Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON

Zenun Kastrati(✉), Ali Shariq Imran, Sule Yildirim-Yayilgan, and Fisnik Dalipi

Faculty of Computer Science and Media Technology, GjøVik University College, GjøVik, Norway
{zenun.kastrati,ali.imran,sule.yayilgan,fisnik.dalipi}@hig.no

Abstract. Analysing users' behaviour and social activity for investigating suspects is an area of great interest nowadays, particularly investigating the activities of users on Online Social Networks (OSNs) for crimes. The criminal activity analysis provides a useful source of information for law enforcement and intelligence agencies across the globe. Current approaches dealing with the social criminal activity analysis mainly rely on the contextual analysis of data using only co-occurrence of terms appearing in a document to find the relationship between criminal activities in a network. In this paper, we propose a model for automated social network analysis in order to assist law enforcement and intelligence agencies to predict whether a user is a possible suspect or not. The model uses web crawlers suited to retrieve users' data such as *posts*, *feeds*, *comments*, etc., and exploits them semantically and contextually using an ontology enhancement objective metric SEMCON. The output of the model is a probability value of a user being a suspect which is computed by finding the similarity between the terms obtained from the SEMCON and the concepts of criminal ontology. An experiment on analysing the public information of 20 Facebook users is conducted to evaluate the proposed model.

Keywords: Ontology · Online social networks · Facebook · SEMCON

1 Introduction

In recent years, the usage of Online Social Networks (OSNs) has increased rapidly throughout all layers of society. Law enforcement and intelligence agencies analyse traces of digital evidence in order to solve crimes and capture criminals whom are also OSNs users during their investigation activities. Particularly analysing contents shared by users on social networks such as Facebook, Twitter and LinkedIn are of interest. Several approaches of analysis aiming at extracting useful information, modelling users profile, and understanding users behaviour and social activity have been proposed [1].

Analysing users behaviour and social activity for investigating suspects is also an interesting area of research, particularly investigating the activities of users on OSN for crimes. The criminal activity analysis provides a useful source of information for law enforcement and intelligence agencies across the globe.

G. Meiselwitz (Ed.): SCSM 2015, LNCS 9182, pp. 148–157, 2015.
DOI: 10.1007/978-3-319-20367-6_16

Some agencies are now using social media as a crime-solving tool [2]. Digital traces from social media such as Facebook is gaining fast acceptance for use as evidence in courts [3]. According to a survey by LexisNexis in 2012 [4], there are more than 950 law enforcement professionals with federal, state, and local agencies in United States whom use social media, particularly Facebook and YouTube, to obtain evidence to deepen their criminal investigation. Other similar criminal cases have been reported recently where digital evidence from OSNs is used as support for digital investigation [5,6].

Criminal activity analysis consists of different stages such as data processing, transformation, analysis, and visualization. Many of these stages are done manually. Thus, it takes much time and human effort to extract the required evidence from the massive amount of information.

Recently some research has been done to automate the social criminal activity analysis to help law enforcement and intelligence agencies discover the criminal networks. In this light, a framework for the forensic analysis of user interaction in OSNs is proposed in [7]. The framework enables searching for actor activities and filtering them further for temporal and geographical analysis. The authors in [8] proposed a framework that consists of major components of a network analysis process: network creation, network partition, structural analysis, and network visualization. Based on this framework, the authors developed a system called *CrimeNet Explorer*. The system has structural analysis functionality to detect subgroups from a network, identifying central members of subgroups, and extracting interaction patterns between subgroups. The authors in [9,10] used data mining approach for analyzing criminal groups. They used data mining in multiple social networks data to discover criminal networks.

However current approaches to automating the social criminal activity analysis have some limitations. They mainly rely on the contextual analysis using only co-occurrence of terms appearing in a document to find the relationship between criminal activities in a network [8]. Moreover, some of the approaches perform experiment using no real-world datasets [9].

In this paper, we try to fill this gap by proposing a framework for automated social network analysis. This framework will assists law enforcement and intelligence agencies to predict efficiently and effectively whether a user is a possible suspect or not. This is achieved by exploiting users' *posts*, *feeds* and *comments*, semantically and contextually using SEMCON [11]. SEMCON is a context and semantic based ontology enhancement model developed at our lab originally for the purpose of enriching an ontology from posts of multimedia documents.

The rest of the paper is organized as follows. In Sect. 2 we illustrate in detail our proposed model. Section 3 describes the setting for experimental procedure whereas Sect. 4 illustrates the experimental results and their analysis. Lastly, in Sect. 5 we sketch conclusions and future work.

2 Proposed Model and Methodology

The proposed model, illustrated in Fig. 1, aims at performing the analysis of social networks profiles. Information such as *posts*, *feeds* and *comments* are extracted

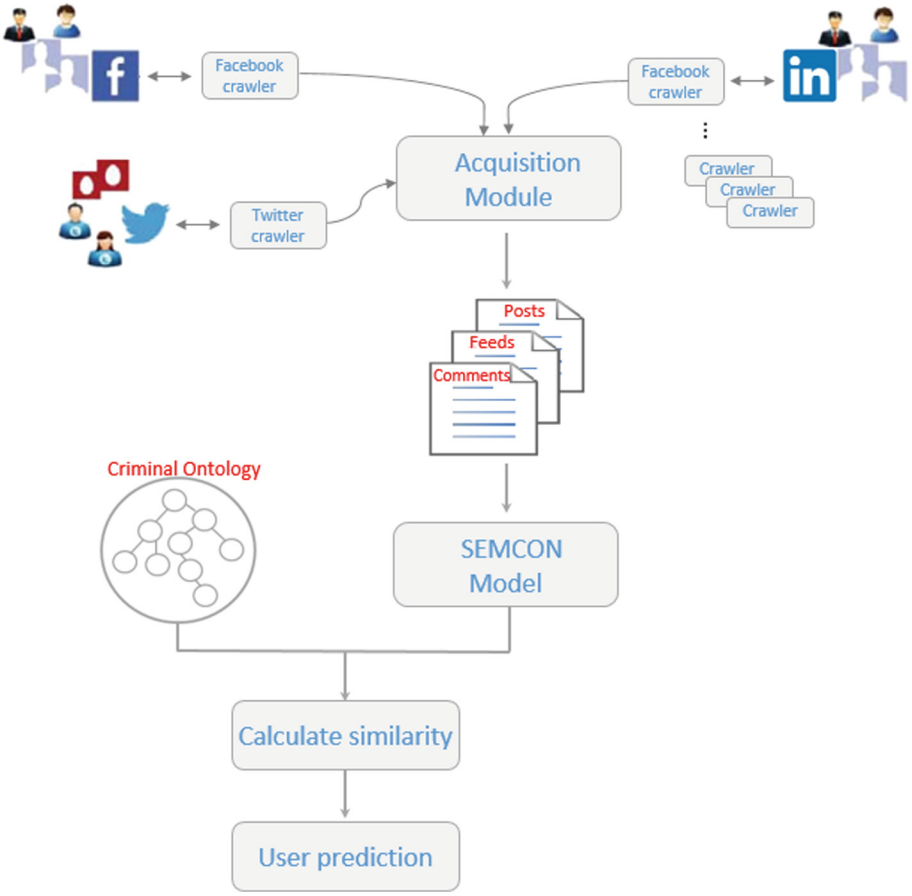

Fig. 1. Flow chart of the proposed model

and analysed, considering in particular both the context and the semantics of terms used by users. The model is explained in the following sections.

2.1 Acquisition Module

The module use web crawlers suited to retrieve and manage data coming from particular social networks such as Facebook, Twitter, LinkedIn, etc. In our case we have used Facebook crawler for managing Facebook posts. Facebook crawler is based on the Facebook Graph APIs and Facebook Query Language (FQL). To fetch Facebook messages and making queries, this paper uses RestFB [12] which is a simple and flexible Facebook Graph API client written in Java. The crawler uses an opaque string called Facebook access token that identifies a user, application, or page and can be used by the application to make graph API calls. In this work, the Facebook crawler is dedicated to fetch only *posts*, *feeds* and

comments of a user. Facebook imposes some limitations on the number of *posts*, *feeds* and *comments* retrievable through its APIs according to the data access policy of Facebook. It does not allow to retrieve the information of more than 25 *posts*, *feeds* and *comments* per user. The restrictions on the maximum number of retrievable information are overcome by using specific parameters which enables to filter and page through the connection data.

2.2 SEMCON Module

The information fetched by Acquisition Module is used as input to the SEMCON Module. The SEMCON module treats each *post*, *feed* and *comment* basically as an independent document-passage and it performs the following steps.

Initially a morpho-syntatic analysis using TreeTagger [13] is performed where the partitioned passages are tokenized and lemmatized. The potential terms that are obtained as a result can either be a noun, verb, adverb or adjectives. These are different parts-of-speech (POS) of a language. It is a well-known fact that nouns represent the most meaningful terms in a document [14], thus, our focus is on extracting only common noun terms t for further consideration.

The next step is the calculation of the observation matrix. The observation matrix is formed by calculating the frequency of occurrences of each term t, its font type (*bold*, *underline*, *italic*) and its font size (*title*, *level* 1, *level* 2) as given in Eq. 1.

$$O_{i,j} = \sum_{i \in t} \sum_{j \in p} (Freq_{i,j} + Type_{i,j} + Size_{i,j}) \tag{1}$$

where, t and p indicate the set of terms and passages, respectively. $Freq_{i,j}$ denotes the frequency of occurrences of term t_i in passage, p_j, $Type_{i,j}$ denotes font type of term t_i in passage p_j, and $Size_{i,j}$ indicates font size of term t_i in passage p_j.

The observation matrix is used as input to compute the contextual and semantic similarity between two terms.

Term to term contextual score (S_{con}) is calculated using the cosine similarity metric with respect to the passages, and it is given in Eq. 2.

$$S_{con}(t_i, t_j) = \frac{t_i \cdot t_j}{\| t_i \| \| t_j \|} \tag{2}$$

A term square matrix is used to store the contextual(S_{con}) values among all extracted terms t.

The next step is the computation of the semantic score (S_{sem}). The semantic score is calculated using the Wu&Palmer algorithm [15] and the score is computed using the Eq. 3.

$$S_{sem}(t_i, t_j) = \frac{2 * depth(lcs)}{depth(t_i) + depth(t_j)} \tag{3}$$

where t_i and t_j indicate terms extracted from the passage, *depth(lcs)* indicates least common subsumer of t_i and t_j, *depth*(t_i) and *depth*(t_j) indicate the path's

depth of t_i and t_j, respectively. Go through the all terms, we take all possible pairs and compute the semantic score $S_{sem}(t_i, t_j)$, for each pair t_i and t_j, where t_i, $t_j \in C$ and C is the set of terms extracted from the corpus.

The overall correlation between two terms t_i and t_j extracted from the the passage is computed using the contextual and semantic score. Mathematically, the overall score is given in Eq. 4.

$$S_{overall}(t_i, t_j) = w * S_{con}(t_i, t_j) + (1 - w) * S_{sem}(t_i, t_j) \tag{4}$$

where S_{con} is the contextual score, S_{sem} is the semantic score and w is a parameter with value set as 0.5 in our case, based on the empirical analysis from the data set. The overall score is in the range (0,1]. The overall score is 1 if two extracted terms are the same.

2.3 User Prediction Module

The prediction of a user as a suspect or not depends on the similarity score between the terms extracted from the user' *posts*, *feeds* and *comments* via SEMCON module and concepts extracted by the criminal ontology. The higher the score, the closer the user is considered as a suspect user.

The similar calculation is performed using the cosine similarity measurement. More formally, it is given in Eq. 5.

$$Similarity(O_c, u_i) = \frac{\overrightarrow{O}_c \times \overrightarrow{u}_i}{\| \overrightarrow{O}_c \| \cdot \| \overrightarrow{u}_i \|} \tag{5}$$

where, O_c indicates concepts extracted from the criminal ontology and u_i indicates terms extracted by the user postings.

The output of the system is a probability value P, of a user being a suspect s. If the P_s is greater than a specified threshold t then the user is labelled as a suspect.

3 Experimental Setting

We have performed the investigation of suspects using the public users' *posts*, *feeds* and *comments*. The facebook crawler is established to collect the data for the period starting from 1 January till 31 December 2014. The posts are extracted from news and media.

The posts from social networks contain usually noisy text, e.g. null values, therefore we filtered out only the posts which comply with the standard rules of orthography, syntax and semantics. After this process, we created a corpus which consists of 198 posts published by 20 users. The average number of posts per user is 10. The total number of terms used is 8493 with an average of 43 terms for each post. Finally, from these terms we identified and extracted 1042 nouns (singular and plural). The detailed information for each user is shown in Table 1.

Table 1. The corpus data

User	# of Posts	# of Terms	# of Nouns
1	11	929	55
2	3	121	11
3	12	301	58
4	5	130	25
5	8	376	56
6	4	1550	140
7	11	366	48
8	9	383	51
9	16	600	58
10	11	270	40
11	12	313	46
12	6	117	21
13	21	567	102
14	9	344	46
15	12	336	43
16	8	317	42
17	12	494	58
18	9	307	44
19	10	298	42
20	8	374	56
Total	198	8493	1042

We have also created a criminal ontology shown in Fig. 3. Basically it is used to predict if a user is a suspect by comparing its concepts with the terms outputted by the SEMCON as described in Sect. 2.3. However, the criminal ontology may also be used for visualization of criminal information by displaying concise overviews of its concepts and their hierarchical relations using treemaps.

4 Results and Analysis

In order to evaluate a user being as a suspect or not, we have performed an experiment on 20 Facebook users by analysing their public postings. For each user, we initially computed an overall score by aggregating the semantic and contextual score for each term (noun) extracted. The overall scores of terms are used to find the similarities of the terms with the criminal ontology concepts. Figure 2 illustrates the terms score obtained by SEMCON for the *User* #13 as depicted in Table 1. As can be seen from the graph, the contextual score indicated by the blue curve is much lower then the semantic score denoted by

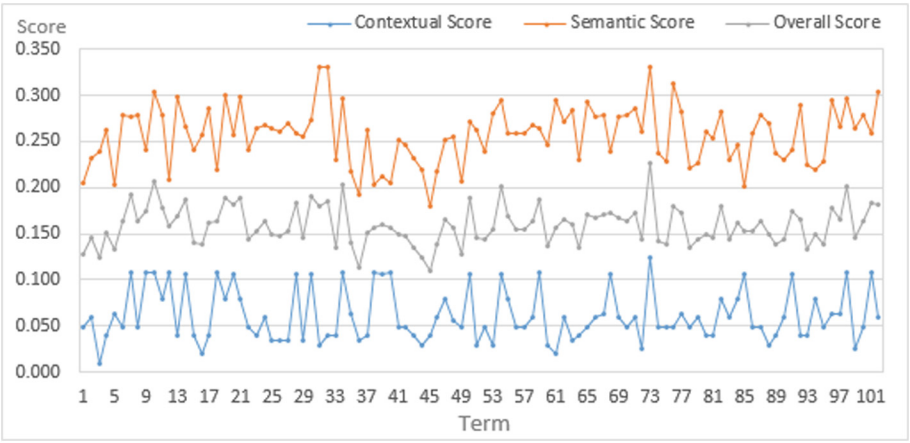

Fig. 2. Scores obtained by SEMCON for a user in investigation

Table 2. The probability of users being suspects

User	Probability	User	Probability
1	0.990	11	0.992
2	0.727	12	0.000
3	1.000	13	1.000
4	0.892	14	0.772
5	0.578	15	0.784
6	0.820	16	0.800
7	1.000	17	0.799
8	0.000	18	0.000
9	0.933	19	1.000
10	1.000	20	0.800

the orange curve. This may have happened due to the fact that the user has posted or commented in different topics. However, terms used in these posts have high semantic correlation with each other.

In the next step we found out how likely that a user is a suspect. This is achieved by comparing the user overall score obtained by SEMCON module and the scores of concepts of criminal ontology. User suspicion is represented by a probability value. The obtained probabilities for users being a suspect are shown in Table 2. The probability value 0.000 represents the users whose posts does not contain any of the criminal ontology concepts. Thus, these users are considered as unsuspected users. The probability value 1.000 indicates the users whose posts contain some of the concepts of the criminal ontology, i.e. *gun*, *rifle*, *shooting*, *threat* and *death*. These users are considered to be highly suspected users.

Table 3. Categorization of user prediction

	Unsuspected	Moderate suspected	Highly suspected
User	8, 12, 18	2, 4, 5, 6, 14	1, 3, 7, 10
		15, 16, 17, 20	11, 13, 19

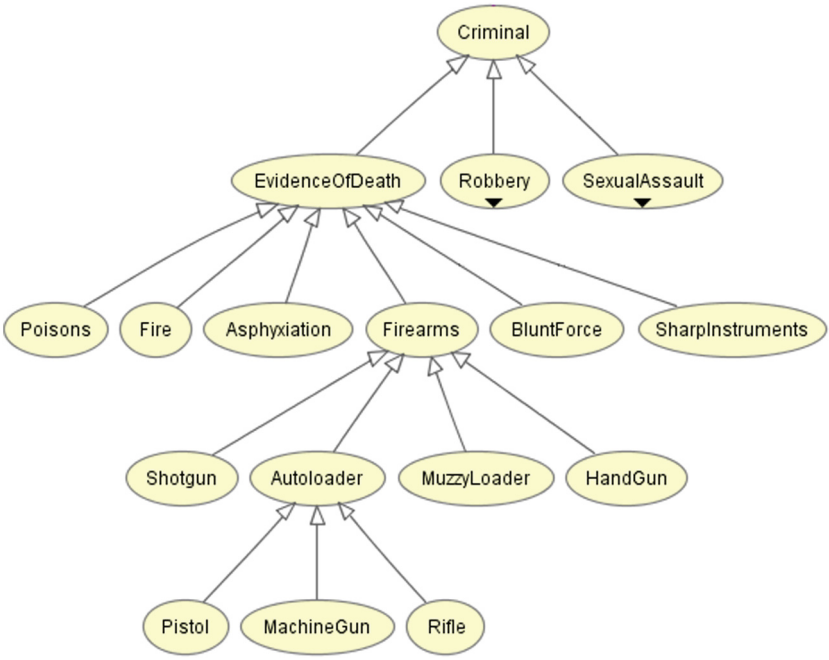

Fig. 3. A part of criminal ontology

Based on the obtained probability results, we can identify three major categories of users; users classified as unsuspected users, moderate suspected users and the highly suspected users. More precisely, if the probability score of user exceeds a given positive threshold value (in our case 0.90) we classify his/her as being a highly suspected user; if his/her probability score is 0.00 we label him/her as unsuspected user, otherwise he/she is considered as being a moderate suspected user. The labelling of users in particular categories is shown in Table 3.

5 Conclusion and Future Work

In this paper, we have proposed a new approach to investigating if a user is a suspect by analysing the OSNs data. We used Facebook as a case study of OSNs and Facebook user' *posts*, *feeds* and *comments* have been the object of the

study. The approach employs the SEMCON to provide a semantic and contextual data-mining analysis for automatically monitoring users' activity through textual analysis. We initially built a domain ontology called criminal ontology. The prediction of a user as a suspect or not is computed by finding the similarity score between terms extracted from user' *posts*, *feeds* and *comments* via SEMCON module and the concepts extracted by the criminal ontology.

From the experiment conducted by analysing the postings published within a year by 20 users, we identified three categories of users: unsuspected, moderate suspected and highly suspected users. The categorization of users can assist law enforcement and intelligence agencies to narrow the investigation, identify and focus only on suspected users in order to prevent or solve crimes.

In the future we plan to further extend our proposed approach. Terms obtained from the SEMCON model will be used to build a user ontology. The user ontology can be used to create a history based user activity profile which may actually put light on otherwise invisible relations between a particular social network user and his/her network, and the dependence among a user's various activities. It also can be used by the law enforcement personnels for deeper investigation in order to search for suspicious user activities and filtering them for temporal and geographical analysis.

References

1. Adamic, L., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**, 211–230 (2001)
2. Knibbs, K.: In the online hunt for criminals, social media is the ultimate snitch. http://www.digitaltrends.com/social-media/the-new-inside-source-for-police-forces-social-networks/. Accessed on 11th February 2015
3. The New York Criminal Law Blog: Assault Fugitive Who Was Found Via Facebook Is Back In NY. http://newyorkcriminallawyersblog.com/2010/03/assault-criminal-who-was-found-via-facebook-is-back-in-ny.html. Accessed on 16th October 2014
4. LexisNexis: Social media use in law enforcement agencies. http://www.lexisnexis.com/government/investigations/. Accessed on 16th October 2014
5. McGuire, M.: Technology, Crime and Justice: The Question Concerning Technomia. Routledge, New York (2012)
6. Binham, C., and Croft, J.: Twitter fuels debate over super-injunctions. Financial Times (2011)
7. Abdalla, A., Yayilgan, S.Y.: A review of using online social networks for investigative activities. In: Meiselwitz, G. (ed.) SCSM 2014. LNCS, vol. 8531, pp. 3–12. Springer, Heidelberg (2014)
8. Xu, J., Chen, H.: CrimeNet explorer: a framework for criminal network knowledge discovery. ACM Trans. Inf. Syst. (TOIS) **23**, 201–226 (2005)
9. Fard, A.M., Ester, M.: Collaborative mining in multiple social networks data for criminal group discovery. In: Proceedings of International Conference on Computational Science and Engineering, Vancouver, Canada (2009)
10. Shang, X., Yuan, Y.: Social network analysis in multiple social networks data for criminal group discovery. In: Proceedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discover, Sanya, China (2012)

11. Kastrati, Z., Imran, A.S., Yayilgan, S.Y.: SEMCON: semantic and contextual objective metric. In: Proceedings of the 9th IEEE International Conference on Semantic Computing, Anaheim, California, USA (2015)
12. RestFB: RestFB facebook graph API. www.restfb.com. Accessed on 10th February 2015
13. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing (1994)
14. Li, H., Tian, Y., Ye, B., Cai, Q.: Comparison of current semantic similarity methods in wordnet. In: International Conference on Computer Application and System Modeling, vol. 4, pp. 4008-4011 (2010)
15. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, pp. 133–138 (1994)