INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2386 - 2396
RECEIVED : 25.10.2017 PUBLISHED : 09.11.2017
Nov 09, 2017

PERSONALIZED WEBSEARCH IN DATAMINING

[1]Mrs. D. Ananthanayaki., M.C.A., M.Phil.,
Assistant Professor Department Of Computer Science, Selvamm Arts and Science College (Autonomous), Namakkal.
[2]S.B.Ajai Vignesh,
M. Phil Scholar, Department Of Computer Science, Selvamm Arts and Science College (Autonomous), Namakkal.

ABSTRACT

World wide web(WWW) is very popular and commonly used internet's information retrieval service. Now a days commonly used task on internet is web search. User get variety of related information for their queries. To provide more relevant and effective results to user, Personalization technique is used. Personalized web search refer to search information that is tailored specifically to a person's interests by incorporating information about query provided. Two general types of approaches to personalizing search results are modifying user's query and re ranking search results. Several personalized web search techniques based on web contents, web link structure, browsing history ,user profiles and user queries. The proposed paper is to represent survey on various techniques of personalization.

Keywords: Re-ranking, Tailored, Survey.

1. INTRODUCTION

GENERAL BACKGROUND

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

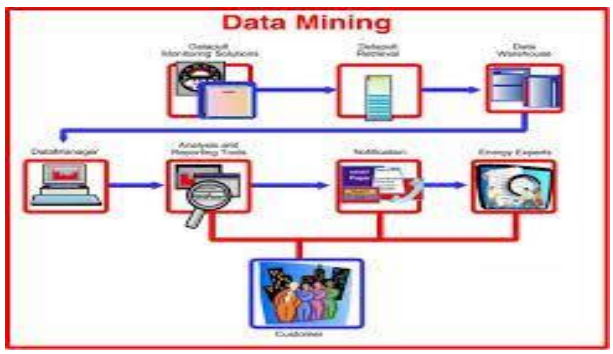


Fig 1.1 Data Mining

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2386 - 2396
RECEIVED : 25.10.2017 PUBLISHED : 09.11.2017
Nov 09, 2017

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

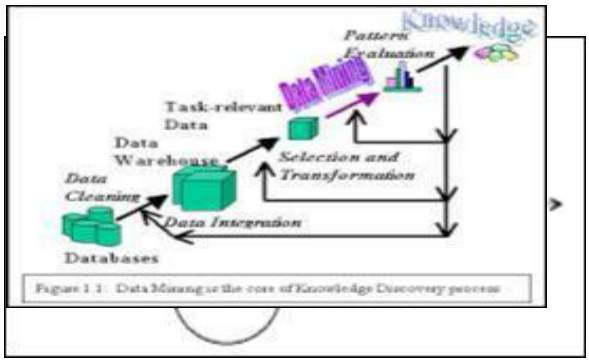


Fig 1.2 Data Mining Process

SCOPE OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database - for example, finding linked products in gigabytes of store scanner data - and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly.

Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions. More columns. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2386 - 2396
RECEIVED : 25.10.2017 PUBLISHED : 09.11.2017
Nov 09, 2017

The most commonly used techniques in data mining are:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1**). Sometimes called the k-nearest neighbor technique.**

Rule induction: The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms.

WEB MINING

Web mining - is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

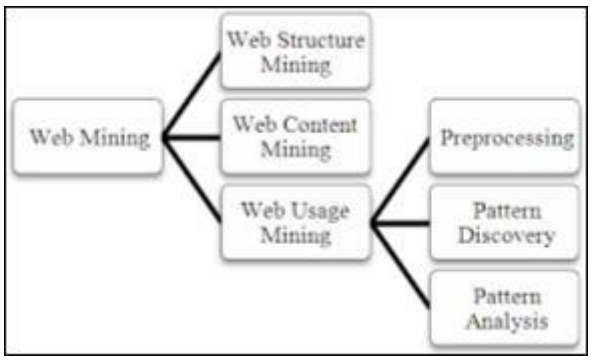
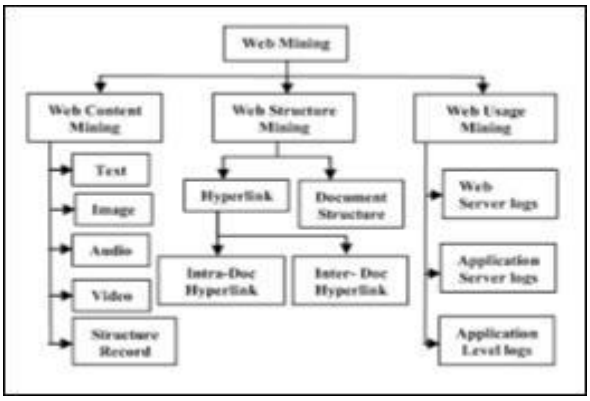


Fig 1.3 Web Mining

Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time. Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2386 - 2396
RECEIVED : 25.10.2017 PUBLISHED : 09.11.2017
Nov 09, 2017

Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.



Web Usage Mining

Web usage mining by itself does not create issues, but this technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. The obtained data will be analyzed, and clustered to form profiles;

The data will be made anonymous before clustering so that there are no personal profiles. Thus these applications de-individualize the users by judging them by their mouse clicks. De-individualization, can be defined as a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits. Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user's interests. This process could result in denial of service or a privilege to an individual based on his race, religion or sexual orientation, right now this situation can be avoided by the high ethical standards maintained by the data mining company. The collected data is being made anonymous so that obtained data and the obtained patterns cannot be traced back to an individual. It might look as if this poses no threat to one's privacy, however additional information can be inferred by the application by combining two separate unscrupulous data from the user.

Web Structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds: Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2386 - 2396
RECEIVED : 25.10.2017 PUBLISHED : 09.11.2017
Nov 09, 2017

the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB, Meta Crawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site to transform a web site to become a database. Web mining is an important component of content pipeline for web portals. It is used in data confirmation and validity verification, data integrity and building taxonomies, content management, content generation and opinion mining.

Web Mining versus Data Mining

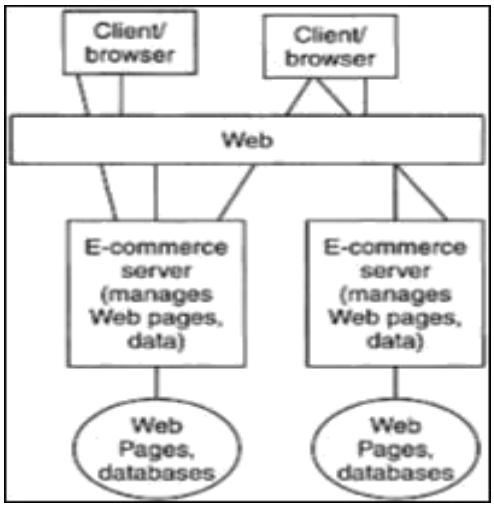
When comparing web mining with traditional data mining, there are three main differences to consider:

Scale – In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.

Access – When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated (non-user) access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler (a) promises not to overload the site, and (b) has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful/polite during the crawling process, to avoid causing any problems for the webmaster.

Structure – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

WEB SECURITY



Web usage mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled e-commerce to do personalized marketing, which eventually results in higher trade volumes. Government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of mining applications can benefit society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

TECHNIQUE USED IN THESIS – RANKING METHODOLOGY

Ranking of query results is one of the fundamental problems in information retrieval (IR), the scientific/engineering discipline behind search engines. Given a query q and a collection D of documents that match the query, the problem is to rank, that is, sort, the documents in D according to some criterion so that the "best" results appear early in the result list displayed to the user. Classically, ranking criteria are phrased in terms of relevance of documents with respect to an information need expressed in the query.

Ranking is often reduced to the computation of numeric scores on query/document pairs; a baseline score function for this purpose is the cosine similarity between term frequency–inverse document frequency vectors representing the query and the document in a vector space model, probabilities in a probabilistic Information Retrieval model. A ranking can then be computed by sorting documents by descending score.

OBJECTIVE OF THESIS WORK

The main objective of the proposed methodology is a privacy-preserving personalized web search framework User customizable Privacy-preserving Search, which can generalize profiles for each query according to user-specified privacy requirements.

The proposed method aims to provide an inexpensive mechanism for the client to decide whether to personalize a query in user customizable privacy-preserving search. This decision can be made before each runtime profiling to enhance the stability of the search results while avoid the unnecessary exposure of the profile. And also the proposed work aims at providing protection against a typical model of privacy attack, namely eavesdropping. The proposed model is knowledge and Session bounded ie., The background knowledge of the adversary is limited to the taxonomy repository R. And both the profile H and privacy are defined based on R. and none of previously captured information is available for tracing the same victim in a long duration. In the other hand, the eavesdropping will be started and ended within a single query session.

PROBLEM DEFINITION

The first step in the software development life cycle is the identification of the problem. As the success of the system depends largely on how accurately a problem is identified. The project is protect user privacy in profile-based PWS, researchers have to consider two contradicting effects during the search process. On the one hand, they attempt to improve the search quality with the

personalization utility of the user profile. On the other hand, they need to hide the privacy contents existing in the user profile to place the privacy risk under control. A few previous studies suggest those users are willing to compromise privacy if the personalization by supplying user profile to the search engine yields better search quality.

SCOPE OF THESIS WORK

The overview of the thesis will results in the finding search query results that is system will contains personalized web search can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods are straightforward they simply impose bias to clicked pages in the user's query history. Although his strategy has been demonstrated to perform consistently and considerably it can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques Profile-based methods can be potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances.

LITERATURE REVIEW

Xindong Wu, Xingquan Zhu et al [1] stated Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Yuefeng Li, Algarni A et al [2] describes it is a big challenge to guarantee the quality of discovered relevance features in text documents for describing user preferences because of large scale terms and data patterns. Most existing popular text mining and classification methods have adopted term-based approaches. However, they have all suffered from the problems of polysemy and synonymy. Over the years, there has been often held the hypothesis that pattern-based methods should perform better than term-based ones in describing user preferences; yet, how to effectively use large scale patterns.

Tseng, V.S, Cheng-Wei Wu et al [3] stated Mining high utility itemsets (HUIs) from databases is an important data mining task, which refers to the discovery of itemsets with high utilities (e.g. high profits). However, it may present too many HUIs to users, which also degrades the efficiency of the mining process. To achieve high efficiency for the mining task and provide a concise mining result to users, we propose a novel framework in this paper for mining closed+ high utility itemsets (CHUIs), which serves as a compact and lossless representation of HUIs.

METHODOLOGY

Previous works on profile-based PWS mainly focus on improving the search utility. The basic idea of these works is to tailor the search results by referring to, often implicitly, a user profile that reveals an individual information goal. In the remainder of this section, we review the previous solutions to PWS on two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 6 (2017) PAGES 2386 - 2396
RECEIVED : 25.10.2017 PUBLISHED : 09.11.2017
Nov 09, 2017

PRIVACY PROTECTION IN PWS SYSTEM

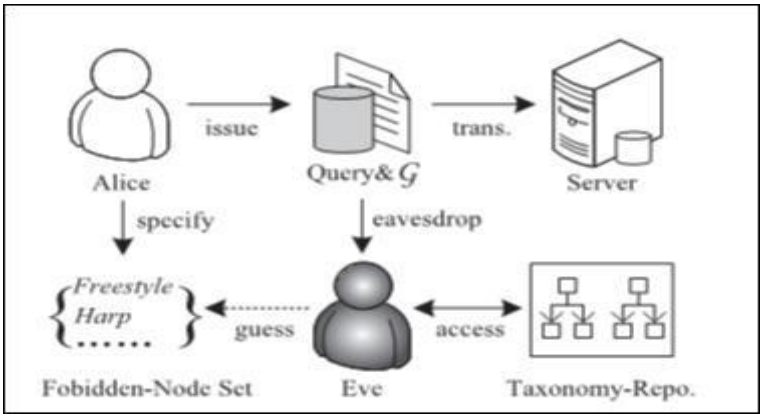
Generally there are two classes of privacy protection problems for PWS. One class includes those treat privacy as the identification of an individual, as described in the other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server. Typical works in the literature of protecting user identifications (class one) try to solve the privacy problem on different levels, including the pseudo identity, the group identity, no identity, and no personal information. Solution to the first level is proved to fragile [11].

USER PROFILE

Consistent with many previous works in personalized web services, each user profile in UPS adopts a hierarchical structure. Moreover, our profile is constructed based on the availability of a public accessible taxonomy, denoted as R, which satisfies the following assumption. The repository R is a huge topic hierarchy covering the entire topic domain of human knowledge. That is, given any human recognizable topic t, a corresponding node (also referred to as t) can be found in R, with the sub-tree subtre (t, R) as the taxonomy accompanying.

ATTACK MODEL

The proposed work aims at providing protection against a typical model of privacy attack, namely eavesdropping. As shown in Fig. 3, to corrupt Alice's privacy, the eavesdropper Even successfully intercepts the communication between Alice and the PWS-server via some measures, such as man-in-the-middle attack, invading the server, and so on. Consequently, whenever Alice issues a query q, the entire copy of q together with a runtime profile G will be captured by Eve.



Knowledge bounded. The background knowledge of the adversary is limited to the taxonomy repository R. Both the profile H and privacy are defined based on R. Session bounded. None of previously captured information is available for tracing the same victim in a long duration. In other words, the eavesdropping will be started and ended within a single query session.

DRAWBACKS OF EXISTING METHODOLOGY

The existing system works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The online phase handles queries as follows:

When a user issues a query qi on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile Gi satisfying the privacy requirements. The generalization process is guided by considering two conflicting metrics, namely the

personalization utility and the privacy risk, both defined for user profiles. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search. The search results are personalized with the profile and delivered back to the query proxy.

- No capability to capture a series of queries.
- User profile is categorized into single node in the tree structure only.
- Past query based suggestion is not given to user.

EXPERIMENTAL RESULTS AND DISCUSSION

IMPLEMENTATION SOFTWARE

The empirical system is designed and implemented by using the Microsoft visual studio .net as a front end tool. And the coding language used is C#.net. Microsoft SQL Server used as a back end tool. Visual studio is an integrated development environment which is used in this thesis for designing the thesis experiments.

Features of C#.Net

C# is a Microsoft‘s new language designed for its new platform ".NET". It is fully object oriented language like java and is the first component-oriented language. Because it contains integral supports for writing the software components. C# is designed for building robust, reliable and durable components to handle real world application. The C# language specification stated the objectives and features of C#:

- It is simple, modern, general purpose and object oriented programming language.
- This provides a support for the software analysis principles such as strong type checking, array bounds checking, detection of attempts to use uninitialized variables and automatic garbage collection.
- It is useful for developing software components which are suitable for deployment in the distributed environments. This supports internationalization.

Characteristics of C#

- Garbage Collection**:** the memory management feature leads all managed objects. Garbage collection is a feature .NET. The C# uses it during the runtime.
- Indexes: C# has indexes which help to access value in a class with an array like syntax programs.
- Exception Handling: .NET standardizes the exception handling across languages. C# offers the conditional keyword to control the flow and make the code more readable.

Business Intelligence Development Studio

The most important parts of the BI Dev Studio are:

- **Solution Explorer:** Solution Explorer is where managing y solution and projects. All objects are created and managed in this window

- **Window tabs:** The Window tabs allow to quickly switching between designer windows. A tab will be displayed for each object or file that is currently open.
- **Designer window:** The Designer window is where edit and analyze your objects. Creating a new object or double-clicking on an object in Solution Explorer will open that object's specific designer, allowing modifying and interacting with the object.
- **Designer tabs:** Many objects have different aspects that can be edited or interacted with. These aspects are indicated by tabs within the Designer window.
- **Properties window:** The Properties window is a context-sensitive window that displays properties for the currently selected item. This is a general concept in Visual Studio and applies to any type of operation performed within the studio.

- **BI menus:** The area on the main menu bar between the Debug menu and the Tools menu is where will find context-sensitive menus specific to Analysis Services objects.

CONCLUSION

The proposed system is a client-side privacy protection framework called UPS for personalized web search. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. It proposed a greedy algorithm, namely GreedyIL, for the online generalization. The experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency of our solution. The main benefits are capability to capture a series of queries, User profile is categorized into multiple nodes in the tree structure and past query based suggestion is given to user.

This study proposed the different approaches that have been implemented for personalizing web search. There is tremendous growth in the approaches taken to represent, construct and employ user profiles. These enabling techniques are key to providing user with accurate, personalized information services. As personalized search has different effectiveness for different kinds of queries, we believed that queries should not be handled in same manner with regards to personalization. It also provided cross re-ranking algorithm for online generalization.

In this proposed system, building the user profile hierarchically with user interest if user specifies sensitivity for any topic then that are not allowed appearing in generalized user profile. After submitting query q, we retrieve the documents similar to query using conventional approach. These documents are then grouped together. The relevance method used in this framework is simple and fast to evaluate and will also check users' last searches to get the relevant query meaning.

FUTURE ENHANCEMENT

At present, the project presented a client-side privacy protection framework called UPS for personalized web search. For future work, the thesis will try to resist adversaries with broader background knowledge, such as richer relationship among topics (e.g., exclusiveness, sequentially, and so on), or capability to capture a series of queries (relaxing the second constraint of the adversary)

from the victim. It will also seek more sophisticated method to build the user profile, and better metrics to predict the performance (especially the utility) of UPS.

Further propose specific techniques to capture and exploit two types of implicit feedback information: Identifying related immediately preceding query and using the query and the corresponding search results to select appropriate terms to expand the current query, and exploiting the viewed document summaries to immediately re-rank any documents that have not yet been seen by the user. Using these techniques, develop a client-side web search agent on top of a popular search engine. Experiments on web search show that our search agent can improve search accuracy.

REFERENCES

1) Baeza-Yates.R, and Ribeiro-Neto.B, "Modern Information Retrieval". Addison-Wesley, June 1999.
2) Cristianini N. and Shawe-Taylor J., An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
3) Fergus.R, Fei-Fei.L, Perona.P, and Zisserman.A. Learning object categories from google's image search.
4) Hand D., Mannila H. and Smyth P., Principles of Data Mining, MIT Press, 2001.
5) Strachey, Christopher (June 1959). "Time Sharing in Large Fast Computers". Proceedings of the International Conference on Information processing, UNESCO. paper B.2.19: 336–341..
6) Larose D.T., Discovering knowledge in data: an introduction to data mining, Wiley-Interscience, 2005.
7) Mitchell T.M., Machine learning, McGraw-Hill, 1997.
8) Pal S.K. and Mitra P., Pattern Recognition Algorithms for Data Mining, CRC Press, 2004.
9) Pankaj Jalole ,"An Integral approach to software engineering",Narosa publishing Home-3rd Edition.
10) Smith, David Mitchell. "Hype Cycle for Cloud Computing", 2013Gartner. Retrieved 3 July 201