International Research Journal of Applied and Basic Sciences
© 2013 Available online at www.irjabs.com
ISSN 2251-838X / Vol, 4 (9): 2436-2441
Science Explorer Publications

A Proposed Algorithm for Spam Filtering Emails by Hash Table Approach

Ali Shafigh Aski*

Department of Computer Engineering, Amol Branch, Islamic Azad University, Amol, Iran.

* ***Corresponding Author email:***alishafigh@gmail.com

ABSTRACT: The number of received emails is constantly increasing; as a result, much time is daily spending on filtering and organizing emails. Therefore, a system is needed which is optimally used to classify and manage emails. Different systems and algorithms have been developed by researchers to automatically classify emails as well as decrease spam. In current study, an algorithm is proposed to classify emails and minimize spam using nearest Neighbor classifier (K-NN). This approach involves low computational load in relatively high rate relying on a Hash table as well as a flag varying in the range of {1, 0}.

Keywords:Electronic mails (emails), Classification, Nearest Neighbor classifier

INTRODUCTION

Currently, spam is considered as a most important problem for internet users. To send emails cost approximately zero and this multiplied spam through internet. Various methods have been considered to eliminate this phenomenon including black and white list, competency-based techniques (Blanzieri and melgani, 2006) and content-based filtering (Brighton and Mellish, 2002)Spam is not only considered as an internet threat but also a social threat.

AOL and MSN report that 2.4 million spamsare daily received in their clients' mailboxes. That is, daily emails sent to AOL are 80% spam. To estimate number of spam is relatively difficult in a global extent; but evidence indicates relative statistics of its high amount. According to European Union, European countries annually receive 10 billion spams(Young and elfayoumy, 2007).

Many large software companies including Microsoft, creator of Outlook Express, implement settings in their software to filter spam; this causes complications for the user. Most users avoid this; on the other hand, manually filtering emails is time-consuming for users. Therefore, a system is required to automatically minimize spam.

Most models developed to minimize spam rate have involved machine learning algorithms. Various systems have been developed to automatically classify emails including systems based on decision-making rules (Bu yu and zong-ben xu, 2008),Bayesian classifiers (Clark j et al,. 2003) support vector machines (Cohen,1996) (Self km, 2003), neural networks (Cui b and mondal et al.,2005)(Self km, 2003) and sample-based methods (Lai cc, 2007) some of which include considerable results. Many studies have been conducted on different aspects of identifying and filtering textual spam on which our suggested method relies.

There are three important strategies in textual emails; considered algorithms can be run for filtering as they are extracted. These strategies include: 1) word characteristics; 2) clause characteristics and 3) structural characteristics.

According to above three factors, current study attempts to extract text characteristics, to classify texts and identify words which are most abundant in Hash table by nearest Neighbor classifier (k-nn) approach and finally do filtering.

The following will study text classification of emails, extracting characteristics of a mail, nearest Neighbor classifier, comparing multiple techniques and suggested algorithm

Classification of Texts

Currently, different classifying techniques have been used to treat spam due to large amount of received emails and spam by users; they can be classified into three groups:

1) rule-based techniques: these techniques attempt to find patterns; thus keywords available in the main body of digital mail and header information and sender address; 2) content-based techniques: they do classification from words available in the header and main text of email; they most use

machine learning methods and 3) sample-based techniques: which do classification based on similarity of received email to correct email sample and spams stored in the memory.

According to conducted studies, the most helpful and widely used technique is machine learning algorithms (Lai cc, 2007) that is a completely automatic process which perform a proper classification by learning a series of pre-classified documents (training data) and characteristics of available classes. The process is explained as follows:

Extraction of Characteristics

To distinguish spams, emails are initially reviewed by extracting characteristics approach. To extract characteristics, various methods have been employed, including Naïve Bayes structure (Mckenna e and smyth b, 2000), support vector machine and tf – id ←, as the most important of them (Oren n, 2002).TF - IDF is a statistical method by which significance of a word is evaluated for a text document.

Significance of a word is related to number of times it is repeated in a text document; but main index is accepted to identify its significance in total documents. Tf - idf (Equation 1) is usually used for internet explorers and to find the most proper available documents.

$$tf = \frac{ni}{\sum_k nk} \qquad \text{Equation 1}$$

where, tf indicates number of times a word is used in a document and determines significance of a word, t_i in a document.

In above formula, n_i is number of times word t_i is occurred and denominatorindicates total occurrence of all selected words of the document. Idf also determines total significance of a word calculated as follows:

$$idf = \log \frac{|D|}{|(t_i cd_i)|} \qquad \text{Equation 2}$$

where, $|D|$ determines all available documents and $|(t_i cd_i)|$ is number of documents in which the word t_i has been used.

Results from tf-idf , Extracting Terms of an email

According to studies conducted on a text email containing 10 lines and some repetitive words using the software SQL Server, it is concluded that index obtained from tf – idfhas good output to an acceptable level. Table 1 shows file values and Table 2 indicates results obtained from this experiment.

Table 1. values of text file in SQL

No.	Text
1	Test exam text
2	Test house text
3	Text Text house
4	nobody where schema
5	nobody where schema
6	nobody where schema
7	Dog Dog Dog
8	Cat
9	Dog Dog Cat
10	Dog Cat Cat

In above file, the word 'schema' has been calculated; according to following formula, results related to Table 2 are obtained:

$$TFIDF\ of\ \ a\ term\ \ T = (frequency\ of\ \ T) * \log(\neq Rows\ in\ input), (\neq Rows\ \ having\ \ T)$$

Table 2. results obtained from text file in SQL

Data	Out put
Select 3*(log (10.0/3.0)	3.191811297779
Frequency	3

Nearest Neighbor Classifier (k-nn)

Various methods have been developed to classify emails; the most common classifier is nearest neighbor method(Sahami , et al,. 1998).(Sahami , et al,. 1998)well described the method. The most important reasons to select the method include easy implementation, speed, local classification and no need for negative samples. This algorithm performs as follows: to classify a new email and determine whether it is Ham or Junk, the email is compared to all existing emails; then, according to K value, K emails similar to experimental emails are selected from each class. Then, average similarity or distance is calculated for each class and an experimental email is assigned to a class which has the most average similarity or nearest distance. The classifier k-nn belongs to sample-based learning algorithms. Similarity of received email or the distance between received email and all existing emails is usually calculated by Euclidean distance as Equation 3 shows:

If $P=(P_1,P_2,....,P_n)$, $q=(q_1,q_2,...,q_n)$

Then $$\rightarrow d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2}$$ Equation 3

Comparing Several Techniques for Spam Filtering

Table 3 shows a summary of advantages and disadvantages or productivity and limitations of several spam filtering methods. Conducted studies indicate that machine learning techniques are more flexible than other methods:

Table3. comparing several techniques related to spam filtering

Methods	Advantages	disadvantages
Content-based filtering	If there is a sign or a pointer of spam, there is no reason to announce spam. The rate of correct email blocking will decrease, even if ISP uses Real time block list.	In some cases system needs to refine its data.
Machine learning techniques	It is highly efficient and consistent as well as difficult to trap. It relies on text classifiers, such as techniques including TF-IDF, Naïve Bayes, N-germ, SVM, Boosting It is highly accurate. It automatically learns new methods and updated techniques produced by spammer. It is highly consistent to changes in spam mood.	Relatively high experimental data is needed. It is weak against emails which do not contain text. It needs many characteristics. It is better consistent with special functions in which user engages. When it is generally considered as final solution, it is not considerably accurate. It needs high calculation effort.
Support vector machine	It has a mood vector in very high sizes. It is able to work with more than 30,000 characteristics.	It is not consistent with non-linear Inseparable models. It needs many training times.
Filtering based on personal and group settings	It is highly consistent and updates contacts who send suspected emails. It eradicates all spams based on group and institutional rules	It uses the protocol SMTP; obviously, SMTP is outdated and it is not recommended. It is highly sensitive to random movements and replacements; it is not yet reliable to generalize the system.

Suggested Algorithm

Primary Works

The first thing to implement email classifying system is to read email, analyze body and header information requiring pre-process.

Remove prepositions (at, as, on …), references (and, or, with …), conjunctive verbs (were, was …), pronouns (I, you …), adverbs of time and place (after, sometimes …) and demonstratives (this, that, it …) and other words which do not influence on stem and determining content of the mail, such as (?/ \ : ; ! ...) .

Verbs and its different tenses including past, present, etc. are extensively wide; thus, verb base is extracted. Here, word stem is considered.

Despite mentioned strategies, the number of characteristic vectors is high; in this case, another method can be added; that is, to remove words which are less than 4 letters. In next step, each mail converts to certain number of words for which iterations are calculated and maintained as follows:

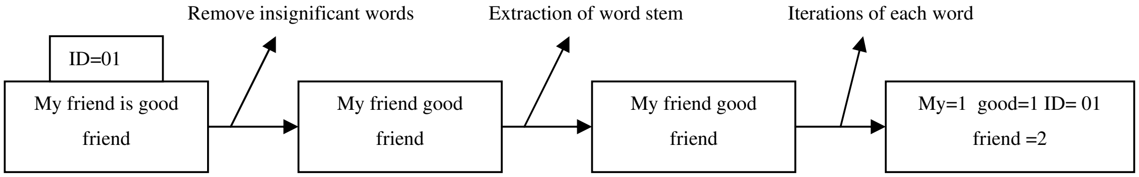

Figure 1: extraction of characteristics and iterations of each word from the first received email

Then, they are stored in database and a record is created.Using above method for an email containing the text ‘my friend is good friend’, Table 4 gives:

Table 4. calculating iterations of words and storing them in Rec.No 01 table

No	Word	Repeat
1	my	1
2	good	1
3	friend	2

According to Figure 2 and Table 5, for an email containing the text ‘this good idea is good for your friend’ we have:

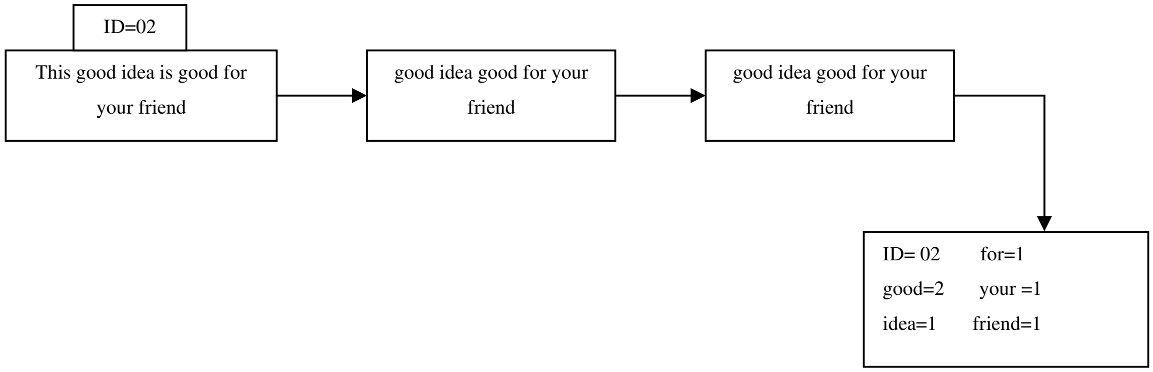
ID=02
This good idea is good for your friend
good idea good for your friend
good idea good for your friend
ID= 02 for=1
good=2 your =1
idea=1 friend=1

Figure 2. extraction characteristics and iterations of a word from the second received mail

Table 5. Calculating iterations of words and storing them in Rec.No02 table

No	Word	Repeat
1	Good	2
2	Idea	1
3	For	1
4	Your	1
5	Friend	1

As email is received in the mailbox, changes are made including elimination of insignificant words, extraction of word stems and counting word iterations. The value of word iterations is determined in a certain record and that record is stored in a table within database (Tables Rec. No 1, Rec. No 2).

Then, values of all records Rec. No_i (Rec. No_1, Rec. No_2… Rec. No_n) are stored in a single table called Recs. Counter to obtain an accurate statistics of words. Table 6 show values for Recs. Counter:

Table 6. Calculating values for word iterations and storing them in Recs. Counter table and attaching Flag field to considered table

flag	No	Word	Repeat	ID
0	1	Good	3	1,2
0	2	Friend	3	2

Here, the most abundant words were stored in Rec. Counter table. The most abundance can occur in different numerical ranges; for example, number of words iterated more than 3 times in the table. In suggested system, all tables are attached to each other as pointers and changes are made as real time. The next step is to transfer values of Rec. Counter table to a Hash table. As said before, Hash table is considerable in terms of speed and accuracy and saving memory; thus, it is an open-linear method.

Table 7. counter table- Recs. Counter

	No.	Word	Repeat	ID
0	1	good	3	1,2
0	2	Friend	3	2

Table 8: identification of received emails by system having ID field

Wno.	1	2	. . .	n
Value	3	3	. . .	n_k
flag	0	0	. . .	0 OR 1
ID	1,2	2	. . .	n

Flag value of the sixth point changed during an hour period. Then, the point is removed and the next value is placed in an empty point of the table and the table is updated.

Table 9. calculating word iterations in emails and transferring considered email or emails to spam folder

Wno.	1	2	. . .	6	. . .	n
Repeat Value	3	3	. . .	4		n_k
flag	0	0	. . .	1		1OR 0
ID	1,2	2	. . .	۵		n

As Table 8 shows, Hash table has a field called flag changing in {0, 1}. Almost all junks enter the system during 12 p.m. to 7 a.m. System is set to initially adjust all flag values of Hash table to zero. Using training emails available in database, Hash table is reviewed in a one-hour period (24-7), for example from 12 p.m. to 1a.m., 1 a.m. to 2 a.m. and ID(s) of which flag values vary from zero to one in considered period are labeled as junk and removed from Hash table and transferred to spam folder. For example, flag value ofID = 5 stored in Hash table (Table 9) changed to 1; as a result, according to training data, it was removed from Hash table and transferred to spam folder. According to Table 7, Table 8 and Table 9, mechanism of the system can be seen.

CONCLUSION AND FUTURE WORKS

The current study suggested a system which scans all emails, extracts their characteristics by a nearest neighbor classifier and a total Hash table and classifies emails using an abundance factor of iterated words. The system is able to extract characteristics whether in header or the body. It is recommended to use better and more efficient classifiers and to examine emails which do not only contain text in the future.

Figure 3. shows total schema of suggested system:

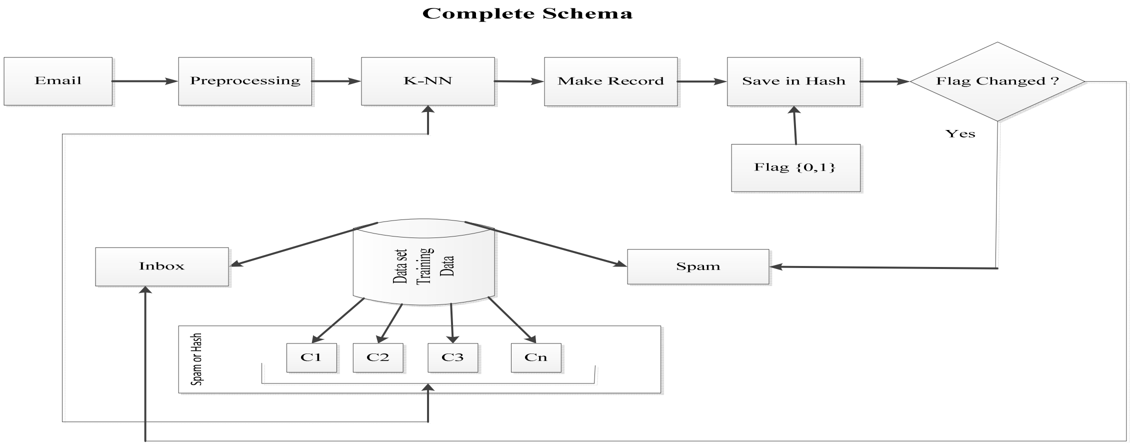

Figure 3. total schema of suggested system

REFERENCES

Blanzieri e, melgani f. 2006. An adaptive svm nearest neighbor classifier for remotely sensed imagery. In geoscience and remote sensing symposium, 2006. Igarss 2006. Ieee international conference on (pp. 3931–3934). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4242156.

Brighton h, mellish c.2002. Advances in instance selection for instance-based learning algorithms. Data mining and knowledge discovery, 6(2), 153–172.

Bu yu, zong-ben xu.2008. A comparative study for content based dynamic spam classification using four machine learning algorithms, knowledge-based systems, 21, 355-362.

Clark j, koprinska i, poon j. 2003. A neural network based approach to automated e-mail classification. In web intelligence, 2003. Wi 2003. Proceedings. Ieee/wic international conference on (pp. 702–705). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1241300.

Cohen w.1996. Learning rules that classify email.in proceedings of the aaai spring symposium on machine learning in information access.

Cui b, mondal a, shen j, cong g, tan kl. 2005. On effective e-mail classification via neural networks. In database and expert systems applications (pp. 85–94). Retrieved from http://www.springerlink.com/index/bww3dqkx2w7jd1yc.pdf.

Drucker h, wu d, vapnik vn. 1999. Support vector machines for spam categorization. Neural networks, ieee transactions on, 10(5), 1048–1054.

Lai cc. 2007. An empirical study of three machine learning methods for spam filtering. Knowledge-based systems, 20(3), 249–254.

Mckenna e, smyth b. 2000. Competence-guided case-base editing techniques , in : proceedings of advances case-based reasoning , 5th european workshop , ewcbr 2000, trento , italy, pp. 186-197 .

Oren n. 2002. Reexamining tf. Idf based information retrieval with genetic programming. In proceedings of the 2002 annual research conference of the south african institute of computer scientists and information technologists on enablement through technology (pp. 224–234).

Sahami m, dumais s, heckerman d, horvitz e. 1998. A bayesian approach to filtering junk e-mail. In learning for text categorization: papers from the 1998 workshop (vol. 62, pp. 98–105).

Self km.2003. Challenge-response anti-spam systems considered harmful. Website, 2004, http://kmself.home.netcom.com/rants/challenge-response.html.

Yang y, elfayoumy s. 2007. Anti-spam filtering using neural networks and baysian classifiers. In computational intelligence in robotics and automation, 2007. Cira 2007. International symposium on (pp. 272–278). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4269929.

Zhang j, jin r, yang y, hauptmann ag. 2003. Modified logistic regression: an approximation to svm and its applications in large-scale text categorization. In machine learning-international workshop then conference- (vol. 20, p. 888).