

جامعة السودان للعلوم والتكنولوجيا
كلية الدراسات العليا
كلية الحاسوب وتقانة المعلومات

التعرف على الرسائل الإلكترونية غير المرغوب فيها
بواسطة المحتوى

رسالة مقدمه لنيل درجة الماجستير في علوم الحاسوب

إعداد

موسى عبدالفتاح أحمد عبدو

إشراف

د. محمد الحافظ مصطفى

2008

DEDICATION

To the soul of my father and my mother

To my beloved wife

To my children Abdul Fattah, Aws, Marah and Muna

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor Dr. Mohammad Al Hafiz Mustaf for his continual, guidance, detailed comments on nearly every page of this thesis. His valuable suggestions and unlimited support throughout the study were the major factors in bringing this work to life.

I also would like to thank Jury members Dr. Howaida Ali and Dr. Ahmad Khaled for their valuable comments and suggestions.

Special respect and thanks to my friend Dr. Nael Salman for his enthusiasm and encouragement.

I am grateful to the president of Palestine Technical University and all of my colleagues there, especially my friends in Department of Computer Science who contributed in one way or another to conduct this study.

My heartfelt appreciation is extended to my parents, my brothers, and my sisters, for their support and prayers. Without the support of my brother Ibrahim I would have not been able to do this study.

Last but not least, words are not enough to express my thanks to my wife and my children for their encouragement, support, patience and the endless love. I hope that they will forgive me for my long absence.

ABSTRACT

The dramatically increasing number of email users, and the increasing number of free email providers, like yahoo, hotmail, gmail, increase the number of unwanted emails which is known as 'Spam emails'.

The huge number of spam emails received daily by users account, made the necessity of existence of some sort of automated spam filters to detect and remove these unwanted emails. Several researchers have started working on automated techniques and tools that can be used to classify emails automatically into wanted (legitimate) or unwanted (spam) emails.

Most of these filters are based on naïve Bayesian method. This thesis introduces a new automated filter based on naïve Bayesian. The basic idea of this filter is to give each word appears in emails a probabilistic value, this value indicates its probable belonging to spam. As there are many common words appear in spam as well as legitimate messages with the same rate, the proposed filter has a preprocessing component which removes all common words. The researcher carefully collected these common words.

In the training phase a set of 1300 emails (legitimate and Spam) has been used. In this phase the weight of every uncommon word is estimated as the probability of a given word in spam email divided by the probability of the same word in legitimate email.

In classification, a Bayesian classifier uses the weight table generated in the training phase to classify a given email as spam or legitimate.

The proposed filter has been tested on a dataset of 400 emails, 200 of them are Spam and 200 of them are legitimate, the proposed algorithm succeeded in detecting 90% of the spam messages.

المستخلص

ان الزيادة الكبيرة في عدد مستخدمي البريد الالكتروني، وزيادة عدد مزودي خدمة البريد الالكتروني المجاني، أدت إلى زيادة عدد الرسائل غير المرغوب فيها والتي تعرف باسم سبام (Spam).

ان العدد الكبير من الرسائل غير المرغوب فيها التي تصل يوميا الى حسابات المستخدمين جعل من الامر المهم وجود بعض أنواع التصفية الأوتوماتيكية لاكتشاف وإيقاف هذه الرسائل غير المرغوب فيها. ان الكثير من الباحثين بدأو بعمل هذه الأنواع من التصفية الأوتوماتيكية والتي تستخدم لتصنيف الرسائل الالكترونية الى مرغوب فيها وغير مرغوب فيها.

معظم هذه المرشحات تعتمد على طريقة ونظرية يبيز في الاحتمالات (Bayesian Theory).

الفكرة الرئيسية في هذا المرشح هو اعطاء كل كلمة تظهر في الرسائل الالكترونية قيمة احتمالية، وهذه القيمة تبين احتمالية انتماء الكلمة الى الرسائل النصية غير المرغوب فيها، كما ان هناك العديد من الكلمات الشائعة التي تظهر في الرسائل النصية المرغوب فيها وغير المرغوب فيها وبنفس النسبة. المرشح المقترح يحتوي على خطوة مسبقة التي تقوم بإزالة هذه الكلمات الشائعة التي تم تجميعها من قبل الباحث بعناية.

وفي مرحلة التدريب تم استخدام 1300 رسالة (مرغوب فيها وغير مرغوب فيها). في هذه المرحلة تم حساب وزن كل كلمة غير شائعة بالعلاقة التالية : ناتج قسمة احتمال الكلمة في الرسالة غير المرغوب فيها على احتمال الكلمة في الرسالة المرغوب فيها.

في مرحلة التصنيف يتم استخدام جدول الاوزان الناتج من مرحلة التدريب لتصنيف الرسالة الى مرغوب فيها وغير مرغوب فيها.

تم فحص البرامج المكتوبة من قبل الباحث ب 400 رسالة، 200 رسالة غير مرغوب فيها و 200 رسالة مرغوب فيها وتمكن البرنامج من تصنيف ما نسبته 90% من الرسائل.

Table of contents

Subject	Page
Dedication	II
Acknowledgments	III
Abstract in English	IV
Abstract in Arabic	V
Table of contents	VII
List of Tables	X
List of Figures	XI
List of Equations	XII
CHAPTER 1 INTRODUCTION	1
1.1 Overview of Spam	1
1.2 Thesis Outline	2
CHAPTER 2 DATA MINING	3
2.1 Introduction	3
2.2 Data Mining Steps	3
2.3 Data Mining Functions	4
2.3.1 Clustering	4
2.3.2 Association Rules	5
2.3.3 Sequential patterns	5
2.3.4 Classification	6
2.3.4.1 Decision Tree Algorithm	6
2.3.4.2 Neural networks	7
2.3.4.3 Naive Bayes Algorithm	9
CHAPTER 3 NAIVE BAYES ALGORITHM	10
3.1 Naive Bayesian algorithm	10
3.2 Why Bayesian Filtering	13
CHAPTER 4 SPAM	15
4.1 Spam definition	15

Subject	Page
4.2 Spam History	16
4.3 Facts and Figures about Spam	16
4.4 The Necessity of Spam Filtering	18
4.5 Spam Costs	18
4.6 kinds of Spam	19
4.6.1 Commercial Spam	19
4.6.2 Pornographic spam	19
4.6.3 Viruses	19
4.7 Email Environment	20
4.7.1 IMAP Protocol	20
4.7.2 POP3 Protocol	21
4.7.3 SMTP Protocol	21
4.7.4 HTTP Protocol	21
4.7.5 Incoming EMail Server	21
4.7.6 Outgoing Mail Server	22
4.8 How Spammers Collect Email Lists	22
4.8.1 Harvesting	23
4.8.2 Buying	23
4.9 Anti Spamming	24
4.9.1 White Lists	24
4.9.2 Black Lists	24
4.9.3 Header-based algorithms	24
4.9.4 Content-based algorithms	25
CHAPTER 5 BAYESIAN BASED SAM DETECTOR	28
5.1 The Proposed Algorithm	28
5.1.1 Preprocessing Stage	29
5.1.2 Training Stage	30
5.1.3 Classification Stage	31
5.2 Implementation	32

Subject	Page
5.2.1 Classifier Construction	32
5.3 Illustrative Examples	45
5.4 Testing	54
CHAPTER 6 CONCLUSION & FUTURE WORK	60
6.1 Conclusion	60
6.2 Future improvements	61
References	62
Appendix A	65

List of Tables

Table	Page
5.1 50% Spam & 50% legitimate in the training data set	33
5.2 spam and 75% legitimate in the training data set	33
5.3 25% legitimate and 75% spam in the training data set	33
5.4 The output of classification stage	37
5.5 The spam table after the common words removal	45
5.6 The legitimate table after the common words removal	46
5.7 The frequency file build by calculating the frequencies	46
5.8 The file after eliminating weights <1	46
5.9 Words in Real example1	48
5.10 Words in Real example2	49
5.11 Words in Real example3	51
5.12 Words in Real example4	53
5.13 Spam email categories	55
5.14 Error percentage in each category of spam emails	56
5.15 The percentage of false positive emails	57
5.16 The percentage of false negative emails	57
5.17 The total error rate	58
5.18 The values for TP, FP, FN	59

List of Figures

Figure	Page
2.1 The process of clustering	5
2.2 Decision tree for the weather dataset	7
2.3 Unit neuron	8
2.4 Simple neural network	9
3.1 Posteriori probabilities	12
4.1 The percentage spam emails sent daily	17
4.2 The fields percentage of spam	17
4.3 Typical email environment	20
5.1 The preprocessing stage	28
5.2 The training stage	28
5.3 The Proposed algorithm - classification stage	28
5.4 Classification stage	32
5.5 Common words removal Algorithm	34
5.6 Word weight calculation algorithm	36
5.7 Classification algorithm	37
5.8 Percentage of spam emails types	56
5.9 Error percentage according to emails type	56
5.10 The percentage of false positive emails	57
5.11 The percentage of false negative emails	57
5.12 The total error rate	58

List of Equations

Equation	Page
3.1 Conditional probability	11
3.2 Posterior probability	11
3.3 Error making probability	13
3.4 Total probabilities	13
3.5 Total probabilities	13
3.6 Bayes' formula	13
3.7 Probability of spam email	14
4.1 Bayesian method formula	26
5.1 Probability of word given spam email	30
5.2 Probability of word given legitimate email	31
5.3 Weight calculation	31
5.4 Recall calculation	58
5.5 Precision calculation	58