International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 2; Issue: 5; May -2014

An approach for predicting Feedback session based user search goals

H. G. Chetan
4th sem M.Tech. Department of CS&E,
Adichunchanagiri Institute of Technology,
Chikmagalur, Karnataka, India.
Email: chetanhgcool@gmail.com

Chandra Naik G
Asst Professor, Department of CS&E,
Adichunchanagiri Institute of Technology,
Chikmagalur, Karnataka, India.
E-mail:chandru2468@gmail.com

Abstract : Different users may have different search goals when they submit it to a search engine. Many activities on the web are driven by high-level goals of users, such as "plan a trip" or "buy some product". The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. This paper presents results from an exploratory study that focused on analyzing selected search sessions from a search engine log by clustering feedback sessions. URL is considered to prepare pseudo-document to represent feedback session for clustering. Based on clicking the URLs, scoring is done. New Criterion called Classified Average Precision (CAP) is proposed to evaluate the performance of the inferred user search goals. In this paper, we are interested in exploring the role and structure of user goals in web search.

General Terms : ***Pseudo document, scoring.***

Keywords : ***Feedback- session, URL, clustering, CAP.***

1. INTRODUCTION

A Web is a collection of inter-related files on one or more Web servers. Web mining is the application of data mining technique; it is used extract knowledge from Web data. Web data is Web content data (text, image, record), Web structure data (hyperlinks, logs) and Web usage data (http logs, app server logs). In web search applications, 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. Analysing and exploring regularities in Weblog records (consist of URL's, time interval, click sequence and etc.,) for electronic commerce, enhance the quality and delivery of internet information services to the end user, and improve Web server system performance. In this paper we use the Web usage mining data.

Web server usually registers a log entry, or Weblog entry for every access of a Web page. It includes the URL requested the IP address from which the request originated, and a timestamp. Based on the Weblog records, we have to construct the feedback session. Because Weblog data provide information about what kind of users will access what kind of Web pages. This session consists of URL's and click sequence and it focus on user search goals. Only using a feedback session we do not understand the user search goals exactly. Based on the feedback session, construct the pseudo document for analysing the accurate result. This pseudo document consists of keywords of URL's in the feedback session. This is called as enriched URL's. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have a high similarity in comparison to one another but are very dissimilar to object in other clusters. The enriched URL's are clustered and form a pseudo document. After constructing the pseudo document the Web search results are restructured based on the documents collection detail.

2. LITERATURE REVIEW

R. Jones and K.L. Klinkner [8] proposed a method to detect search goal and mission boundaries for automatic segmenting query logs into hierarchical structure. User may issue number of queries to search engine in order to accomplish information need/tasks at a variety of granularities. Their method identifies whether a pair of queries belongs to the same goal or mission and does not consider search goal in detail.

Uichin Lee and Zhenyu Liu proposed work is based on the Web query assigned by the user's analysis the goal [2], the goal identification is used to improve quality of search results. In existing system which use the manual query log investigation to identify the goals. This proposed system use automatic goal identification process. The human-subject study strongly indicates the automatic query goal identification. It can use two tasks like as past user click behavior and anchor link distribution for goal identification combining these two tasks can identify 90% goal accurately.

Preceding studies comprehends mainly interest on manual query-log investigation to recognize Web query goals. U. Lee et al. [7] studied the "goal" at the back based on a user's Web query, so that this goal can be used to get better the excellence of a search engine's results. Their proposed method identifies the user goal automatically with no any explicit feedback from the user.
Zamir et al. [9] used Suffix Tree Clustering (STC) to identify set of documents having common phrases and then create cluster based on these phrases or contents. They used documents snippets instead whole document for clustering web documents. However, generating meaningful labels for clusters is most challenging in document clustering. So, to overcome this difficulty, in [3], a supervised learning method is used to extract possible phrases from search result snippets or contents and these phrases are then used to cluster web search results.

T. Joachim's [4] approach is automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking. By

using support vector machine (SVM) approach, for learning ranking functions in information retrieval. T. Joachim's et al. [5] contribution is on examining the reliability of implicit feedback generated from click-through data in World Wide Web search. The author approached strategy to automatically generate training examples for learning retrieval functions from observed user behaviour. The user study is intended to examine how users interrelate with the list of ranked results from the Google search engine and how their behaviour can be interpreted as significance judgments. Implicit feedback can be used for evaluating quality of retrieval functions [6].

3. OVERVIEW OF THE SYSTEM ARCHITECTURE

We define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user's particular desire to obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience. In web search applications, queries are submitted to search engines to represent the information needs of users. 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 apple is submitted to a search engine, some users want to locate the homepage of an Apple Computer, Inc. California, while some other may want to know about the Apple movie which is musical fiction film, or some may want to know about the fruit. Therefore, it is necessary and potential to capture different user search goals in information retrieval.

The system architecture has two parts. In the upper part, all the feedback sessions of a query are first extracted from user click-through logs and mapped to pseudo-documents. Then, user search goals are inferred by clustering these pseudo-documents and depicted with some keywords. Since the exact number of user search goals is unknown in advance, several different values are tried and the optimal value will be determined by the feedback from the bottom part.

In the bottom part, the original search results are restructured based on the user search goals inferred from the upper part. Then, evaluate the performance of restructuring search results by proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the upper part. The system architecture is as shown in figure 1.

3.1 Feedback Sessions

Feedback session consists of 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.

3.2 Pseudo-documents

The URLs consists of additional textual contents, using this we are extracting the titles and snippets of the returned URLs and these are appearing in the feedback session**.** The building of a pseudo-document includes two steps. They are described as follows:

3.2.1 Representing the URLs in the feedback session:

In this step 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. Finally, each URL's title and snippet are represented by a Term Frequency-Inverse Document Frequency (TF-IDF) vector respectively, as in

$$T_{ui} = [t_{w1}, t_{w2}, ..., t_{wn}]^T$$

$$S_{ui} = [s_{w1}, s_{w2}, ..., s_{wn}]^T \quad (1)$$

where T_{ui} and S_{ui} are TF-IDF vectors of the URL's title and snippet, respectively. ui means the i^{th} URL in the feedback session and w_j(j= 1,2, . . . , n) is the j^{th} term appearing in the enriched URLs. Here, a "term" is defined as a word or a number in the dictionary of document collections. t_{wj} and s_{wj} represent the TF-IDF value of the j^{th} term in the URL's title and snippet, respectively.

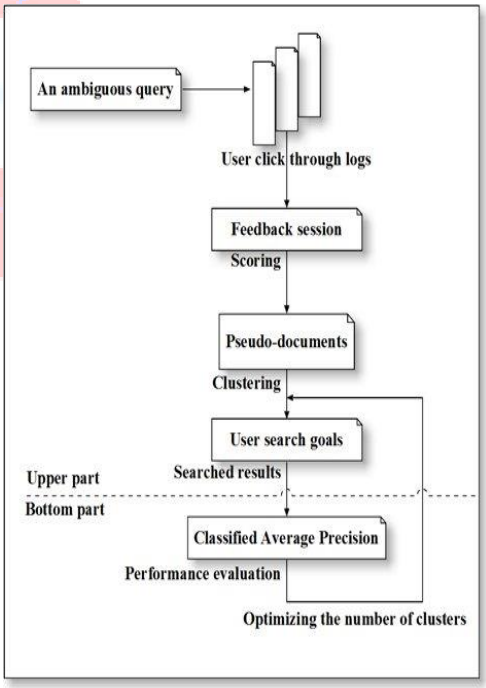

Fig.1. System architecture of the feedback session based user search goal.

Considering that URLs' titles and snippets have different significances, we represent the enriched URL by the weighted sum of T_{ui} and S_{ui} , namely

$$F_{ui} = w_t T_{ui} + w_s S_{ui} = [f_{w1}, f_{w2}, ..., f_{wn}]^T \quad (2)$$

where F_{ui} means the feature representation of the i^{th} URL in the feedback session, and w_t and w_s are the weights of the titles and the snippets, respectively. First w_s value is set by 1. Then, stipulate that the titles should be more significant than the snippets. Based on (2), the feature representation of the URLs in the feedback session can be obtained. It is worth noting that although T_{ui} and S_{ui} are TF-IDF features, F_{ui} is not a TF-IDF feature. This is because the normalized TF feature is relative to the documents and therefore it cannot be aggregated across documents. In our case, each term of F_{ui} (i.e., f_{wj}) indicates the importance of a term in the i^{th} URL.

3.2.2 Forming pseudo-document based on URL representation:

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. Let F_{fs} be the feature representation of a feedback session, and $f_{fs}(w)$ be the value for the term w. Let F_{ucm}(m= 1, 2, . . .M) and F_{ucl}(l = 1, 2, . . . , L) be the feature representations of the clicked and unclicked URLs in this feedback session, respectively. Let $f_{ucm}(w)$ and f_{ucl} (w) be the values for the term w in the vectors. We want to obtain such a F_{fs} that the sum of the distances between F_{fs} and each F_{ucm} is minimized and the sum of the distances between F_{fs} and each F_{ucl} is maximized. Based on the assumption that the terms in the vectors are independent, we can perform optimization on each dimension independently, as shown in

$$F_{fs} = [f_{fs}(w1), f_{fs}(w2),..., f_{fs}(wn)]^T$$
$$F_{fs}(w) = \arg\min\{\sum_M [f_{fs}(w) - f_{ucm}(w)]2 - \lambda\sum_L [f_{fs}(w)]2\}, f_{fs}(w) \in I_c \quad (3)$$
$$[\mu f_{cu}(w) - \sigma f_{uc}(w^2), \mu f_{uc}(w) + \sigma f_{uc}(w)]$$

and I_c be the interval

$$[\mu f_{cu'}(w') - \sigma f_{uc'}(w), \mu f_{uc'}(w') + \sigma f_{uc'}(w)]$$

where $\mu f_{cu}(w)$ and $\mu f_{cu}(w')$ represent the mean and mean square error of $f_{cu}(w)$ respectively, and $\mu f_{uc}(w)$ and $\mu f_{uc'}(w')$ represent the mean and mean square error of $f_{cu'}(w)$ respectively. If $I_c \subseteq I_c$ or $I_c \subseteq I_c$, we consider that the user does not care about the term w. In this situation, we set ffs(w) to be 0, as shown in

$$ffs(w) = 0, I_c \subseteq I_c \text{ or } I_c \subseteq I_c \quad (4)$$

λ is a parameter balancing the importance of clicked and unclicked URLs. When λ in (3) is 0, unclicked URLs are not taken into account. On the other hand, if λ is too big, unclicked URLs will dominate the value of ffs(w). In this paper, we set λ to be 0.5.

It is worth noting that people will also skip some URLs because they are too similar to the previous ones. In this situation, the "unclicked" URLs could wrongly reduce the weight of some terms in the pseudo-documents to some extent. However, our method can address this problem. Let us analyse the problem from three cases.

Case 1 (the ideal case): one term appears in all the clicked URLs and does not appear in any unclicked ones. In this case, people skip because the unclicked URLs do not contain this important term. The weight of the term in the pseudo-document will be set to the highest value in I_c in (3).

Case 2 (the general case): one term appears in both the clicked URLs and a subset of the unclicked ones. In this case, some unclicked URLs are skipped because they are irrelevant and some are skipped because of duplication. The weight of the term will be reduced to some extent; however, it will not be set to zero and it is still included in I_c according to (3). Therefore, skipping because of duplication does not affect too much in this case.

Case 3 (the bad case): one term appears in both the clicked URLs and almost all the unclicked ones. In this case, people skip because of duplication. I_c could contain I_c and the weight of the term will be set to zero according to (4). However, when this case happens, both the clicked and the unclicked URLs is almost about one single subject and the term is no longer distinguishable. Therefore, even if people skip some unclicked URLs because of duplication, our method can still assign reasonable weight of the term in most cases.

Up to now, the feedback session is represented by F_{fs}. Each dimension of F_{fs} indicates the importance of a term in this feedback session. F_{fs} is the pseudo-document that we want to introduce. It reflects what users desire and what they do not care about. It can be used to approximate the goal texts in user mind.

3.3. Inferring Pseudo-documents:

The proposed pseudo-documents, will infer user search goals. In this section, we will describe how to infer user search goals and depict them with some meaningful keywords. As each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document. Pseudo-documents by K-means clustering are simple and effective. Since exact number of user search goals is unknown, by setting K to five different values for each query, and perform clustering based on these five values, respectively. The terms with the highest values in the centre points are used as the keywords to depict user search goals. The 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 and thus can represent user information needs more effectively.

With the proposed pseudo-documents, we can infer user search goals. In this section, we have described how to infer user search goals and depict them with some meaningful keywords. As in (3) and (4), each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document is F_{fs}. The similarity between two pseudo documents is computed as the cosine score of F_{fsi} and F_{fsj} as follows:

$$Sim_{i,j} = cos(F_{fsi}, F_{fsj}) = \frac{F_{fsi} * F_{fsj}}{|F_{fsi}| * |F_{fsj}|} \quad (5)$$

And the distance between two feedback sessions is

$Dis_{i,j} = 1 - Sim_i$ (6)

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 (i.e., 1, 2 . . . 5) 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 centre point of a cluster is computed as the average of the vectors of all the pseudo-documents in the cluster, as shown in

$$F_{centri} = \sum_{k=1}^{Cl} \frac{F_{fsk}}{C_i}, (F_{fsk} \subset Clusteri) \quad (7)$$

where $F_{centeri}$ is the i^{th} cluster's centre and C_i is the number of the pseudo-documents in the i^{th} cluster. $F_{centeri}$ is utilized to conclude the search goal of the i^{th} cluster.

Finally, the terms with the highest values in the centre 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 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.

3.4. Evaluation Search Result:

If user search goals are inferred properly, the search results can also be restructured properly, since restructuring web search results is one application of inferring user search goals. Therefore, we propose an evaluation method based on restructuring web search results to evaluate whether user search goals are inferred properly or not. In this section, we propose this novel criterion "Classified Average Precision" to evaluate the restructure results. Based on the proposed criterion, we also describe the method to select the best cluster number.

In order to apply the evaluation method to large-scale data, the single sessions in user click-through logs are used to minimize manual work. Because from user click-through logs, we can get implicit relevance feedbacks, namely "clicked" means relevant and "unclicked" means irrelevant. A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks. AP is the average of precisions computed at the point of each relevant document in the ranked sequence, as shown in

$$AP = \frac{1}{N+} \sum_{r=1}^{N} rel(r) \frac{R_r}{r} \quad (8)$$

Where N+ is the number of relevant (or clicked) documents in the retrieved ones, r is the rank, N is the total number of retrieved documents, rel() is a binary function on the relevance of a given rank, and R_r is the number of relevant retrieved documents of rank r or less. We can compute AP as:

$$\frac{1}{4} * \left(\frac{1}{2} + \frac{2}{3} + \frac{3}{7} + \frac{4}{9} \right) = 0.510$$

However, AP is not suitable for evaluating the restructured or clustered searching results. The proposed new criterion for evaluating restructured results is described in the following.

The URLs in the single session are restructured into two classes where the un-boldfaced ones are clustered into class 1 and boldfaced ones are clustered into class 2. We first introduce "Voted AP (VAP)" which is the AP of the class including more clicks namely votes. If the numbers of the clicks in two classes are the same, we select the bigger AP as the VAP. Assume that one user has only one search goal, then ideally all the clicked URLs in a single session should belong to one class. And a good restructuring of search results should have higher VAP.

However, VAP is still an unsatisfactory criterion. Considering an extreme case, if each URL in the click session is categorized into one class, VAP will always be the highest value namely 1 no matter whether users have so many search goals or not. Therefore, there should be a risk to avoid classifying search results into too many classes by error. We propose the risk as follows:

$$Risk = \frac{\sum_{i,j=1(i<j)}^{m} d_{ij}}{c^2} \quad (9)$$

It calculates the normalized number of clicked URL pairs that are not in the same class, where m is the number of the clicked URLs. If the pair of the i^{th} clicked URL and the j^{th} clicked URL are not categorized into one class, d_{ij} will be 1; otherwise, it will be 0.

$$c_m^{\ 2} = \frac{m(m-1)}{2}$$

is the total number of the clicked URL pairs. Based on the above discussions, we can further extend VAP by introducing the above Risk and propose a new criterion "Classified AP," as shown below

$$CAP = VAP * (1 - Risk)^{\gamma} \quad (10)$$

From (10), we can see that CAP selects the AP of the class that user is interested in (i.e., with the most clicks/votes) and takes the risk of wrong classification into account. The term γ is used to adjust the influence of *Risk* on CAP that can be learned from training data. Finally, we utilize CAP to evaluate the performance of restructuring search results.

Considering another extreme case, if all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0; however, VAP could be very low. Generally, categorizing search results into fewer clusters will induce smaller Risk and bigger

VAP, and more clusters will result in bigger Risk and smaller VAP. The proposed CAP depends on both of Risk and VAP.

4. CONCLUSION

As the Web and its usage continues to grow, so grows the opportunity to analyse Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper, we propose a novel approach for user search goals using feedback session and pseudo document. First we construct a feedback session to analysis the user search goal from the Weblog record. It cannot provide the accurate result. This proposed system includes the pseudo document to provide the accurate results. Based on the pseudo document we have to restructure the Web search results.

5. ACKNOWLEDGMENT

The authors gratefully acknowledge support from the Adichunchanagiri Institute of Technology, Chikmagalur, through its strategic initiative. The authors also acknowledge their Head of the Department, who gave guidelines for preparing this work. Last but not least the authors acknowledge support from their parents and from their friends.

6. REFERENCES

[1] "New algorithm for inferring user search goals with feedback session" Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng

[2] Uichin Lee and Zhenyu Liu proposed "Automatic Identification of User Goals in Web Search"

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005.

[8] 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 Management (CIKM '08), pp. 699-708, 2008.

[9] O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to web search results. Computer Networks, 31(11-16), pp.1361- 1374, 1999.

[10] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[11] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 41-50, 2008.

[12] D. Shen, J. Sun, Q. Yang, and Z. 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.

[13] 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.

[14] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.

[15] 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.

BIOGRAPHY

Chetan H.G is currently pursuing M.Tech in the Department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikmagalur. He received B.E. degree in 2012 from SDMIT, Ujire. His research interests are Data Mining and Image Processing. He has undertaken Feedback session based user search goals prediction as final year M.Tech. Project.

Chandra Naik G is presently working as an Assistant Professor in the Department of CS&E, AIT, Chickmagalur, Karnataka, India. He is having 3 years of teaching experience. His areas of interest are Digital Communication,Cryptography, Network security, Networking