Attributed Identity Resolution for Fraud Detection and Prevention

Article · April 2009

DOI: 10.1109/ICC.2009.6

CITATIONS
0

READS
200

2 authors, including:

John Talburt
Association for Computing Machinery
127 PUBLICATIONS **446** CITATIONS

SEE PROFILE

Attributed Identity Resolution for Fraud Detection and Prevention

John Talburt and *Chia-Chu Chiang

Department of Information Science
*Department of Computer Science
University of Arkansas at Little Rock
2801 South University Avenue
Little Rock, Arkansas 72204-1099, USA
E-mail: {cxchiang|jrtalburt}@ualr.edu

Abstract

Many knowledge discovery in databases (KDD) could be applied to crime investigation. Unfortunately, most of them are ineffective and only provide postmortem results for the crime investigation. As we detect the suspicious events, the crime has been committed. Since many crime- and terrorism related activities often involve identity frauds, we have a very strong motive why not to develop some new technique for identity fraud detection and prevention. This paper presents how our research results can be applied to help the prevention of identity fraud. We believe if we can detect identity fraud early, we can prevent if from occurring and hence protect people from identity theft.

Keywords: fraud investigation, identity authentication, identity resolution, knowledge discovery in databases, text mining

1. Introduction

Identity fraud is pervasive and is the fastest growing crime in today's society. International and domestic terrorism often involves identity fraud which needs to be dealt with by the national security authorities. Bose [2] classifies identity into three categories: (1) attributed identity, (2) biographical identity, and (3) biometric identity. A person's attributed identity is usually recorded in his/her identification card such as driver license and passport. The identification card includes the person's name, social security number, birthdate, height, weight, nationality, and etc. A person's biographical identity can be achieved from a person's social activities recorded in a public or private domain. A person's biometric identity includes a person's physical characteristics such as DNA and fingerprints which are usually unique to him or her. A person wrongly taking someone else's identity for the purpose of committing a crime is considered as identity fraud.

Chen et al. [3] categorize crimes into eight categories: traffic violations, sex crime, theft, fraud, arson, organized crime, violent crime, and cyber crime. These eight different crime types all require law enforcement agencies to conduct identity authentication of a person who commits a crime. The authentication of identity should be done in a real-time manner. If identity fraud is detected or suspected, a legitimate action should be immediately taken by the law enforcement agent. According to the classifications of eight different crime types, Chen et al. [3] present a research framework with KDD technologies relevant to solving a certain type of crime. What is surprising us in the proposed research framework is that text mining is not in the class of KDD technologies solving the fraud crime type including identity fraud. We believe text mining is critical and is great relevance to the fraud prevention and investigation research.

In this paper, we will present our research results on text mining we have done before. We then discuss how our research results can be applied to help identity authentication for identity fraud detection, prevention, and investigation. The idea is to obtain a person's identity information from the public domain and combine the person's attributed identity for identity authentication. In addition to the technical discussions, we also point out that it is important not to violate the personal privacy during the applications of the KDD technologies.

This document is structured as follows. Section 2 presents related work. Section 3 briefly overviews our research results we have accomplished in the projects partially by NSF and ACXIOM. Section 4 discusses how our research results can be further used for identity authentication. Future direction and research is included in Section 5. The paper is summarized in Section 6.

2. Literature survey

KDD (Knowledge Discovery in Databases) techniques have played a critical role in improving counter-terrorism and crime-fighting capabilities of law enforcement agencies and national security authorities. The KDD techniques [1, 6] including data mining, adaptive machine learning, and cluster classification and analysis are used to detect suspicious activities in terms of fraud patterns built in a database. Most of the technologies count on a volume of data to be available for fraud detection. There are two major issues using the existing KDD technologies. The technologies are ineffective because there can be a significant lag between the fraud event and its detection. The advances of these technologies must be considered in terms of timing. Second, the results through these traditional KDD technologies are postmortem which are not effective in fraud detection and prevention. We believe that if we can provide early detection in identity authentication, we might help prevent the fraud crime effectively.

3. Research results overview

Text mining is an important data mining technique for extracting entities with the attributes (names, addresses, phone numbers, etc.) and their relationships presented in a public domain such as incident reports, open news, and web sites. However, before we proceed to do the research on national security issues, we should be aware that there exist laws and regulations governing the privacy and civil liberties of the public [5] that could impact the way how the data are collected and analyzed. There are several suggestions listed below,

- identities collected from the public or public domains must ensure anonymity;
- specific identities collected, if required, must be by court order; and
- personal data collected, analyzed, and disseminated protected under the laws of civil liberties.

We have conducted two researches using text mining techniques in collaboration with a national security authority and local company. In the first project[1], we mined text data including first name, middle name, last name, city, and date of birth of a deceased person from public obituary web sites publishing announcements [7]. Using the text mining technique described in the article [7], a case study was conducted for evaluating the technique [8]. The experimental results show that the extracted data scored 95 confidence were validated by the Authoritative Source 80% of the time. There are situations where the data was complete, but a match against the Authoritative Source did not occur. We found three possible reasons for this problem to occur. First, the Authoritative Source might not be a complete database as we conducted the evaluation of the mining technique. The second one is that some announcements in the web sites did not truly reflect actual deceased persons. The third reason is that the entities against the Authoritative Source are selected as the "best match" identification. In many situations, entities with different attributes may all point to the same entity. The ambiguities or incompleteness of entities need to be resolved through an entity resolution technique.

In the second project[2], we developed a technique to create an entity-relationship model among entities with the attributes [4, 10-11]. An entity resolution technique is applied to resolve entities with ambiguous or incomplete attributes [10]. For example, suppose we are interested in two entities, "William Doe" and "Mary Doe". From the Internet, we extract the following information,

Table 1. Candidates for "William Doe" and "Mary Doe" [10]

Bill Doe, 123 Oak	Mary Ellen Doe, 678 Willow
Bill S. Doe, 789 Hickory	Mary Doe, 654 Elm
William Q. Doe, 456 Pine	Mary Doe, 789 Hickory

In Table 1, using our entity resolution technique, the results turn out that the correct entities are "Bill S. Doe, 789 Hickory" and "Mary Doe, 789 Hickory" who share a common address.

There is still a problem existing in the above single reference entity resolution technique. How are we so sure that the two entities are the correct entities just by validating the address? A multiple reference technique on entity resolution is applied to validate the relationship among more than one entity. Again, the same example in Table 1 is used here. If we can find out other entities sharing the same relationships with "Bill S. Doe, 789 Hickory" and "Mary Doe, 789 Hickory", then we can be more confident about the interested subjects. For instance, from the Internet, we also found out that "John Doe" also resides in the same address with "Bill S Doe" and "Mary Doe" which increases the possibility that "Bill S. Doe, 789 Hickory" and "Mary Doe, 789 Hickory" are the correct entities we are looking for.

4. Fraud detection and prevention

The purpose of fraud detection and prevention is to stop identity fraud before the crime actually takes place. Like what we mentioned before, existing KDD

[1] This research is partially supported by a grant from the ACXIOM Corporation in Little Rock, Arkansas, USA.

[2] This research is partially supported by the National Science Foundation (NSF) grant IIS-0635655.

technologies build a profile of customer behavior and compare a model of patterns of fraudulent activity. The technologies heavily rely on deviation detection which might not effectively detect frauds in a real-time manner. To combat identity fraud real-time, an innovative product named URU [9] has been developed for customer identity verification on line. The idea of the development of URU is very similar to our technology. For each customer identity verification, URU returns the identity matched, partially matched, or unmatched against their reference database. URU uses multiple database for reference and ours only uses one reference database for identity authentication. URU also supports identity authentication using biometric attributes such as voice. One feature that URU does not support yet is the confidence level of the entity identity checked against their reference databases. What happens if more than one reference database conflict to each other on the results? What is the confidence level on the partially matched identity? Different from URL, our technology takes the confidence level into considerations. In addition, URU only considers attributes of the corresponding entity for identity verification. It is not clear how URL resolves ambiguous entities. Our tool also considers the relationships of the interested entity with other entities for entity resolution.

5. Future direction and research

A list of research work is considered to improve our tool,

- the technique must be multilingual;
- the technique must support on-line and real-time;
- identity verification against multiple reference databases; and
- identity verification using biometric attributes.

6. Summary

Identity fraud mostly attributes to all types of crimes including terrorism. Traditional KDD technologies have the problems in stopping identity fraud before it actually happens. We present a technology for identity fraud detection and prevention to stop identity fraud before it takes place. The idea is not new. A tool called URU has been developed for this purpose. In this paper, we present a technique for the same purpose which provides more accuracy on the results of attributed and biographical identity verification using the multiple resolution technology.

7. References

[1] D. W. Abbott, I. P. Matkovsky, and J. F. Elder IV, "An Evaluation of High-End Data Mining Tools for Fraud Detection," *IEEE Systems, Man, and Cybernetics*, Vol. 3, October 1998, pp. 2836-2841.

[2] R. Bose, "Intelligent Technologies for Managing Fraud and Identity Theft," *Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)*, Las Vegas: Nevada, USA, April 10-12, 2006, pp. 446-451.

[3] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime Data Mining: A General Framework and Some Examples," *Computer*, April 2004, pp. 50-56.

[4] Chia-Chu Chiang, John Talburt, Ningning Wu, Elizaberth Pierce, Chris Heien, Ebony Gulley, and JaMia Moore, 'Partially Parsing Unstructured Formats for Text Mining,' *Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008)*, Las Vegas: Nevada, USA, April 7-9, 2008. (TO APPEAR)

[5] J. S. Cook and L. L. Cook, "Social, Ethical, and Legal Issues of Data Mining," In *Data Mining: Opportunities and Challenges*, J. Wang, Ed., Idea Group Publishing, Hershey, PA., USA, pp. 395-420.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, Vol. 17, No. 3, Fall 1996, pp. 37-54.

[7] Charles W. Ford, Chia-Chu Chiang, Hao Wu, Radhika R. Chilka, and John Talburt, 'Confidence on Approximate Query in Large Datasets,' *Proceedings of the IEEE International Conference on Information Technology Coding and Computing (ITCC 2004)*, April 5-7, 2004, Las Vegas: Nevada, USA, pp. 480-484.

[8] Charles W. Ford, Chia-Chu Chiang, Hao Wu, Radhika R. Chilka, and John Talburt, 'Text Data Mining: A Case Study,' *Proceedings of the IEEE International Conference on Information Technology Coding and Computing (ITCC 2005)*, Vol. 1, April 11-13, 2005, Las Vegas: Nevada, USA, pp. 122-127.

[9] C. J. Gahan, "URU – Online Identity Verification," *BT Technology Journal*, Vol. 22, No. 1, January 2004, Springer, Netherlands, pp. 43-51.

[10] John R. Talburt, Ningning Wu, Elizabeth Pierce, Chia-Chu Chiang, Chris Heien, Ebony Gulley, JaMia Moore, 'Entity Identification in Documents Expressing Shared Relationships,' *Proceedings of the 11th World Scientific and Engineering Academy and Society International Conference on SYSTEMS (WSEAS ICS 2007)*, July 23-25, 2007, Agios Nikolaos, Crete Island, Greece, Vol. 2, pp. 223-228.

[11] Ningning Wu, John Talburt, Chris Heien, Nick Pippenger, Chia-Chu Chiang, Elizabeth Pierce, Ebony Gulley, JaMia Moore, 'A Method for Entity Identification in Open Source Documents with Partially Redacted Attributes,' *Proceedings of Fifth Annual Mid-South College Computing Conference (MSCCC 2007)*, March 30-31, 2007, Monroe: Louisiana, USA, pp. 138-144.