SEUC-Search Engines Based on User Preferences with Click-Through Data

Dr. M. Arun Manicka Raja, Ms.T.Sumitha, Dr. N.R. Rejin Paul

Associate Professor,Department of Computer Science and Engineering,
RMK College of Engineering and Technology, Chennai.
Assistant Professor,Department of Computer Science and Engineering,
RMK Engineering College, Chennai
Associate Professor,
Department of Computer Science and Engineering,
RMK College of Engineering and Technology, Chennai.
arunmanickarajam@gmail.com, Sumitharmk90@gmail.com, nrrejinpaul@gmail.com

Abstract. In general different people have different search goals while searching in an engine. A search engine helps people reach their needed information. Click-through rate (CTR) is the ratio of users who click on a certain link to the total number of users who view a page. But on a broader concept it becomes difficult to actually tally each user's query and give them the required solution. This paper here is to overcome the above problem. It is to conclude the user's search goals by analyzing the search engine logs. Majorly the previous works try to fit in all user' s search goal in a single way which is actually difficult because there is no guarantee that each user may have same search goals. The problem arises when the search goals of the users vary amongst one another and to overcome this here we use a framework model SEUC to mine the user preferences through feedback sessions. This will help in retrieval of information based on the user query. The feedback sessions are actually obtained by the click-through data of the user.

KEYWORDS: Mining; Click-Through data(CTR); Search engine; SEUC.

1 Introduction

As the web is expanding dramatically the size of the result obtained from the search query in a search engine is also increasing in number. This makes the task of choosing the required information for the user a tedious one. According to a study it is actually found that the average length of the search query 2.35 terms which is very difficult to be understood and to show the exact result. Likely even the users may not wish to change the query because it becomes a difficult process on them to do so. To overcome all these in this paper the user needs are cluster based on their queries ensured to supply them with the needed and relevant result. Personalized systems address the overload problem by building, managing and representing information that can customize individual users. This customization may take the form of filtering the irrelevant information and obtaining the additional information of the likely interest for user. In modern Web, as the amount of information available can cause information overloading, and the demand for personalized approaches for the relevant information access increases. For example, if a user who is planning to visit India might issue the query "hotel," and click on the search results for finding hotels in India. From click through of the query "hotel," user can learn
the content preference (e.g., "room rate" and "facilities").

In this paper, the model (SEUC- Search Engines based on User preference Click through data) is actually developed and designed to gather the user queries and with the help of a software, it separates the content and the user click-through data. Further this data is mined and each user profile is updated with the mined data so that when a user searches in a search engine then his preference is displayed first. Thus this is a practical approach to capture a user's interests for search personalization by analyzing the user's click-through data, and user's concept preferences had showed that it is more useful than methods which based on page preferences.

2 Related Works

In the novel fusion technique [1] which combined closely associated and related documents tend to be relevant for a same query to achieve consistent and needed improvements. This method involves three steps: clustering, re-ranking, and fusion.

E. Agichtein et al [2] showed that incorporating the user behavior data could significantly improve the ordering of top results in the real web search setting. The alternatives are analyzed for incorporating the feedback into ranking process and explore the contributions of the user feedbacks. Results are evaluated of large scale over 3,000 queries & 12 million user

interactions with a popular web search engine. Incorporating implicit feedback could augment other features by improving the accuracy of a competitive web search ranking algorithm with 31 percent related to the original performance

H. Chen and S. Dumais [3] proposed a novel approach for inferring user search goals by analyzing the search engine query logs. The goal of personalized IR is to return the search results that have a better match to the user intent. Firstly, they proposed a framework to discover the different user search goals for a query to cluster the proposed feedback sessions where the feedback sessions are getting constructed from the user's click-through logs and can efficiently show the information needs of the users. Secondly, they proposed an approach to generate the pseudo -documents for better representing of the feedback sessions for clustering.

K.W.T Leung et al., [4][5][6] have introduced an effective approach that captures the user's preferences in order to give personalized suggestions for query by using two new strategies , one being the online technique that could extract the concepts from those web snippets of the search result obtained from a query and to use that concepts to identify related queries for that query and another one, is a new two-phase agglomerative clustering algorithm for generating personalized query clusters for the users.

Liu Sheng et al [7] proposed a personalized search approach was easily extended and the conventional search engine was setup on the client side. The mapping framework automatically maps a set of known user interests onto a group of categories in the Open Directory Project (ODP) and takes an advantage of manually edited data available in ODP for training text classifiers that correspond to, and therefore categorizes and personalizes the search results according to the user interests. In the two sets of controlled experiments, they compare the personalized categorization system (PCAT) with a list interface systems (LIST) which mimics a typical search engine and with nonpersonalized categorization systems (CAT). In both the experiments, PCAT was preferable to LIST in information gathering the types of tasks and for the searches of short queries, and PCAT outperforms CAT in both the information gathering and the finding of types of tasks, and the searches associated with free-form queries.

A. Chowdhury et al [8] the web query assigned by the users analysis the goal and the goal identification is used to improve quality of search results. The human subject strongly studies and indicates the automatic query goal identification. It has used two tasks like as past user click behavior and anchor link distribution for goal identification and combining these two tasks can identify 90 percent goal accurately.

Alessandro Micarelli et al [9] discussed the most popular techniques for collecting the information about the users, for representing, and building user profiles. It shows the contrast between the explicit information techniques with the implicitly collected user info using the browser caches, proxy servers, browser agents, search logs, and desktop agents. It shows the survey report on how each of the profiles is constructed and gave examples of projects that uses each of these techniques.

Ning Zhong et al [10] proposed a personalized ontology model for knowledge representation and reasoning over user profiles. This model analyses the ontological user profiles from both world knowledge base and the user local instance repositories[11]. The ontology model was evaluated by comparing it against the benchmark models in the web information gathering.

3 PROPOSED SYSTEM

System architecture defines the process of a model in detail and helps us to get a clear overview of the functions performed by that model. The system architecture for the above proposed model is as follows.

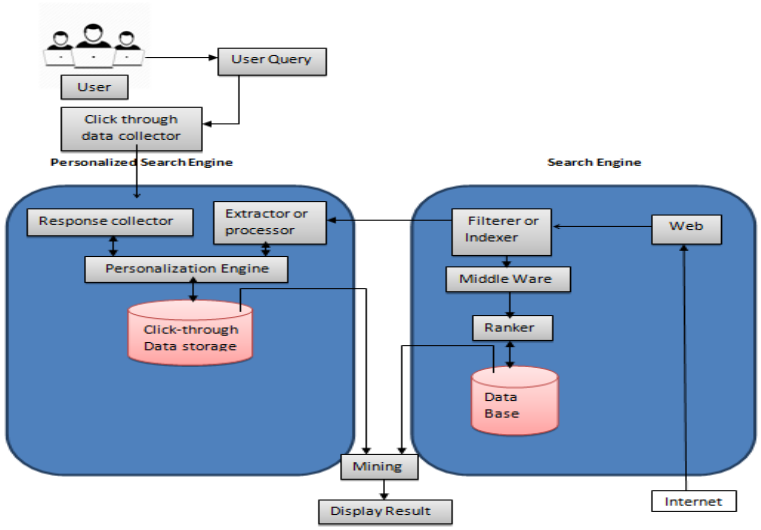


Fig. 1 Architecture diagram with click through data

Detailing the above structure,

USER: It refers to the person who is actually working or searching in the engine. He / She are the end user of our proposed model.

USER QUERY: It is the question or the search goal of the user. It defines on what the user wants to get from the search engine. It specifies the constraints and also the needed statements to get the required information.

3.1 Personalized Search Engine

a) CLICK-THROUGH DATA COLLECTOR: It is an data collector to get the information about the user. Here the click-through data of each and every user is collected. It helps in further division of that data into the needed responses and extract or the content.

b) RESPONSE: The responses deals with the observation of what the user thinks about any specific thing. It helps in gathering the user preferences in common on all domain and helps in personalizing each user's likes and dislikes into the search engine.

c) EXTRACT OR CONTENT: The extract or content is where the domain relating to the response is recorded. It also helps in recording the most wanted and needed information of that user.

d) PERSONALIZATION ENGINE: It is the engine that actually performs the personalization and collection of each user preference to gather it and store in the database for further process. Now the data are collected and stored in the database for segregation based on user needs and different users.

3.2 Search Engine

a) WEB: It is used to get us the needed information. It is World Wide Web and hence with the help of the internet it helps in gathering all the related information of that statement to give us the broader options on choosing the needed one.

b) FILTERER OR INDEXER: It filters the information based on the keywords present in the statement query given. It does show all the related info but it indexes it in some specific order before the actual display.

c) MIDDILE WARE: It is software that gathers the information collected from the filterer or indexer to send it for ranking purpose. The information may be filtered based on needs but may not be ranked on priority. So this software collects the information to send it for the ranking purpose.

d) RANKER: Ranking helps in making it possible to evaluate the complex information according to certain criteria or constraint. Thus, an Internet search engine may rank the pages in the web which it finds most needed by an estimation of their relevance and making it possible for the user in quick selection of the web pages they are likely to want to see. Now these data are collected and stored in database.

3.3 Mining

The data from both the data bases are collected together to perform the necessary mining process in it. The mining that is used to get the preference of the user through the click through data is WEB MINING. Web mining is process of using the data mining techniques for automatic discovering and extracting the information from the web documents and some services. There are generally three classes of information that can be obtained by the web mining and they are web activity, from server logs and web browser activity tracking. After the web mining, the data are gathered and displayed to each user based on needs. Their review is collected for further modification of the results if any and this result is displayed for the user.

4 Implementation and Result

Here,
Q represents **Query.**
U(Q) represents **user query.**
C represents **clickthrough data collector.**
C1, C2,…..,Cn represents different **clickthrough datas.**
R(Q) represents the **result** for the above Query.

ALGORITHM:
Step 1: Initially, the user enters the query to be search i.e, Q. The above query is to be searched is known as the user Query, U(Q). Therefore, **Q U(Q).**
Step 2: U(Q) is stored in the clickthrough data collector, C. i.e, **U(Q) C.**
Step 3: From the Clickthrough datas, C1,C2,….Cn, the responses for these datas are ollected and reranked. These response contents are stored and extracted when needed from the personalized search engine database. i.e, **C C1,C2,…..Cn.**
Step 4: In Search engine, the filter or Indexer Filters the user query, U(Q), from the database and sends it to the Extractor of the personalized search engine.
Step 5: In this step, using Web Mining, the resultant Clickthrough data, R(Q), is reranked and displayed to the user for the given User Query, U(Q). Thus, **C R(Q).**

4.1 Required Support Software

Required support software in this work is:
- Sun Microsystem's Java Development Kit (JDK) 2.
- Windows 98/2000/NT/higher
- Oracle 8i

4.2 Input Data

Input data to data mining engine is divided into two parts: from user and from database or data warehouse. From user, minimum support, minimum confidence and interested category are provided to data mining engine. The minimum support and minimum confidence shall be a percentage value in the range of 0% to 100%. Interested category shall be either 0 for all categories or exist in database or data warehouse where data mining engine retrieves data.

4.3 Output Data

For standalone mode of data mining engine, the output data device is computer screen where all association rules are displayed. If the data mining engine is used by web based tools, there is no output data medium or device for data mining engine itself, but web based tools will use web browser on computer screen to display association rules to the user.

4.4 Experimental Analysis

The following graph shows the comparison in time taken between the hashing technique used in this model and the previous hashing models. It shows that the time taken for hashing to get the frequent item sets in this model is much less when compared to that of other models and hence turns out to be effective. The x-axis denotes the datasets implemented in previous and present work while the y-axis denotes the time taken to give the information required.

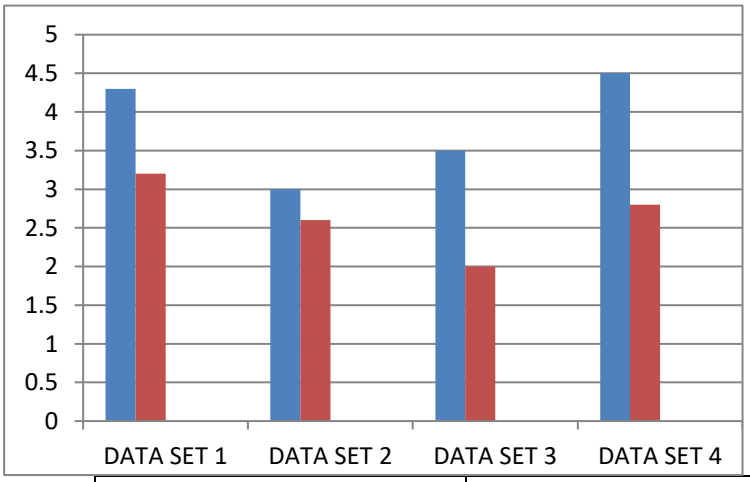
4.5 Example of Proposed Work

Consider the following example specified in the table for performing the function.

Transaction	**Item-set**
T1	Amyoyo syndrome, Facecal encephalopathy, sbi, ctf, cut and paste
T2	Amyoyo syndrome, Facecal encephalopathy, sbi
T3	sbi, ctf, cut and paste
T4	Facecal encephalopathy, sbi, cut and paste
T5	Facecal encephalopathy, sbi, ctf

Let Amyoyo syndrome, Facecal encephalopathy, sbi, ctf, cut and paste be I1,I2,I3,I4,I5

VERTICAL FORMAT:



Item-set	**Transaction**

I1	**T1,T2**
I2	**T1,T2,T4,T5**
I3	**T1,T2,T3,T4,T5**
I4	**T1,T3,T5**
I5	**T1,T3,T4**

General grouping is done in second level. But we can also specify minimum support count if needed.

SECOND LEVEL:

Item-set	**Transaction**
I1,I2	**T1,T2**
I1,I3	**T1,T2**
I1,I4	**T1**
I1,I5	**T1**
I2,I3	**T1,T2,T4,T5**
I2,I4	**T1,T5**
I2,I5	**T1,T5**
I3,I4	**T1,T3,T5**
I3,I5	**T1,T3,T4**
I4,I5	**T1,T3**

With minimum support count to be equal or more than 3 we perform third level;

Item-set	**Transaction**
I2,I3,I4	**T1,T5**
I2,I3,I5	**T1,T4**
I2,I4,I5	**T1,T3**

This process of grouping is done with the help of hashing and hence once grouped the user can easily get the required item-set or information needed. Hashing plays important role in grouping of the item-sets.

This way the process is made easier and more efficient. Each level is drawn with the corresponding hash tables to show the linked list or the linking between the multiple transactions and its ID. We have used the item set to show the relevance of the data , and also to get the preference of the user through this. It is known fact that the transaction requires more memory than the item set which is the main reason for preferring this idea. Thus it shows that this algorithm performs in efficient way and supports various tasks.

5.CONCLUSION:

The experimental results demonstrate that this method is more efficient as it easily tags corresponding needed preference of user without the primary and secondary clustering problems. This way has been made helpful in identifying the user preference and the user goals to demonstrate their need in future when they log in to the search engine. The vertical format representation makes it easy to manipulate the data. Hence this tends out to be more effective and useful model.

6.REFERENCES:

[1] A.Leuski and J. Allan, "Improving interactive retrieval by combining ranked list and clustering", *Proceedings of RIAO*, pp. 665-681, 2000

[2] E. Agichtein, E. Brill, and S. Dumais, “Improving web search ranking by incorporating user behavior information”, in Proc. of ACM SIGIR Conference, 2006.
[3] 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
[4] K. W.-T. Leung, W. Ng, and D. L. Lee, “Personalized concept-based clustering of search engine queries”, IEEE TKDE, vol. 20, no. 11, 2008 .
[5] K. W.-T. Leung, W. Ng, and D. L. Lee, “Personalized concept-based clustering of search engine queries”, IEEE TKDE, vol. 20, no. 11, 2008.
[6] K. Leung, D. Lee, W. Lee ,” Personalized Web Search with Location Preferences”, ICDE Conference 2010, 978-1-4244-5446-4/10/@ 2010 IEEE.
[7] Ma, Z., Pant, G., and Liu Sheng, ” Interest-based personalized search”, ACM Trans. Inform. Syst. 25, 1, Article 5 (February 2007), 38 pages. DOI = 10.1145/1198296.1198301.
[8] 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)*.
[9] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli, “User Profiles for Personalized Information Access”, the Adaptive Web, LNCS 4321, pp. 54 – 89, 2007 © Springer-Verlag Berlin Heidelberg 2007.
[10] Xiaohui Tao, Yuefeng Li, and Ning Zhong, “A Personalized Ontology Model for Web Information Gathering” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO.4, APRIL 2011, 1041-4347/11 _ 2011 IEEE Published by the IEEE Computer Society.
[11] S. Di, "Deep Interest Network for Taobao advertising data Click-Through Rate Prediction," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 741-744, doi: 10.1109/CISCE52179.2021.9445990.