Improved Accuracy and User Satisfaction by Inferring User Search Goals based on Implicit and Explicit Feedback

[1]Prof.Ratna Raja Kumar, [2]Snehankita Badgujar

[1,] HOD of Department of Computer Engineering of Moze College, Pune

[2]Student of Moze College of Engineering, Pune

Abstract: In web search engine user submits query for getting information in easiest way. Different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience.

In this paper, we propose a novel approach to infer user search goals by analyzing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion "Voted Average Precision (VAP)" to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from search engine to validate the effectiveness of our proposed methods.

Index Terms—User search goals, feedback sessions, pseudo-documents, restructuring search results, classified average precision (CAP), explicit feedback

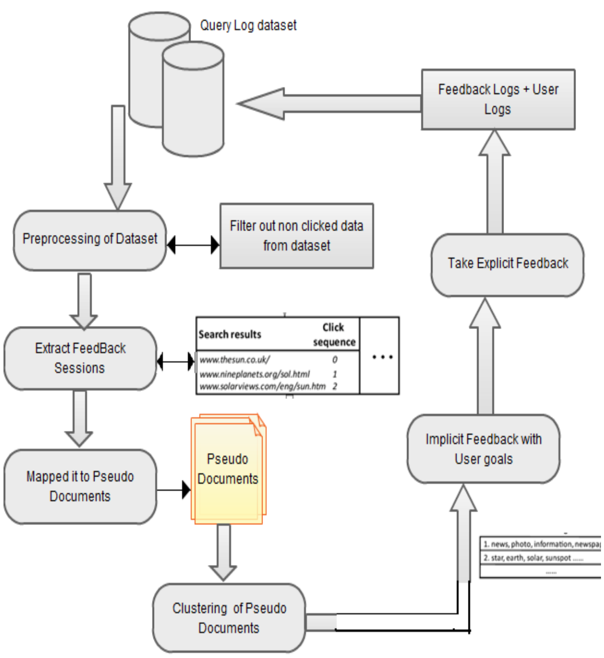
1. Introduction

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. 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, as shown in Fig. 1. Therefore, it is necessary and potential to capture different user search goals in information retrieval. 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. Some advantages are
summarized as follows. First, we can restructure web search results [6], [18], [20] according to user search goals
by grouping the search results with the same search goal; thus, users with different search goals can easily find what
they want. Second, user search goals represented by some keywords can be utilized in query recommendation [2], [5],
[7]; thus, the suggested queries can help users to form their queries more precisely. Third, the distributions of user
search goals can also be useful in applications such as reranking web search results that contain different user search goals. Enhanced goals specific search result .Forth Effective search result using users explicit feedback.

Fifth we can restructure result by displaying websites related to explicit feedback first and then implicit feedback
and then others 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. I the first class, people attempt to infer user goals and intents by predefining some specific classes and performing `

2. FRAMEWORK OF OUR APPROACH

Fig. 2 shows the framework of our approach. Our framework consists of two parts divided by the dashed line.
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 we do not know the exact number of user search goals 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, we evaluate the performance of restructuring search results by our 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



3. III Mapping Pseudo-Documents

In the first step, we first enrich the URLs 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,eachURL'title and snippet are represented by a Term Frequency-Inverse

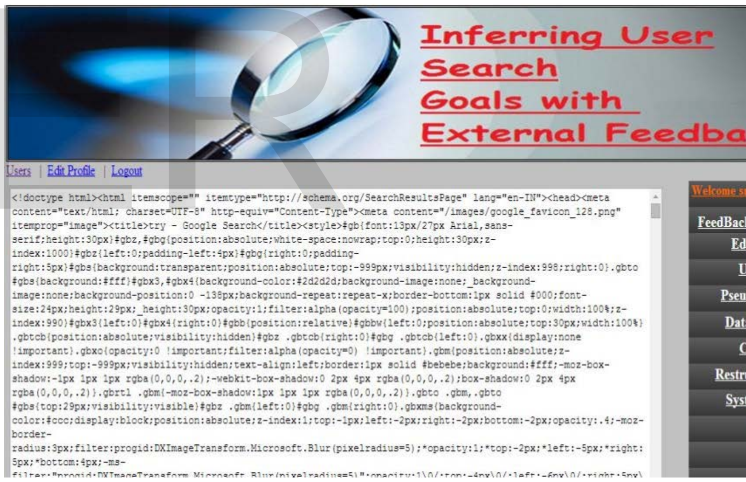


Fig Forming pseudo document

4. Creating Data Extraction

ISSN 2229-5518

title	url	
Ram Trucks - Pickup Trucks, Work Trucks & Vans	http://www.ramtrucks.com/	Official site of Ram Trucks, America. Come see our late trucks and commercial vehi
Random-access memory - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/Random-access_memory	Random-access memory is random-access memory dev written in roughly the same .
Computer Memory | RAM - Newegg.com	http://www.newegg.com/Memory/Category/ID-17	Find RAM, computer mem Crucial, Mushkin at Neweg shipping and top-rated cust
Crucial.com - DDR3, DDR2 and DDR RAM memory upgrades	http://www.crucial.com/usa/en/memory-info	Compatible upgrades. Guar either of our easy-to-use too know your system's specs.
What is Random Access Memory (RAM)?	http://www.webopedia.com/TERM/R/RAM.html	RAM (pronounced ramm) i memory and is the most cor

Fig : Data Extraction

Data extraction creating by pseudo document. We are dividing data in Title, URL, and Content .We are using XML code reader for extracting titles, URL and content .

5. INFERRING USER SEARCH GOALS BY CLUSTERING PSEUDO-DOCUMENTS

With the proposed pseudo-documents, we can infer user search goals. In this section, we will describe how to infer user search goals and depict them with some meaningful keywords. each feedback session is represented by a pseudo-document and the feature representation of the Pseudo-document .

We cluster pseudo-documents by Hierarchical K-means clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, hierarchical clustering scheme produces a sequence of clusterings in which each clustering is nested into the next clustering in the sequence and perform clustering based on these five values, respectively. The optimal value will be determined through the evaluation criterion .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 .

The evaluation of user search goal inference is a big problem , since user search goals are not predefined and there is no ground truth. Previous work has not proposed a suitable approach on this task. Furthermore, since the optimal number of clusters is still not determined when inferring user search goals, a feedback information is needed to finally determine the best cluster number,. Therefore, it is necessary to develop a metric to evaluate the performance of user search goal inference objectively. Considering that 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.

Cluster1

official, site, hardest, working, trucks, come, latest, models, work, commercial,
http://www.ramtrucks.com/

Cluster2

memory, form, computer, device, llows, items, written, roughly, find, from, mushkin, offer, best, f
http://en.wikipedia.org/wiki/Random-access_memory
http://www.newegg.com/Memory/Category/ID-17
http://www.webopedia.com/TERM/R/RAM.html
http://computer.howstuffworks.com/ram.htm
http://www.computerhope.com/jargon/r/ram.htm
http://www.bestbuy.com/site/computer-cards-components/computer-memory/abcat0506000.c?ic

Cluster3

compatible, find, right, upgrades, with, advisor, tool, know, technology, brings, globl, industry, fir
http://www.crucial.com/usa/en/memory-info
http://www.rammount.com/
http://www.allmusic.com/album/ram-mw0000198052

Cluster4

restaurant, group, includes, menus, list, locations, oregon,
http://www.theram.com/

Fig : Clustering of result

6. CAP Evaluation

I will go with Heirarchical k-means clustering The similarity between two pseudo-documents is computed as the cosine score of F_{fsi} and F_{fsj} , as follows: Where F_{fs} is feature representation of pseudo document Pd

$$Sim_{i,j}=\cos(F_{fsi}, F_{fsj})$$
$$= \frac{Ffsi.Ffsj}{|Ffsi||Ffsj|}$$

And the distance between two feedback sessions is $Dis_{ij}=1-Sim_{ij}$

The center point of a cluster is computed as the average of the vectors of all the pseudo-documents in the cluster, as shown in

ISSN 2229-5518

$F_{centeri}= \frac{\sum_{k=1}^{ci} Ffsk}{Ci}$, ($F_{fsk} \in_{clusteri}$) Where

$F_{centeri}$ is the ith cluster's center and c_i is the number of pseudo documents Pd in the ith cluster

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

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

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 Rr is the number of relevant retrieved documents of rank r or less.

"Voted AP (VAP)" which is the AP of the class including more clicks namely votes. For example, the VAP of the restructured search results in Fig. 7b is the AP of class 1, calculated by:

VAP = 1/3*(1/1+2/2+3/6) =0:833. If the numbers of the clicks in two classes are the same, we select the bigger AP as the VAP.

However, VAP is still an unsatisfactory criterion. Therefore, there should be a risk to avoid classifying search results into too many classes by error. We propose the risk as follows:

$$Risk= (1+x)^n = \frac{\sum_{i,j=1\ (i<j)}^{m} dij}{C2m}$$

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 ith clicked URL and the jth clicked URL are not categorized into one class, dij will be 1; otherwise, it will be 0.

C2m = m (m-1)/2 is the total number of the clicked URL pairs.

.

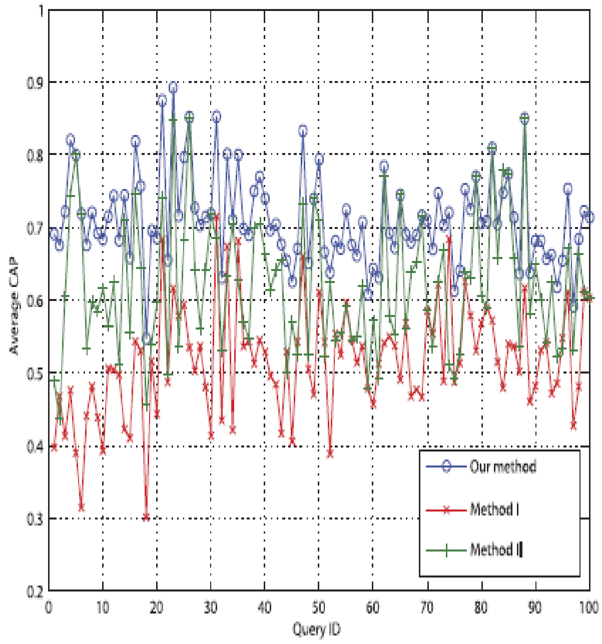
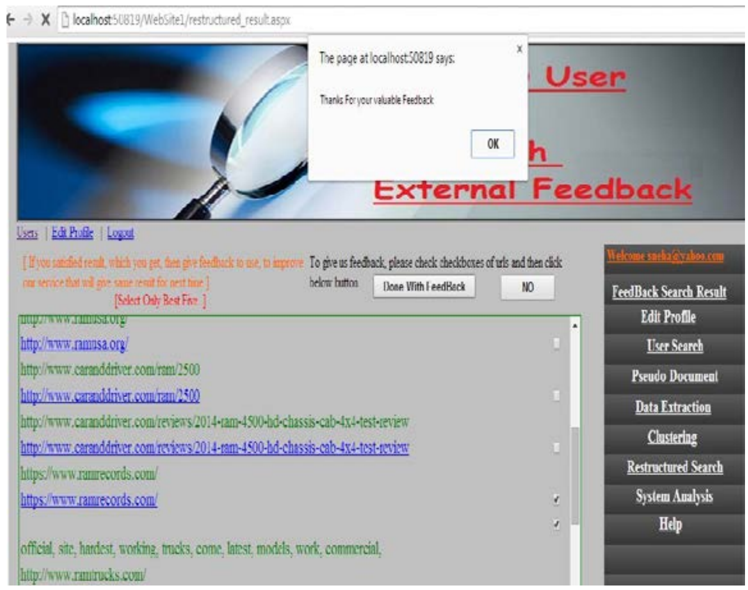


Fig :The Existing chart of CAP comparison of three methods for 100 most ambiguous queries.

7. Web Restructure Result



We are taking explicit feedback from Users means we exactly come to know which are important websites for users .as well as there will

ISSN 2229-5518

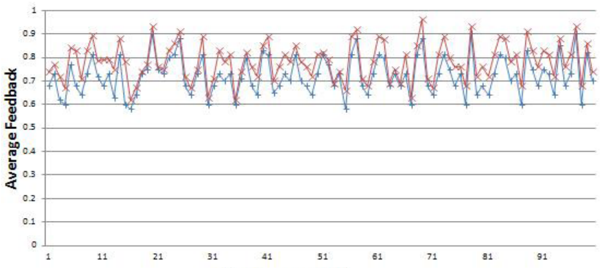
be implicit feedback of users we can count by CAP

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 unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference .Thus users can find what they want conveniently.

8.**Analysis**

We infer user search goals for a query by clustering its feedback sessions. User search goals
are represented by the center points of different clusters. Since each dimension of the feature
vector of a center point indicates the importance
of the corresponding term, we choose those
keywords with the highest values in the feature
vector to depict the content of one user
search goal. fig 2 gives some examples of depicting user search

Fig 13 :Proposed Graph



X axis –query id
Y axis – average feedback
Improved search precision due to Explicit (inclusive implicit) Feedback

The Fig. 11 above depicts the improved search results due to an explicit feedback over
implicit feedback results. The precision of search reluts get improved by getting explicit
feedback counted in search process. For getting the result of restructured we added implicit
feedback and explicit feedback.

We can get intuitive results of our search goal inference. Taking the query as an example,
since VAP with single class session .The restructured search results is the highest when there
are totally three clusters (i.e., three lines) corresponding to lamborghini and each cluster is represented by four keywords. From the keywords car, history, company, overview, we can
find that this part of users are interested in the history of Lamborghini. From the keywords
new, auto, picture, vehicle, we can see that other users want to retrieve the pictures of new
Lamborghini cars. From the keywords club, oica, worldwide, Lamborghiniclub, we can find
that the rest of the users are interested in a Lamborghini club. We can find that the inferred user search goals of the other queries are also meaningful. This confirms that our approach
can infer user search goals properly and depict them with some keywords meaningfully.

Method	Mean Avearge VAP	Mean Avearge CAP	Mean Avearge Risk
our method	0.93	0.93	0
method I	0.755	0.632	0.224
method II	0.68	0.584	0.196
method III	0.742	0.611	0.243

Proposed VAP Table

8. Conclusion

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. 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. . 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

ISSN 2229-5518

online. Thus, users can find what they want conveniently.

REFERENCES

1. .A New Algorithm for Inferring User Search Goals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng Bates, M.J. Information Search Tactics. *Journal of the American Society for Information Science*, 30, July 1979, 205-214.

2.]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.

3. 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.

4. 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.
5. 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.

6. 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.
7. 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.

8. Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(3), 1999. Originally published as DEC Systems Research Center Technical Note, 1998.

9. 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.

10. 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.

11. X. Li, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346, 2008.

12. 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.