
UNIVERSITI PUTRA MALAYSIA

# INTERROGATIVE ELEMENTS AND VERB-NOUN RANKING FOR CRIMINAL CHATTING FORENSICS

SITI HANOM BINTI MARJUNI

FSKTM 2011 12

# INTERROGATIVE ELEMENTS AND VERB-NOUN RANKING FOR CRIMINAL CHATTING FORENSICS

**by**

**SITI HANOM BINTI MARJUNI**

**Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Doctor of Philosophy**

**October 2011**

© COPYRIGHT UPM

**DEDICATION**

…whom who share their love, soul and strength

*Zainal Abidin Bin Abu Hassan*

and great children:

*Nurul Fathihah*
*Muhammad Arif*
*Nurul Aina*
*Muhammad Aiman*
*Ammar Firdaus*

© COPYRIGHT UPM

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

**INTERROGATIVE ELEMENTS AND VERB-NOUN RANKING FOR CRIMINAL CHATTING FORENSICS**

**By**

**SITI HANOM BINTI MARJUNI**

**October 2011**

**Chairman : Professor Ramlan Mahmod, PhD**

**Faculty : Computer Science and Information Technology**

The rapid development in computer and Internet technology through cyber space as well as communicating in the real world globally has brought a tremendous increase in cyber-crimes. Chat is an easy and fast way to communicate interactively without having face to face conversation. It incorporates various types of animation including human emotions which is factored during the chat session.

The challenges faced in this research is that the chatter develops his own language, a language where speed prevails over correct spelling through short form words, thus contributing to greater interactivity which is defined as unstructured and colloquial. Furthermore, chat utterance is built from a simple sentence which normally contains only one clause. Hence without the subject of the object of sentence structures, each of the words gives variable meaning in the criminal conversation.

© COPYRIGHT UPM

Thus, this research is aimed to solve the problem towards finding out the meaning of the text behind the messages. Preprocessing is the cleaning process before proceeding to the actual processes. Criminal identification is the first process which requires three steps. Firstly, tokenization is done to assign each lexical automatically with a corresponding serial number in every suspect's and victim's utterance. The second step is to tag the lexical with the interrogative elements together with Part-of-Speech (POS). In this process, the combination of interrogative elements and verb-noun ranking is considered in the experiment. Thirdly, criminal investigation by using the Protégé criminal ontology is used to investigate all the evidences behind the text of utterances. Finally, the reporting is produced in the Digital Evidence Form (Casey, 2004), as well as the validation and satisfaction of methodology implemented in the research are done by a forensic lawyer.

The chatting corpus consists of 3,098 suspects' and victims' utterances with 16,278 words, collected from nine criminal chatting cases. For criminal identification, two processes of identifying are considered. The identifying is done by the system and an expert. The results obtained from the system and expert show that the criminal identification is almost similar. However, the sign test to get the significance differences between the number of interrogative words extracted by the system and an expert shows that the system has an ability to function as an identifier of the interrogative elements which extracts the verb-noun ranking in criminal forensics. Furthermore, the 40 respondents are measured in interpolation precision. The interpolated precision shows that all of the interrogative elements meet the higher average percentage where the *why* and the *how* represent the highest percentages.



© COPYRIGHT UPM

Furthermore, the *COps* prototype system is produced to investigate the words behind the text. About 128 respondents of three backgrounds of qualification are investigating 5,175 words (31.8%) of words in the criminal chatting corpus and the values of recalls and precisions are measured. The interpolated precision shows that the backgrounds of respondents play a key role in the experiment of criminal investigation. Finally, the criminal chatting evidence as well as the validation of the methodology implemented in the research is carried out by a forensic lawyer.

© COPYRIGHT UPM

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**KEDUDUKAN ELEMEN INTEROGATIF DAN KATA KERJA-KATA NAMA BAGI FORENSIK PERBUALAN JENAYAH (*CHATTING*)**

**Oleh**

**SITI HANOM BINTI MARJUNI**

**Oktober 2011**

**Pengerusi : Profesor Ramlan Mahmod, PhD**

**Fakulti : Sains Komputer dan Teknologi Maklumat**

Perkembangan pesat dalam komputer dan teknologi Internet melalui ruang siber seiring dengan komunikasi dalam dunia nyata di seluruh dunia telah mengakibatkan peningkatan yang dahsyat di dalam jenayah siber. *Chat* adalah satu perantara komunikasi yang mudah dan pantas secara interaktif tanpa mengadakan perbualan bersemuka. Ia memasukkan pelbagai jenis animasi termasuk emosi manusia semasa sesi perbualan.

Cabaran yang dihadapi dalam penyelidikan ini adalah "*chatter*" membangunkan bahasanya sendiri, bahasa di mana kepantasan mengatasi ketepatan ejaan menerusi singkatan perkataan. Oleh itu, ia menyumbang kepada interaktiviti yang ditakrifkan sebagai ayat yang tidak berstruktur dan ditakrifkan sebagai bahasa pasar. Tambahan pula, pertuturan perbualan tersebut dibina daripada ayat mudah di mana biasanya ia hanya mengandungi satu klausa sahaja. Dengan itu, tanpa subjek bagi sesuatu objek



© COPYRIGHT UPM

di dalam struktur ayat tersebut dan setiap perkataan memberi pelbagai maksud di dalam perbualan jenayah.

Oleh yang demikian, penyelidikan ini bertujuan untuk menyelesaikan masalah dalam mencari makna teks yang berselindung di sebalik mesej tersebut. Prapemprosesan adalah proses pembersihan sebelum meneruskan proses sebenar. Pengenalpastian jenayah adalah proses pertama yang memerlukan tiga langkah. Pertama, *tokenization* dilakukan untuk menentukan setiap leksikal secara automatik dengan menyelaraskan nombor siri pada setiap perbualan suspek dan mangsa. Langkah kedua adalah dengan melabel leksikal dengan elemen interrogatif bersama dengan *Part-of-Speech* (POS). Dalam proses ini, kombinasi antara elemen interogatif dan kedudukan kata kerja-kata nama diambil kira di dalam eksperimen. Ketiga, penyiasatan jenayah dengan menggunakan ontologi jenayah Protégé digunakan untuk menyiasat kesemua bukti yang terlindung di sebalik pertuturan teks. Akhirnya, laporan dikeluarkan dalam Borang Bukti Digital (Casey, 2004), serta pengesahan dan kepuasan metodologi yang digunakan untuk penyelidikan ini dibuat oleh seorang peguam forensik.

Korpus perbualan mengandungi 3,098 rekod pertuturan suspek dan mangsa dengan 16,278 patah perkataan yang dikumpul daripada sembilan kes perbualan jenayah. Untuk pengenalpastian jenayah, dua proses pengenalpastian dilakukan. Pengenalpastian dilakukan oleh sistem dan seorang pakar. Keputusan yang diperolehi daripada sistem dan pakar menunjukkan bahawa pengenalpastian adalah hampir sama. Walau bagaimanapun, ujian tanda untuk mendapatkan perbezaan jumlah antara perkataan interogatif yang diekstrak oleh sistem dan pakar menunjukkan bahawa sistem mempunyai keupayaan berfungsi sebagai pengenal pasti ke atas elemen interogatif yang mengekstrak kedudukan kata kerja-kata nama



© COPYRIGHT UPM

dalam penjenayahan forensik. Selanjutnya, 40 responden telah diukur dalam *interpolation precision*. *Interpolated precision* menunjukkan bahawa kesemua element interogatif memenuhi peratusan purata yang tinggi di mana kenapa (*why)* dan bagaimana (*how)* mewakili peratusan yang tertinggi.

Tambahan pula, sistem prototaip sistem *COps* dihasilkan untuk menyiasat perkataan-perkataan di sebalik teks. Sebanyak 128 responden daripada tiga latar belakang kelayakan membuat penyiasatan terhadap 5,175 perkataan (31.8%) daripada perkataan di dalam korpus perbualan penjenayahan dan nilai *recalls* dan *precisions* adalah dikira. *Interpolated precision* menunjukkan bahawa latar belakang responden memainkan peranan yang penting di dalam eksperiman penyiasatan penjenayahan. Akhir sekali, bukti perbualan jenayah serta pengesahan metodologi yang dilaksanakan di dalam penyelidikan disahkan oleh seorang peguam forensik.

© COPYRIGHT UPM

## ACKNOWLEDGEMENTS

Bismillahhirohmanirohim in the name of *Allah*, the most merciful and most compassionate. Praise to be *Allah S.W.T.* for giving me the strength, patience and motivation to complete this research.

My deepest appreciation and gratitude is dedicated to the research committee lead by Professor Dr. Ramlan Mahmod for his knowledge, motivation, continuous encouragement and constant guidance. Many thanks to co-supervisors Professor Dr. Abd. Azim Abd. Ghani and Professor Dr. Abdullah Mohd. Zin for their expertise and intellectual experiences. I am also thankful to Dr. Fatimah Sidi for her guidance and help towards the end of the research. To Dr. Aida Mustapha for her ideas in the early stages of research are most appreciated. Many thanks to Mr. Mohamed Nasharudin and Mdm. Petrina from the Cambridge English for Life for their expertise in Interrogative Elements and Part-of-Speech tagging for this research. Last but not least, to Mrs. Norriza Hussin for her intellectual experiences as a forensic lawyer.

I am grateful to Jabatan Perkhidmatan Awam, Malaysia for the study leave, and scholarship to pursue this research. Special thanks for their moral support, caring and encouragement to Dr. Puteri Suhaiza directly or indirectly during my completion of study.

My deepest thanks to my husband, children, beloved mother and sisters for their understanding, support and patience.

© COPYRIGHT UPM

## APPROVAL

I certify that an Examination Committee has met on _____ to conduct the final examination of Siti Hanom binti Marjuni on her Doctor of Philosophy thesis entitled "Interrogative Elements and Verb-Noun Ranking For Criminal Chatting Forensics" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Mohd. Hasan Selamat
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Norwati Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Hjh. Fatimah Dato Ahmad, PhD
Professor
Dean
Centre for Graduate Studies
Universiti Pertahanan Nasional Malaysia (UPNM)
(Internal Examiner)

Emeritus Prof. Dr. Frederick Charles Piper
Royal Holloway (University of London)
Information Security Group, Egham, Surrey
TW20 OEX, United Kingdom
(External Examiner)

_____

**PhD**
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:



© COPYRIGHT UPM

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

**Ramlan Mahmod, PhD**
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

**Abdul Azim Abd. Ghani, PhD**
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

**Abdullah Mohd Zin, PhD**
Professor
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
(Member)

© COPYRIGHT UPM

**BUJANG BIN KIM HUAT, PhD**
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

## DECLARATION

I declare that the thesis is my own work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institutions.

**SITI HANOM BINTI MARJUNI**

Date: 17th October 2011

© COPYRIGHT UPM

## TABLE OF CONTENTS





© COPYRIGHT UPM

© COPYRIGHT UPM

© COPYRIGHT UPM