International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Supporting Privacy Protection in Personalized Web Search with Secured User Profile

Archana Ukande[1], Nitin Shivale[2]

[1]ME CSE- Final Year, BSIOTR Wagholi (Pune), Maharashtra, India

[2]Assistant Professor, CSE Department., BSIOTR Wagholi (Pune), Maharashtra, India

Abstract: ***Web search engines (e.g. Google, Yahoo, Microsoft Live Search, etc.) are widely used to find certain data among a huge amount of information in a minimal amount of time. These useful tools also pose a privacy threat to the users. Web search engines profile their users on the basis of past searches submitted by them. In the proposed system, we can implement the String Similarity Match Algorithm (SSM Algorithm) for improving the better search quality results. To address this privacy threat, current solutions propose new mechanisms that introduce a high cost in terms of computation and communication. Personalized search is promising way to improve the accuracy of web search. However, effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. Indeed, these concerns have become one of the main barriers for deploying personalized search applications, and how to do privacy-preserving personalization is a great challenge. In this we propose and try to resist adversaries with broader background knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we can hide the user search results. By using this mechanism, we can achieve the privacy.***

Keywords: Privacy protection, personalized web search, utility, risk, profile

1. Introduction

In this paper we present a novel protocol specially designed to protect the users' privacy in front of web search profiling. In this we propose and try to resist adversaries with broader background knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we can hide the user search results. In the Existing System, Greedy IL and Greedy DP algorithm, it takes large computational and communication time.

2. Introduction

In this paper we present a novel protocol specially designed to protect the users' privacy in front of web search profiling. In this we propose and try to resist adversaries with broader background knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we can hide the user search results. In the Existing System, Greedy IL and Greedy DP algorithm, it takes large computational and communication time.

For generalize the retrieved data by using the background knowledge. Through this we can resist the adversaries. Privacy protection in publishing transaction data is an important problem. A key feature of transaction data is the extreme sparsity, which renders any single technique ineffective in anonymizing such data. Among recent works, some incur high information loss, some result in data hard to interpret, and some suffer from performance drawbacks. This paper proposes to integrate generalization and compression to reduce information loss. However, the integration is non-trivial. We propose novel techniques to address the efficiency and scalability challenges.

2.1 Problem Statement

In the Existing System, they presented 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. They proposed two greedy algorithms namely GreedyDP and GreedyIL, for the online generalization. It achieves quality search results while preserving user's customized privacy requirements. It also improves effectiveness and efficiency. But in the Existing system, it uses only the generalization concept. It degrades the performance of existing system. For this we are going implement and extend the process by using some other properties such as exclusiveness and to make a system capable to capture a series of queries. In the Existing System, it has a high cost in terms of computation and communication. Existing System have three system architectures. In these three components has been used. There are server, client and proxy. Client information's are shared to the proxy. In the proposed system, information's has exclusiveness. It cannot be shared to the privacy. When the searched information's are generalized and then only information's are stored in the history. Only hided information's are stored into the history. String Similarity Match Algorithm (SSM Algorithm) is better than the greedy algorithm. It achieves more accuracy in search results.

2.2 Definition of terms

2.2.1Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Generally, 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 tools for analyzing data. It allows users to analyze data categorize it, and summarize the relationships identified. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

2.2.2What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

2.2.3 How data mining work?

Data mining provides the link between transaction and analytical systems,. Data mining software analyses 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. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

2.3 Data mining consists of five major elements:

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

2.4 Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **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) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **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.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

3. System Architecture

Existing System:
Algorithm Used—Greedy Information Loss Algorithm (Greedy IL)
In the Existing System, each user has to undertake the following procedures.
1. Offline profile construction,
2. Offline privacy requirement customization,
3. Online query-topic mapping, and
4. Online generalization.

Normally, user posts the query and retrieves the information from the server. In several systems, information is loosed due to the algorithm inefficiency. In this, Greedy IL algorithm minimizes the information loss during retrieving the information's. The advantage of GreedyIL over GreedyDP is more obvious in terms of response time. This is because GreedyDP requires much more computation of DP, which incurs lots of logarithmic operations. The problem worsens as the query becomes more ambiguous. For instance, the average time to process GreedyDP for queries in the ambiguous group is more than 7 seconds. In contrast, GreedyIL incurs a much smaller real-time cost, and outperforms GreedyDP by two orders of magnitude. GreedyIL displays near-linear scalability, and significantly outperforms GreedyDP.

3.1. Algorithms for Proposed System

Step1:Detecting & removal of unwanted symbols
Step2: compute similarity calculation for user given word and word in database
Step3: In that similarity calculation, extract the features in the dataset.
Step4: Then estimate the ASCII difference for user given word and words in database
SteP5: The estimate the similarity values.
Step6:Then retrieve the most relevant documents based on the similar values

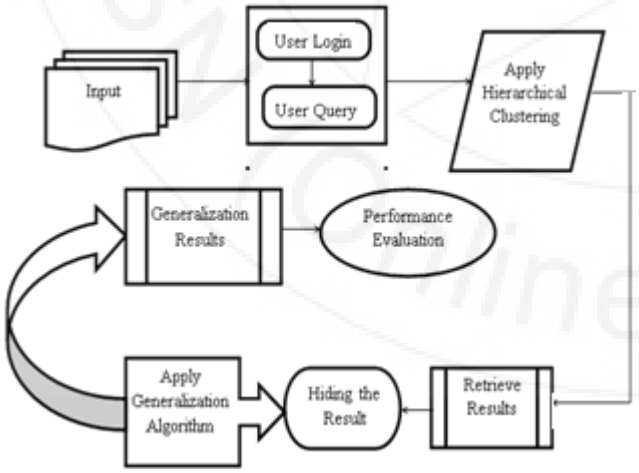

Figure 1: System Architecture

4. Existing System

In the Existing Work, a client-side privacy protection framework called UPS for personalized web search was proposed. UPS could theoretically be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The context allowed users to stipulate 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. In this they proposed two greedy algorithms, namely GreedyDP and GreedyIL, for the online generalization. In this for query mapping process it has various steps to compute the relevant items.

Most works on anonymization focus on relational data where every record has the same number of sensitive attributes. There are a few works taking the first step towards anonymizing set-valued or transactional data where sensitive items or values are not clearly defined. While they could be potentially applied to user profiles, one main limitation is that they either assume a predefined set of sensitive items that need to be protected, which are hard to done in the web context in practice, or only guarantee the anonymity of a user but do not prevent the linking attack between a user and a potentially sensitive item.

Another approach to provide privacy in web searches is the use of a general purpose anonymous web browsing mechanism. Simple mechanisms to achieve a certain level of anonymity in web browsing include: (i) the use of proxies; or (ii) the use of dynamic IP addresses.

4.1 Disadvantages

It has demonstrated the ineffectiveness or privacy risks of naive anonymization schemes. The utility of the data is limited to statistical information and it is not clear how it can be used for personalized web search. For retrieving the user query results, it takes high computational and communication time and also cost. Proxies do not solve the privacy problem. This solution only moves the privacy threat from the web search engine to the proxies themselves. A proxy will prevent the web search engine from profiling the users, but the proxy will be able to profile them instead. The renewal policy of the dynamic IP address is not controlled by the user but the network operator.

5. Proposed System

Web search engines (e.g. Google, Yahoo, Microsoft Live Search, etc.) are widely used to find certain data among a huge amount of information in a minimal amount of time. However, these useful tools also pose a privacy threat to the users: web search engines profile their users by storing and analyzing past searches submitted by them. In the proposed system, we can implement the clustering algorithms for improving the better search quality results. It is retrieved by using the String Similarity Match Algorithm (SSM Algorithm) algorithm. To address this privacy threat, current solutions propose new mechanisms that introduce a low cost in terms of computation and communication. In this paper we present a novel protocol specially designed to protect the users' privacy in front of web search profiling. In this we propose and try to resist adversaries with broader background

knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we can hide the user search results. In the Existing System, Greedy IL and Greedy DP algorithm, it takes large computational and communication time.

Advantages

- It achieves better search results.
- It achieves the privacy results when applying the background knowledge to the user profiling results.
- It has less computational time and communicational time.
- It achieves better accuracy when compared with the Existing Works.

6. Conclusion

Privacy protection in publishing transaction data is an important problem. A key feature of transaction data is the extreme sparsity, which renders any single technique ineffective in anonymizing such data. Among recent works, some incur high information loss, some result in data hard to interpret, and some suffer from performance drawbacks. This paper proposes to integrate generalization and compression to reduce information loss. However, the integration is non-trivial. We propose novel techniques to address the efficiency and scalability challenges.

Our proposed system gives better quality results and gives more efficiency. Privacy is too good when compared with the Existing system. In the Existing System, only generalization technique is used. Our String matching algorithm gives more accuracy when compared with the Greedy IL algorithm. Generalization and suppression technique achieves better privacy when compared with the existing system.

7. Future Enhancements

In Future Work, we can implement the hierarchical divisive approach for retrieving the search results. It will gives better performance when compared with our proposed System.

References

[1] (1996). Health Insurance Portability and Accountability Act of (HIPPA) [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/ understanding/index.html
[2] P. Agouris, J. Carswell, and A. Stefanidis, ``An environment for contentbased image retrieval from large spatial databases," ISPRS J. Photogram. Remote Sens., vol. 54, no. 4, pp. 263_272, 1999.
[3] M. Atallah and K. Frikken, ``Securely outsourcing linear algebra computations," in Proc. 5th ASIACCS, 2010, pp. 48_59.
[4] M. Atallah and J. Li, ``Secure outsourcing of sequence comparisons," Int. J. Inf. Security, vol. 4, no. 4, pp. 277_287, 2005.
[5] M. Atallah, K. Pantazopoulos, J. Rice, and E. Spafford, ``Secure outsourcing of scienti_c computations," Adv. Comput., vol. 54, pp. 216_272, Feb. 2001.
[6] D. Benjamin and M. Atallah, ``Private and cheating-free outsourcing of algebraic computations," in Proc. Conf. PST, 2008, pp. 240_245.
[7] E. Candès, ``The restricted isometry property and its implications for compressed sensing," Comptes Rendus Mathematique, vol. 346, nos. 9_10, pp. 589_592, 2008.
[8] E. Candès, J. Romberg, and T. Tao, ``Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory, vol. 52, no. 2, pp. 489_509, Feb. 2006.
[9] E. Candès and T. Tao, ``Decoding by linear programming," IEEE Trans. Inf. Theory, vol. 51, no. 12, pp. 4203_4215, Dec. 2005.
[10] E. Candès and T. Tao, ``Near-optimal signal recovery from random projections: Universal encoding strategies," IEEE Trans. Inf. Theory, vol. 52, no. 12, pp. 5406_5425, Dec. 2006.
[11] E. Candès and M. Wakin, ``An introduction to compressive sampling," IEEE Signal Proc. Mag., vol. 25, no. 2, pp. 21_30, Mar. 2008.
[12] (2009). Security Guidance for Critical Areas of Focus in Cloud Computing, [Online]. Available: http://www.cloudsecurityalliance.org
[13] A. Divekar and O. Ersoy, ``Compact storage of correlated data for content based retrieval," in Proc. Asilomar Conf. Signals, Syst. Comput., 2009, pp. 109_112.
[14] D. Donoho, ``Compressed sensing," IEEE Trans. Inf. Theory, vol. 52, no. 4, pp. 1289_1306, Apr. 2006.
[15] C. Dwork, ``Differential privacy," in Proc. ICALP, 2006, pp. 1_12.
[16] C. Dwork, ``The differential privacy frontier (extended abstract)," in Proc. TCC, 2009, pp. 496□502.