Android based Preference Mining using clickthrough data

Ms Deepika Bhatia*
Department of Computer Sc. & Engg.,
G.H.Raisoni College of Engineering
Nagpur, India
deepika.1885@rediffmail.com

Ms Smita Nirkhi
Department of Computer Sc. & Engg.,
G.H.Raisoni College of Engineering
Nagpur, India
smita811@gmail.com

Dr Preeti Bajaj
Department of Electronic Sc.& Engg.,
G.H.Raisoni College of Engineering
Nagpur, India
principal@ghrce.edu.in

Abstract*— Search engines are effective tools for discovering knowledge from the bulk of information available on the Internet. However, while search engines can return highly relevant results for a query, the user has to understand the results page by page [1] in order to understand the concepts hidden in the results. Our approach is to give results to the user according to his/her preferences. The user is provided with location based personalized search. The result returned to user is more efficient as per our clickthrough data collection approach.

***Keywords*—Clickthrough data, Personalised Search, GPS, context aware**

I. INTRODUCTION

Most current search engines, however, return the same results to all users who ask the similar type of query. This is clearly inadequate when the users have different search objectives and interests. For example, for the query “jaguar”, some users may be interested in Web pages about “jaguar" as a car, while other users may want information related to “jaguar" as an animal. In fact, current Web search engines return mostly pages about jaguar as a car, making it difficult for users to retrieve pages about jaguar as an animal. We can easily discover that many queries such as “java”, apple”, semantically related words may be interpreted by different users differently. We should also note that this problem is more than a problem of query semantics; even if a query is interpreted by users in the same way, users may still be looking for different aspects of the subject (e.g., one may be interested in Java compilers while others may be interested in Java language notes). Thus, delivering the same search results for the same query is not satisfactory. Personalization is based on observed patterns, and resulting probabilities. There are a number of patterns that search engines are likely to track, and that will permit search engines to calculate probabilities of increasing either reliability or speed of use. Various types of personalization are:-

1) Location Based Personalization **:-**

It serves personalized results based on user’s current location (e.g. movies). Example:-Search engines have a good idea where user is (based on GPS settings on our mobile phone), when a search query is entered. So someone searching for "movies" from a location in Nagpur is very likely looking for a movie location within the radius of their inferred location.

2) Interface Based Personalization: **-**

It serves personalized results based on the interface user is using. Example: if a search engine knows user is performing a search for "hospitals" from a Nokia phone, they can typically understand that user is looking for a hospital within a given radius of current location, and could reorganize the results with the closest hospitals first.

3) Query History Based Personalization:-

It serves personalized results based on a logical sequence of keyword queries. Example: if user performed a search for "apple" 15 minutes prior, and now searches for "apple", personalized results should show fruits results above computer results.

4) Individual User behavior Based Personalization:-

Serves personalized results based on user’s actions in the past. Example: has the user clicked on this site in the past in response to a similar query? Did user abandon it immediately, or stay on it and interact with it? User’s previous interaction with the site offers insight into whether it is a better match this time or not.

Our approach is based upon the following viewpoints. The differences between our work and existing works are as follows:-

- Existing earlier work [7], [8] require the users' to manually define their location preferences explicitly (with latitude-longitude pairs, as similar in case of desktop environment). Our proposed method does not require users to explicitly provide their location interests manually.
- Our method automatically extracts both the user's content and location preferences, which are automatically extracted from the user's clickthrough data [9] without requiring any extra input from the user.

The rest of the paper is organized as follows. We review the related work in Section II. In Section III, we present our method for clickthrough data collection for content and location ontologies. In Section IV, we present the experimental evaluation of our approach and provide snapshots to show how this benefits the personalized search. We classify the users and queries in our experiments into different groups according to different user’s interests. Section V and VI concludes the paper and also suggests future advancements. References are given at the end of the proposed future advancements. Out approaches are more efficient and also provide personalized results.

II. RELATED WORK

Preference mining is a challenging problem as evident in the recent work in [Joachims 2002b; Deng et al. 2004; Joachims et al. 2005] ([3] [4]). Earlier algorithms are based on some strong assumptions on how users scan the search results in a strict order and then deduce the relative preferences, which may not be correct in reality. For example, Joachims' algorithm assumes that users find search results strictly from top to bottom. However, it is possible that a user skips several results without examining them carefully. As a result, Joachims' assumption is too simplistic to predict all correct preference pairs to accurately reflect users' needs.

Because of geography's important role in search requests [5], and the significant commercial potential of such queries (e.g., for hotels or businesses etc.), local search, i.e., methods aimed at giving improved answers to geographic search requests were the main issues. Various approaches range from integration of yellow pages to answer simple but lucrative queries (e.g., restaurants), to a more detailed analysis of queries, page content, and site and link structure. Geo search applications can use a standard keyword interface and extract geographic terms from queries, employ graphic interfaces such as interactive maps, or use the current location of a mobile user. User's interest in what types of geographic queries (informational, navigational etc.) have been studied. Study has been done about what sites users visited as a result of a geo query, how different geographic terms were used by the same user, and what non-geographic terms are associated with geographic terms. All these aspect were taken into account without user's intentions about a particular query word.

Gan et. al [5] suggested that search queries can be classified into two types, content (i.e., non-geo) and location (i.e., geo). Examples of geographic queries are .hotels USA, Indian historical sites.. A classifier was built to classify geo and non-geo queries, and also the properties of geographical queries were studied in detail.

III. PROPOSED APPROACHES

In this paper, we tackle the problem of search engine adaptation by considering three main research approaches as given below:-

1) Clickthrough data collection approach:**-** The data source we investigate is clickthrough data [2], which can be formally represented as a triplet (*q*, *r*, *c*), where *q* is the input query, *r* is the result list of links (link 1,….,*link n*), and *c* is the set of links that the user has clicked upon. Table1 illustrates an example of clickthrough data for the query "government Kannada jobs in Delhi". In the table, the two links, L3 and L5, are in bold, indicating that they have been clicked on by the user. The advantage of using clickthrough data to analyze a user's preferences is that it does not intervene or distracts the user's interaction with the searching process through the middleware. The data can be collected by a search engine without extra burden on the user. Thus, clickthrough data are much easier to collect and more abundant than explicit feedback [Bartell et al. 1994] that requires the user's explicit ratings. Table I shows the links which are clicked and not clicked by the user. Also it gives the title abstracts and URLs of web pages.

TABLE I
Clickthrough data collected for the query Kannada government jobs Delhi

Links	**Search results with title, abstracts and URLs of Web pages**
L1	Kannada-jobs in Delhi-Times Jobs.com::http: // delhi.timesjobs.com/jobs/kannada-jobs-in-delhi
L2	IT & engineering jobs in India: faculty in Delhi Kannada Society's…:: http ://itjobsdelhi.blogspot.com /2009/06/faculty-in-delhi-kannada-societys.html
L3 (clicked)	**http://www.google.co.in/url?q=http://www.sarkari-saukri.in**
L4	http://www.google.co.in/url?q=http://bangalore.click .in/ classified/jobs/ placement-consultants/ government-jobs-india-state-central-
L5 (clicked)	**http://www.google.co.in/url?q=http://123oye.com/newindex/kannada-typist-jobs-in-delhi/**

2) Preference mining:**-**It discovers user's preferences of search results from clickthrough data. For example, for a particular query, Q, if a user chooses to click a search result, link *A*, but skips another link *B*, preference mining algorithms aim to discover the user's preferences from the click- through data, e.g., the user prefers search result link *A* to link *B* for query Q. Clickthrough data is a search engine log that records for each query the result list presented to the user by the search engine(here middleware used as Google) as well as the links clicked on by the user.

3) Weighted page ranking concept:**-**which optimizes the ranking function of a search engine according to the user's preferences.

IV. EXPERIMENTAL EVALUATION

This paper addresses search engine personalization. We present a new approach to mining a user's preferences on the search results from clickthrough data and using the discovered preferences to improve search quality. We used an approach which is based on the practical assumption that the search results clicked on by the user show the user's preferences, but it does not draw any conclusions about the results on which the user did not click on. Also user does not follow any predefined order in reading the search results or does not click on all relevant results. Our extensive online experiments demonstrate that our approach discovers better results. Our personalization approach is effective in practice.

We asked four groups of students from various departments at our college, namely Computer Science, Electronics, Electrical and Embedded to use our personalized search engine. Each group had five students. We assumed the following about our groups:

Users from different departments have different interests but users within the same department share the same interests. The students from electronics department are interested in equipments and electrical items. Students from computer science department are interested in computer related information and so on. As far as the personalization method is concerned, the four groups of students can be considered as four logical users and the personalization methods tries to adapt the metasearch engine to deliver the best results to the respective group of users.

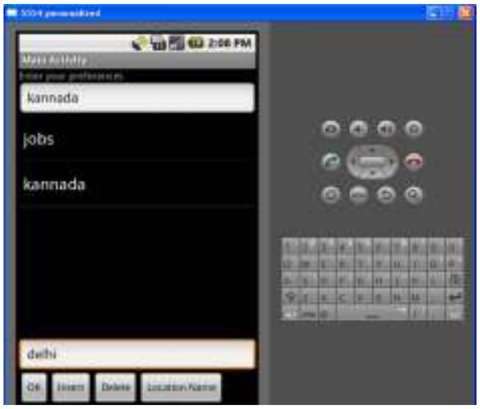
Using more than one student in each group ensures that the experimental results are not affected by a few peculiar actions made by few users. To collect the clickthrough data, each of the four groups of students submits to the metasearch engine 20 queries related to their interests. The metasearch engine at the beginning adopts a default ranking function to deliver results. The default ranking function combines the retrieved results from the underlying search engines in a round-robin manner. Table II shows some statistics of the clickthrough data collected.

TABLE II

Statistics of our Clickthrough data set.

Department	**Computer**	**Electronic**	**Electrical**	**Embedd-ed**
No. of queries	20	20	20	20
No. of clicks	3	5	6	9
Avg. clicks per query	6.66	4.0	3.33	2.2.2

The following snapshots explain the actual implementation of our modules. The user types the context initially he/she interested in. After that the search word is given to the middleware (here Google). The results according to locations are given to the user. Our approach will give the personalized results to the user if the user searches for the same query word again after many searches. It will give more effective and efficient search to the user. In figure 1, the user enters the search categories. Here the context is Kannada, jobs in Delhi location.



Enter your preferences

kannada

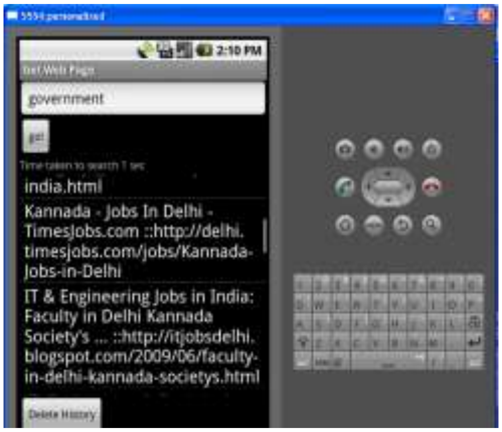
jobs

kannada

delhi

Figure 1.Input context

Figure 2 below shows the entered query word with results as per location. Here the query word is government entered by the user. If a user clicks on the URL, interested in, that clickthrough (Figure 3) data gets stored as personalized preferences for future use.



government

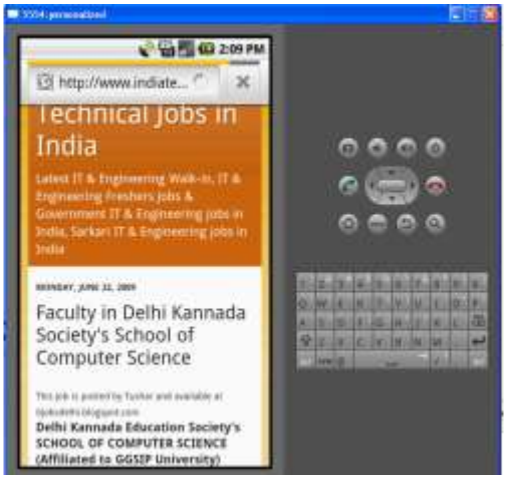
Time taken to search 1 sec

india.html

Kannada - Jobs In Delhi - TimesJobs.com ::http://delhi.timesjobs.com/jobs/Kannada-Jobs-in-Delhi

IT & Engineering Jobs in India: Faculty in Delhi Kannada Society's ... ::http://itjobsdelhi.blogspot.com/2009/06/faculty-in-delhi-kannada-societys.html

Figure 2.User search request



http://www.indiate...

Technical Jobs in India

Latest IT & Engineering Walk-in, IT & Engineering Freshers Jobs & Government IT & Engineering Jobs in India, Sarkari IT & Engineering Jobs in India

MONDAY, JUNE 22, 2009

Faculty in Delhi Kannada Society's School of Computer Science

This job is posted by Tushar and available at itjobsdelhi.blogspot.com

Delhi Kannada Education Society's SCHOOL OF COMPUTER SCIENCE (Affiliated to GGSIP University)

Figure 3.Clickthrough data collection

V. CONCLUSION

Personalized Web search service will play an important role on the Web search. This paper focuses on utilizing clickthrough data to improve Web search. While it is possible to improve the efficiency of search through each of the personalization methods as discussed above, they work best when operated in conjunction with one another, acting as a check and balance mechanism. When used in conjunction, the inferences truly become more probable, and lead to better search results. Google, Yahoo, and MSN all use various means of personalization.

VI. FUTURE ADVANCEMENTS

As for the future work, we plan to study the effectiveness of other kinds of concepts such as people names, time and user’s mood/behavior patterns for personalization. We will also investigate methods to exploit a user's content and location preference history, stored as clickthrough data, to predict regular user patterns or behaviors for enhancing our future search. In our current work, we are concerned about the users whose clickthrough data was recorded and collected. And only queries issued and pages clicked on by the specific users are considered. Therefore, it would be interesting to adapt our framework to new users, queries and Web pages. Also clustering can be done for new users and queries efficiently to improve the web search.

REFERENCES

[1] Constructing Concept Relation Network and its Application to Personalized Web Search Kenneth Wai-Ting Leung EDBT 2011, March 22–24, 2011, Uppsala, Sweden.

[2] Mining User Preference Using Spy Voting for Search Engine Personalization Wilfred Ng et al ACM Transactions on Internet Technologies, Vol. 7, No. 3, August 2007, Pages 1-28.

[3] Joachims, T. 2002a. Evaluating retrieval performance using clickthrough data. In Proc. of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval.

[4] Joachims, T. 2002b. Optimizing search engines using clickthrough data. In Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD02). 133-142.

[5] Qingqing Gan et. al. Analysis of Geographic Queries in a Search Engine Log LocWeb 2008, April 22, 2008, Beijing, China.

[6] Kenneth Wai-Ting Leung et. al. Personalized Web Search with Location Preferences 978-1-4244-5446-4/10/ 2010 IEEE 701 ICDE Conference 2010

[7] S. Yokoji, .Kokono search: A location based search engine,. in Proc. of WWW Conference, 2001.

[8] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, .Hybrid index structures for location-based web search,. in Proc. of CIKM Conference, 2005.

[9] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, .Learning to rank using gradient descent,. in Proc. of ICMLConference,2005.