

IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

SESI PENGAJIAN: 2016/2017

Saya NOR ANIS SYAZANA BINTI ZULKIFLI mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
	اونيور سيتي تيكنيكل مليسيا ملاك

Noris

(TANDATANGAN PENULIS)

Alamat tetap: NO 155 JALAN

MERANTI 1, TAMAN JERAI

FASA 2, 08300 GURUH, KEORAH

Tarikh: 25/8/2017

Nor Azman Bin Mat Ariff

(TANDATANGAN PENYELIA)

NOR AZMAN BIN MAT ARIFF

Nama Penyelia

Tarikh: 25/8/2017

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

NOR ANIS SYAZANA BINTI ZULKIFLI



**This report is submitted in partial fulfillment of the requirements for the
Bachelor of Computer Science (Computer Networking)**

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I hereby declare that this project report entitled

IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

is written by me and is my own effort and that no part has been plagiarized

without citations.



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

STUDENT

Anis

Date: 25/8/2017

(NOR ANIS SYAZANA BINTI ZULKIFLI)

اونيسور سیتی تيكنیکل ماليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

SUPERVISOR

Norazman

Date: 25/8/2017

(EN. NORAZMAN BIN MAT ARIFF)

DEDICATION

This final project is dedicated to my beloved parents Mr. Zulkifli BIn Yahya and Mrs. Salmiah Binti Md Ali for their endless support and helps throughout the year in completing my studies and always pray the best for me.

I also dedicate this dissertation to my friends who have supported me throughout this project. I will always appreciate all they have done, especially to Sarah Syamimi and Amizah Aida for helping me from FYP 1 until FYP 2.

Special dedication to my supervisor, Encik Azman Bin Mat Ariff who has guided and never get tired in supports, helping and motivating me while making the progress for this project.

Last and but not least, do not forget to my evaluator, P.M. Dr. Faizal Bin Abdollah for his positive advices and feedback about this project,

ACKNOWLEDGEMENT

First and foremost, I would like to express my thanks to Allah S.W.T for giving me strength for completing the whole process of this project. Without His blessings, I am sure this project cannot complete.

Special thanks to my supervisor, Enick Azman Bin Mat Ariff for his guidance, constant supervision and kindness in completing this project. Without his guide, I am surely would not know the right trail.

I would also like to thanks to my family members who always support and pray the best for me and constantly give me motivation throughout my project. Besides, thanks to my colleague who have been very helpful during project development.

Thank you.



ABSTRACT

Spam is usually sent in the form of a text message in which specific words in the message can be used by spam blocking software to prevent the message from reaching our Inbox. With image spamming, the text is placed inside the image in an effort to bypass the spam blocking software. Since images are considered a normal part of a recipient's email message and the spam blocking software is mainly designed for text, the spammer is successful in getting the message to reach our Inbox. One of the previous study has used frequent item mining technique in order to extract features. However, the researcher only considered minimum weightage scheme for feature weight assignment. Thus, the main objective of this project is to generate feature vectors using weightage schemes, which is a maximum weightage scheme assignments. For further investigation, an ensemble method also is applied for weightage schemes. Firstly, sift descriptor is used to represent an image. Sift keypoint make the process of clustering and each of keypoint were collected the cluster number. Bag-of-word feature vectors are generated directly. Then, the frequent items are generated from BOW feature vector using of weightage schemes. SVM classifier is used as image spam classifier to train and produce a models for scheme. Lastly, an ensemble method used for models to obtain the best models. The significant contribution for this project is using the weightage schemes of maximum of the frequent item mining (FIM) technique to generate a feature vectors that is capable of detecting image with better.

ABSTRAK

Spam biasanya dihantar dalam bentuk mesej teks di mana kata-kata khusus dalam mesej itu boleh digunakan oleh perisian menyekat spam untuk menghalang mesej daripada mencapai Peti Masuk kami. Dengan spamming imej, teks diletakkan di dalam imej dalam usaha untuk memintas perisian menghalang spam. Oleh kerana imej dianggap sebagai sebahagian daripada mesej e-mel penerima dan perisian penyekatan spam terutamanya direka untuk teks, spammer berjaya mendapatkan mesej untuk mencapai Peti Masuk kami. Salah satu kajian terdahulu telah menggunakan teknik perlombongan item kerap untuk mengekstrak ciri-ciri. Walau bagaimanapun, penyelidik hanya menganggap skema weightage minimum untuk tugas berat ciri. Oleh itu, matlamat utama projek ini adalah untuk menghasilkan vektor ciri menggunakan skema weightage, yang merupakan tugas skema pemberat weight maksimum. Untuk siasatan lanjut, kaedah ensemble juga digunakan untuk skema weightage. Pertama, penyaring deskriptor digunakan untuk mewakili imej. Menapis keypoint membuat proses clustering dan setiap keypoint dikumpulkan nombor kluster. Vektor ciri beg-perkataan dihasilkan secara langsung. Kemudian, item kerap dijana daripada vektor ciri BOW menggunakan skema weightage. Pengelas SVM digunakan sebagai pengelas imej spam untuk melatih dan menghasilkan model untuk skema. Akhir sekali, kaedah ensemble yang digunakan untuk model untuk mendapatkan model terbaik. Sumbangan besar untuk projek ini menggunakan skema berat maksimum teknik perlombongan item kerap (FIM) untuk menghasilkan vektor ciri yang mampu mengesan imej dengan lebih baik.

TABLE OF CONTENT

CHAPTER	SUBJECT	PAGE
	DECLARATION	
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF FIGURES	xi
	LIST OF TABLES	Xii
	LIST OF APPENDICES	xiv
CHAPTER I	INTRODUCTION	
	1.1 Introduction	1
	1.2 Problem Statement	2
	1.3 Project Question	3
	1.4 Project Objective	4
	1.5 Scopes	4
	1.6 Project Contribution	5
	1.7 Thesis Organization	5
	1.8 Conclusion	6
CHAPTER II	LITERATURE REVIEW	
	2.1 Introduction	7
	2.2 Related Work	8
	2.2.1 What is Spam?	8
	2.2.1.1 What is Image Spam?	9
	2.2.1.2 What is Email Spam?	11

2.2.1.2.1 How to prevent Email Spam?	12
2.2.2 Statistic about Cyber security	13
2.2.3 How to prevent Spam?	15
2.2.3.1 Machine Learning	15
2.2.3.2 White List	22
2.2.3.3 Black List	23
2.2.4 Image	23
2.2.5 Method	25
2.3 Previous Researcher	27
2.3.1 A survey of learning-based techniques email spam filtering	27
2.3.2 A survey of image spamming and filtering techniques	27
2.3.3 A review of machine learning approaches to Spam filtering	28
2.3.4 Machine Learning Approaches for Modeling Spammer Behavior	28
2.4 Conclusion	29
CHAPTER III METHODOLOGY	
3.1 Introduction	30
3.2 Project Methodology	31
3.2.1 Scale Invariant Feature Transform (SIFT) Keypoint	33
3.2.2 Scale Invariant Feature Transform (SIFT) Descriptor	35
3.3 Frequent Item Mining (Previous Researcher)	36
3.3.1 BOW Feature Vector and Generated from BOW	36
3.4 Frequent Item Mining (New Research)	37
3.4.1 BOW Feature Vector and Generated from BOW	37

3.5 Project Schedule and Milestone	48
3.5.1 Milestone of project	48
3.5.2 Gantt Chart	49
3.6 Conclusion	50
CHAPTER IV ANALYSIS AND DESIGN	
4.1 Introduction	51
4.2 Problem Analysis	51
4.3 Flow Chart	52
4.3.1 Concept of Flow Chart Diagram	52
4.3.2 Flow Chart Diagram	53
4.4 Use Case Model	54
4.4.1 Use Case Diagram for Data Mining Module	54
4.5 Process Design	59
4.5.1 Sequence Diagram Concept	59
4.5.2 Sequence Diagram for Image Spam Detection	59
4.6 Requirement Analysis	62
4.6.1 Software Requirement	62
4.6.2 Hardware Requirement	62
4.7 Conclusion	65
CHAPTER V EXPERIMENT, ANALYSIS AND RESULT	
5.1 Introduction	66
5.2 Software Development Environment Setup	66
5.3 Process Module	68
5.3.1 Collection of Dataset	69
5.3.2 Pre-processing data	69
5.3.3 Train and Test Data	69
5.3.3.1 Process Create The train.scale.model	70
5.3.3..2 Process Generate the Predict File	71
5.4 Implementation of Project Execution	72
5.5 Installation of required software	76

5.6 Experiment Result	77
5.6.1 Bar Graph for First Run	78
5.6.2 Bar Graph for Second Run	79
5.6.3 Bar Graph for Third Run	80
5.6.4 Bar Graph for Fourth Run	81
5.6.5 Bar Graph for Fifth Run	82
5.6.6 Bar Graph for Sixth Run	83
5.6.7 Bar Graph for Seven Run	84
5.6.8 Bar Graph for Eight Run	85
5.6.9 Bar Graph for Nine Run	86
5.6.10 Bar Graph for Ten Run	87
5.6.11 Bar Graph for Average of Each Feature Vector	88
5.7 Conclusion	89
CHAPTER VI DISCUSSION	
6.1 Introduction	90
6.2 Comparison of Experiment Result	91
6.3 Conclusion	92
CHAPTER VII CONCLUSION	
7.1 Introduction	93
7.2 Project Summarization	93
7.2.1 Observation of Weakness and Strength	94
7.3 Project Contribution	95
7.4 Project Limitation	95
7.5 Feature Works	96
7.6 Conclusion	96
REFERENCES	97

LIST OF FIGURES

FIGURE	TITLE	PAGE
Figure 2.1	Example of real spam images	10
Figure 2.2	Example of emailspam	12
Figure 2.3	Example graph of Statistic	14
Figure 2.4	Machine Learning attribute	15
Figure 2.5	Example of Feature Selection	17
Figure 2.6	Example of SVM	19
Figure 2.7	Example of Neural Network	20
Figure 2.8	Example of Naive Bayes	20
Figure 2.9	Example of Regression	21
Figure 2.10	Example of Decision TreeFigure	21
Figure 2.11	Process of prediction	22
Figure 2.12	Example of SIFT Keypoint	24
Figure 3.1	Block diagram to generate models	32
Figure 3.2	Example of keypoint	34
Figure 3.3	Example of keypoint	34
Figure 3.4	Imagefor Minimum Weightage 1	37
Figure 3.5	Imagefor Minimum Weightage 2	38
Figure 3.6	Imagefor Minimum Weightage 3	39
Figure 3.7	Imagefor Minimum Weightage 4	40
Figure 3.8	Imagefor Minimum Weightage 5	41
Figure 3.10	Image for Maximum Weightage 1	43
Figure 3.11	Image for Maximum Weightage 2	44
Figure 3.12	Image for Maximum Weightage 3	45
Figure 3.13	Image for Maximum Weightage 4	46
Figure 3.14	Image for Maximum Weightage 5	47
Figure 4.1	Flow Chart Diagram	53
Figure 4.2	Use Case Diagram for data mining	54

	module	
Figure 5.1	The Software Development Environment Setup Architecture	67
Figure 5.2	Three Module involve in Frequent Item Mining	68
Figure 5.3	Process Train and Test	
Figure 5.4	Process Create The train .scale.model	70
Figure 5.5	Process Generate the Predict File Figure	70
Figure 5.6	Run program .bat	71
Figure 5.7	Example of allRuleSpam	73
Figure 5.8	Example of file trainBOW	74
Figure 5.9	Example of allRuleLegitimate	74
Figure 5.10	Example of file	75
Figure 5.11	Deployment Diagram	76
Figure 5.12	Result for First Run	77
Figure 5.13	Result for Second run	78
Figure 5.14	Result for Third run	79
Figure 5.15	Result for Fourth run	80
Figure 5.16	Result for Fifth run	81
Figure 5.17	Result for Sixth run	82
Figure 5.18	Result for Seven run	83
Figure 5.19	Result for Eight run	84
Figure 5.20	Result for Nine run	85
Figure 5.21	Result for Ten run	86
Figure 5.22	Result of average for each Feature Vector	87
		88

LIST OF TABLE

TABLE	TITLE	PAGE
Table 1.1	Summaries of Problem Statement	2
Table 1.2	Summaries of Project Question	3
Table 1.1	Summaries of Project Objectives	4
Table 3.1	Example of BOW Feature Vector	36
Table 3.2	Frequency from Minimum Weightage in image 1	37
Table 3.3	Frequency from Minimum Weightage in Image 2	38
Table 3.4	Frequency from Minimum Weightage in Image 3	39
Table 3.5	Frequency from Minimum Weightage in Image 4	40
Table 3.6	Frequency from Minimum Weightage in Image 5	41
Table 3.7	Example of BOW Feature Vector	42
Table 3.8	Frequency from Maximum Weightage in Image 1	43
Table 3.9	Frequency from Maximum Weightage in Image 2	44
Table 3.10	Frequency from Maximum Weightage in Image 3	45
Table 3.11	Frequency from Maximum Weightage in Image 4	46
Table 3.12	Frequency from Maximum Weightage in Image 5	47
Table 3.13	Milestone of project	48
Table 4.1	Summaries of Problem Analysis	51
Table 4.2	Narrative Use Case for Get ProcessDataset	55
Table 4.3	Narrative Use Case for Pre-processing Data	56
Table 4.4	Narrative Use Case for Train Machine Learning	57
Table 4.5	Narrative Use Case for Predict	58
Table 4.5	Software Requirement	62
Table 4.6	Personal computer Specification	64
Table 5.1	Result and average for BOW, Minimum and Maximum	77
Table 5.2	Result for First Run	78
Table 5.3	Result for Second Run	79
Table 5.4	Result for Third Run	80
Table 5.5	Result for Fourth Run	81
Table 5.6	Result for Fifth Run	82
Table 5.7	Result for Sixth Run	83

Table 5.8	Result for Seven Run	84
Table 5.9	Result for Eight Run	85
Table 5.10	Result for Nine Run	86
Table 5.11	Result for Ten Run	87
Table 5.12	Result for Average of each Feature Vector	88
Table 6.1	Average Result for All Feature	91
Table 7.1:	Weakness and Strength of Project	94



CHAPTER 1

INTRODUCTION

1.1 Introduction

In spite of the fact that a considerable measure of programming has been created as of late to square spam from your email customer, there is another type of spam that is intended to get around spam blockers and into your Inbox. The new type of spamming is known as picture spamming and uses pictures rather than content to sidestep spam blocking programming.

Spam is normally sent as an instant message in which particular words in the message can be utilized by spam blocking programming to keep the message from achieving your Inbox. With picture spamming, the content is put inside the picture with an end goal to sidestep the spam blocking programming.

Since pictures are viewed as a typical piece of a beneficiary's email message and the spam blocking programming is for the most part intended for content, the spammer is fruitful in getting the message to come to your Inbox.

Spam channels are intended to peruse particular words that decide if the spam ought to be sifted through. The messages are then sent to a spam envelope where the beneficiary can see them or not.

Later spam channels have fixed security by taking out the spam messages out and out and enhancing the criteria that decides a message as spam. Accordingly, spammers thought of an approach to consolidate content with pictures to get around the fixed security of spam channels.

1.2 Problem Statement

The problem statement that has been identified is summarized in table 1.1 below

Table 1.1 Summaries of Problem Statement

PS	Problem Statement
PS1	The effect of image spam

PS1: The effect of image spam.

The examples of the problem is bandwidth used for free and the email users will waste time to delete spam emails or the image spam will attach for virus and also malware.

1.3 Project Question

In this research, there is three Project Question (PQ) that needs to be answered in this project. The summary of project question is shown in Table 1.2

Table 1.2 Summaries of Project Question

PQ	Project Questions
PQ1	What are the purposes on this project?
PQ2	Which method is the best to detect an image spam?
PQ3	What is the method that used in this project?

PQ1: What are the purposes on this project?

To investigate the performance of the classifier using the feature vector generated weightage schemes and maximum.

PQ2: Which method is the best to detect an image spam?

To identify the performance which is the best between a single classifier.

PQ3: What is the method that used in this project?

To identify the performance of ensemble method when combined of these models.

1.4 Project Objective

Based on this research, there are three Project Objectives (PO) that are developed as follows in table 1.3.

Table 1.1 Summaries of Project Objectives

PQ	Project Objectives
PO1	To investigate the performance of the classifier using the feature vector generated from weightage schemes, and maximum.
PO2	To identify the performance which is the best between a single classifier
PO3	To identify the performance of ensemble method when combined of these models.

1.5 Scopes

The scope of this projects is to detect an image spam by using Frequent Itemset Mining Techniques. The main goal of the project is to protect the image of the attack by spam. When image spam is filtered, it can be more secured.

1.6 Project Contribution

Image spam sifting used to guarantees messages you do need endure. Spam keeps vital and true blue correspondence from contacting the target group. Spam separating hinders the garbage and keeps it from hitting the inbox in any case. At that point it can aggregates with information security and email controls since Spam sifting helps organizations with staying consistent and a la mode on security.

By doing this project, I can know what a suitable technique that can detect an image spam. Besides, an image spam also can classify by the dataset. In this case, it can classify an image spam to build a graph or table to show the result after do the experiment.

1.7 Thesis Organization

Chapter 1: Introduction

In this chapter it will discuss the background of project. It also include the problem statement, project question, project scope, project contribution, thesis organization and conclusion.

Chapter 2: Literature Review

In this chapter it will discuss about previous work that are related to the project and review of previous researcher. This chapter also include the solution for the project.

Chapter 3: Project Methodology

In this chapter it will discuss about the methodology that will be used in these projects. This chapter also will explain on the methodology for every step that has been used and the project milestone.

Chapter 4: Analysis and Design

In this chapter it will discuss about the analysis of the project and the design of the project that had been used in the experiment.

1.8 Conclusion

The conclusion for this project is using the two (2) weightage which is Bag Of Word and Minimum schemes and the Maximum of the frequent item mining (FIM) technique to generate a feature vectors that is capable of detecting image with better.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

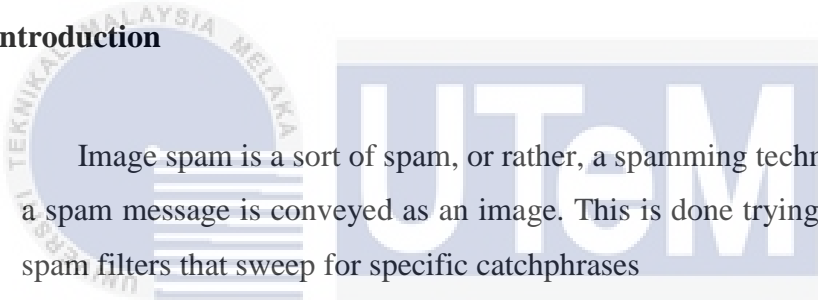


Image spam is a sort of spam, or rather, a spamming technique, in which a spam message is conveyed as an image. This is done trying to circumvent spam filters that sweep for specific catchphrases

Image spam is junk email that replaces content with pictures as a methods for tricking spam channels. Picture conveyance works by implanting code in a HTML message that connections to a picture record on the Web. Image spam is a bigger deplete on system assets than content spam since picture documents are bigger than ASCII character strings. Bigger records require more data transmission and, as a result, cause more prominent corruption of exchange rates.

2.2 Related Work

In this section it will explain about the work that related with the project. It includes all the information about spam, image spam, email spam, web spam, sms spam, social spam, the statistic of cyber crime, machine learning and the related work and others that relate with the project.

2.2.1 What is Spam?

Spam is thought to be electronic garbage mail or garbage newsgroup postings. A few people characterize spam much more for the most part as any unsolicited email. Be that as it may, if a departed sibling finds your email address and sends you a message, this could barely be called spam, despite the fact that it is spontaneous. Genuine spam is for the most part email publicizing for some item sent to a mailing list or newsgroup.

"Spam" is an acronym gotten from the words "spiced" and 'ham'. It's superfluous or spontaneous messages sent over the Internet, regularly to a substantial number of clients, for the motivations behind promoting, phishing, spreading malware, and so forth.

In view of past specialist, There exist different meanings of what spam (likewise called garbage mail) is and how it varies from true blue mail (additionally called non-spam, real mail or ham). The most limited among the well known definitions describes spam as "spontaneous mass email" (Androustopoulos et al. 2000b; SPAMHAUS 2005). In some cases the word business is included, however this expansion is far from being obviously true. The TREC Spam Track depends on a comparative definition: spam is "spontaneous, undesirable email that was sent