CRIMINAL DATA CLASSIFICATION IN CYBERCRIME INVESTIGATION USING WORDNET AND MODIFIED NAÏVE BAYES CLASSIFIER

P.BHUVANESWARI[1], U.ROHIT KUMAR[2], S.KEERTHANA[3], S.SANJAY AKASH[4]

Assistant Professor, Department of Information Technology[1]

Student, Department of Information Technology[2,3,4]

Hindusthan Institute of Technology, Coimbatore, India.

ABSTRACT:

Crime is one of the biggest and dominating problem in our society and its prevention is an important. Task. Daily there are huge numbers of crimes committed frequently. This require keeping track of all the crimes and maintaining a database for same which may be used for future reference. The current problem faced are maintaining of proper dataset of crime and analyzing this data to help in predicting and solving crimes in future.

KEYWORD: Data mining, crime investigation, Wordnet, criminal communities.

1. INTRODUCTION

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing and computer processing.

Mining of data is a method of dealing with expansive data indexes to perceive outlines and set up an association to handle issues through information examination. The devices used, allow endeavours to accept future examples. Data mining is a procedure to analyze data from an informational collection to change it into a reasonable structure for additional utilization. It predicts future patterns and also enables the organization to make the learning driven decision. Generally utilized strategies for mining of data are artificial neural networks, decision tree, rule induction, nearest neighbor method and genetic algorithm. They are applied in many fields. One such interesting application is crime investigation.

2.LITERATURE SURVEY

1. A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace

One of the problems often associated with online anonymity is that it hinders social accountability, as substantiated by the high levels of cybercrime. Although identity cues are scarce in cyberspace, individuals often leave behind textual identity traces. In this study we proposed the use of stylometric analysis techniques to help identify individuals based on writing style. We incorporated a rich set of stylistic features, including lexical, syntactic, structural, content-specific, and idiosyncratic attributes.We also developed the Writeprints technique for

identification and similarity detection of anonymous identities. Writeprints is a Karhunen-Loeve transforms-based technique that uses a sliding window and pattern disruption algorithm with individual author-level feature sets. The Writeprints technique and extended feature set were evaluated on a testbed encompassing four online datasets spanning different domains: email, instant messaging, feedback comments, and program code.

2. Mining Association Rules between Sets of Items in Large Databases

In Data Mining, the usefulness of association rules is strongly limited by the huge amount of delivered rules. To overcome this drawback, several methods were proposed in the literature such as item set concise representations, redundancy reduction, and post processing. However, being generally based on statistical information, most of these methods do not guarantee that the extracted rules are interesting for the user. Thus, it is crucial to help the decision-maker with an efficient post processing step in order to reduce the number of rules. This paper proposes a new interactive approach to prune and filter discovered rules.First, we propose to use ontologies in order to improve the integration of user knowledge in the post processing task. Second, we propose the Rule Schema formalism extending the specification language proposed by Liu et al. For user expectations. Furthermore, an interactive framework is designed to assist the user throughout the analyzing task. Applying our new approach over voluminous sets of rules, we were able, by integrating domain expert knowledge in the post processing step, to reduce the number of rules to several dozens or less. Moreover, the quality of the filtered rules was validated by the domain expert at various points in the interactive process.

3. Mining criminal networks from unstructured text documents

Digital data collected for forensics analysis often contain valuable information about the suspects' social networks. However, most collected records are in the form of unstructured textual data, such as e-mails, chat messages, and text documents. An investigator often has to manually extract the useful information from the text and then enter the important pieces into a structured database for further investigation by using various criminal network analysis tools. Obviously, this information extraction process is tedious and error prone.Moreover, the quality of the analysis varies by the experience and expertise of the investigator. In this paper, we propose a systematic method to discover criminal networks from a collection of text documents obtained from a suspect's machine, extract useful information for investigation, and then visualize the suspect's criminal network. Furthermore, we present a hypothesis generation approach to identify potential indirect relationships among the members in the identified networks. Most collected digital evidence are often in the form of textual data, such as e-mails, chat logs, blogs, webpages, and text documents.

3.SYSTEM ANALYSIS

3.1 Existing System

In previous work, used a framework to analyze chat logs for crime investigation using data mining and natural language processing techniques. The proposed framework extracts the social network from chat logs and summarizes conversation into topics. The crime investigator can use information visualizer to see the crime-related results. To test the validity of our proposed framework, we worked in a joint effort with the cybercrime unit of a Canadian law enforcement agency.

3.2 Drawbacks

a. No classifier used to classify the crime data.

b. The existing framework not performs the better feature extraction from the chat logs.

c. Clique detector, Concept miner, and information visualizer. These are not enough to give the accurate result for crime detection from the chat logs.

3.3 Proposed Method

Crime is one of the biggest and dominating problem in our society and its prevention is an important. Task. Daily there are huge numbers of crimes committed frequently. This require keeping track of all the crimes and maintaining a database for same which may be used for future reference. The current problem faced are maintaining of proper dataset of crime and analyzing this data to help in predicting and solving crimes in future.

3.4 Objective

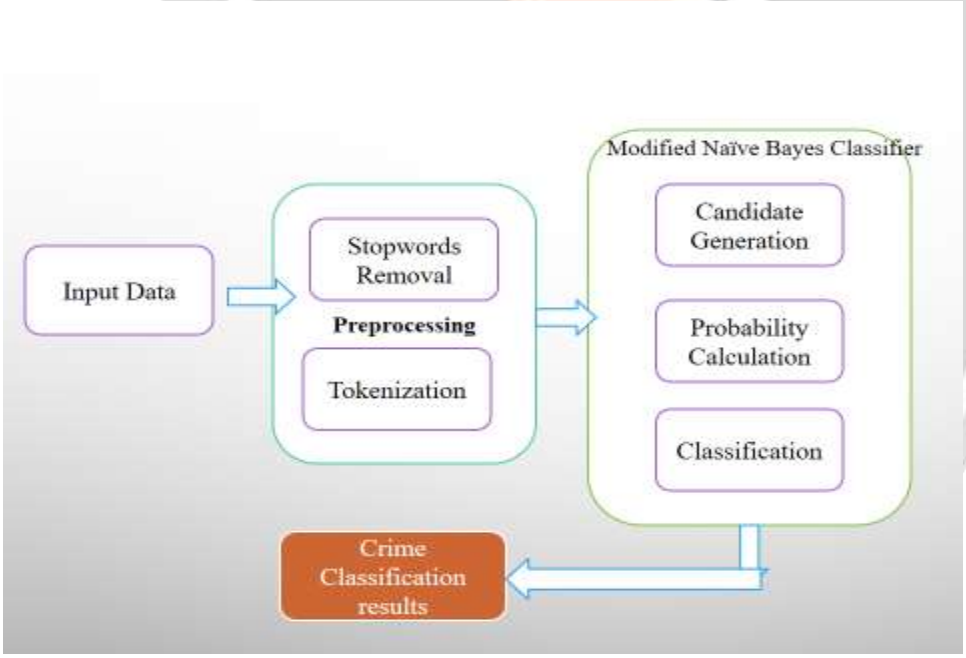
The objective of this project is to analyze dataset which consist of numerous crimes and predicting the type of crime which may happen in future depending upon various conditions.In this paper we propose an approach for crime prediction and classification using data mining. Here we use Naïve Bayes Classifier for crime prediction and classification.

3.5 Advantage

- Better results can be expected for crime data classification.
- Feature extraction compared to existing.
- Naïve bayes classifier provides higher classification using cross validation.

4.SYSTEM DESIGN

4.1 Block diagram



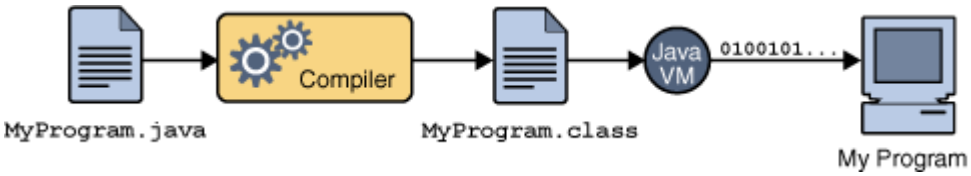
5. HARDWARE REQUIREMENTS:

- Processor : Dual Core
- Ram : 2GB
- Hard Disk : 160 GB Space

6. SOFTWARE REQUIREMENTS:

- Operating System : Windows 7 / 8
- Language : Java, J2EE
- Developing Tool : Netbeans 7.2.1
- Technologies : JSP, Servlet
- Backend : MySql Server

6.1 SOFTWARE DIAGRAM



An overview of the software development process

7.RESULT

A practical model based on the Naive Bayes classifier is proposed with novel methodologies applied for the criminal prediction problem. The incident-level crime data are generated synthetically by the model itself, otherwise which is hard to obtain in practice. The proposed model is practical due to the simplicity caused by the independence assumption of the Naive Bayes.The work in this project mainly revolves around predicting the type of crime which may happen if we know the location of where it has occurred. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation.Moreover, the model is moderately fit with the independence assumption in that the model achieves reducing the suspect list with 80 % rate. The experimental results showed that the proposed model can be used in criminology with its averagely 78.05 % success rate in order to help security forces to find the criminal of the incidents. One more significant of the proposed model is its ability to take the acquaintances into decision-making process.

REFERENCE

[1] Prefuse: Information Visualization Toolkit. Accessed: Sep. 1, 2010. [Online]. Available: http://prefuse.org/download/

[2] A. Abbasi and H. Chen, ‘‘Writeprints: A stylometric approach to identitylevelidentification and similarity detection in cyberspace,’’ ACM Trans. Inf. Syst., vol. 26, no. 2, Mar. 2008.

[3] R. Agrawal, T. Imieliński, and A. Swami, ‘‘Mining association rules between sets of items in large databases,’’ in Proc. ACM SIGMOD Int. Conf. Manage. Data, Washington, DC, USA, Jun. 1993, pp. 207–216.

[4] R. Al-Zaidy, B. Fung, and A. M. Youssef,‘‘Towards discovering criminal communities from textual data,’’ in Proc. ACM Symp. Appl. Comput., Taichung, Taiwan, Mar. 2011.

[5] E. Alfonseca and S. Manandhar, ‘‘Anunsupervised method for general named entity recognition and automated concept discovery,’’ in Proc. Int. Conf. General WordNet, 2002, pp. 34–43.

[6] R. Al-Zaidy, B. C. M. Fung, A. M. Youssef, and F. Fortin, ‘‘Mining criminal networks from unstructured text documents,’’ Digit. Invest., vol. 8, nos. 3–4, pp. 147–160, Feb. 2012.

[7] M.-H. Antoni-Lay, G. Francopoulo, and L. Zaysser, ‘‘A generic model for reusable lexicons: The genelex project,’’ Literary Linguistic Comput., vol. 9, no. 1, pp. 47–54, 1994.

[8] V. R. Carvalho and W. W. Cohen, ‘‘Learning to extract signature and reply lines from email,’’ in Proc. Conf. Email Anti-Spam, Mountain View, CA, USA, Jul. 2004.

[9] J. Schroeder, J. Xu, H. Chen, and M. Chau, ‘‘Automated criminal link analysis based on domain knowledge,’’ J. Amer. Soc. Inf. Sci. Technol., vol. 58, no. 6, pp. 842–855, Apr. 2007.

[10] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann, ‘‘Uncovering the dark Web: A case study of Jihad on the Web,’’ J. Amer. Soc. Inf. Sci. Technol., vol. 59, no. 8, pp. 1347–1359, Jun. 2008.