

DEDICATION

To my father

Dr. Elnageeb Adam Gammereldin

and to my mother

Miss. Widad Ahmed Yahia

To my brother and my sisters

To my teachers

To my friends

Salma Elnageeb

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my supervisor Dr. Mohamed Elhafiz Mustafa for his continual, guidance detail comments on nearly every page on this thesis. His valuable suggestions and unlimited support through the study were the major factors in bringing this work into existence.

Special respect and thanks to my teachers and my best friends for support and help and prayers.

Last but not least , words are not enough to express my thanks to my parents, sisters and brother for their encouragement ,support, patience and endless love.

Abstract

The massive increase of spam is posing a very serious threat to email which has become an important means for communication. Not only it annoys users, but it also consumes much of the bandwidth of the Internet. Current spam filters are based on the contents of the email one way or the other. In this thesis we present a social network-based spam detection method in which the core idea is using social network measurements as feature to be used by classifier. Two separate classification models have been designed and tested. The first is k-Nearest-Neighbor Classifiers (KNN) classifier and the second is Locally weighted learning (LWL). The experimental results have shown a great favour of using KNN model for spam detection. However, it classifies many legitimate as spam which may annoy the email user. Hence we recommend this model to be applied where the acceptance of a spam message is more danger than legitimate messages rejection. While the classification result of LWL is better than KNN result. It is clear that KNN has advantage of detecting all spammer.

المستخلص

الزيادة الهائلة فى أعداد رسائل البريد الإلكتروني المزعجة شكلت تهديدا خطيرا للمستخدمى البريد الإلكتروني الذي أصبح وسيلة مهمة للإتصال. والمشكلة ليست فقط فى إزعاج المستخدمين، ولكنها تستهلك الكثير من عرض النطاق الترددي للإنترنت. تستند المرشحات التى تستخدم حاليا للحد من الرسائل غير المرغوب فيها على تحليل المحتوى. ي قترح البحث التعرف على مرسلى الرسائل أنفسهم بالتعرف على سلوكهم. لهذا الغرض استخدم البحث الشبكات الاجتماعية للكشف عن مرسلى الرسائل غير المرغوب فيها. الفكرة الأساسية هي بناء شبكة اجتماعية ومن ثم تحليلها. تم بناء نموذجين منفصلين للتصنيف. الأول هو المصنف كى الجار الاقرب (K-Nearest-Neighbor (KNN والثاني هو تعليم الوزن المحلى (Locally Weighted learning (LWL). وقد أظهرت النتائج التجريبية ان التصنيف بإستخدام نموذج كى الجار الاقرب (K-Nearest-Neighbor (KNN للكشف عن الرسائل غير المرغوب فيها نجح فى الكشف على كل مرسلى الرسائل غير المرغوب فيها . ومع ذلك ، فإنه يصنف العديد من المرسلين الشرعيين على انهم مزعجين و غير شرعيين. هذا يؤدي الى ازعاج مستخدمى البريد الإلكتروني لما فية من رفض لرسائل شرعية. ومن هنا فإننا ننصح أن يكون تطبيق هذا النموذج حيث قبول رسالة البريد المزعجة هو أكثر خطرا من رفض الرسائل المشروعة خطأ. في حين أن النتيجة تصنيف تعليم الوزن المحلى (Locally Weighted learning (LWL أفضل من نتيجة كى الجار الاقرب (K-Nearest-Neighbor (KNN. فمن الواضح أن كى الجار الاقرب (K-Nearest-Neighbor (KNN يتميز بانه نجح بالكشف عن كل مرسلى الرسائل المزعجة.

Table of contents

Title	Page No
Dedication	I
Acknowledgments	II
Abstract in English	III
Abstract in Arabic	IV
Table of contents	V
List of figures	VIII
List of Tables	IX
1 INTRDUCTION	
1.1Introduction	1
1.2Objective and Contribution	1
1.3Thesis structure	2
2 SOCIAL NETWORK ANALYSIS	
2.1Introduction	3
2.2The Social Network Perspective	3
2.3Fundamental Concepts in Network Analysis	3
2.4Distinctive Features of Network Theory and Measurement	6
2.5Graph and Matrices	7
2.5.1 Using graph in social network?	7
2.5.2Directed Graphs	8
2.5.3Matrices	10
2.5.4Matrices for digraphs	10
3 SPAM	
3.1Introduction	13
3.2Spam Definition	13
3.3The Necessity of Spam Filtering	13
3.4Anti Spamming	14
3.4.1 Content-based Spam Detection Methods	14
3.4.2 Sender-based Spam Detection Methods	14
3.4.3 Social Network Structure Analysis-based Spam Detection	16
3.5 Spam statistics for 2011from Kaspersky anti-Spam Lab	17
3.5.1Spam by region	18
3.5.2Spam by category	18
4 DATA MINING AND LAZY LEARNING ALGORITHMS	
4.1Introduction	20
4.2Data Mining	20
4.3Data Mining Steps	20
4.4Data Mining Function	21
4.5Lazy Classifiers	22
4.5.1k-Nearest-Neighbor Classifiers	22
4.5.2Locally weighted learning	24
4.6Feature Subset Selection	26
5_PROPOSED SYSTEM	
5.1Introduction	28

5.2Data Set	29
5.3Preprocessing Stage	29
5.3.1Email log	29
5.3.2Prepare emails headers to build adjacent network	30
5.3.3Construct adjacent matrix and social network	30
5.3.4Calculate social network measurement	30
5.3.5Feature selection algorithm	32
5.3.5.1 Testing Feature subset using KNN algorithm	32
5.3.5.2 Testing Feature subset using LWL algorithm	40
6.4Classification Stage	45
6.4.1KNN model testing results	45
6.4.2LWL model testing results	47
6.5The Conclusion of Classifier Testing	49
7 CONCLUSION AND FUTURE WORKS	
7.1Conclusion	50
7.2Future Works	51
8 REFERENCES	
Reference	52
9 APPENDIX	
Appendix A	55
Appendix B	67

List of figures

Title	Page No
Figure 2.1 Graph for “lives near” relationship for six children	10
Figure 3.1: Spam by region in 2011	18
Figure 3.2: Spam by category in 2011	19
Figure 4.1: Basic Relief algorithm	26
Figure 5.1: The proposed system structure	28
Figure 5.2: Relation of experiment precision and dataset ranked by reliefF <u>neighbor k=1. And testing using KNN classifier k=1, 3 and 5</u>	34
Figure 5.3: Relation of experiment precision and dataset ranked by reliefF <u>neighbour K=5. And testing using KNN classifier k=1, 3 and 5</u>	35
Figure 5.4: Relation of experiment precision and dataset ranked by ReliefF <u>neighbour K=1. And testing using KNN classifier k=1, 3, 5, 7, 9 and 12</u>	38
Figure 5.5: Relation of experiment precision and dataset ranked by relief <u>neighbour K=5. And testing using KNN classifier k=1, 3, 5, 7, 9 and 12</u>	39
Figure 5.6: Relation of experiment precision and dataset ranked by ReliefF <u>neighbour K=1, 5 and 10. And testing using LWL classifier</u>	42
Figure 5.7: Relation of experiment precision and dataset ranked by reliefF <u>neighbour K=1, 5 10, 15, 17 and 20. And testing using LWL classifier</u>	43
Figure 5.8: The percentage of testing email accounts KNN model	46
Figure 5.9: The total error rate at KNN model	47
Figure 5.10: The percentage of false positive email accounts at LWL model	48
Figure 5.11: The total error rate at LWL model	48

List of tables

Title	Page No
Table 2.1 Example of sociomatrix for directed graph : friendship at the beginning of the year for six children	11
Table 2.2 Example of incidence for directed graph : friendship at the beginning of the year for six children	12
Table 5.1 Summary of experiment precision for dataset rank by reliefF neighbor $K=1, 5$ and 10. And testing using KNN classifier $k=1, 3$ and 5	33
Table 5.2: Summary of experiment precision for dataset rank by reliefF neighbour $K=1$. And testing using KNN classifier $k=7, 9$ and 12	33
Table 5.3: Summary of experiment precision for dataset rank by reliefF neighbour $K=1$. And testing using KNN classifier $k=5$	36
Table 5.4: Summary of experiment precision for dataset rank by reliefF neighbour $K=5$. And testing using KNN classifier $k=7, 9$ and 12	37
Table 5.5: Summary of experiment precision for dataset rank by reliefF neighbor $K=5$. And testing using KNN classifier $k=5$	37
Table 5.6: Summary of experiment precision for dataset rank by reliefF neighbour $K=1, 5$ and 10. And testing using LWL	40
Table 5.7: Summary of experiment precision for dataset rank by reliefF neighbour $K=15, 17$ and 20. And testing using LWL	41
Table 5.8: Summary of experiment precision for dataset rank by reliefF neighbour $K=17$. And testing using LWL classifier	41
Table 5.9: The result of attribute set for experiment that has best precision	44
Table 5.10: confusion matrix of testing result of KNN with $K=5$	46
Table 5.11: Show the total error rate which means the summation of false positive and false negative at KNN model	46
Table 5.12: Matrix of testing result of LWL	47
Table 5.13: Show the total error rate which means the summation of false positive and negative at LWL model	48
Table 5.14 : conclusion of KNN and LWL classifiers	49