Journal of Applied Science and Agriculture, 9(20) Special 2014, Pages: 1-6

AENSI Journals

Journal of Applied Science and Agriculture

ISSN 1816-9112

Journal home page: www.aensiweb.com/JASA

Modus Operandi Extraction in Police Report

Mohd Pouzi Hamzah, Syarifah Fatem Na'imah Binti Syed Kamaruddin

School of Information and Applied Mathematics, Universiti Malaysia Terengganu, 21030, Kuala Terengganu, Terengganu, Malaysia

ARTICLE INFO

Article history:
Received 30 September 2014
Received in revised form
17 November 2014
Accepted 25 November 2014
Available online 13 December 2014

Keywords:
Information Retrieval, Modus Operandi, Malay Language

ABSTRACT

Background: This paper analyzes the modus operandi extraction in Malay language using information retrieval technique. The similarity of modus operandi will be measured according to the cosine of the angle between query and document vectors. Our research focuses on phrase identification to find out the rules of forming modus operandi from free-text of police reports. Modus operandi is the most important aspect in a police investigation. The valuable data that relevant to the user's information needed in free-text report is a challenge and difficult to be detected or tracked by the police. It would be better if modus operandi from text reports such as location of crime and crime behavior could be identified automatically. This paper also shows that modus operandi extraction can improve the performance of similarity measurement using precision, recall and F-measures. It evaluates the accuracy of modus operandi extraction on each report from ten different police reports. The best F-measure is 0.960 with recall and precision at 1.0 and 0.933 respectively for retrieval performance of modus operandi extraction.

To Cite This Article: Mohd Pouzi Hamzah, Syarifah Fatem Na'imah Binti Syed Kamaruddin., Modus Operandi Extraction in Police Report. ***J. Appl. Sci. & Agric.,*** *9(20): 1-6, 2014*

INTRODUCTION

Criminal database has been rising in number as there are increasing reports made by the public, hoping that these can be solved. Nevertheless, investigation is not an easy task and the authorities have to come out with a modus operandi based on their analysis from available reports. Therefore, modus operandi is an important element in order to understand criminal's mode of action (Chau and Xu, 2002).

Modus operandi is a term used to describe someone's habits or manner of working, their method of operating or functioning (Douglas, Burgess *et al*. 2011). Modus operandi can help the police to find a technique of criminal's work or its characteristic patterns (Manning and Raghavan, 2008). In investigation, the police also can acquire the behavior of the criminal by analyzing the modus operandi (Douglas and Burgess, 2011).

From the analysis of modus operandi, similarities are automatically identified among different police reports (Fosdick, 1915; Hazelwood, 2004). All reports will be compared to each other to discover similarities in modus operandi such as suspect behavior, methods of operation, location and other factors in the collected data. In police investigation, similarities of modus operandi are the most important tasks to link a series of crime (Valerie and Baldé 2005). This has become necessary with the implementation of the information retrieval software which is modus operandi extraction system. Information retrieval software is software that helps to find material of an unstructured nature that satisfies an information need within a large collection (Salton, 1983; Manning and Raghavan, 2008). Finally, the crimes or suspects will be traced according to similarities in modus operandi and provide of crime pattern.

Further, this paper discusses about the extraction of modus operandi and analysis of similarity. Result and evaluation will also be discussed at the end of the paper.

1. The Corpus (Police Reports):

The texts are collection of Malay texts that are taken from the Police Diraja Malaysia (PDRM) corpus. Corpus is a large volume of unstructured documents (Manning and Raghavan, 2008). The corpus is collected from daily reports and common texts. The corpus is tokenized, and its words are tagged according to knowledge base. The system will automatically tag proper words and numerical words, such as date, time and place where these words were not found in the knowledge base. They are tagged with a coarse tagset consisting of four different tags which are noun, verb, adverb and adjective. Examples of adverbs are yesterday, outside and always. 643 words

Corresponding Author: Mohd Pouzi Hamzah, School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia.
E-mail: mph@umt.edu.my

Journal of Applied Science and Agriculture, 9(20) Special 2014, Pages: 1-6

are tagged from a sample of report at this early stage. Table 1 shows the tags and their corresponding frequencies in the corpus.

Table 1: The tags distribution.

Tag Name	Frequency in Corpus	Probability
Noun(N)	366	0.5692
Verb(V)	80	0.1244
Adjective(ADV)	12	0.0187
Adverb(ADV)	182	0.2830
Others	3	0.0047
Sum	643	1

2. Methodology:

This method will be proposed to solve the crime problem. The police will track the suspects by searching the similarity of modus operandi in each case.

Stage 1: Pre-processing phrase:

There are two steps which are tokenization and tagging of words.

Tokenization:

Tokenization is the task of "chopping" up a sentence into pieces, called tokens, and at the same time eliminates certain characteristics, such as punctuation (Manning and Raghavan, 2008). An example of tokenization process can be simulated based on sentence in Fig. 1. The sample words after tokenizing are shown in Table 2.

e.g.: "Pada 21/09/2011 jam lebih kurang 9.00 pagi semasa berada dirumah sewa alamat Lot 2193 Kampung Banggol Jalan Mengabang Telipot 21030 Kuala Terengganu,ketika itu saya baru bangun dari tidur, saya mencari telefon bimbit tiada."

Fig. 1: Sample of sentence.

Table 2: Words after tokenization process.

Words	
pada	Banggol
21	Jalan
09	Mengabang
2011	Telipot
jam	21030
lebih	Kuala
kurang	Terengganu
9	ketika
00	itu
pagi	saya
semasa	baru
berada	bangun
di	dari
rumah	tidur,
sewa	saya
alamat	mencari
Lot	telefon
2193	bimbit
Kampung	tiada

Tagging:

Words will be checked from the Malay dictionary to tag their classes which are noun, verb, adjective and adverb (Omar, 1993). Examples of terms that have gone through the tagging processes are shown in table 3.

Table 3: Words after tagged.

Tagged Word			
Words	Tag	Words	Tag
pada	4	Jalan	1
21	1	Mengabang	1
09	1	Telipot	1
2011	1	Mengabang	1
jam	1	Telipot	1
lebih	4	21030	1
kurang	4	Kuala	1
9	1	Terengganu	1

Journal of Applied Science and Agriculture, 9(20) Special 2014, Pages: 1-6

00	1	ketika	4
pagi	1	itu	1
semasa	4	saya	1
berada	2	baru	3
di	4	bangun	1
rumah	1	dari	4
sewa	1	tidur,	2
alamat	1	saya	1
Lot	1	mencari	2
2193	1	telefon	1
Kampung	1	bimbit	1
Banggol	1	tiada	2
1=Noun 2=Verb 3=Adjective 4=Adverb			

A General algorithm for tagging using the knowledge base:

1) Read the first sentence to the end
2) Check the word in the table compound words
3) If step (2) failed to get the words,
a. Check the word in the table of reduplicate words
4) If step (3) failed to get the word,
a. Check the word in the table of "kata luar biasa"
5) If step (4) failed to get the words,
a. Check the word in the table of "kata Pandu"
6) If step (5) failed to get the word,
a. Check the word in the table of root words
7) If step (6) failed to get the word,
a. Check the word in the table of affix words
8) If step (7) failed to get the words,
a. Check the first letter of the word, if the first letter of the word has a capital letter, the list of candidates is a noun
9) If step (8) failed to get the words,
a. Check whether the word is a numerical, if the word is numerical, the list of candidates is a noun
10) If step (9) failed to get the words,
a. The word will not be tagged

End

Stage 2: Modus operandi extraction:

This stage will formulate rules that can be used to extract modus operandi from free texts. A few combinations of phrases will be selected as modus operandi. This stage also evaluates the performance of the modus operandi retrievals. There are three components (or other suitable term) of modus operandi which are:

Noun + Verb
Verb + Noun
Noun + Adjective

Adverb is not a part of the components because most adverbs are common words with no semantic and do not aggregate relevant information to the task. Only the meaningful and significant data or elements only will be extracted from the corpus (Mohd and Ali, 2011). An example of this process can be simulated based on sentences in Fig. 2. The table 4 shows a new element of corpus after further extraction. The Fig. 3 shows the overall of information retrieval system process from stage 1.

> e.g.: "Semasa saya sudah berada di dalam kedai, tiba-tiba datang seorang lelaki dari arah belakang saya dan menutup mulut saya sambil mengacukan pisau ke leher saya. Dua orang lagi rakan perompak itu mengambil semua wang yang berada di dalam laci cabinet hitam"

Fig. 2: Sample of sentences.

Table 4: Modus operandi extraction.

ID	Modus	Rules
1	saya berada kedai tiba-tiba datang lelaki arah saya menutup mulut saya mengacukan pisau leher saya	NOUN + VERB + NOUN + VERB + VERB + NOUN + VERB + NOUN + VERB + NOUN + NOUN + VERB + NOUN + NOUN + NOUN
2	2 orang rakan perompak itu mengambil wang berada laci kabinet hitam	NOUN + NOUN + NOUN + NOUN + NOUN + VERB + NOUN + VERB + NOUN + NOUN + ADJECTIVE

Journal of Applied Science and Agriculture, 9(20) Special 2014, Pages: 1-6

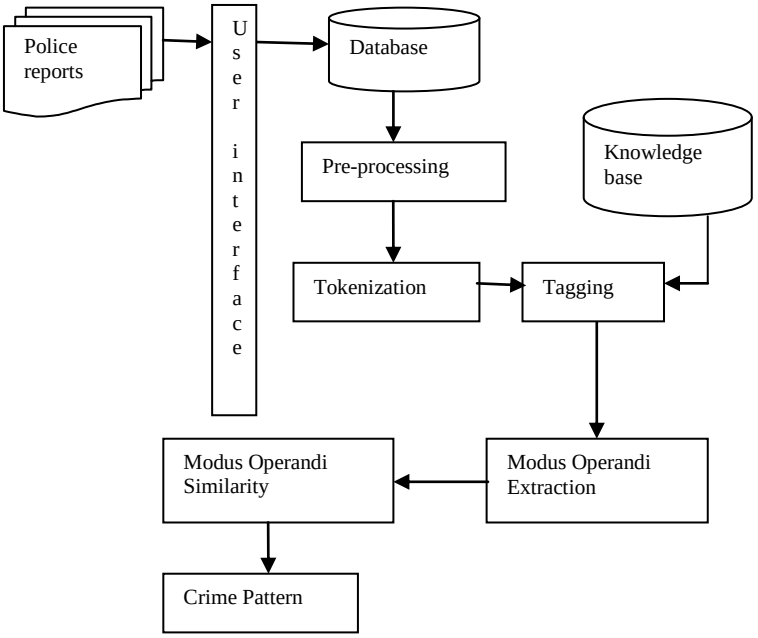

Fig. 3: Process in information retrieval system.

3. Retrieval Performance Evaluation:

In an experiment that was conducted, there are 3 important aspects of measurement which are recall, precision and F-measure. Recall and precision are assessments of measurement that have been used in information retrieval since before (Manning and Raghavan, 2008). To evaluate the performance of modus operandi extraction, precision and recall rate are used (Wong and Ziarko, 1987). The standard IR evaluation is used to evaluate the algorithm. The precision gives the metric percentage of the number of relevant documents retrieved to the documents retrieved.

Precision rate is the proportion of the retrieved documents which are relevant (Manning and Raghavan, 2008). Additionally, precision can be increased by narrowing down the queries.

$$\text{Precision} = \frac{\text{Number of documents retrieved and relevant}}{\text{Total documents retrieved from collection}}$$

Recall rate is the proportion of all relevant documents that have been retrieved (Manning and Raghavan, 2008). In order to raise recall, the queries should be broadened.

$$\text{Recall} = \frac{\text{Number of documents retrieved and relevant}}{\text{Total relevant documents from collection}}$$

The F-measure of the system is defined as the weighted harmonic mean of its precision and recall

$$\text{F-Measures} = \frac{2\text{Precision.Recall}}{\text{Precision} + \text{Recall}}$$

Before evaluation process, the system will be automatically extract the modus operandi and then matching the similarity between documents. Lastly, the system will display and rank the result obtained from evaluation as shown in Fig. 4.

RESULT AND DISCUSSION

A set of 10 police narrative reports in Malay language has been used as the test of corpus. In order to measure the effectiveness of the system, standard measures of evaluation, namely precision, recall, and the F-measure are used to test the system. This is done manually by the experimenter.

The system calculates the weight for each term in the database, represents the similarity of a set of modus operandi between each record and ranks them for investigation. Every report in the database is then rank-ordered according to the similarity scores. From the similarity score, it can show any crime pattern in the investigation such as the pattern can show whether any criminals used the similar modus operandi for several crimes of the same crime type. The system has a high ability to link crimes accurately to suspects according to their rank scores. The system has been tested by using 10 data reports.

Journal of Applied Science and Agriculture, 9(20) Special 2014, Pages: 1-6

The recall, precision and F-measure for each of the police reports were computed and their averages were summarized in Table 6. The best F-measure is 0.960 with recall and precision at 1.0 and 0.933 for retrieval performance of modus operandi extraction respectively. The experiment results show that the proposed system performs slightly better in retrieving modus operandi compared to the previous research.

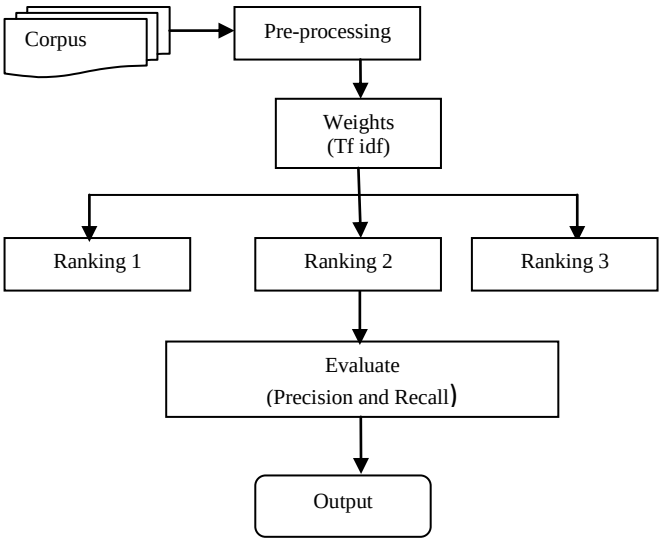

Fig. 4: Process to rank and evaluate the results.

The experimental results in Table 5 show results obtained after calculating precision for each query using method 1 and method 2. Method 2 produces better results compared to method 1. Based on Table 5, eight queries received positive effects and only two queries were negatively impacted. The Fig. 5 represents 13.86% of queries were increased.

Method 1- Evaluate without modus operandi extraction
Method 2- Evaluate with modus operandi extraction

Table 5: Precision for each query.

Query	Method 1	Method 2	% increment
1	0.2722	0.2876	5.6576
2	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000
5	0.1864	0.2159	15.8262
6	0.4031	0.3919	-2.7785
7	0.3120	0.4793	53.6218
8	0.0993	0.0991	-0.1712
9	0.1544	0.2570	66.4508
10	0.0000	0.0000	0.0000
Average	**0.14274**	**0.17308**	**13.86067**

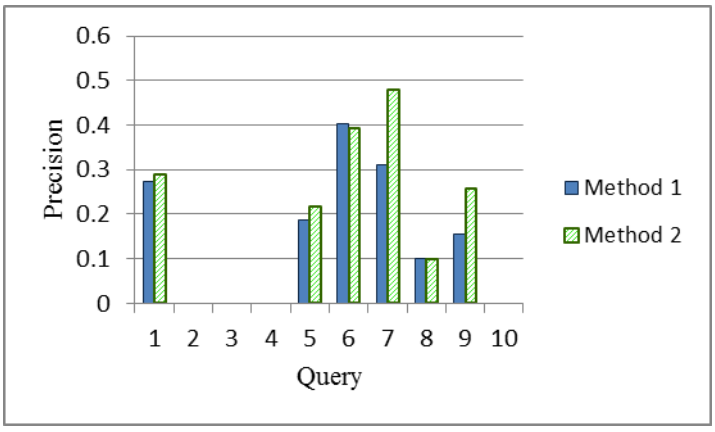

Fig. 5: Comparison of precision between queries.

Table 6: Results of precision, recall and f-measure for 10 queries.

Query	Number of correct entities extracted by system	Number of total entities extracted by system	Number of entities extracted by human	Precision	Recall	F-Measures
1	12	13	15	0.923	0.800	0.857
2	12	13	12	0.923	1.000	0.960
3	11	13	11	0.846	1.000	0.917
4	12	13	12	0.923	1.000	0.960
5	7	8	7	0.875	1.000	0.933
6	5	7	6	0.714	0.833	0.769
7	14	15	17	0.933	0.824	0.875
8	7	8	7	0.875	1.000	0.933
9	11	12	12	0.917	0.917	0.917
10	10	12	11	0.833	0.909	0.869
Average				0.876	0.928	0.890

4. Conclusion:

In this paper we have described modus operandi extraction in information retrieval research on Malay police reports. Based on the result in table 6 and table 5, we can accept the hypothesis that modus operandi extraction can improve retrieval effectiveness significantly.

REFERENCES

Chau, M., J.J. Xu, *et al*., 2002. Extracting meaningful entities from police narrative reports. Proceedings of the 2002 annual national conference on Digital government research. Los Angeles, California, Digital Government Society of North America: 1-5.

Douglas, J., A.W. Burgess, *et al*., 2011. Crime Classification Manual: A Standard System for Investigating and Classifying Violent Crimes, Wiley.

Fosdick, R.B., 1915. "The Modus Operandi System in the Detection of Criminals." Journal of the American Institute of Criminal Law and Criminology, 6(4): 560-570.

Hazelwood, R.R., *et al*., 2004. "Linkage analysis: modus operandi, ritual, and signature in serial sexual crime." Aggression and Violent Behavior, 9(3): 307-318.

Manning, C.D., P. Raghavan, *et al*., 2008. Introduction to Information Retrieval, Cambridge: Cambridge University Press

Mohd, M. and N.M. Ali, 2011. An Interactive Malaysia Crime News Retrieval System. Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on: 220-223.

Omar, A.H., 1993. Word Classes in Malay. Essays on Malaysian Linguistic. Malaysia, Dewan Bahasa dan Pustaka, Ministry of Education Malaysia, Kuala Lumpur, 13: 162-174.

Salton, G., 1983. Introduction to modern information retrieval. New York, McGraw-Hill.

Valerie Pottie Bunge, H.J. and a. T.A. Baldé, 2005. "Crime and Justice Research Paper Series Exploring Crime Patterns in Canada."

Wong, S.K.M., W. Ziarko, *et al*., 1987. "On modeling of information retrieval concepts in vector spaces." ACM Trans. Database Syst., 12(2): 299-321.