The detection of potentially illegal activity on financial discussion boards using information extraction

Edward Knott and Majdi Owda
School of Computing, Mathematics & Digital Technology
The Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK
Telephone (+44) (0) 161 247 1520
Email: Edward.Knott@stu.mmu.ac.uk
Email: M.Owda@mmu.ac.uk

Abstract.
The Internet poses as a real and tangible forum for illegal financial activity to flourish. This paper presents a novel prototype system for a financial share discussion detection system, to highlight potentially unlawful practices. Information is extracted from financial discussion boards, where templates hosting scenarios of known illegal activities are used to detect any potential misdemeanors. From an analysis of a single day's trading, it was observed that of the 3000 comments extracted, 0.2% of these were deemed suspicious and required the investigation of a discussion board moderator. The man-power required to perform this task manually over the course of a year could be a prohibitive. The initial work underscores the importance and need of an automated crime detection system, using financial discussion boards as its key extraction component.

Keywords: Information Extraction, Financial Discussion Boards, Fraud Detection, Crime Prevention, and Computer Forensics

1 Introduction

Online financial discussion boards (FDBs) grant users commentary and subsequent discussion opportunities centering on shares, stocks, common funds, business and political issues. Such forums, in the main, are not moderated by external third parties and loosely self-moderated via the forums' users themselves; whether it be a user reporting a comment as inappropriate, for instance. This form of un-moderated communication is open to abuse and could play a significant part in the aiding and abetting of financial misconduct.

Information Extraction (IE) has been sanctioned in various fields in recent years, notably: Web Knowledge Bases [1], Text Mining [2] and bioinformatics [3]. It appears that very little research has been conducted with specific reference to IE via FDBs for the analysis of potentially illegal activity. The solution presented in this paper could significantly impact the way FDBs are regulated in the future. The paper

will outline why a proposed system is needed and how it has the potential to tackle fraudulent activity born out of seemingly innocuous exchanges on FDBs.

Parallel systems exist; two of which are of note to critically consider. H, Limanto et al. created a system that involves "...an information extraction engine for web-based forums" [4]. Their engine analyses HTML files crawled from web forums. The crawler extracts the information about posts where users can query the system based on the information which is extracted and stored in a database. M, Costantino et al. examined a financial information extraction system which is currently under development at the University of Durham. [5] The system uses pre-defined templates to extract information based on financial activities which may have direct influence on share-prices. These financial activities are split into three groups based on what activity is being discussed. [5]

Section 2 describes the type of fraud that may occur within FDBs. Section 3 will introduce IE. Section 4 will outline the proposed system architecture. Section 5 will display the results and Section 6 will conclude the research.

2 Finance Fraud & Analysis

Sections 2.1 and 2.2 outline statistics of financial fraud and how FDBs may play a contributory factor. Key unlawful financial activities include, market manipulation, high yield investment fraud, ponzi schemes, advanced fee schemes, pyramid schemes, foreign currency fraud, broker embezzlement, hedge fund related fraud and late day trading. [6]

2.1 Finance Fraud

Securities fraud "...covers a wide range of illegal activities, all of which involve the deception of investors or the manipulation of financial markets." [7] Figure 1 below illustrates statistics based on the number of securities fraud cases between 2005 and 2009 in the United States of America (USA).

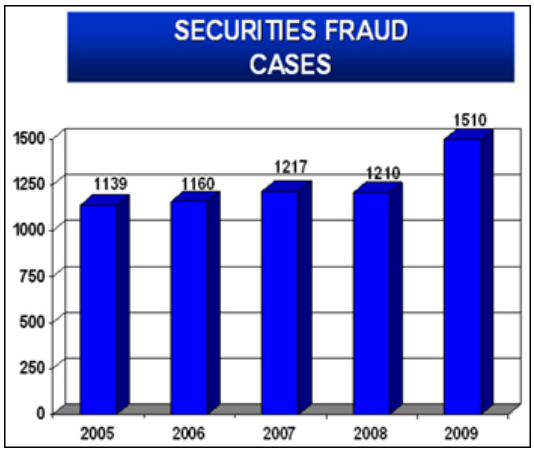

Figure 1 - Securities Fraud Pending Cases 2005-2009 (USA) [6]

Statistics indicate that the numbers of pending cases are steadily on the increase, *(average number of pending cases over a 5 year period is 1247).*

2.2 Finance Discussion on Message Boards

Approximately 5000 [7] comments are posted daily on one of the most active FDBs in the UK. To moderate such volume is manually prohibitive and the proposed system aims to aid a moderator in scanning FDBs automatically, returning any results that may indicate potentially illegal activity.

3 Information Extraction

IE is a type of information retrieval where a user can define specific information to be extracted from documents (i.e. using a set of criteria, usually text, as opposed to images and videos). It can be associated with any method whose purpose is to extract information from documents and/or web pages. Chelba and Mahajan defined IE as a text filtering and template filling process, segments of text are to be filled into a specific number of slots which forms a template or frame. [8]

IE has two basic approaches; knowledge engineering and automatic training. First, the knowledge engineering approach is based on having a knowledge engineer who develops rules and knowledge that have the ability to solve problems in the real world. Appelt and Israel believe that the knowledge engineering based approach is most effective when resources such as lexicons and rule writers are available. [9] Secondly, the automatic training approach does not require a human to write rules for the IE system, instead, it only requires someone who knows well the domain, and then the task is to annotate a corpus of texts for the information being extracted. IE will play an important role in developing the financial discussion detection system (FDDS).

4 Prototype Architecture

The architecture in figure 2 below shows the processes of the proposed system.

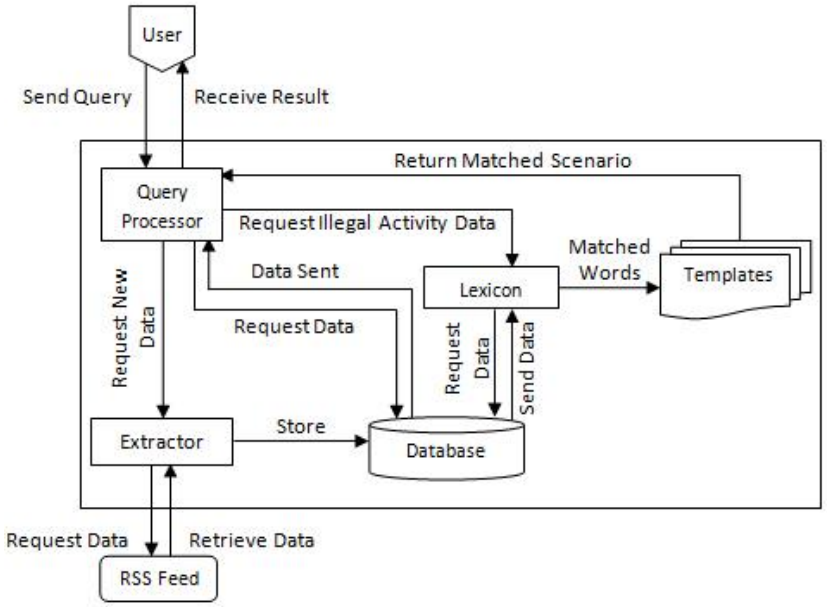

Figure 2 - Component diagram for proposed system

A user sends a query to retrieve illegal activities. The query processor then utilises the extractor to retrieve comments on FDBs and stores them into a database. Once the query processor begins analysing comments, it uses a lexicon in order to populate predefined templates for potentially illegal activities. Once a template has been filled, a response is sent to the query processor which is then sent to the user for analysis. The components used in the diagram above are explained below.

4.1.1 RSS Feeds

FDBs that possess Really Simple Syndication Feed (RSS Feed) technology have the potential to be used with the FDDS. RSS feeds are Extensible Markup Language (XML) files. The system extracts information through the tags in the file. This allows for extraction of all information from tags that contain the relevant information. Figure 3 illustrates raw XML information from RSS feeds which is to be extracted. For security and privacy measures, the information has been masked. The red boxes indicate the information which is to be extracted and the single black box indicates irrelevant information. In order to extract information from FDBs, approval had to be ob-

tained from each website. This has been successfully obtained and documented. All information stored in the database is for experimental purposes only.

Figure 3 - RSS Feed - XML

4.1.2 Extractor

The extractor component is the connection between the FDB's RSS feed and the database. The extractor extracts comments from the RSS feed which are stored into the MySQL database.

4.1.3 MySQL Database

The database behind the proposed system had been created in MySQL allowing for complex SQL commands to be performed rather than using a Microsoft Access Database which has limited features. The database has two purposes which include:

1. Storing comments that have been extracted using the extractor
2. Storing the lexicon (combination of known words) – The words will be used to form information extraction templates.

4.1.4 Lexicon

The lexicon contains the words associated with the templates. As each word is matched from the lexicon, the relevant scenario highlights the word that has been matched.

4.1.5 Types of Scenarios

Some scenarios from the proposed system include: pump and dump, insider information and securities fraud. Words associated with each scenario are presented in a separate template. This template is then matched up with comments extracted from the FDBs which are saved to the database.

4.1.6 Templates

The templates are created on the back of the research performed on each type of scenario described. Words relating to each scenario are then stored into the lexicon

ready to be used with the system. The greater the number of highlighted words returned, the greater the probability that the user may be discussing potentially illegal activity.

4.1.7 Query Processor

This section of the system processes the communication between the user and the database through SQL queries.

4.1.8 Information Extraction

As each comment is extracted from the RSS feeds, it is temporarily stored as an object and certain information (i.e. stock code) is extracted from the objects which are then stored into the database.

4.2 Users

Users have the ability to perform a number of different search functions within the FDDS:

1. Searching comments using a specific stock code.
2. Searching comments on a specific username.
3. Searching comments using a set of user defined keywords including the functionality of AND or OR options.
4. Display comments from all FDBs.
5. Searching potentially illegal activity using a scenario on a specific stock code.
6. Searching potentially illegal activity using a scenario on a specific username.
7. Display potentially illegal activity using a scenario from all FDBs.

Administrators of the FDDS have the ability to maintain the system and perform the following tasks:

1. Add/Modify/Remove scenarios.
2. Retrieve latest comments from FDBs.
3. Review keywords that have been used.
4. Review database statistics which include functions to reset the database.

4.3 Financial Discussion Detection System (FDDS)

This component is the user interface and allows the user to perform the following three types of functions:

- Configuration Manager: templates, database, user history, retrieval of latest comments.
- Potentially Illegal Activity: Use templates/scenarios on a specific user, a specific stock code, or on all FDBs.

- Search Comments: Search comments on a specific user, a specific stock code, or on all FDBs.

5 Implementation

The proposed system is user friendly. In three simple clicks, results on potentially illegal activity will be collected and returned. Users select a scenario to work with, and then select a stock code to search activity on. Results returned, highlight any words associated with the selected scenario. Figure 4 below illustrates the three step process:

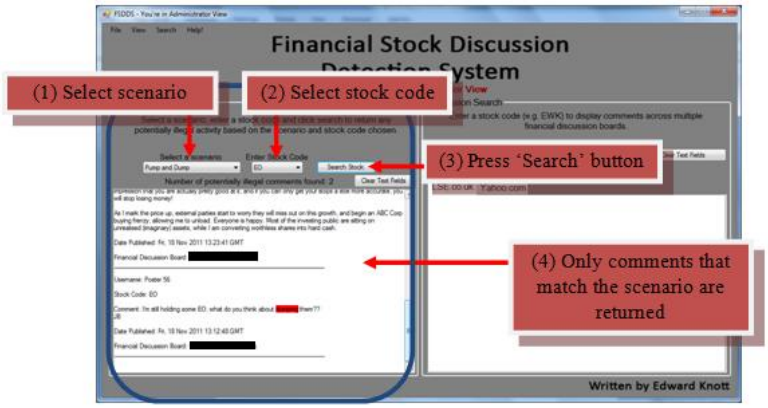

Figure 4 - Screenshot of proposed system

6 Results

Statistics from three different scenarios carried out on the same 3000 comments are displayed below in figure 5. The results show that from the comments analysed, all three scenarios contain potentially illegal activity.

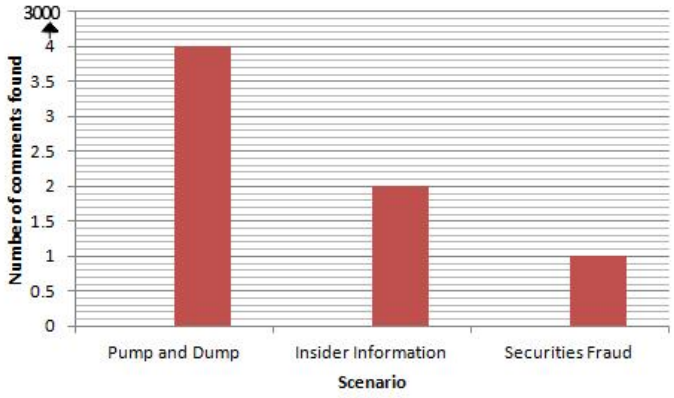

Figure 5 - Results from three scenarios

7 Conclusion

The paper highlights the creation of a novel prototype system for a FDDS. The prototype architecture uses IE as its main component to extract comments from FDBs to fill predefined templates for the analysis of any potential fraudulent scenarios. Results returned indicated that from 3000 comments made in one single day in one FDB alone, 7 were flagged as potentially discussing illegal activity. Over the course of one year, the FDDS could flag over 2,500 such comments from one single FDB, therefore, we have to pose the question, *why haven't FDBs been moderated before?*

References

[1] M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. & Slattery, S. Craven, *Learning to extract symbolic knowledge from the World Wide Web.*, 1998.

[2] R. & Bunescu, R. Mooney, "Mining knowledge from text using information extraction.," pp. 3-10.

[3] R., Ge, R., Kate, R., Marcotte, E., Mooney, R., Ramani, A. & Wong, Y. W. Bunescu, "Comparative experiments on learning information extractors for proteins and their interactions.," 2005.

[4] Hanny Yulius Limanto et al., "An Information Extraction Engine for Web Discussion Forums," Nanyang Technological University, Singapore, ACM 1-59593-051-5/05/0005, 2005.

[5] Marco Costantino, Richard G Morgan, Russell J Collingham, and Roberto Garigliano, *Natural Language Processing and Information Extraction: Qualitative Analysis of Financial News Articles*, February 1997.

[6] The Federal Bureau of Investigation (FBI). ([n.d.]) FBI - Securities Fraud Awareness & Prevention Tips. [Online]. http://www.fbi.gov/stats-services/publications/securities-fraud [Accessed on 13th January 2012]

[7] Interactive Investor. Community | Interactive Investor. [Online]. http://www.iii.co.uk/community/ [Accessed on 19th January 2012]

[8] C. & Mahajan, M Chelba, "Information Extraction suing the structured language model," in *Proceedings of the 2001 Conference on Empiricla Methods in Natural Language Processing*, Pittsburgh, PA, USA, 2000.

[9] E. Appelt and D. Israel, *Introduction to Information Extraction*, 1999.