Information Systems Department
Faculty of Computer & Information Sciences
Ain Shams University

Efficient Email Classification Technique Based on Semantic Methods

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences.

By

Eman Mohamed Bahgat

B.Sc. in Computer and Information Sciences,
Teaching Assistant at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Under Supervision of

Prof. Dr. Ibrahim Moawed
Professor
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Egypt

Dr. Walaa Khaled
Associate Professor
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Egypt

Dr. Sherine Rady
Associate Professor
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Egypt

2018

Acknowledgements

Firstly, thanks to Allah Almighty for helping me in accomplishing this thesis.

I would like to thank my supervisors; Prof. Dr. Ibrahim Fathy, Dr. Sherine Rady, Dr. Walaa Gad at the Faculty of Computer and Information Sciences, Ain Shams University, for their guidance, support, encouragement, valuable comments and advices that helped me in finishing this work. It was my pleasure and honor to work with them.

I am also grateful to all my colleagues for their continuous support and help. Finally, with all my appreciation and love, i would like to express my profound gratitude to all my family especially my parents, brothers and husband who always provided me with continuous encouragement and prayers throughout the research time. This accomplishment would not have been possible without them.

Abstract

Emails have become one of the major applications in daily life. It is one of the most popular ways of communication due to its easy accessibility, low sending cost and fast message transfer. The continuous growth in the number of email users has led to a massive increase of useful emails, in addition to unsolicited emails. The latter are known as spam emails which appear as a severe problem affecting the users' and computer network performances. Managing and classifying the huge number of emails is an important challenge. Email filtering approach is the solution to manage such big size, in addition to isolate spam emails.

Recently, most of the approaches introduced in the literature to solve the huge number of spam emails. Filtering syntactic features handles the high dimensionality of emails.

This thesis proposes an email filtering approach based on the semantic methods. A framework is proposed, which consists of two phases. The first one uses classification techniques, where the body of email messages is analyzed and the terms are extracted from email body. Weights are assigned to terms (features) that can help to identify emails as spam or ham (i.e clean). An adaptation to this structure is proposed to reduce the extracted number of features, in which only meaningful terms are regarded by consulting an English dictionary.

In the second phase, WordNet is introduced as an ontology to apply different semantic based similarity measures to reduce the number of features, space and time complexities. Moreover, to get the minimal

optimal features set, feature dimensionality reduction is integrated. Two feature selection techniques are used: the Principal Component Analysis (PCA) and the Correlation Feature Selection (CFS) are evaluated for such purpose.

Experimental results have been conducted and the proposed framework and methods have been tested on the standard benchmark Enron Dataset. It is a large public email database collection. SVM and Logistic Regression classifiers recorded the best accuracy values of 96%, followed by the Naïve Bayes with 92.3% accuracy value.

Integrating semantics and feature selection, the classifier Logistic Regression achieved the highest accuracy value of 95%. Followed by the Naïve Bayes and SVM having similar results of 94% accuracy value. It has been shown that when integrating the feature selection, the average recorded accuracy for the all used classifiers is enhanced reaching all above 90%. This happens with more than 90% feature space reduction. The experimental results also showed that CFS feature selection technique had better results compared to PCA.

Consequently, the proposed framework and the conducted experiments showed that the proposed work has a highly significant performance in terms of accuracy and time compared to other related work. The integration of the semantic concepts and feature reduction approaches added important benefits to enhancing the computational performance and the accuracy of classification.

Table of Contents

Abstract ii

List of Figures vii

List of Tables ix

List of Abbreviations x

1 Introduction 1

1.1 Overview 1

1.2 Problem definition 4

1.3 Research objective 4

1.4 Contributions 5

1.5 Thesis organization 5

2 Related work 7

2.1 Email filtering and Document classification 7

2.1.1 Origin-based filtering 8

2.1.2 Content-based filtering 9

2.1.3 Comparative Study 11

2.1.4 Ensemble classification 12

2.2 Feature selection in email filtering 12
2.3 Semantics in document classification 15

3 Background **17**
3.1 Classification methods 17
3.1.1 Naive Bayes 18
3.1.2 Support Vector Machine 19
3.1.3 J48 Classifier 21
3.1.4 Logistic regression 23
3.1.5 Random Forest 24
3.1.6 Radial basis function 25
3.2 Semantic representation 26
3.2.1 WordNet Ontology 26
3.2.2 Semantic relations 26
3.2.3 Semantic similarities measures 27
3.2.3.1 Path-based measure 27
3.2.3.2 Information Content measure 29
3.2.4 Relatedness measures 29
3.3 Feature selection methods 30
3.3.1 Principal component analysis 30
3.3.2 Correlation feature selection 31

4 Email filtering using classification techniques **33**
4.1 Architecture of the proposed approach 33
4.1.1 Email pre-processing 34
4.1.2 Email representation 35

4.1.3 Classification 36
4.2 Experiments and evaluation 36
4.2.1 Dataset and development environment 36
4.2.2 Evaluation measures 37
4.3 Experimental results 39
4.4 Summary 44

5 Integrating semantic and feature selection methods in Email filtering 45

5.1 The enhanced architecture 45
5.1.1 Email pre-processing 47
5.1.2 Feature weighting 47
5.1.3 Feature reduction 48
5.1.3.1 Semantic-based reduction 49
5.1.3.2 Feature weights update 50
5.1.3.3 Feature selection 52
5.2 Experiments and evaluation 52
5.3 Experimental results 52
5.3.1 Study the effect of semantic similarity measures 53
5.3.2 Selecting the most important feature set 58
5.3.3 Comparison to related works 59
5.4 Summary 63

6 Conclusion and future work 64

6.1 Conclusion 64
6.2 Future work 66

List of Figures

1.1 Recent statistics indicating the percentage of spam emails versus non-spam emails [3] 2

1.2 Spam emails costs for business [2] 3

2.1 Example of the header of an Email [4] 9

3.1 Margin and support vector in SVM classifier 19

3.2 Pseudo code for the J48 algorithm 21

3.3 An example of concept chain for football keyword in WordNet 27

4.1 Proposed email filtering architecture of the first approach 34

4.2 Accuracy for different classifiers for the all features . . 41

4.3 Accuracy for different classifiers for the reduced feature set 42

4.4 Accuracy value of proposed work compared to related work in [22] 43

4.5 Accuracy value of proposed work compared to related work in [23] 43

5.1 Modified architecture by integrating semantic and Feature selection components 46

5.2 Pseudo-code for semantic-based reduction 50

5.3 The flowchart for feature weights update process 51

5.4 Accuracy performance of different classifiers for WUP similarity measure versus number of features threshold 54

5.5 Accuracy performance of different classifiers for Path similarity measure versus number of features threshold 54

5.6 Accuracy performance of different classifiers for LCH similarity measure versus number of features threshold 55

5.7 Accuracy performance of different classifiers for Resnik similarity measure versus number of features threshold 55

5.8 Accuracy performance of different classifiers for HSO similarity measure versus number of features threshold 56

5.9 Illustrative example for semantic reduction framework using path similarity 57

5.10 Comparing accuracy performance when using feature selection techniques (PCA and CFS) 59

5.11 Accuracy performance using CFS compared to related work in [22] 60

5.12 Accuracy performance of different classifiers using different dataset sizes 63

List of Tables

4.1 Confusion matrix of performance measures 38

4.2 Performance of 6424 feature set 40

4.3 Performance of 3636 reduced feature set 41

5.1 Performance of path similarity measure at threshold 0.4 58

5.2 Comparing the accuracy value and time using different methods of email classification 62

List of Abbreviations

NB	Naive Bayes
SVM	Support Vector Machine
RBF	Radial Basis Function
RF	Random Forest
KNN	k-Nearest Neighbor
ANN	Artificial Neural Networks
MLP	Multi-Layer Perceptron
PCA	Principal Component Analysis
CFS	Correlation Feature Selection
IG	Information Gain
WUP	Wu and Palmer
LCH	Leacock and Chodorow
IC	Information Content
LCS	Least common subsume
HSO	Hirst and St-Onge
TF-IDF	Term Frequency- Inverse Document Frequency

List of publications

1. E. M. Bahgat, S. Rady, and W. Gad, "**An e-mail filtering approach using classification techniques**" in the 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), Springer, pp. 321–331, November, 2015.

2. E. M. Bahgat and I. F. Moawad, "**Semantic-based feature reduction approach for e-mail classification**" in the 2nd international Conference on Advanced Intelligent Systems and Informatics (AISI 2016), Springer, pp. 53–63, October, 2016.

3. E. M. Bahgat, S. Rady, W. Gad, I. F. Moawad, "**Efficient Email Classification Approach Based on Semantic Methods**" in Ain Shams Engineering Journal, vol. 9 Elsevier, pp. 3259-3269 SJR: 0.6, 2018.

Chapter 1

Introduction

1.1 Overview

Electronic email has become one of the most important applications for computer users. According to a Cyberoam report [1], the average number of spam messages sent every day has reached 54 billion messages. Fig.1.1 shows recent statistics showing the percentage of spam email accounts, which means that almost half of the emails is spam. Such explosive growth of spam emails leads to severe problems for users. The sender of spam email does not target the recipient personally, but the spam invades users without their consent and fills their email box. In addition to the time consumed in checking and deleting spam emails, they overload the network bandwidth by useless data packets. All these factors cause an increase to the operating costs, affect work productivity and privacy. Moreover, it harms the network infrastructure and the recipient's device if the email is pernicious. According to the Radicati Research Group Inc.[2], the spam emails cost businesses over $20.5 billion dollars every year, and could even increase to $257 billion dollars annually, as shown in Fig. 1.2. Therefore, a big challenge in business is to manage the huge number of emails efficiently. As a solution to such challenge, an email filtering approach is significantly required.

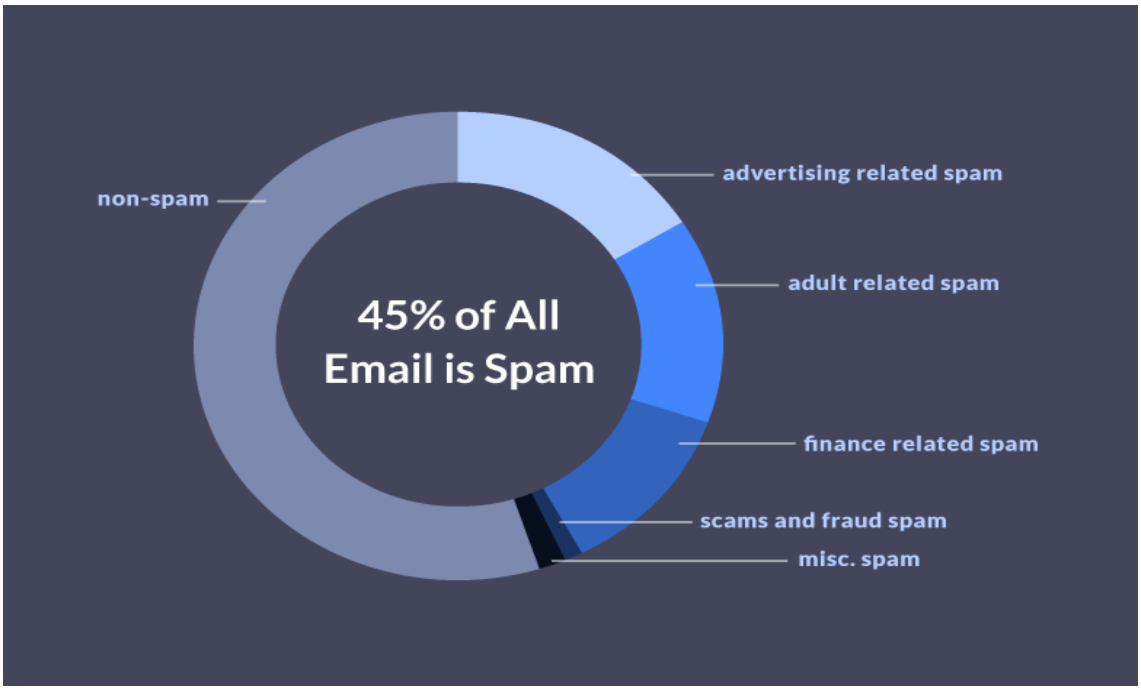

Figure 1.1: Recent statistics indicating the percentage of spam emails versus non-spam emails [3]

The basic format of email consists of two parts:

- Header of the email: Includes the sender, the receiver of email address, the subject, and the date.
- Body of the email: Includes text, images, and other multimedia data.

Common email filters filter incoming email automatically. The filtering process results in a set of categories or classifications. Example is spam filtering application which filters incoming emails into spam and ham(i.e clean). An Email Filtering is always required to identify and detect the spam emails and to dispose the huge number of spam emails efficiently. In literature, the filtering process is decided and executed based on two common methods: The email origin or header method [4] (i.e. source) and the email content based method [5].

Origin-based filtering focuses on header part. It monitors the source of the e-mail, which is stored in the domain name and address

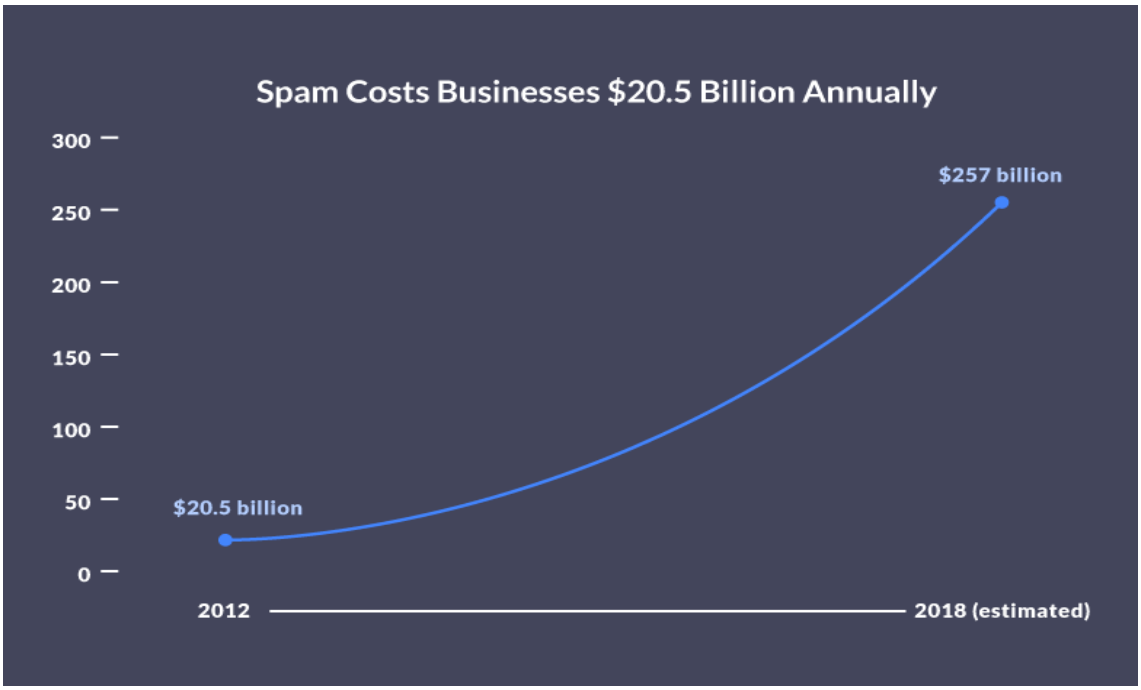

Figure 1.2: Spam emails costs for business [2]

of the sender device. Such filtering preserves two types of emails; white-list and black-list.

- White-list: keeps only the emails from reliable or trusted list of email sources through.
- Black-list: contains lists of known spammers. It keeps track of any email source except emails with matching IP addresses or email source from the blacklist of spammers.

In those types, when receiving a new email the source of this email is compared with the preserved lists or database to know how it is classified (i.e. spam history). However, the disadvantage of such technique is that the spammers regularly change the email source address and IP and hence, spam emails cannot be identified.

On the other hand, the content-based filtering reviews the email content depending on a proposed analysis technique [5]. It extracts features from the email body and classifies email into one of two classes ham or spam. Examples for the common classification methods are