Here we focus on inferring user search goals for a particular query. Therefore, the single session containing only one query is introduced, which distinguishes from the conventional session. Meanwhile, the feedback session is based on a single session, although it can be extended to the whole session.

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. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the un-clicked ones before the last click should be a part of the user feedbacks. Fig.1 shows an example of a feedback session and a single session. In Fig. 1, the left part lists 10 search results of the query "the sun" and the right part is a user's click sequence where "0" means "un-clicked." The single session includes all the 10 URLs, while the feedback session only includes the seven URLs in the rectangular box. The seven URLs consist of three clicked URLs and four un-clicked URLs in this example.

Since users scan the URLs one by one from top to down, we can consider that besides the three clicked URLs, the four un-clicked ones in the rectangular box have also been browsed and evaluated by the user and they should reasonably be a part of the user feedback. Inside the feedback session, the clicked URLs tell what users require and the un-clicked URLs reflect what users do not care about. It should be noted that the un-clicked URLs after the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not.

Each feedback session can tell what a user requires and what they do 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 results	Click sequence
www.thesun.co.uk/	0
www.nineplanets.org/sol.html	1
www.solarviews.com/eng/sun.htm	2
en.wikipedia.org/wiki/Sun	0
www.thesunmagazine.org/	0
www.space.com/sun/	0
en.wikipedia.org/wiki/The_Sun_(newspaper)	3
imagine.gsfc.nasa.gov/docs/science/know_l1/sun.html	0
www.nasa.gov/worldbook/sun_worldbook.html	0
www.enchantedlearning.com/subjects/astronomy/sun/	0

Figure 1: A feedback session in a single session

Search results	Click sequence	Binary vector
www.thesun.co.uk/	0	0
www.nineplanets.org/sol.html	1	1
www.solarviews.com/eng/sun.htm	2	1
en.wikipedia.org/wiki/Sun	0	0
www.thesunmagazine.org/	0	0
www.space.com/sun/	0	0
en.wikipedia.org/wiki/The_Sun_(newspaper)	3	1

Figure 2: The binary vector representation of a feedback session

"0" in click sequence means "un-clicked." All the 10 URLs construct a single session. The URLs in the rectangular box construct a feedback session.

Since feedback sessions vary a lot for different click-through and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. Some representation method is required to describe feedback sessions in a more efficient and coherent way. There can be many kinds of feature representations of feedback sessions. For example, Fig. 2 shows a popular binary vector method to represent a feedback session. Same as Fig. 1, search results are the URLs returned by the search engine when the query "the sun" is submitted, and "0" represents "un-clicked" in the click sequence. The binary vector [0110001] can be used to represent the feedback session, where "1" represents "clicked" and "0" represents "un-clicked." However, since different feedback sessions have different numbers of URLs, the binary vectors of different feedback sessions may have different dimensions. Moreover, binary vector representation is not informative enough to tell the contents of user search goals. Therefore, it is improper to use methods such as the binary vectors and new methods are needed to represent feedback sessions.

For a query, users will usually have some vague keywords representing their interests in their minds. They use these keywords , i,.e goal texts to determine whether a document can satisfy their needs. However, although goal texts can reflect user information needs, they are latent and not expressed explicitly. Therefore, we introduce pseudo-documents as surrogates to approximate goal texts.

3.1 Pseudo-Documents to Infer User Search Goals

Pseudo-documents can be built in two steps:

1) **Representing the URLs in the Feedback Session**
In the first step, the URLs are enriched with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the feedback session. In this way, each 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.

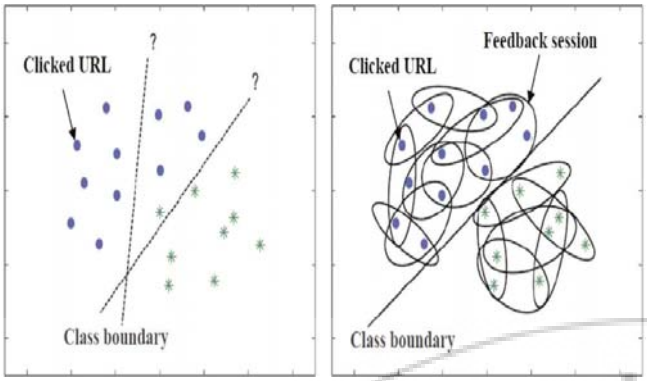

Figure 3: Representing URL's

2)**Forming pseudo-document based on URL representations**

In order to obtain the feature representation of a feedback session, both clicked and un-clicked URLs are combined in the feedback session. It is worth noting that people will also skip some URLs because they are too similar to the previous ones. In this situation, the "un-clicked" URLs could wrongly reduce the weight of some terms in the pseudo-documents to some extent. However, our method can address this problem.

3.2 Clustering pseudo-documents to infer user search goals

Pseudo-documents are clustered by K-means clustering [13] 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. The optimal value will be determined through the evaluation criterion presented.

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

Finally, the terms with the highest values in the center points are used as the keywords to depict user search goals. Note that an additional advantage of using this keyword-based description is that the extracted keywords can also be utilized to form a more meaningful query in query recommendation [2], [4], [5] and thus can represent user information needs more effectively.

Moreover, since we can get the number of the feedback sessions in each cluster, the useful distributions of user search goals can be obtained simultaneously. The ratio of the number of the feedback sessions in one cluster and the total number of all the feedback sessions is the distribution of the corresponding user search goal.

4. Advantages of Clustering Feedback Sessions

1) *Resampling using feedback sessions* . If we view the original URLs in the search results as original samples, then feedback sessions can be viewed as the "processed" 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. If we cluster the search results with these noisy ones, the performance of clustering will degrade greatly. However, feedback sessions actually "resample" the URLs and exclude those noisy ones. Furthermore, the resampling by feedback sessions brings the information of user goal distribution to the new samples. For instance, most URLs in the search results of the query "the sun" are about the sun in nature while most feedback sessions are about the newspaper. Therefore, the introduction of feedback sessions provides a more reasonable way for clustering.

2) *Feedback session, a combination of several URLs.* It can reflect user information need more precisely and there are plenty of feedback sessions to be analyzed.

For example, in Fig.3, 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. The large ellipse in Fig. 3 represents 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. However, if un-clicked URLs are taken into account to construct feedback sessions, they will contain more information and be more efficient to be clustered.) Since the number of the different clicked URLs may be small, if we perform clustering directly on the points, it is very difficult to segment them precisely. However, supposing that most users have only one search goal, it is much easier to segment the ellipses. From another point of view, feedback sessions can also be viewed as a pre-clustering of the clicked URLs for a more efficient clustering. Moreover, the number of the combinations of the clicked URLs can be much larger than the one of the clicked URLs themselves.

5. Conclusion

The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. Due to its usefulness, many works about user search goals analysis such as query classification, search result reorganization, and session boundary detection have been investigated. However finding suitable predefined search goal classes is very difficult and impractical using these approaches. Since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well. To overcome these drawbacks ,a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. Feedback sessions can reflect user information needs more efficiently. when users submit one of the queries, the search engine uses feedback sessions and return the results that are categorized into different groups according to user search goals, allowing users find what they require more conveniently.

References

[1] 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.
[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recom-mendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.
[3] R. Jones and K.L. Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Manage-ment (CIKM '08), pp. 699-708, 2008.
[4] 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.
[5] 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.
[6] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.
[7] H. Chen and S. Dumais, "Bringing Order to the Web: Auto-matically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.
[8] X. Wang and C.-X 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.
[9] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. 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.
[10] D.Kavitha, K.M.Subramanian, Dr.K.Venkatachalam " SURVEY ON INFERRING USER SEARCH GOAL USING FEEDBACK SESSION" (IJARCET) Volume 2, Issue 12, December 2013,pp/3231-3237.
[11] Joachims. T, 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
[12] Beeferman. D and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD'00), pp. 407- 416, 2000.
[13] Andrew Moore: "K-means and Hierarchical Clustering –Tutorial Slides" http://www2.cs.cmu.edu/~awm/tutorials/kmeans.html
[14] Max Welling, "Support Vector Machines", University of Toronto, Toronto, M5S 3G5 Canada

Author Profile

Y. Sai Krishna received the B. Tech degree in Computer Science and Engineering from JNTU Hyderabad in 2012. He is pursuing M. Tech in Computer Science and Engineering from JNTU Hyderabad. His research interests include data mining.

N. Swapna Goud received the B. Tech degree from Vijay Rural Engineering College and M. Tech degree from C.V.S.R College of Engineering and Technology. She is an associate professor in the Department of Computer Science and Engineering, Anurag Group of Institutions. Her research interests include data mining.