Web Data Management for Mobile Users

Zheng Huo, Jing Zhao, Xiangmei Hu
huozheng123@gmail.com

1. Introduction

Mobile devices are becoming increasingly popular as a means of information access while on-the-go. With the emergence of web access friendly mobile devices, the number of mobile users who will access the web using their mobile devices is expected to increase drastically in the near future. Meanwhile most Web data are stored in millions of deep web data sources which can be accessed by desktop and also mobile users, however, the mobile users have other needs, or maybe they can't access deep web data conveniently as desktop users, such as the terminal have small screen, and the input capabilities is not as strong as desktop users. Sometimes mobile users' information needs are more location sensitive than desktop users. So the challenge is how to provide useful and convenient services for mobile users. In our survey of this topic, we found several questions to be solved, and we proposed an initial framework for web data management for mobile users.

2. Features of mobile users

There have bee several large scale examinations for user search behavior through search engine logs for both computer and mobile search. The result of this analysis have been used to improve performances of mobile users' access to the deep web.

Shorter queries - As analysis shows, the query length of the mobile users is shorter than desktop users. For computer-based search, the average number of words per query is 2.93 and the average number of characters per query is 18.72. The length of conventional mobile phone queries is the shortest of all the mediums, with an average query consisting of 2.44 words and 15.89 characters. The shorter query terms can be easily understood since the limitation of input function in mobile devices.

Information needs - Mobile users look for very different topics than standard desktop web users. Researchers find that the most popular mobile topics are local services and travel & commuting.

Location of mobile users - There is strong evidence indicating that location-based searches are popular among mobile searchers. By taking into account of users' location information, we can provide more personalized services.

Small screen size - This makes it difficult or impossible to see text and graphics dependent on the standard size of a desktop computer screen. So what kind of integrated interface is suitable for mobile users is a challenging problem.

Lack of windows - On a desktop computer, the ability to open more than one window at a time allows for multi-tasking and for easy revert to a previous page. On mobile devices, only one page can be displayed at a time, and pages can only be viewed in the sequence they were originally accessed.

Computing and memory limits - Most of them have slow computing speed and small storage capacity which restricts spatial search calculations, routing operations and the creation of a user specific “mobile” map.

Type limitation of accessible pages - Many sites that can be accessed on a desktop cannot on a mobile device. Many devices cannot access pages with a secured connection, Flash or other similar software, PDFs, or video sites, although recently this has been changing.

Lower speed - On most mobile devices, the speed of service is very slow, often slower than dial-up Internet access.

Compressed pages - Many pages, in their conversion to mobile format, are squeezed into an order different from how they would customarily be viewed on a desktop computer.

Size of messages limits - Many devices have limits on the number of characters that can be sent in an email message.

Expensive cost - the access and bandwidth charges levied by cell phone networks are much, much higher than those for fixed-line internet access.

3. Main framework

We will introduce some web data integration issues in this part, this is specially for mobile users based on the behavior analysis and features of mobile users above. Figure 1 shows the main part of the deep web data integration modules for mobile users. Following are the functions of each part.

WDS discovery- Discovering accessible web databases in the web.

Interface clustering- This part classifies web data sources according to their domains.

Interface analysis- Analyzing and extracting the schema information in query interfaces.

WDS profile (interface)- Meta information about WDS query interface, including attribute, type, etc.

Interface integration- Integrating interfaces of several WDS to a global integrated interface

Domain selection- Specifying a suitable domain for users.

Query predicate match- Matching queries submitted to the easy query interface to the integrated interface

WDS selection- Selecting suitable web data sources for users

Query translation- Translating user queries to local queries.

WDS connection- Submitting queries to WDS

WDS content analysis- Analyze content of WDS

WDS profile (content)- Meta data of WDS, including scale of a Web database, distribution of values in each attributes.

Result extraction- Getting results from web pages

Result annotation- Finishing semantic annotation of the results

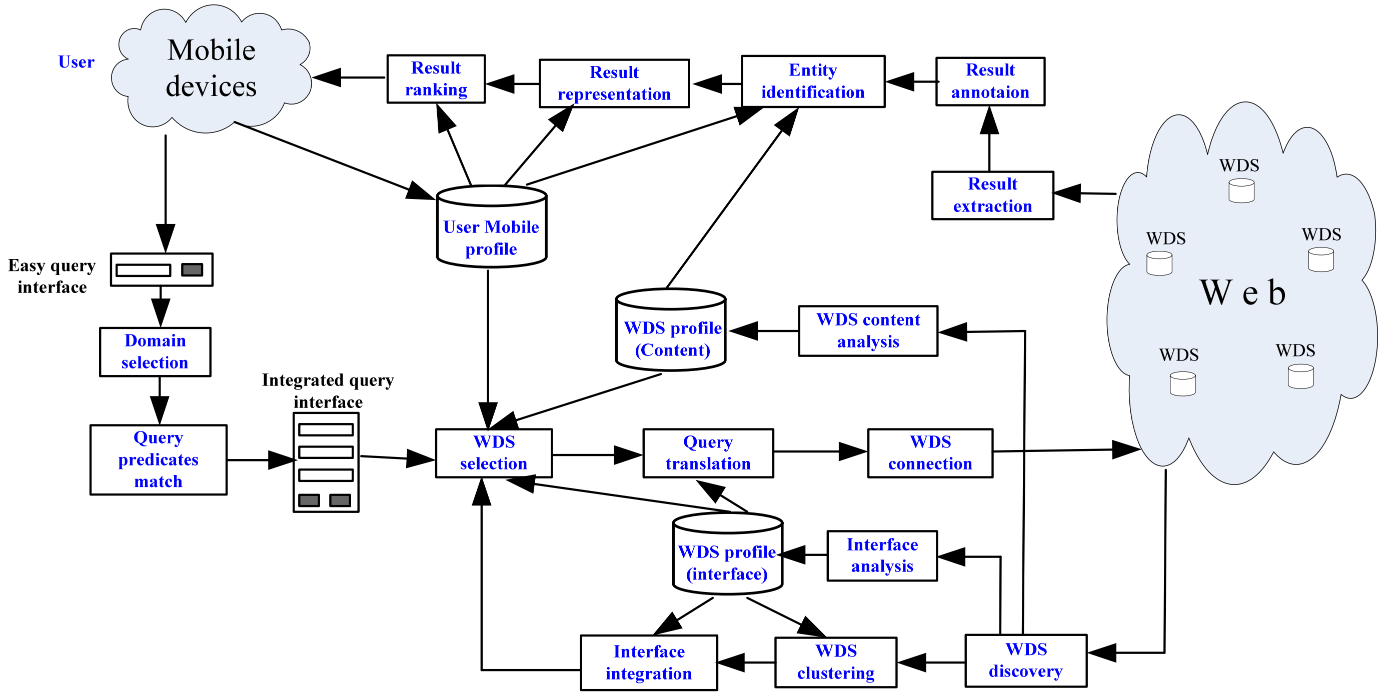

Fig.1 Main Framework

Entity identification- identifying records that are describing the same real word entity.

Result representation- Showing the results, including contents and layout of the presentation

Result ranking- Ranking the results according to the user context.

User mobile profile- information about users, including the screen size of user devices, computing resources and the location of users.

In the framework, when a mobile user input a query term to a easy query interface, the query is sent to the search domain selection part, which will select the most related domain according to the users' query terms. Then the query is sent to a integrated query interface. From the integrated interface, the query is sent to a traditional deep web data integration steps, such as ,WDS selection ,query translation and so on. What is special for mobile users is that, before the process of deep web data integration issues, the user mobile profile is introduced. As we showed above, the user mobile profile is key context for mobile searching, since it stores information about the users—location of users, computing ability, screen size of user devices etc. So the information can be used in the query processing, we will introduce it in detail in next section. After processing of the query, the result extraction will extract results from various web pages and send the collected results to the result annotation. The results are combined with some semantic meanings in this part. After entity identification, result representation and result ranking, the final results is sent to the users. In order to display more information in a small screen in a mobile device, it is always useful to have a result cluster in the system. The result cluster will cluster the results in hierarchy, each hierarchy is about a same topic, this is not yet included in the architecture.

Following are some topics on the frameworks above. Not every module of the framework is discussed in detail, since some of the modules are already been maturely researched. Topics from 3.1 to 3.4 are mainly concerned on web data integration issues, and the following four topics focused more on mobile issues.

3.1 Web database selection

(a) WDB selection based on content

When a query is submitted to an integrated interface, it needs to be passed to the Web data sources (WDSs) represented by the integrated interface. If the number of WDSs for this integrated interface is small, the query can be passed to all of them. However, if the number is large, it may be inefficient to invoke these WDSs for each query. Metasearch engines involving text search engines, only scattered work has been reported when deep web WDSs with structured data are involved. Metasearch engines involves only text documents and the representative of each search engine contains terms and some statistics for each term. In contrast, WDBs involve three types of attributes, i.e., textual, categorical and numerical attributes. Categorical attributes usually have a small number of distinct values and they are usually implemented as a selection list or a group of checkboxes or radio buttons on a search interface. The former can be considered as a special case of the latter when the structured data have just one textual attribute. The representative for each type of attribute may be different. We aimed at when the query interface has various attributes or some attributes is missed, it is necessary to find representations of WDBs and find useful methods based on this.

(2) WDB selection based on the location of mobile users

When it comes to mobile users, the location information is an important information. Ranking WDBs according to users' location can help users to get what they want.

(3) Service area identification of WDBs

We need to identify the service areas of a WDB. In order to match a mobile user's location with the information provided by a WDB, it is desirable to find out the intended service areas of WDBs. Some WDBs have very narrow service areas, some have multiple service areas, and some even have national or international coverage. We plan to study how to identify the service area of specialized local WDBs.

3.2 Entity identification across multiple deep web data sources

Entity identification is to determine if two or more records retrieved from different data sources actually correspond to the same real world entity. This is critical in several application scenarios in deep Web data integration. For example, in comparison shopping, it makes sense to compare the prices of two product records only if the two product records correspond to the same one. A general method for determining whether two records R1 and R2 are matched consists of two steps. First, values in corresponding attributes from R1 and R2 are matched. Specifically, for each attribute A, a similarity between R1[A] and R2[A] is computed. Second, the similarities between value pairs under all attributes are aggregated to determine whether R1 and R2 are matched.

Many researches have been done on entity identification, here, there are some initial thoughts on new method of doing this. For attribute value matching, we plan to develop a library of domain specific string matching functions.

3.3 Geo information on web pages

As we have analyzed above, mobile users is always “on-the-go” when they access to the web, another important feature is that mobile users search for location based information much more frequently than desktop users. One interesting issue is the problem of associating an address to each result or web page. Many web pages are associated with an organization or a unit of an organization. As a result, the address of the organization or the unit, whichever is more directly related to the page, can be considered as the address of the page, or the geo information of the web page, when the page itself does not contain an address, we can check if there are other implicit information which may contain geo information. We are interested in determining what address each page should be associated to. We can extract and index location information embedded in these resources, so it is easy for mobile users to receive the right location information from the web pages.

3.4 Search result extraction wrapper generation and maintenance

After the query is evaluated, the retrieved search result records are embedded in dynamically generated response pages. Specifically, there are two tasks – one is result extraction which is to extract the SRRs from the response pages and the other is result annotation which is to assign semantic meanings to the data units/instances within each SRR. The second task in turn consists of two subtasks, the first one is data alignment which aligns/groups data units from different SRRa on the same result page according to their semantics and the second one is data annotation which assigns a semantic label to each group of data units. As different search engines usually organizes and displays their SRRs differently and the SRRs returned by different search engines, even from the same domain, often consist of different types of information, different result extraction, data alignment and data annotation rules are needed for different search engines. Because millions of search engines are present on the Web and they frequently change their result display formats, highly automated solutions are needed to generate and maintain these wrappers.

For result extraction, we plan to carry out research in two directions. The first is to improve visual-feature based solution so that response pages where SRRs are organized into multiple columns and multiple sections can be handled accurately and the time needed to perform the extraction can be significantly reduced. The second is to combine visual-features and non-visual-features in a way that can maximize their contributions to accurate result extraction. For data alignment and data annotation, we plan to find solutions for the problems caused by attributes with multiple values or nearly identical values. We also plan to create a library of patterns for some common values such as email, telephone number, address, etc.

3.5 Location sensitive retrieval

For mobile search, a relevant result must match the user query by content and is close to the user’s location. We plan to take WDS’s service areas into consideration when performing search engine selection.

Furthermore, for results returned from local WDSs, we will try to identify the address associated with each result and perform location-sensitive result merging. A mobile user is more likely to prefer products/services close to his current location. The locations of mobile users can be determined by the mobile service provider when the mobile devices are in use.

A typical scenario is like this, suppose a mobile user is searching for information of the nearest restaurant, he not only needs the way to the restaurant or how the reach the nearest restaurant, he also wants more information about the restaurant, such as, services, prices or other guests' opinions. In this situation, traditional location servers or web servers can not provide services like this, so it is our motivation to do the research on web data integration for mobile users.

3.6 Search result clustering

As many analyses shows, mobile users tend not to "click" the search results, because it is not convenient for mobile users to "click into" a result, also, since the limitation of navigation systems in mobile devices, users do not always concern the search results which have bad ranking results. So it is important to improve the search result clustering methods.

Some researches have been done on search result clustering. In [4] they proposed a method to tackle the problem of mobile search using search result clustering, which consists of organizing the results obtained in response to a query into a hierarchy of labeled clusters that reflect the different components of the query topic.

By clustering the results, one single "page" on mobile device can display more information, it can improve users' experience in mobile search.

3.7 Concise snippet generation

Mobile devices usually have a small display screen, limiting the amount of information that can be displayed. Many investigation shows that, mobile users do not always "click" the search results, since the communication networks is slower than desktop users. So if shorter snippets for search results can be generated, then more results can be displayed on each screen, leading to better user experience. In this project, we are interested in reducing the size of the snippet returned from a WDS without compromising the effectiveness of the snippet in helping the user determine the usefulness of the result.

3.8 Result representation

Mobile devices have much more different user interface than desktop devices. Designing an effective mobile search users interface is challenging, as interacting with the results is often complicated by the lack of available screen space and limited interaction methods. In [7], the author proposed a method which can automatically compute categories to present the user with an overview of the result set.

4 Conclusions

In this paper, we figured out some research point on deep web data integration for mobile users. We are more concerned on the Geo information extraction on the web pages and the location sensitive retrieval during the process. The availability of location-driven data, location-enabled devices, and location application is guaranteed to expand the opportunities that exist in the combination of mobile users and the web. We proposed an initial framework based on the metasearch method, it will be optimized and extended in the future in order to deal the problems of location sensitive processes which is more concerned nowadays.

References

[1] Kamvar, M., Kellar, M., Patel, R., and Xu, Y. 2009. Computers and iphonesand mobile phones, oh my!: a logs-based comparison of search users on different devices. In Proceedings of the 18th international Conference on World Wide Web (Madrid, Spain, April 20 -24, 2009). WWW '09.

[2] Kamvar, M. and Baluja, S. 2008. Query suggestions for

mobile search: understanding usage patterns. In Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy, April 05 -10, 2008). CHI '08. ACM, New York, NY, 1013-1016.

[3] Maryam Kamvar, Shumeet Baluja, Deciphering Trends in Mobile Search, Computer, v.40 n.8, p.58-62, August 2007

[4] Claudio Carpineto, Sefano Mizzaro, Mobile Information Retrieval with Search Results Clustering: Prototypes and Evaluations, ASIS 2008

[5] Church, K. and Smyth, B. 2009. Understanding the intent behind mobile information needs. In Proceedings of the 13th international Conference on intelligent User interfaces (Sanibel Island, Florida, USA, February 08 -11, 2009). IUI '09. ACM, New York, NY, 247-256.

[6] Christopher Jones, Christopher Jones, Location based Advertising, M-bussiness 2002

[7] Heimonen, T. and Käki, M. 2007. Mobile findex: supporting mobile web search with automatic result categories. In Proceedings of the 9th international Conference on Human Computer interaction with Mobile Devices and Services (Singapore, September 09 -12, 2007). MobileHCI'07, vol. 309. ACM, New York, NY, 397-404.

[8] Kamvar, M. and Baluja, S. 2008. Query suggestions for mobile search: understanding usage patterns. In Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy, April 05 -10, 2008). CHI '08. ACM, New York, NY, 1013-1016.

[9] W. Liu, X. Meng, and W. Meng. A Survey of Deep Web Data Integration. Chinese Journal of Computers, Vol.30, No.9, pp.1475-1489, September 2007.

[10] Y. Lu, H. He, Q. Peng, W. Meng, and C. Yu. Clustering E-Commerce Search Engines based on their Search Interface Pages using WISE-Cluster . Data & Knowledge Engineering (DKE) Journal, Vol.59, No.2, pp.231-246, November 2006.

[11] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu. Annotating Structured Data of the Deep Web. IEEE 23rd International Conference on Data Engineering (ICDE), 2007.