جامعة السودان للعلوم والتكنولوجيا
كلية الدراسات العليا
كلية الحاسوب وتقانة المعلومات

التعرف على الرسائل الإلكترونية غير المرغوب فيها بواسطة المحتوى

رسالة مقدمـــــه لنيل درجة الماجستير في علوم الحاسوب

إعـــــــــداد

موسى عبدالفتاح أحمد عبدو

إشـــــــراف

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، أديا 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 وايقاف هذه الرسائل غير المرغوب فيها. ان الكثير من الباحثين بدأو بعمل هذه الانواع من التصفية الأوتوماتيكية والتي تستخدم لتصنيف الرسائل الالكترونية الى مرغوب فيها وغير مرغوب فيها.

معظم هذه المرشحات تعتمد على طريقة ونظرية بييز 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 بازالة 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

List of Tables

List of Figures

List of Equations