TEXT SPAM MESSAGES CLASSIFICATION USING ARTIFICIAL IMMUNE SYSTEM (AIS) ALGORITHMS

Nurul Fadhilah Binti Sulaiman
(Matric No. 3130271)

Thesis submitted in fulfillment for the degree of
MASTER OF SCIENCE

Faculty of Science and Technology
UNIVERSITI SAINS ISLAM MALAYSIA
Nilai

June 2016

AUTHOR BIOGRAPHY

Nurul Fadhilah binti Sulaiman (3130271) was born on the August, 5 1990. She is currently residing at No. 50, Felcra Keruak, 22010 Jerteh, Terengganu. She previously was a student of Universiti Sains Islam Malaysia (USIM) and obtained Bachelor of Computer Science (Information Security and Assurance) from the Faculty of Science and Technology. Presently she studies Master of Science within the same institution. In this Masters journey, she managed to publish two papers related to this field titled "Integrated Mobile Spam Model Using Artificial Immune System Algorithms" in Knowledge Management International Conference 2014 (KMICe2014) and another paper, "A New SMS Spam Detection Method Using Both Content-Based and Non Content-Based Features", in 2015 2nd International Conference on Communication and Computer Engineering (ICOCOE'2015). She can be contacted via fadhilahsulaiman90@gmail.com.

ACKNOWLEDGEMENTS

Praise is upon Allah the Most Gracious Most Merciful for giving me the strength to face all challenges and complete this study.

Firstly, I would like to express my deep and sincere gratitude to my supervisor, Dr. Mohd Zalisham Jali because help me a lot during completing this thesis by giving thoughtful ideas, good resources and references, explain clearly what should i do and give good advice that can improve my skills in learning and writing. He also provided encouragement, enthusiasm and a lot of support until I managed to finish my thesis.

I also would like to thank to my beloved family especially my mother, Ruhani Binti Mamat and my father, Sulaiman Bin Embong for helping me a lot during my study in USIM. Their support gives me courage to face challenge and make me become a mature person and I hope that I have strength to face real life by their supporting.

Lastly, thanks to all my friends and colleagues in USIM for their helping and giving good idea and discussion about this research. Finally, I pray to Allah SWT to accept this work as a sincere deed.

ABSTRAK

Masalah mesej spam agak membimbangkan terutama bagi pengguna telefon mudah alih kerana statistik menunjukkan peningkatan isu-isu yang berkaitan dengan mesej teks spam setiap tahun, walaupun banyak usaha telah diperkenalkan bagi mengurangkan risiko spam. Penghantar spam memilih SMS sebagai sasaran utama mereka untuk spam kerana SMS dikatakan sebagai komunikasi penting di kalangan pengguna. Masalah seperti algoritma yang tidak cekap, kesedaran pengguna dan spam yang berisiko tinggi masih mendominasi dan mencabar. Selain itu, jenis SMS spam yang dihantar oleh penghantar spam memberikan kami persoalan mengenai jenis mesej yang kebanyakannya dihantar oleh mereka. Setelah dinyatakan cabaran tersebut, penyelidikan kami memberi tumpuan kepada fasa kedua iaitu klasifikasi (atau dikenali sebagai kelompok). Objektif utama kajian ini adalah untuk mengkaji lagi hubungan antara "Artificial Immune System (AIS)" dan "Biology Immune System (BIS) berkaitan dengan pengesanan, klasifikasi dan penentuan tahap bagi mesej spam, untuk mencadangkan satu kaedah bagi mempertingkatkan kelompok mesej spam menggunakan gabungan "Clonal Selection" dan "Immune Network Theory" dan akhir sekali untuk menjalankan dan menilai algoritma yang telah dicadangkan. Model pengurusan spam yang diilhamkan daripada BIS yang dinamakan "Integrated Mobile Spam Model" (IMSM) telah diperkenalkan. Model ini terdiri daripada tiga fasa iaitu pengesanan, klasifikasi dan penentuan tahap, dan setiap fasa hanya menggunakan algoritma AIS diilhamkan daripada BIS. BIS mempunyai keupayaan untuk melindungi dan mempertahankan badan kita daripada bakteria atau virus yang menyerang kami, maka teori tersebut juga boleh digunakan untuk telefon mudah alih bagi melindungi kita daripada mesej spam. Pengelasan adalah proses untuk mengumpulkan mesej spam kepada beberapa kumpulan. Dengan melakukan fasa ini, ia membantu kita untuk mengenal pasti kumpulan mana yang mempunyai kejadian mesej spam yang lebih tinggi dan sentiasa dihantar oleh penghantar spam selain dapat membantu dalam fasa penentuan tahap bagi menentukan tahap bahaya untuk mesej spam. Sebuah algoritma baru yang dinamakan "Hybrid Immune Clonal Network Algorithm" (HICNA) telah dicadangkan untuk kelompok mesej spam dan algoritma ini adalah gabungan "Clonal Selection" dan "Immune Network Theory". Tiga fasa terlibat dalam algoritma ini; fasa pertama adalah mengimbas mesej spam menggunakan kata kunci yang biasa manakala fasa kedua menggunakan kata kunci yang luar biasa. Pertimbangan pakar diperlukan untuk fasa terakhir bagi memastikan semua mesej spam dikumpulkan kepada kumpulan-kumpulan yang dikenal pasti. Beberapa eksperimen telah dijalankan untuk menguji prestasi dan kesahihan algoritma menggunakan sumber set data yang berbeza dan juga mengenal pasti kebolehgunaan dalam proses pengesanan. Hasil kajian kami menunjukkan bahawa kami berjaya memenuhi semua tiga objektif kajian ini dan algoritma cadangan kami boleh memberikan hasil yang lebih baik dan boleh diterima pakai untuk kelompok mesej spam kepada beberapa kumpulan. Di samping itu, ia menunjukkan keupayaan AIS algoritma untuk proses kelompok.

KATAKUNCI: Artificial Immune System (AIS), Clonal Selection, Immune Network Theory, Klasifikasi, Kelompok, SPAM, SMS.

ABSTRACT

The problem of spam messages is quite worrying especially for mobile users because statistics show increasing issues albeit many efforts have been introduced to reduce the risk of spam. Spammers chose SMS as their main target for spamming because SMS is considered as an important communication among them. Problems such as inefficient algorithm, users awareness and high risk of spam are still dominating and challenging. Besides, the varieties of SMS spam sending by spammers giving us a question on the types of messages that are mostly sent by them. Having stated the aforementioned challenges, this research focuses on the second phase which is the classification (or known as clustering). The main objectives of this research are to study the relationship between Artificial Immune System (AIS) and Biology Immune System (BIS) related to spam detection, classification and severity determination, to propose an enhance method for clustering spam messages using the combination of Clonal Selection and Immune Network Theory and lastly to conduct and evaluate the proposed algorithms. A spam management model inspired from the ideology of BIS named Integrated Mobile Spam Model (IMSM) is introduced. This model consists of three phases which are detection, classification and severity determination, and each phase uses only AIS algorithms inspired from BIS. BIS has the capability to protect and defend the body from bacteria or virus that attacks us, so this theory can be applied to the mobile phone to protect from spam messages as well. Classification is the process to cluster spam messages into several groups. By doing this phase, it helps us to identify which group of spam messages that has higher occurrence and is always sent by spammers besides can help in the severity determination phase to determine the level of danger for spam messages. A new algorithm named "Hybrid Immune Clonal Network Algorithm" (HICNA) is proposed for clustering spam messages and this algorithm is a combination of Clonal Selection and Immune Network Theory. Three phases involved in this algorithm; phase one is scanning the spam messages using common keywords while phase two is using uncommon keywords. Expert judgement is needed for the last phase to ensure all spam messages are clustered into identified groups. A number of experiments have been conducted to test the performance and validity of the algorithm using different source of datasets and also to identify its usability in the detection process. The research results show that three defined objectives were fulfilled and the proposed algorithm gives better results in clustering spam messages into several groups. In addition, it shows the capability of AIS algorithm for the clustering process.

KEYWORDS: Artificial Immune System (AIS), Clonal Selection, Immune Network Theory, Classification, Clustering, SPAM, SMS.

MULAKHKHAS AL-BAHTH

مشكلة الرسائل غير المرغوب فيها مقلق خاصة لمستخدمي الهواتلأن الإحصاءاتتظهر على زايدة القضايا المتعلقة برسائل غير مرغوب فيها سنويا، وإن كان كثير من الجهود استحدثت للحد من خطر الرسائل غير المرغوب فيها، الاطر اختيار SMS كهدف رئيسي للسبام لأن يقال SMS كخطاب هاما فيما بينهم. مشاكل مثل عدم كفاءة الخوار زمية، وعي المستخدمين ومخاطر عالية من البريد المز عج لا تزال المسيطرة وتكن الصعبة. بجانب ذلك، نوعا من البريد المزعج SMS المرسلة من الاطر يعطينا سؤال حول أنواع الرسائل التي يتم إرسالها معظم منهم. وبعد أن ذكر التحديات المذكورة آنفا، وهذا نموذج إدارة البريد المز عج مستوحاة من فكر Biology Immune System (BIS) يدعى Integrated Mobile Spam Model (IMSM)قد عرض. هذا المودم يكون على ثلاث مراحل وهي كشف وتصنيف وتحديد الخطورة، كل مرحلة ويستخدم خوار زميات Artificial Immune System (AIS) الوحيدة مستوحاة من BIS. BIS لديها القدرة على حماية والدفاع جسمنا من البكتيريا أو الفيروسات التي تهاجم لنا، لذلك يمكن تطبيق نظريتها إلى الهاتف المحمول لحمايتنامن الرسائل غير المر غوب فيها كذلك. ويركز بحثنا على المرحلة الثانية وهو تصنيف (أو المعروف باسم التكتل). التصنيف هو عملية تجميع رسائل البريد المز عج الى عدة مجموعات . عن طريق القيام هذه المرحلة، أنها تساعدنا على تحديد أي مجموعة من الرسائل غير المر غوب فيها التي لديها أعلى حدوثها ويتم إرسالها دائما من الاطر وإلى جانب آخر تساعد في مرحلة تحديد شدة لتحديد مستوى الخطر لرسائل البريد المزعج. تقوم خوارزمية جديدة باسم "Hybrid Immune Clonal Network Algorithm (HICNA)" يقترح لتجميع رسائل البريد المزعج وهذه الخوارزمية هي مزيج من اختيار نسيلي ونظرية الشبكة المناعية. ثلاثة مراحل المشاركة في هذه الخوارزمية. المرحلة الأول بالالتقاط الرسائل غير المر غوب فيها باستخدام كلمات رئيسية مشتركة مع المرحلة والثانية يستخدم كلمات غير مألوفة. هناك حاجة إلى حكم الإنسان للمرحلة الأخيرة لضمان تتجمع كل الرسائل غير المرغوب فيها إلى مجموعات محددة. وقد أجريت العديدة من التجارب لاختبار أداء وصحة الخوار زمية باستخدام مصادر مختلفة من قواعد البيانات وأيضا تحديد قابليته للاستخدام في عملية الكشف. تظهر نتائج بحثنا إلى أن تمكنا من تحقيق كل الأهداف الثلاثة في هذا البحث وأنظمتنا المقتر حة تعطي أفضل نتائج وتكون قابلة للتطبيق لتجميع رسائل البريد المزعج الى عدة مجموعات. بالإضافة إلى ذلك، فإنه يدل على قدرة خوارزمية AIS لعملية التجميع.

الكلماتالرئيسية: Artificial Immune System (AIS), Clonal Selection, Network Theory Immune, SMS,SPAM , تصنيف, المجموعات,

TABLE OF CONTENTS

LIST OF TABLES

CONTENTS PAGE

LIST OF FIGURES

CONTENTS PAGE

LIST OF APPENDICES

CONTENTS PAGE

LIST OF ALGORITHMS

ABBREVIATION

aiNet	= Artificial Immune Network (an immune system based algorithm which combines the pre-processing and clustering procedures)
AIN	= Artificial Immune Network
AIS	= Artificial Immune System
AUC	= Area Under Curve
BIS	= Biology Immune System
CI	= Computational Intelligence
CIA	= Confidentiality, Integrity and Availability
CLONACH	= Clonal Selection Algorithm for Learning Concept Hierarchies
DCDD	= Dynamic Concept Drift Detection
DIT	= Dublin Institute Technology
DT	= Decision Tree
DTL	= Danger Theory based Learning
FFNN	= Feed Forward Neural Networks
FN	= False Negative
FP	= False Positive
GAHC	= Guided Agglomerative Hierarchical clustering
GCAIN	= Guided Clustering and Artificial Immune Network
GT	= Grumble Text
HAC	= Hierarchical Agglomerative Clustering
HICNA	= Hybrid Immune Clonal Network Algorithm
ICOCOE	= International Conference on Communication and Computer Engineering
IMSM	= Integrated Mobile Spam Mobile
KMICE	= Knowledge Management International Conference
k-NN	= k-Nearest Neighbour
LMT	= Logistic Model Tree
MHC	= Major Histocompatibility Complex
Msg	= Message
NB	= Naïve Bayes
PAM	= Partitioning Around Medoids
PCA	= Principle Component Analysis
ROC	= Receiver Operating Characteristics
SKMM	= Suruhanjaya Komunikasi dan Multimedia
SMS	= Short Message Service
SOM	= Self-Organizing Maps (SOM)
SVM	= Support Vector Machine
TF-IDF	= Term frequency-inverse document frequency
TN	= True Negative
TP	= True Positive
UCS	= Supervised Classifier System
URL	= Uniform Resource Locator
U.S	= United State
WEKA	= Waikato Environment for Knowledge Learning