GNITED MINDS
Journals

Journal of Advances in Science and Technology

Vol. VIII, Issue No. XV, November-2014, ISSN 2230-9659

SOLITUDE CONSERVING DATA MINING: EXPANSIONS AND DIRECTIONS

www.ignited.in

Journal of Advances in Science and Technology
Vol. VIII, Issue No. XV, November-2014, ISSN 2230-9659

Solitude Conserving Data Mining: Expansions and Directions

Shivali Yadav

Research Scholar, Jodhpur National University, Rajasthan

Abstract – This article first describes the Solitude concerns that arise due to data mining, especially for national security applications. Then we discuss Solitude-conserving data mining. In particular, we view the Solitude problem as a form of inference problem and introduce the notion of Solitude constraints. We also describe an approach for Solitude constraint processing and discuss its relationship to Solitude-conserving data mining. Then we give an overview of the developments on Solitude-conserving data mining that attempt to maintain Solitude and at the same time extract useful information from data mining. Finally, some directions for future research on Solitude as related to data mining are given.

Keywords: Data Mining, Solitude, Solitude Constraint, Solitude Controller

--------------------------♦-----------------------------

INTRODUCTION

There has been much interest recently on applying data mining for counter-terrorism applications (Thuraisingham, 2003a, 2003b). For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. While all of these applications of data mining can benefit humans and save lives, there is also a negative side to this technology, since it could be a threat to the Solitude of individuals. This is because data mining tools are available on the Web or otherwise and even naive users can apply these tools to extract information from the data stored in various databases and files, and consequently violate the Solitude of individuals. As we have stressed in other papers (see Thuraisingham, 2003a), to carry out effective data mining and extract useful information for counter-terrorism and national security, we need to gather all kinds of information about individuals. However, this information could be a threat to individuals' Solitude and civil liberties (Thuraisingham, 2002). Solitude is getting more attention partly because of counter-terrorism and national security. Recently we have heard a lot about national security in the media. This is mainly because people are now realizing that to handle terrorism; the government may need to collect information about individuals. This is causing a major concern with various civil liberties unions. The challenge is to carry out data mining and yet maintain Solitude. This topic is known as Solitude-conserving data mining.

REVIEW OF LITERATURE:

Solitude-conserving data mining has emerged due to the following reasons. First, there are legal requirements for protecting data. Second, there are liabilities from inadvertent disclosure of data. Third, organizations need to share information with its partners, but do not want to provide certain types of data when they do so (Vaidya, et al., 2006). The Health Insurance Portability and Accountability Act (HIPAA), enacted by the U.S. Congress in 1996, required the establishment of national standards for electronic health care transactions and provides for the security and Solitude of individually identifiable health information. The Solitude discussed in the HIPAA rules refers to information Solitude, or the prevention of disclosure of personal information (Moskop, Marco, Larkin, Geiderman, & Derse, 2005). In addition to HIPAA, there are requirements for protecting children's online Solitude. The Children's Online Solitude Protection Act (COPPA) requires that organizations that collect or maintain personal information to

1) Provide notice on the website of what information is collected, and

2) To obtain verifiable parental consent for its collection, use, or disclosure ("Children's Online Solitude Protection Act," 1998).

Preventing individual information disclosure has become increasingly important due to the number and size of data breaches during the last five years. There are countless examples of inadvertent disclosure of data. For example, in May, 2006, the Social Security numbers of about 26.5 million U.S. veterans were stolen in a random burglary from a VA employee's house where a laptop was stolen (Torres, 2007). In Edmonton, Canada, a security breach occurred where children had their medical information stolen. The medical information of 270 children was stored on a small flash drive (also known as a memory stick or thumb drive) and had been placed in an employee's purse, which subsequently was stolen. The flash drive contained children's personal health numbers, names, dates of service, and diagnoses (Unknown, 2007). As many as 200,000 credit and debit card numbers were compromised due to a security breach at TJX Companies, a Framingham, Massachusetts-based company (Abelson, 2007). This particular breach has resulted in multiple cases of fraudulent activity with the stolen numbers as well as at least one case of identity theft (Abelson, 2007).

1. SOLITUDE ENHANCED DATA MANAGEMENT SYSTEMS FOR SOLITUDE CONSTRAINT PROCESSING:

Our approach is to augment a database management system (DBMS) with a Solitude controller. Such a DBMS is called a Solitude enhanced DBMS. The Solitude controller will process the Solitude constraints. The question is, what are the components of the Solitude controller and when do the constraints get processed? We take an approach similar to the approach proposed by Thuraisingham, Ford, Collins and O'Keeffe (1993) for security constraint processing. In our approach, some Solitude constraints are processed during database design and the database is partitioned according to the Solitude levels. Then, some constraints are processed during database updates. Here, the data is entered at the appropriate Solitude levels. Because the Solitude values change dynamically, it is very difficult to change the Solitude levels of the data in the database in real time. Therefore, some constraints are processed during the query operation. Note that processing constraints in real time will be time consuming. We need some research in this area.

2. SOLITUDE CONSTRAINT PROCESSING AS A FORM OF SOLITUDE-SENSITIVE DATA MINING:

Solitude constraints and discussed the design of a data management system that processes Solitude constraints. Our research has been influenced a great deal by our prior research on the inference problem. Essentially we view the Solitude problem as a variation of the inference problem. As stated earlier, there has been a lot of research on Solitude-conserving data mining. That is, researchers are developing approaches to carry out successful data mining and at the same time ensure some level of Solitude. Next we will survey the developments on Solitude-conserving data mining. In this section we will explore the relationship between Solitude constraint processing and Solitude-conserving data mining. Essentially, we can view Solitude constraint processing as one type of Solitude conserving data mining. The data mining task here is to pose queries to the database and make associations and correlations between the data. The correlations and associations become an issue if they are sensitive, classified or private. Solitude constraints are rules that assign Solitude values to the data. Solitude constraint processing essentially prevents an adversary from mining (i.e. posing queries) and extracting associations that are sensitive or private. We take an all-or-nothing approach.

3. DATA MINING CHALLENGES:

Lee and Siau (2001) listed seven requirements and challenges associated with data mining. They noted that data mining can be a complicated and difficult process. For example, data mining must be able to handle different types of data. Data does not always exist in textual format. Multimedia data, spatial and hypertext data may also be mined, but specific mining techniques must be developed to handle those data types. Currently, most data mining techniques are designed for alphanumeric data only. Secondly, data mining algorithms must be able to handle data in an efficient and scalable manner. Data mining algorithms must be predictable regardless of the size of the dataset. Third, data mining must handle noisy or missing data within a dataset and still be able to produce an accurate representation of the data in the form of a model. There is a significant quality aspect involved and required when attempting to perform data mining activities. Next, end users must be able to perform data mining tasks without having an extensive knowledge of data mining algorithms. In other words, data mining tools should allow the user to explore the data on his or her own, without having to know exactly what he or she is looking for. In fact, much of data mining is exploratory in that the user does not necessarily know exactly what he or she is looking for. However, when data mining results are presented, they should be easily understood (Lee & Siau, 2001). The good thing is that some data mining software packages are free and easy to use. Packages such as Weka and Orange are freely available, have online tutorials, and are open source so one can develop their own algorithms if necessary. Orange is especially easy to use since it is written in Python and many companies have an IT staff person who knows Python. If organizations concentrate largely on getting the PPDM process and related policy developed and organized correctly, the technical implementation should not be prohibitively difficult. There are also commercial data mining packages, but they generally do not have Solitude-conserving tools in them (yet). The best recommendation is to run PPDM pilot studies using low-cost or free data mining tools before investing in large scale PPDM solutions.

CONCLUSION:

There are many barriers to implementing Solitude-conserving data mining (PPDM) techniques; there are almost too many issues that must be addressed prior to implementation. It also appears that there are significant policy, process, and technological issues that must be addressed. For SMEs, adoption of PPDM technologies may be prohibitively expensive. Furthermore, SMEs usually do not have on-site staff with expertise in Solitude, data mining, and Solitude-preservation techniques. Large organizations do have the capacity to engage in PPDM techniques, especially large financial institutions and medical institutions where meeting Solitude legislation requirements are extremely important.

REFERENCES:

1. Thuraisingham, B. (2003a). Web data mining: Technologies and their applications to business intelligence and counter-terrorism. Florida: CRC Press.

2. Thuraisingham, B. (2003b). Data mining for counter-terrorism.. To appear in S. Sivakumar & H. Kargupta (Eds.), Next generation data mining. AAAI Press

3. Thuraisingham, B. (2002). Data mining, national security, Solitude and civil liberties. SIGKDD Explorations, 4, 1-5..

4. Vaidya, J., Clifton, C., & Zhu, M. (2006). Solitude Conserving Data Mining. New York: Springer.

5. Moskop, J., Marco, C., Larkin, G. L., Geiderman, J., & Derse, A. (2005). From Hippocrates to HIPAA: Solitude and Confidentiality in Emergency Medicine - Part I: Conceptual, Moral, and Legal Foundations. Annals of Emergency Medicine, 45(1), 53-59.

6. Children's Online Solitude Protection Act, 15 U.S.C. 6501-6508 § 1301-1308 (1998).

7. Torres, E. (2007). Man arrested in theft of 1.8 million Social Security numbers. The Orange County Register. Retrieved from http://www.ocregister.com/news/kim-numbers-affairs-1924451-security-social#.

8. Unknown. (2007). Children's patient info stolen from Edmonton hospital. CBC News. Retrieved from http://www.cbc.ca/canada/edmonton/story/2007/11/13/glenrose-breach.html

9. Abelson, J. (2007, January 25, 2007). TJX breach snares over 200,000 cards in region, The Boston Globe. Retrieved fromhttp://www.boston.com/business/globe/articles/2007/01/25/tjx_breach_snares_over_2000 00_cards_in_region/

10. Thuraisingham, B., Ford, W., Collins, M., & O'Keeffe, J. (1993). Design and implementation of a database infernce cntroller. Data and Knowledge Engineering Journal, 11, 271-93.

11. Lee, S. J., & Siau, K. (2001). A review of data mining techniques. Industrial Management &Data Systems, 101(1), 41-46.