CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

MONTHIYA SAPAN

A THEMATIC PAPER SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER DEGREE OF SCIENCE (INFORMATION TECHNOLOGY MANAGEMENT) FACULTY OF GRADUATE STUDIES MAHIDOL UNIVERSITY 2017

CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

MONTHIYA SAPAN 5837645 EGIT/M

M.Sc. (INFORMATION TECHNOLOGY MANAGEMENT)

THEMATIC PAPER ADVISORY COMMITTEE: SOTARAT THAMMABOOSADEE, Ph.D., TAWEESAK SAMANCHUEN, Ph.D.,

ABSTRACT

Different types of internet attack have currently increasing exponentially. One of internet attacks that has been used for many years. Currently, Phishing and number of internet users who have been attacked by phishing has also been increasing; this trend has causes a large scale of losses to victims. This research studies contents in phishing email only. Text classification system was applied for analyzing phishing email contents based on the specified eight features, including studying campaign messages that appeared in phishing emails and determining the words used in those messages. The dataset of this study is provided by www.419scam.org and the results were used to create a decision tree. The overall model performance is greater than 80% when binary occurrence is used as an indicator. The decision making rules are further analyzed facilitated by the association rules discovery method to determine the relation of features for creating the final phishing determination model. When analyzing the relationship of features, the relation rule was obtained and the emails that included Messages which notified the recipients that the e-mail is confidential, Messages which rushed the recipients to take an immediate action, and Message which asked for help, were considered is be Phishing E-Mail. This research could help in analyzing email contents and determining whether there is a risk of them being phishing emails. This could be a part of reducing risk of being attacked by email phishing. In the future, it is therefore suggested the research should be extended to analyzing other email components such as the domain reliability and files attached in email.

KEY WORDS: PHISHING EMAIL/ CRAFTING TEXT/ DATA MINING/ TEXT MINING/ DECISION TREE/ MODULAR MODEL

63 pages

แบบจำลองสำหรับจำแนกข้อความหลอกลวงแบบหน่วยย่อยตามเนื้อหาสำหรับการตรวจจับอีเมล์ฟิชชิ่ง

CONTENT-BASED MODULAR CRAFTING TEXT CLASSIFICATION MODEL FOR PHISHING EMAIL DETECTION

มณฑิยา สาปาน 5837645 EGIT/M

วท.ม. (การจัดการเทคโนโลยีการสารสนเทศ)

คณะกรรมการที่ปรึกษาสารนิพนธ์ : โยฑศ์รัตต์ ธรรมบุษดี, Ph.D., ทวีศักดิ์ สมานชื่น, Ph.D.,

บทคัดย่อ

การโจมตีจากภัยคุกคามต่างๆบนโลกอินเทอร์เน็ตในปัจจุบัน นับว่ามีความรุนแรงที่ทวีคูณมากขึ้นเรื่อยๆ และมีรูปแบบการโจมตีที่หลากหลายมาก Phishing ก็ถือว่าเป็นภัยคุกคามรูปแบบหนึ่งที่มีมานานและในปัจจุบันยังคงพบว่ามีผู้ที่ถูกโจมตีด้วย Phishing เป็นจำนวนมากขึ้นทุกปี ซึ่งสร้างความเสียหายให้แก่ผู้ตกเป็นเหยื่อเป็นอย่างมาก ในงานวิจัยเล่มนี้ผู้วิจัยได้ทำการศึกษาเฉพาะเนื้อหาข้อความภายใน Phishing E-Mails เท่านั้นโดยใช้กระบวนการ Text Classification System ในการวิเคราะห์เนื้อหาข้อความภายใน Phishing E-Mail ตาม Features ทั้ง 8 Features ที่ได้กำหนดไว้ โดยเก็บรวบรวมข้อมูล Phishing E-Mails จาก www.419scam.org ผลลัพธ์ที่ได้จะเป็นกฎที่ช่วยในการตัดสินใจ และเมื่อนำมาวัดประสิทธิภาพความแม่นยำของตัวชี้วัดพบว่า ตัวชี้วัดที่มีความแม่นยำสูงที่สุดคือ Binary Term Occurrences ซึ่งมีค่าความแม่นยำกว่า 80% หลังจากนั้นจะนำกฎการตัดสินใจที่ได้ไปวิเคราะห์ต่อด้วยการใช้ Association Rules เพื่อวิเคราะห์หาความสัมพันธ์ของแต่ละ features ซึ่งจะนำไปสู่การสร้างกฎสุดท้ายที่ใช้ในการพิจารณา Phishing E-Mail ดังตัวอย่างกฎที่ได้เช่น ถ้าอีเมล์ปรากฏข้อความตาม Feature ข้อความที่มีการแจ้งว่าอีเมล์นี้ต้องเป็นความลับเท่านั้น, ข้อความที่มีลักษณะเน้นย้ำให้รีบดำเนินการทันที และข้อความที่มีการขอความช่วยเหลือ ถือว่าเป็น E-Mail Phishing เป็นต้น ซึ่งงานวิจัยชิ้นนี้จะช่วยในการวิเคราะห์ข้อความภายใน E-Mail ว่ามีความเสี่ยงที่จะเป็น E-Mail Phishing หรือไม่ และเป็นส่วนหนึ่งที่จะลดความเสี่ยงจากการถูกโจมตีโดย E-Mail Phishing อีกด้วยในอนาคตควรจะขยายขอบเขตงานวิจัยด้วยการนำองค์ประกอบอื่นๆภายใน E-Mail เช่น ความน่าเชื่อถือของ Domain, ประเภทไฟล์ที่แนบมาใน E-Mail มาพิจารณาเพิ่มเติม

63 หน้า