Emails classification by data mining techniques

Mohammed A.Naser, Athar H.Mohammed

Department of Computer, College of Sciences for Women, University of Babylon

Abstract

Electronic mail messages have become increasingly important and widespread method of communication because of its time speed ,where the amount of email messages received per day can range from tens for a regular user to thousands for companies. Nowadays; The spam mail messages are mainly a illegal form where thousands of users are easily reachable. This type of emails has become a serious problem. In this paper was proposed a method of classifying emails using the concept of association rules.

الخلاصة

رسائل البريد الإلكترونية أصبحت طريقة اتصال مهمة وواسعة الانتشار جدا نتيجة لسرعتها من حيث الوقت، حيث إن كميةَ ليوم يُمْكِنُ أَنْ تَتراوح مابين العشرات لدى مستخدم اعتيادي إلى الآلاف لدى الشركات. في الوقت رسائل البريد الإلكتروني المستلمة بِا الحاضر؛ رسائل البريد الالكتروني الدعائية (الغير مرغوب فيها) بشكل عام تمثل شكلا غير شرعي أو غير قانوني يستخدم أو يمكن الوصول إليه بسهولة من قبل آلاف المستخدمين. هذا النوع من الرسائل البريدية الإلكترونية أَصبح مشكلة خطيرة. اقترح في هذا البحث طريقة لتصنيف رسائل البريد الإلكتروني باستخدام مفهوم قواعد الارتباط .

1. Introduction

The amount of text documents available in digital form has been growing significantly during the last decades because of the development of new techniques in storage and exchange of information. The appearance and growth of the World Wide Web facilitated the process of spreading and exchanging the information that gave birth the new ways of communication. These include electronic mail which can be stored in text form. In this paper we refer to electronic mail (e-mail for short), which has become increasingly important and widespread method of communication because of its time speed and cheep . It is used not only for personal contacts but also for business, advertising and electronic commerce. The amount of email messages received per day can range from tens for a regular user to thousands for companies [J. Itskevitch,2001].Some other than commercial purposes of spam are to express political or religious opinions, put viruses in the receivers' computer. Spam has become a serious problem. There are other problems associated with spam,messages can have content that is offensive to people and might cause general psychological annoyance, a large amount of spam messages can crash unprotected mail servers, legitimate personal e-mails can be easily lost and more To resolve the problem there is a growing need for emails classification solution[J. Kagstrom,2005].There are many approaches have been used for spam email classification according to information outside of the content of email messages such as black lists[D.Cook et al. ,2006] ,white lists[D.Cook et al. ,2006] ,challenge response systems[C.O'Brien et al. ,2003] and other methods.Also there are other approaches have been applied in spam email classification according to content such as Nayve Bayes [M. Sahami et al. , 1998] , decision trees[S. Appavu et al. 2007] , neural networks[T. Ayodele et al.,2010] ,K-nearest neighbor approach [D.Trudgian et al. , 2004],genetic algorithm[H. Katirai ,1999] and other methods. Association rule mining is used for finding interesting hidden patterns in large transactional databases. Association rules are rules that identify associations between items in transactions. In the recent years many algorithms have been proposed for finding association rules efficiently. Among these algorithms the most known are *apriori* ,and FP-tree growth[X.Wu et al.,2009]. In this paper it is exploit the use of association rules mining in building emails classification system from an emails training set.

2.Training phase

2.1 Emails Collection

In order to train and test the proposed system ,some corpora of spam and legitimate emails had been used, there are several collections of emails publicly available to be used by researchers in the internet . The dataset used in this paper available at link (http://www.cs.umass.edu/~ronb/enron_dataset.html). The emails that are collected range in size from a single sentence to several paragraphs .These emails consist of two classes ,spam emails class ,and ham email class.

2.2 Emails Preprocessing

Email preprocessing is the process of transforming email messages into a representation suitable for developed apriori algorithm. This step consists of the following processes:

A- Tokenization Process

Before an email text message can be processed, it is first split into units called tokens; this process is called tokenization .As this text is tokenizing into tokens usually takes place by using blank space as a delimiter

B- Noise Symbols Elimination Process

This process involve remove all non alphabetic symbols or also called (noise symbol) because they not useful in classification process.An example of noise symbols (; ,",{,},*,^,%,#,@,!)

C- Stop Words Elimination Process

After tokenization of messages and remove noise symbols,the next step is to eliminate the stop words where the stop words are the most frequent words in messages and they do not discriminative of the message contents such as pronouns(he, she, I), prepositions(in, on) and conjunctions(and, or, but).The stop table that used in the proposed classification system contains (237) words Eliminating stop words leads to the email with less length, which results in more efficient processing. Also the proposed classification system is design to remove the terms or Tokens that are shorter than four characters. In the stop words eliminating process, each token reads from text file is check out with stop words table; if token found in the table or it less than four, then will replace it with white space; otherwise, it is write into output text file

D- Suffix Stripping Process

The suffix stripping also called stemming , it is the process of remove suffix from word to generate token stem (token root). Tokens with a suffix will usually have similar meaning ,for example:Connect, Connected, Connecting, Connection ,Connections ,and can all be replaced by their stem (connect). The suffix stripping process will reduce the total number of tokens this mean reduce the size and complexity of the data in the system. In this paper it will be used many rules for doing the stemming process. Firstly ,the token will compare with table that contain words that do not require stemming process for their .After this comparing, if the token is not in this table then it will pass through the rules to remove the suffix from it.

2.3 Feature Selection Process

A main problem in email data analysis is the very high dimensionality of the feature space it can easily reach tens and even hundreds of thousands .Only a small number of the terms are feature tokens determining of email class ,while the rest are noise tokens, that make the result unreliable and increase computational time. A common approach in dealing with this problem is feature selection. Where the feature selection is the process of removing indiscriminative tokens from emails and finding relevant tokens that could help in identifying spam e-mails such a words were extracted from the input dataset to improve classification accuracy and reduce computational complexity.We used the TF.ITF(Term Frequency. Inverse email Frequency) ,The term frequency (TF) is a measure of how many times a term(token) appears in the email. The inverse term frequency (ITF) is a measure of the general importance of the term (token). It is calculated by dividing the total number of emails by the number of emails that contain a specific term. The TF.IDF is calculated by using the following equation:

$$TF.ITF= tf_{t,e} *\log((1+n)/ N)$$

Where :

$tf_{t,e}$ is the frequency of word t in email e.

n is the number of emails used in this experiment .

N is the number of emails where word t occurs.

2.4 Extract Rules Of Classification.

The concept of Association rule is to discover the strong patterns that are associated with the class labels and then take advantage of these patterns to build the classifier. Once a classifier is built, new emails are categorized into the proper class .Where each email represent a transaction where each transaction consist of the terms of emails .Each transaction is associated with an identifier, called TID that is used next in developed apriori algorithm.

This stage contains three steps will produce rules that are used in email classification system where these steps are:

A-Frequent Itemsets Generation Of Developed Apriori Algorithm

A new mails classification algorithm which proposed by[H. Al-kafagi,2004] is called developed Apriori algorithm. It has the following advantages: it is fast during training stage by pass on the database only once a time.The frequent itemsets is the first step in the developed apriori algorithm is used to find the itemsets that appear frequently from prior knowledge.It use k-itemsets to find (k+1) itemsets. The frequent itemsets construct according to the union of the k-itemsets and intersection of their support. Figure(1) illustrates the developed Apriori algorithm for generate frequent itemsets.

```
Input: set of emails text files E_i ; minimum support value(minsup).
Output: Frequent itemsets

  Begin

        1. L_1= {large1-itemsets }

        2.K=2
```

figure (1) Developed Apriori Algorithm

Where, each item of the itemset is considered as a 1-item candidate itemset, this represented by step (1). Steps (3-7) generate a set of candidate K-itemsets by calling Developed_apriori_gen procedure. Developed_apriori_gen procedure performs the union and intersects operations respectively. Union operation is done by steps (10-12), whereas, step (13) perform intersects operations. The number of TIDs in C is the support of C, which is computed and checked according to minsup threshold, this operation is performed by steps (14-18).

B-Generating Association Rules from Frequent Itemsets

After the frequent itemsets had been discovered, then generated association rules from frequent itemsets . Association rule find the correlation between the items where the items here refer to (words) or (tokens). The association rule consist of two sides the left side and the right side. Where to generate rules used 2-itemsets and more than 2-itemsets. After that generated subsets from the frequent itemsets and put the subsets in left hand side and put the itemsets subtract this subset in right hand side .

For each rule compute the confidence by using the following equation

Confidence(A→B)=support(A ∪B)/support(A)

Where:

***support (A* ∪ *B*)** is the number of transactions containing the itemsets *A* and *B*.

***Support* (*A*)** is the number of transactions containing the itemset *A*.

Because the rules are generated from frequent itemsets, each one automatically satisfies minimum support.

After each Rule generated and computed confidence for it then writing a rule to a text file.

C-Pruning the Set of Association Rules

The purpose of this step is to generate strong association rules from frequent itemsets, which means these rules must satisfy both minimum support value and minimum confidence value. The number of rules that can be generated in the association rule mining phase could be very large. There are two issues that must be addressed in this case. One of them is that such a huge amount of rules could contain noisy information which would mislead the classification process. Another is that a huge set of rules would make the classification time longer. This could be an important problem in applications where fast responses are required.

The pruning method that is study in this paper is keep only those rules that have confidences greater than minimum confidence.

3. Testing Phase

The set of rules that were deducted after the pruning process represent the classifier. This classifier will be used to predict to which classes new emails are belong. Given a new email, the classification process searches in this set of rules for finding those classes that are the closest to be attached with the email presented for categorization. To predict the class of new email ,the new email also passing in tokenization process, stopwords elimination process ,suffix stripping process and compute the frequency for each remaining token. After compute the frequency of tokens, compute the number of rules that the token appeared in the left hand sides and multiply with frequency of tokens, then summation the result of all email tokens. A sum is given to a class related to a new email. This sum is counted for each category for an email. The highest class sum provides the predicted category for an email This process applies for spam class and ham class. To find the class for the new email compare the summation of spam and summation of ham ,the greater number is become the class of new email.

4. Evaluation Experiments

The performance of emails classification method is determined by well known measures used in text classification. These measures are precision and recall and accuracy.Where the **recall** for a class is defined as the percentage of correctly classified messages among all messages belonging to that category, and **precision** is the percentage of correctly classified messages among all messages that were assigned to the class by the classifier. precision and recall[C. Liao et al. ,2004], [E. Clark,2003] which can be computed as follows:

$$Spam\ Precision(SP) = \frac{N_{SS}}{N_{SS} + N_{LS}}$$

$$Legitimate\ Precision(LP) = \frac{N_{LL}}{N_{LL} + N_{SL}}$$

$$Spam\ Recall(SR) = \frac{N_{SS}}{N_{SS} + N_{SL}}$$

$$Legitimate\ \ Recall(LR) = \frac{N_{LL}}{N_{LL} + N_{LS}}$$

Where:
NSS = the number of spam messages correctly classified as spam.
NSL = the number of spam messages incorrectly classified as legitimate.
NLL = the number of legitimate messages correctly classified as legitimate.
NLS = the number of legitimate messages incorrectly classified as spam.

The most popular performance evaluation measure used in classification learning is classifier accuracy which measures the proportion of correctly classifier instances[M. Muztaba et al. ,2005].

$$Accuracy = \frac{N_{SS} + N_{LL}}{N_{SS} + N_{LL} + N_{LS} + N_{SL}}$$

Error rates are calculated as follows:

$$Error = 1 - \ Accuracy$$

. After applying precision and recall equations on number of test emails related to "spam" and "ham" class labels,the accuracy is (87%) and error rate (13%).The results of precision, recall illustrated in the table (1)

Table (1) Classification Measures results

Class Label	Precision %	Recall %
spam	91	87
Ham	82	88

5-Conclusions

While designing and implementing email classification system, the following conclusions are drawn:

1-Using stop words elimination process which discard the words these are not important in the classification process and these words make the classification noise , this process lead to decrease the time of classification.

2- Using suffix stripping process will lead to decrease time .

3-A main problem in email data analysis is the very high dimensionality of the feature space it can easily reach tens and even hundreds of thousands. Only a small number of the terms are feature tokens determining of email class ,while the rest are noise tokens, that make the result unreliable and increase computational time. A common approach in dealing with this problem is feature selection. Using feature selection process by applying (term frequency*inverse term frequency) to determine what words in emails might be more important to be an input to association rules algorithm and this will lead to increase the classification accuracy.

4-Using association rule mining for building classification models is very new. These classification systems discover the strongest association rules in the database and use them to build effective classifiers.Moreover, the rules generated are understandable and can easily be manually updated or adjusted if necessary

References

[C.O'Brien et al. ,2003] C.O'Brien ,and C.Vogel , **"Spam filters: bayes vs. chi-squared; letters vs. words",** Proceedings of the 1st international symposium on Information and communication technologies, Dublin, Ireland, pp. 291-296, 2003.

[D . Trudgian et al. , 2004] D . Trudgian , and Z. Yang, "**Spam Classification Using Nearest Neighbor Techniques",** 2004.

[D.Cook et al. ,2006] D.Cook , and J. Hartnett , "**Catching spam before it arrives: domain specific dynamic blacklists",** Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54, Hobart, Tasmania, Australia, pp. 193-202, 2006.

[H. Al-Kafagi ,2004] H. Al-Kafagi ,**"Pruning Of Apriori s Algorithm Pruning",**Al-Rafiden university colloge,2004.

[H. Katirai ,1999] H. Katirai, **"A Performance Comparison between Genetic Programming & Naïve Bayes", 1999.**

[J. Itskevitch,2001] J. Itskevitch , "**Automatic Hierarchical E-mail Classification Using Association Rules**", M.SC. Thesis, Simon Fraser University, 2001.

[J. Kagstrom,2005] J. Kagstrom, "**Improving Naïve Baysian Spam Filtering**", M.SC. Thesis, Information Technology And Media Department, Mid Sweden University, 2005

[M. Sahami et al. , 1998] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz, "**A Bayesian approach to filtering junk e-mail ",** in AAAI'98 Wkshp. Learning for Text Categorization, Madison, WI, July 27, 1998.

[S. Appavu et al. 2007] S. Appavu alias Balamurugan, and Ramasamy Rajaram, "**Suspicious E-mail Detection via Decision Tree: A Data Mining Approach"** Journal of Computing and Information Technology - CIT 15,, 2,pp 161–169 ,doi:10.2498/cit.1000984,2007.

[T.Ayodele et al., 2010] Taiwo Ayodele, Shikun Zhou, and Rinat Khusainov, "**Email Classification Using Back Propgation Technique",** International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue pp1/2, March/June 2010.

[X.Wu et al.,2009] X. Wu ,and V. Kumar.,"**The Top Ten Algorithms In Data Mining ",**CRC Press,USA, 2009.