ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Competent Cache Mechanism for Consumer Search Using Feedback Program

Palla Pavan Kumar
M.Tech Student
Department of Computer Science and Engineering,
Sai Madhavi Institute of Science And Technology,
Rajahmundry, A.P-533296, India.

Mr. Chinnam Yuva Raju, M.Tech, (Ph.D)
Associate Professor
Department of Computer Science and Engineering,
Sai Madhavi Institute of Science And Technology,
Rajahmundry, A.P-533296, India.

ABSTRACT

Learning the user's search intensions are definitely the major addressable issue for a search engine. We all propose novel criteria to understand the person search goal by extracting the meaning of the issue and generate related keywords which will be retained in the dictionary as annotations. This method will handle the user search goal and which will deliver more optimal search results. This method also produces more accurate results where customer can find the end result as quickly as possible and will decrease the extra time and extra content the user has to go through. This kind of will likely save time and cut down the method costs. We define customer search goals as the data on different aspects of a query that end user groups want to obtain. Information need is an user's particular desire to obtain information to gratify his/her need. User search goals can be considered as the clusters of information needs for a question. The inference and analysis of user search goals can have a lot of advantages in increasing search engine significance and user experience

The customer feedbacks, feedback classes are proposed. Then, we propose a strategy to map reviews sessions to pseudo-documents which can successfully reflect end user information needs. Ranking model can be used to provide the ranking for these products. Simply by providing the ranking for the product the retailers can evidently understand the betterment of goods and may easily find the actual frequently used products. This can be very within increasing the product in the new product development level. Backward algorithm is a methodology of association guideline of data mining, can be used to determine the commonly used products. In addition to this a method is provided to analyze the customer behavior. Here users can pose their questions, in which answers will give by other customers. From their responses we can forecast the person expectations and needs. Since the analysis of clustering is also an important problem, analysis is explained to evaluate the performance new product.

INTRODUCTION

The 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, 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. First, we can restructure web search results 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, 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 re ranking web search results that contain different user search goals.

Nowadays data mining has attracted a great deal of attention in the information industry and in society as a whole, due to the wide availability of large amounts of data and the imminent need for turning such data into useful knowledge and information. The information and knowledge gained can be used for many applications ranging from market analysis, customer retention, fraud detection, to production control and science exploration. Clustering is the most important concept used here. Clustering analyzes data objects without consulting a known class label. The knowledge of customers and market channels is transformed into knowledge assets of the enterprises during the stage of NPD (New Product Development). The priori algorithm in data mining is a methodology of association rule, which is implemented for mining demand chain knowledge from channels and customers.

Introduction to Data Mining:

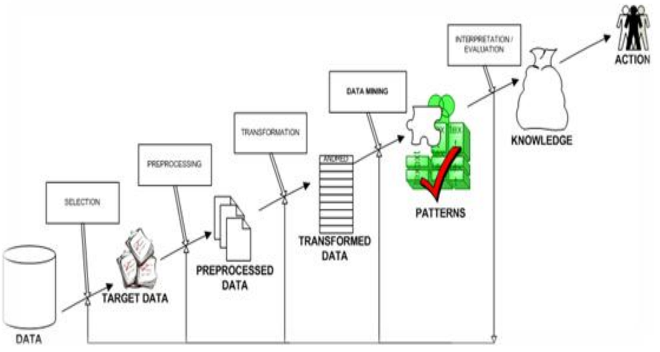


Fig 1.1 Structure of Data Mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Working of Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

Data mining consists of five major elements:

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

Characteristics of Data Mining:

Large quantities of data:

The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.

Noisy, incomplete data:

Imprecise data is the characteristic of all data collection.
Complex data structure: conventional statistical analysis not possible
Heterogeneous data stored in legacy systems

Benefits of Data Mining:

1) It's one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them
2) An analytical CRM model and strategic business related decisions can be made with the help of data

mining as it helps in providing a complete synopsis of customers
3) An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)
4) Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors
5) Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

EXISTING SYSTEM

This session shows the framework of our approach. Our framework consists of two parts divided by the dashed line [Refer Figure 3.1]. 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, the evaluation of the performance of restructuring search results by the evaluation criterion Classified Average Precision (CAP) is done. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the upper part.

PROPOSED SYSTEM

In this project, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a novel approach to infer user search goals for a query by clustering our proposed feedback sessions. The feedback session is defined as the series of both clicked and un-clicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords. Since the evaluation of clustering is also an important problem, we also propose a novel evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. We also demonstrate that the proposed evaluation criterion can help us to optimize the parameter in the clustering method when inferring user search goals.

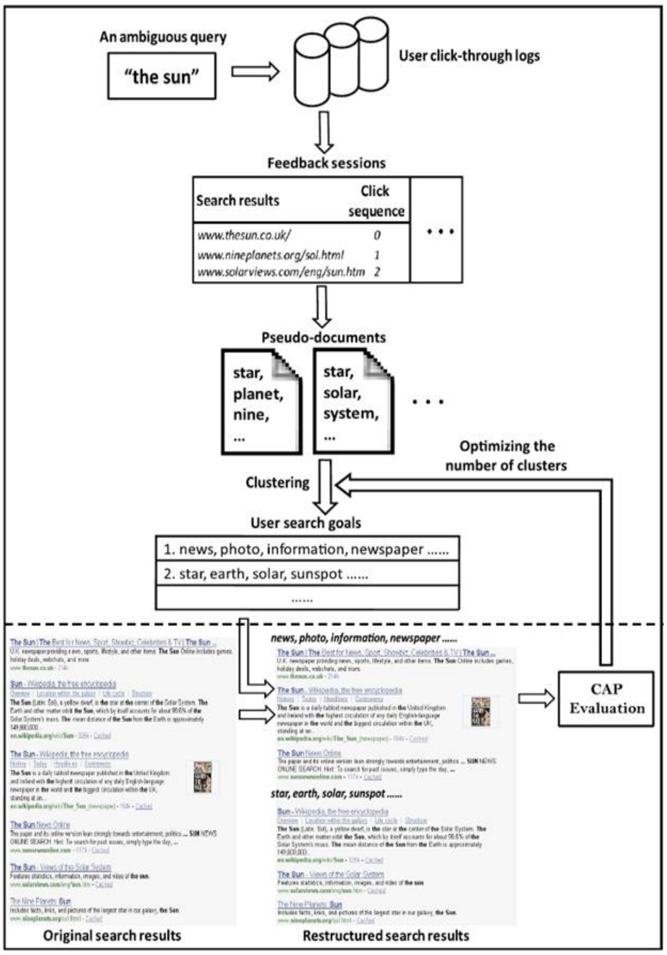
To sum up, our work has three major contributions as follows:

- We propose a framework to infer different user search goals for a query by clustering feedback sessions. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.
- We propose a novel optimization method to combine the enriched URLs in a feedback session to form a pseudo document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail.
- We propose a new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus, we can determine the number of user search goals for a query.

Restructuring Web Search Results

Since search engines always return millions of search results, 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. We will introduce how to restructure web search results by inferred user search goals at first. Then, the evaluation based on restructuring web search results will be described. The inferred user search goals are represented by the vectors in and the feature representation of each URL in the search results can be computed. Then, we can categorize each URL into a cluster centered by the inferred search goals. In this paper, we perform categorization by choosing the smallest distance between the URL vector and user-search-goal vectors. By this way, the search results can be restructured according to the inferred user search goals.

SYSTEM ARCHITECTURE:


An ambiguous query
"the sun"
User click-through logs
Feedback sessions
Search results
Click sequence
www.thesun.co.uk/ 0
www.nineplanets.org/sol.html 1
www.solarviews.com/eng/sun.htm 2
...
Pseudo-documents
star, planet, nine, ...
star, solar, system, ...
...
Optimizing the number of clusters
Clustering
User search goals
1. news, photo, information, newspaper ......
2. star, earth, solar, sunspot ......
......
news, photo, information, newspaper ......
star, earth, solar, sunspot ......
CAP Evaluation
Original search results
Restructured search results

MODULES:

- Feedback Sessions
- Pseudo-documents
- Inferring Pseudo-documents
- Evaluation Search Result

MODULES DESCRIPTION:

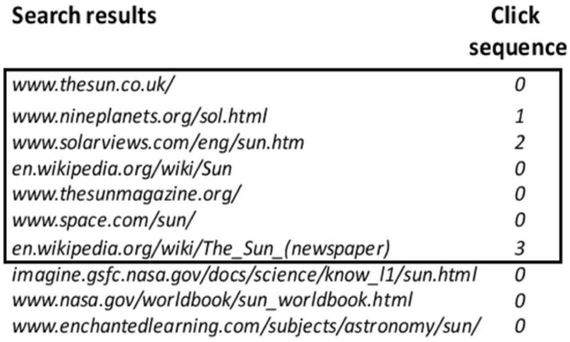
Feedback Sessions

The 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 in this paper is based on a single session, although it can be extended to the whole session. The proposed 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. Fig. 3.1 shows an example of a feedback session and a single session. In Fig. 3.1, 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 "unclicked." The single session includes all the 10 URLs in Fig. 3.1, while the feedback session only includes the seven URLs in the rectangular box. The seven URLs consist of three clicked URLs and four unclicked URLs in this example.

Generally speaking, since users will scan the URLs one by one from top to down, we can consider that besides the three clicked URLs, the four unclicked 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 unclicked URLs reflect what users do not care about. It should be noted that the unclicked 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 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

Pseudo-documents

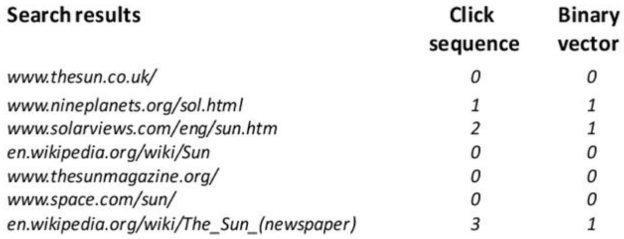
The URLs with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the feedback session. In this way, each URL in a feedback session is represented by a small text paragraph that consists of its title and snippet. Then, some textual processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. 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.



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

Fig. 3.2. A feedback session in a single session. "0" in click sequence means "unclicked." 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-throughs and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. Some representation method is needed to describe feedback sessions in a more efficient and coherent way. There can be any kinds of feature representations of feedback sessions. For example, Fig. 3.3 shows a popular binary vector method to represent a feedback session. Same as Fig. 3.2, search results are the URLs returned by the search engine when the query "the sun" is submitted, and "0" represents "unclicked" in the click sequence. The binary vector [0110001] can be used to represent the feedback session, where "1" represents "clicked" and "0" represents "unclicked." 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.



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

Fig. 3.3. The binary vector representation of a feedback session.

For a query, users will usually have some vague keywords representing their interests in their minds. They use these keywords to determine whether a document can satisfy their needs. We name these keywords "goal texts" as shown in Fig.3.4 . 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. Thus, pseudo-documents can be used to infer user search goals. In this paper, we propose a novel way to map feedback sessions to pseudo-documents, as illustrated The building of a pseudo-document includes two steps. They are described in the following:

1) Representing the URLs in the feedback session.

2) Forming pseudo-document based on URL representations.

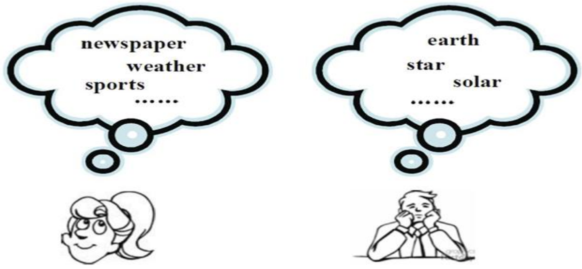


Fig. 3.4. Goal texts. For a query, different users will have differentkeywords in their minds. These keywords are vague and have no order. We name them "goal texts," which reflect user information needs.

Inferring Pseudo-documents

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. As each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document. 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. 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 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.

Evaluation Search Result

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, as shown in Fig. 3.1.Therefore, it is necessary to develop a metric to evaluate the performance of user search goal inference objectively.

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.

ALGORITHMS

CAP

Capturing Feedback Sessions:

For a web search is a series of successive queries to satisfy a single information need and some clicked search results. Here, feedback session consists of both clicked and unclicked URL's and ends with the last URL that was clicked in a single session. Clicked URL's state what users require and unclicked URL's reflect what users do not care about. For inferring user search goals it is more efficient to analyze the feedback sessions than to analyze search results or clicked URL's directly because there are different feedback sessions in user click-through logs.

Building pseudo-documents:

Representing the URL's in feedback session. Each URL's title and snippet are represented by term frequency-inverse document frequency as below,

$$\begin{aligned}\mathbf{T}_{u_i} &= [t_{w_1}, t_{w_2}, \ldots, t_{w_n}]^T, \\ \mathbf{S}_{u_i} &= [s_{w_1}, s_{w_2}, \ldots, s_{w_n}]^T,\end{aligned} \qquad (1)$$

Where T_{ui} and S_{ui} are TF-IDF vectors of the URL's title snippet. *ui* means *i*th URL in the feedback session.

wj(*j*=1,2,..,*n*) is *j*th term appearing in the enriched URL.

$$\mathbf{F}_{u_i} = \omega_t \mathbf{T}_{u_i} + \omega_s \mathbf{S}_{u_i} = [f_{w_1}, f_{w_2}, \ldots, f_{w_n}]^T, \qquad (2)$$

Here, F_{ui} is feature representation of *i*th URL in feedback session. Wt and Ws are weights of title and snippet. Here title should be more significant than snippets. So, the weight of title should be higher.

Forming pseudo-documents based on URL representations:

Here, an optimization method is used to combine both clicked and unclicked URL's in the feedback sessions. Let Ffs be the feature representation of feedback sessions and ffs(w) be the value for term w. Fucm(m=1,2,..,M) and Fucl(l=1,2,..,l) be the representation of clicked and unclicked URL's in the feedback sessions. Fucm(w) and Fucl(w) are the values of term w in vectors. Obtain such a Ffs that sum of distances between Ffs and each Fucm is minimized and sum of distances between Ffs and each Fucl is maximized.

$$\mathbf{F}_{fs} = \left[f_{fs}(w_1), f_{fs}(w_2), \ldots f_{fs}(w_n)\right]^T,$$
$$f_{fs}(w) = \arg\min_{f_{fs}(w)} \left\{ \sum_M \left[f_{fs}(w) - f_{uc_m}(w)\right]^2 - \lambda \sum_L \left[f_{fs}(w) - f_{u\bar{c}_l}(w)\right]^2 \right\}, f_{fs}(w) \in I_c. \quad (3)$$

Let I_c be the interval $[\mu f_{uc}$ *(w)*-σf_{uc} *(w)*,μf_{uc} *(w)*+σf_{uc} *(w)*] and I_c- be the interval [μf_{uc}^{-} *(w)*-σf_{uc}^{-} *(w)*,μf_{uc}^{-} *(w)*+σf_{uc}^{-} *(w)*],where μf_{uc} *(w)* and σf_{uc} *(w)* represent the mean and mean square error of f_{uc}*(w)* respectively, and μf_{uc}^{-} *(w)* and σf_{uc}^{-} *(w)* represent the mean and mean square error of f_{uc}^{-} *(w)*, respectively. If I_c *€* I_c^{-} *or* I_c^{-} *€* I_c, we consider that the user does not care about the term w. In this situation, we set f_{fs}*(w)* to be 0, as shown in

$$f_{fs}(w) = 0, I_c \subseteq I_{\bar{c}} \text{ or } I_{\bar{c}} \subseteq I_c. \quad (4)$$

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 cosinescore of F_{fsi} and F_{fsj},as follows:

$$Sim_{i,j} = \cos\left(\mathbf{F}_{fs_i}, \mathbf{F}_{fs_j}\right) = \frac{\mathbf{F}_{fs_i} \cdot \mathbf{F}_{fs_j}}{|\mathbf{F}_{fs_i}||\mathbf{F}_{fs_j}|}. \quad (5)$$

And the distance between two feedback session is

$$Dis_{i,j} = 1 - Sim_{i,j}. \quad (6)$$

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. 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, as shown in

$$\mathbf{F}_{center_i} = \frac{\sum_{k=1}^{C_i} \mathbf{F}_{fs_k}}{C_i}, \left(\mathbf{F}_{fs_k} \subset Cluster\ i\right), \quad (7)$$

where $F_{centeri}$ is the *i*th cluster's center 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.

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

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:

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

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 ith clicked URL and the *j*th clicked URL are not categorized into one class, d_{ij} will be 1; otherwise, it will be 0. $C^2{}_m$ =(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

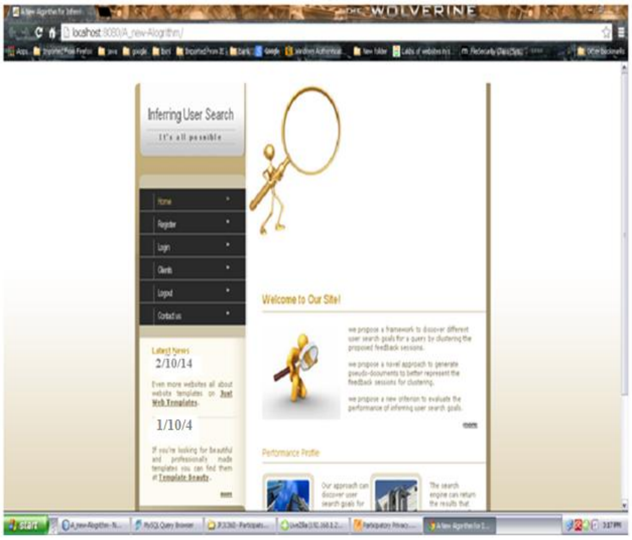
$$CAP = VAP \times (1 - Risk)^{\gamma}. \quad (10)$$

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. And γ is used to adjust the influence of

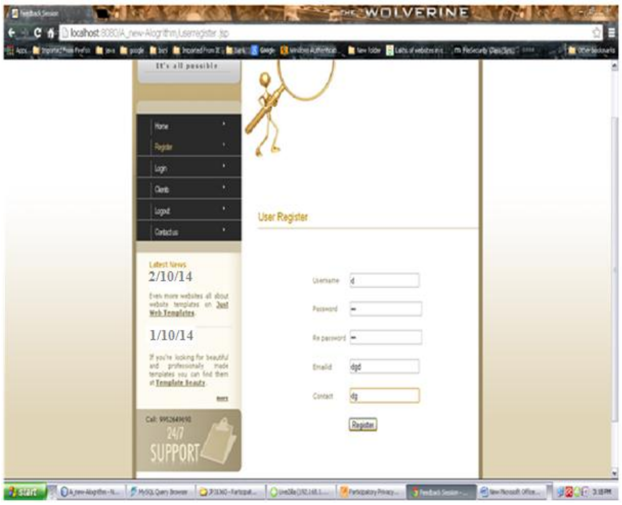
Risk on CAP, which can be learned from training data. Finally, we utilize CAP to evaluate the performance of restructuring search results.

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 less 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.

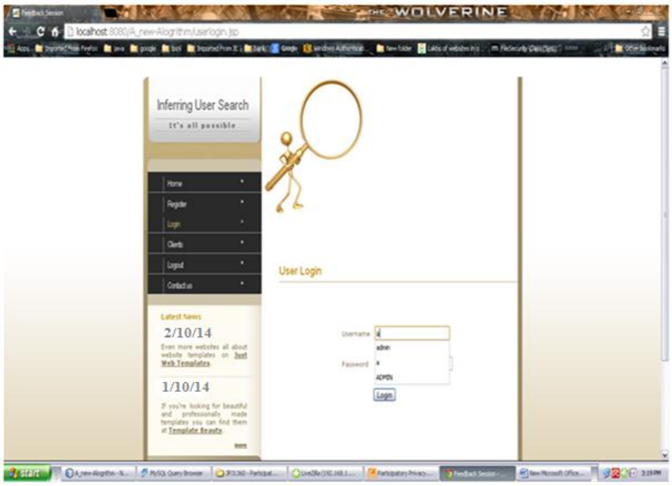
SCREEN SHOTS



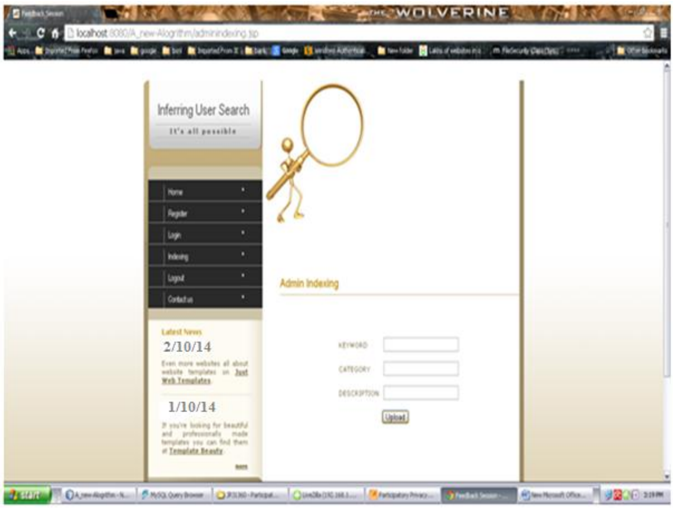
SCREEN 7.1 HOME PAGE



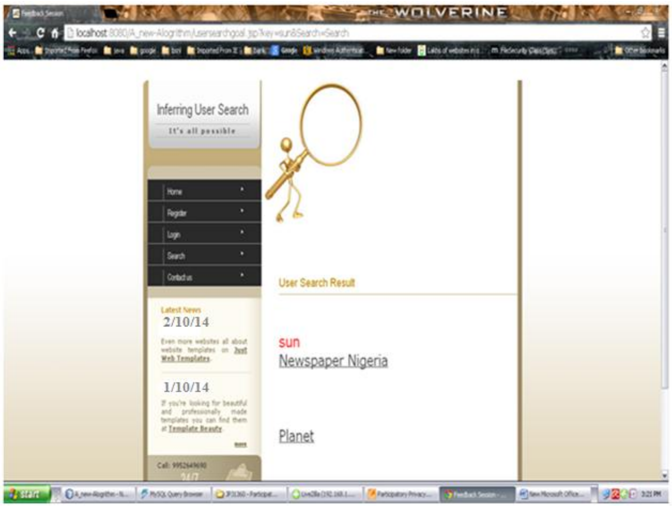
SCREEN 7.2 USER REGISTER PAGE



SCREEN 7.3 USER LOGIN PAGE

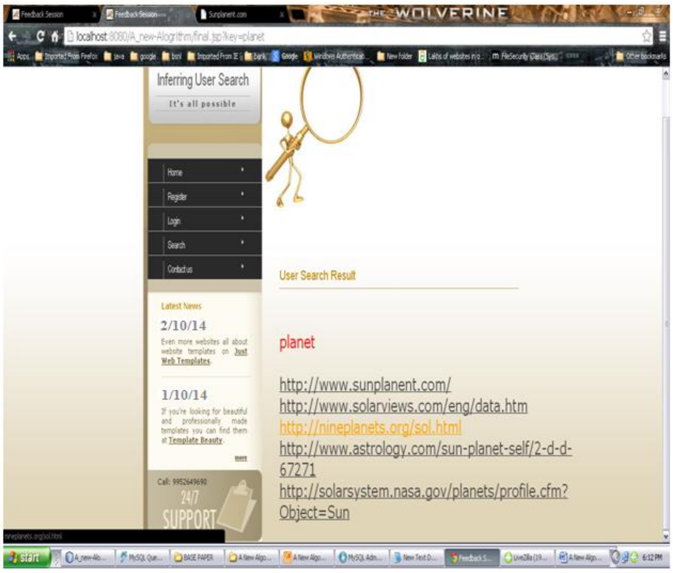


SCREEN 7.4 ADDING INFORMATION



sun
Newspaper Nigeria
Planet

SCREEN 7.5 RESULT FOR THE KEY WORD

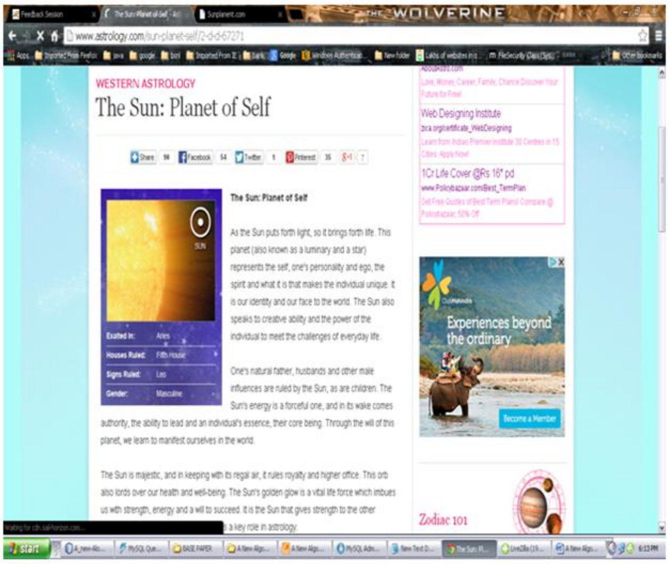

localhost:8080/A_new-Alogrithm/final.jsp?key=planet
Inferring User Search
It's all possible
Home
Register
Login
Search
Contact us
Latest News
2/10/14
Even more websites all about website templates on Just Web Templates.
1/10/14
If you're looking for beautiful and professionally made templates you can find them at Template Beauty.
more
Call: 9952649690
24/7 SUPPORT

User Search Result

planet

http://www.sunplanent.com/
http://www.solarviews.com/eng/data.htm
http://nineplanets.org/sol.html
http://www.astrology.com/sun-planet-self/2-d-d-67271
http://solarsystem.nasa.gov/planets/profile.cfm?Object=Sun

SCREEN 7.6 LINKS FOR THE KEY WORD



The Sun: Planet of Self

The Sun: Planet of Self

As the Sun puts forth light, so it brings forth life. This planet (also known as a luminary and a star) represents the self, one's personality and ego, the spirit and what it is that makes the individual unique. It is our identity and our face to the world. The Sun also speaks to creative ability and the power of the individual to meet the challenges of everyday life.

One's natural father, husbands and other male influences are ruled by the Sun, as are children. The Sun's energy is a forceful one, and in its wake comes authority, the ability to lead and an individual's essence, their core being. Through the will of this planet, we learn to manifest ourselves in the world.

The Sun is majestic, and in keeping with its regal air, it rules royalty and higher office. This orb also lords over our health and well-being. The Sun's golden glow is a vital life force which imbues us with strength, energy and a will to succeed. It is the Sun that gives strength to the other ... a key role in astrology.

SCREEN 7.7 DISCRIPTION FOR THE KEY WORD

CONCLUSION

In this project, 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 un-clicked 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.

FUTURE SCOPE

In this we described a new approach for understanding user search goals using statistical methods to gather the search history and track changes in user's interests. Unlike most previous related work, we focus on the updating of the search history representation using user relevance point of view on familiar words, in order to build and learn different user's interests. The design of an experimental evaluation of our approach requires a large scale of quantitative data on user search sessions and accurate contexts provided by the related queries during a reasonable period of testing a particular search engine. We currently develop an evaluation methodology which includes the construction of such collections test and the definition of accurate performance measures.

The user goals we used to build were based on a three-level deep concept hierarchy. We would like to examine the effect of using fewer or more level. Also, the current concept hierarchy is static, and we would like to evaluate algorithms to dynamically adapt the hierarchy for specific users by merging and/or splitting concepts based upon the amount of user interest. Finally, we would like to combine the user profiles with the document selection process, not just the document re-ranking, to provide a wider set of relevant results to the user rather than just reorganizing the existing results

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.

[2] 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.

[3] 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.

[4] 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.

[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. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’08), pp. 875-883, 2008.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] R. Jones, B. Rey, O. Madani, and W. Greiner, “Generating Query Substitutions,” Proc. 15th Int’l Conf. World Wide Web (WWW ’06), pp. 387-396, 2006.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.