International Journal of Advanced Research in Engineering and Applied Sciences
ISSN: 2278-6252

SECURITY ISSUES IN WEB DATA MINING, NATIONAL SECURITY: A SURVEY

Md Nadeem Ahmed*

Abstract: *Web mining refers to the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. When used in a business context and applied to some type of personal data, it helps companies to build detailed customer profiles, and gain marketing intelligence. Web mining does, however, pose a threat to some important ethical values like privacy and individuality. Web mining makes it difficult for an individual to autonomously control the unveiling and dissemination of data about his/her private life. To study these threats, we distinguish between 'content and structure mining' and 'usage mining.' Web content and structure mining is a cause for concern when data published on the web in a certain context is mined and combined with other data for use in a totally different context. Web usage mining raises privacy concerns when web users are traced, and their actions are analyzed without their knowledge. Database mining can be defined as the process of mining for implicit, previously unknown, and potentially useful information from very large databases by efficient knowledge discovery techniques. Naturally such a process may open up new inference channels, detect new intrusion patterns, and raises new security problems. New security concern and research problems are addressed and identified. Finally a particularly well developed theory, rough set theory, is discussed and some potential applications to security problems are illustrated.*

Keywords: *ethics, individuality, KDD, privacy, web data mining*

*Research Scholar, IFTM University, India

INTRODUCTION

The World Wide Web can be seen as the largest database in the world. This huge and ever-growing amount of data is a fertile area for data mining research. Data mining is the process of extracting previously unknown information from (usually large quantities of) data, which can, in the right context, lead to knowledge. When data mining techniques are applied to web data, we speak of web-data mining or web mining. In accordance with [14], we define web mining as the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services, based on the definition of Etzioni (1996).

The important ethical issue with data mining is that, if someone is not aware that the information/ knowledge is being collected or of how it will be used, he/she has no opportunity to consent or withhold consent for its collection and use. This invisible information gathering is common on the Web. Knowledge discovered whilst mining the web could pose a threat to people, when, for instance, personal data is misused, or is used for a purpose other than the one for which it is supplied (secondary use). This same knowledge, however, can bring lots of advantages. Knowledge discovered through data mining is important for all sorts of applications involving planning and control. There are some specific benefits of web-data mining like improving the intelligence of search engines. Web-data mining can also be used for marketing intelligence by analyzing a web user's on-line behavior, and turning this information into marketing knowledge. It should be noted that ethical issues can arise from mining web data that do not involve personal data at all, such as technical data on cars, or data on different kinds of animals. This paper, however, is limited to web-data mining that does, in some way, involve personal data. We shall only look at the possible harm that can be done to people, which means that harm done to organizations, animals, or other subjects of any kind fall beyond the scope of this study. Since most web-data mining applications are currently found in the private sector, this will be our main domain of interest. So, web-data mining involving personal data will be viewed from an ethical perspective in a business context. We clearly recognize that web-data mining is a technique with a large number of good qualities and potential. Web-data mining is attractive for companies because of several reasons. For instance, to determine who might be a new customer by analyzing consumer data, government records, and any other

useful information. In the most general sense, it can contribute to the increase of profits be it by actually selling more products or services, or by minimizing costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses or on the relationship with customers. The different kinds of web data that are somehow related to customers will then be categorized and clustered to build detailed customer profiles. This not only helps companies to retain current customers by being able to provide more personalized services, but it also contributes to the search for potential customers. So, it is beyond dispute that web-data mining can be quite beneficial to businesses. To make sure that this technique will be further developed in a properly throughout way, however, we shall focus on the possible objections to it. Awareness of all the possible dangers is of great importance for a well-guided development and a well-considered application. Dangers of web data mining lie in the different ways in which privacy is threatened.

DATA MINING AS A SECURITY CONCERN

Thomas Hinke presented an overview of their NASA data mining research. This research does not have any security concerns; in fact it has just the opposite objective to provide content-based metadata (data about data) for the anticipated vast data holding of the Earth Observing System Data and Information System (EOSDIS). This content based metadata would then be used to assist scientists in finding data of interest from the EOSDIS data holdings. However, even though this project has no security focus, it provides a useful example of data mining in a scientific-data domain. The problem addressed by the UAH data mining research is that the amount of data in scientific data archives is growing. Some estimates project that projects such as EOSDIS will ingest up to one terabyte of data per day Scientists need to be able to find data of interest -- proverbial ``needle in a haystack.'' The problem of finding data is difficult since there is a lack of content-based metadata. The typical metadata available in the currently operation Version EOSDIS system is limited to satellite, sensor and date captured. All of these represent non content-based metadata. The only content-based metadata available on some data sets are browse images that provide a low resolution view of one of the channels of the data. However, this cannot be automatically processed. From our discussions with users of data at the Marshall Space Flight Center's Distributed Active Archive Center (the EOSDIS archive), there is a general lack

of agreement as to what metadata is desired. Thus, this precludes the capture of content-based metadata during data ingest.

INFERENCE PROBLEM AND DATA MINING

Data mining (DM) is an attempt to answer the long-standing question "what does all this data mean?". Such investigations are inherently an attempt to automate the "inference problem" in database security. Inference is basically the process of establishing relationships between datasets, the same objective as data mining. That is, given that certain attributes apply to a set of data, we "know" that certain other attributes also apply to that set of data. This is equivalent to stating that one set "implies" the other. Now, in a multi-level secure (MLS) database, we do not want Low-classified data to infer High-classified data. Data mining processes cannot be used to compromise such rules, of course. This is because each DM process must operate at a specified level (i.e. Low) and must have access to the High data in order to "discover" the rule. However, such Low-to-High rules may be "common knowledge" but unknown to the database designer. Data mining could then be used to combine Low information until the tail of the common-knowledge rule is derived. This is the process of inference. Data are put together "in a surprising way" until some common-knowledge rule, relating Low and High data, can be applied.

Fortunately, data mining can be used effectively to enforce security. The most straightforward way is to search for rules relating Low and High data. We need not be concerned with chains of N inferences, merely what conjunction of attributes for a High set may be implied by Low classified attributes for that set. The security officer doing this analysis has some advantages over an attacker, since he/she has access to both the High and Low data. In most systems, there is relatively little High data, so the number of rules relating High data to Low data is much fewer than the total number of possible rules.

DATA MINING AND SECURITY

Data mining is the process of posing a series of appropriate queries to extract information from large quantities of data in the database. Data mining techniques can be applied to handle problems in database security. On the other hand, data mining techniques can also be employed to cause security problems. This position paper reviews both aspects Data mining techniques include those based on rough sets, inductive logic programming, machine learning, and neural networks, among others. Essentially one arrives at some hypothesis,

which is the information extracted, from examples and patterns observed. These patterns are observed from posing a series of queries; each query may depend on the response obtained to the previous queries posed.

Data mining techniques have applications in intrusion detection and auditing databases. In the case of auditing, the data to be mined is the large quantity of audit data. One may apply data mining tools to detect abnormal patterns. For example, suppose an employee makes an excessive number of trips to a particular country and this fact is known by posing some queries. The next query to pose is whether the employee has associations with certain people from that country. If the answer is positive, then the employee's behavior is flagged.

While the previous example shows how data mining tools can be used to detect abnormal behavior, the next example shows how data mining tools can be applied to cause security problems. Consider a user who has the ability to apply data mining tools. This user can pose various queries and infer sensitive hypothesis. That is, the inference problem occurs via data mining. There are various ways to handle this problem. One approach is as follows. Given a database and a particular data mining tool, apply the tool to see if sensitive information can be deduced from the unclassified information legitimately obtained. If so, then there is an inference problem. There are some issues with this approach. One is that we are applying only one tool. In reality, the user may have several tools available to him. Furthermore, it is impossible to cover all ways that the inference problem could occur.

PRIVACY THREATENED BY WEB-DATA MINING

In this section, we shall point out that web-data mining, which involves the use of personal data of some kind, can lead to the disruption of some important normative values. One of the most obvious ethical objections lies in the possible violation of peoples' (informational) privacy. Protecting the privacy of users of the Internet is an important issue. Our understanding of privacy, however, is conceptually fragile. The term 'privacy' is used to refer to a wide range of social practices and domains [13]. In this article, we will not discuss the philosophical and legal discussions on privacy. Here, we will restrict ourselves with an informal (and common) definition of informational privacy. Informational privacy mainly concerns the control of information about oneself. It refers to the ability of the individual to protect information about himself. The privacy can be violated when information concerning an individual is obtained, used, or disseminated, especially if this occurs without

their knowledge There are some differences between privacy issues related to traditional information retrieval techniques, and the ones resulting from data mining. The technical distinction between data mining and traditional information retrieval techniques does have consequences for the privacy problems evolving from the application of such techniques [11]. While in traditional information retrieval techniques one has to 'talk' to a database by specifically querying for information, data mining makes it possible to 'listen' to a database (cf. Holsheimer 1999). A system of algorithms searches the database for relevant patterns by formulating thousands of hypotheses on its own. In this way, interesting patterns can be discovered in huge amounts of data. Tavani (1999a) argues that it is this very nature of data mining techniques that conflicts with some of the current privacy guidelines as formulated by the OECD.

ARGUMENTS IN DEFENCE OF WEB-DATA MINING

All the benefits obviously show that web-data mining is a highly valuable technique, which is being developed and applied on a large and growing scale. However, the threats to some important values tend to be rather serious, and will create tension in the web data mining field. Unfortunately, many professionals applying web-data mining in a business context do not foresee any moral dangers in web-data mining. To gain some insight into current web-data mining practices and the attitude of web data miners to the ethical issues involved, twenty of these professionals were interviewed. These interviews combined with a literature study teach us that people prefer to focus on the advantages of web-data mining instead of discussing the possible dangers. Moreover, they revealed several different arguments to support the view that web-data mining does not really pose a threat to privacy and related values. The arguments given in their defense of almost unlimited use of data mining can be sorted into six arguments, and are enlightening. We shall discuss these arguments briefly to show that these arguments do not justify unlimited use of data mining.

POSSIBLE SOLUTIONS

There are means to solve some problems with respect to privacy in the ethical context of web-data mining. We can distinguish solutions at an individual and at a collective level. With solutions at an *individual level*, we refer to actions an individual can take in order to protect himself/herself against possible harms. For example, using privacy enhancing technologies (PETs), being cautious when providing (personal) information on-line, and checking privacy

policies on web sites. The solutions at a collective level refer to things that could be done by society (government, businesses, or other organizations) to prevent web-data mining from causing any harm. For example, further development of PETs, publishing privacy policies, web quality seals, monitoring web mining activities, legal measures, creating awareness amongst web users and web data miners, and debating the use of profiling. A mixture of technical and non-technical solutions at both the individual and the collective level is probably required to even begin solving some of the problems presented here. But, to what extent can the problems really be solved in both web-data mining categories

REFERENCES

1. Database Security IX Status and Prospects Edited by D. L. Spooner, S. A. Demurjian and J. E. Dobson ISBN 0 412 72920 2, 1996, pp. 391-399.
2. Pawlak, Z. (1990). Rough sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1992.
3. Lin, T. Y. (1994), "Anamoly Detection -- A Soft Computing Approach", *Proceedings in the ACM SIGSAC New Security Paradigm Workshop*, Aug 3-5, 1994, 44-53. This paper reappeared in the Proceedings of 1994 National Computer Security Center Conference under the title "Fuzzy Patterns in data".
4. Lin, T. Y. (1993), "Rough Patterns in Data-Rough Sets and Intrusion Detection Systems", *Journal of Foundation of Computer Science and Decision Support*, Vol.18, No. 3-4, 1993. pp. 225- 241. The extended version of "Patterns in Data-Rough Sets and Foundation of Intrusion Detection Systems" presented at the First Invitational Workshop on Rough Sets, Poznan-Kiekrz, September 2-4. 1992.
5. M.J.A. Berry and G.S. Linoff. *Mining the Web: TransformingCustomer Data*. John Wiley & Sons, New York, 2002.
6. R. Clarke. 'Profiling' and Its Privacy Implications. *Privacy Law & Policy Reporter*, 1: 128, 1994.
7. R. Clarke. Platformfor Privacy Preferences: A Critique. *Privacy Law & Policy Reporter*, 5(3): 46–48, 1998.
8. B. Custers. Data Mining and Group Profiling on the Internet. In A. Vedder, editor, *Ethics and the Internet*, pages 87–104.

9. O. Etzioni. The World Wide Web: Quagmire or Gold Mine? *Communications of the ACM*, 39(11): 65–68, 1996.
10. D.R. Fordham, D.A. Riordan and M. Riordan. Business Intelligence. *Management Accounting*, 83(11): 24–29, 2002.
11. J.S. Fulda. Data Mining and Privacy. In R. Spinello and H. Tavani, editors, *Readings in CyberEthics*, pages 413–417. Jones and Bartlett, Sudbury MA, 2001.
12. D.G. Johnson. *Computer Ethics*, 3rd. edition. Prentice-Hall, Upper Saddle River New Jersey, 2001.
13. J.F. Johnson. Immunity from the Illegitimate Focused Attention of Others: An Explanation of our Thinking and Talking about Privacy. In A. Vedder, editor, *Ethics and the Internet*, pages 49–70. Intersentia, Antwerpen Groningen Oxford, 2001.
14. R. Kosala, H. Blockeel and F. Neven. An Overview of Web Mining. In J. Meij, editor, *Dealing with the Data Flood: Mining Data, Text and Multimedia*, pages 480–497. STT, Rotterdam, 2002.