International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 5, No. 4, 2016, Page: 52-63, ISSN: 2296-1739

www.elvedit.com

Adaptive Spam Email Detection using Support Vector Machines (SVMs)

Authors

Javad Hosseinkhani *jhkhani@gmail.com*
Department of Computer Engineering/ Islamic Azad University, Damavand Branch, Iran *Damavand, Iran*

Mohsen Nematollahi *mohsen8295@gmail.com*
Department of Computer, Zahedan Branch, Islamic Azad University, Zahedan, Iran *Zahedan, Iran*

Mohammad Akhlaghpour *mohammadakhlaghpour@yahoo.com*
Department of Computer Engineering/ Islamic Azad University, Isfahan (Khorasgan) Branch, Iran *Isfahan, Iran*

Horiyeh Heydari Sadegh *hheydari_sadegh@yahoo.com*
Department of Computer Engineering/ Baft Branch, Islamic Azad University, Baft, Iran *Baft, Iran*

Zeynab Sayad Arbabi *z_s_arbabi@yahoo.com*
Department of Computer Engineering/ Zahedan Branch, Islamic Azad University, Zahedan, Iran *Zahedan, Iran*

Ehsan Ahrari *Ehsan.ahrari.1369@gmail.com*
Department of Computer Engineering/ Islamic Azad University, Damavand Branch, Iran *damavand, Iran*

Abstract

Today, many spam attempts to make difficulty with email connections. In this article we try to expose a way regarding spam identification based on Support Vector Machines (SVMs). Based on this method on delivery email three steps should be occur first of all a reoperation then flowing data. In operation step the user is sending an email preprocess is done by data miner system. The number of training information apply with window based solution will be selected with default, W=100, the first 100 data would be used as training category. Each delivery email input to SVM to be sorted in to 2 predetermined categories named: Non spam, and Spam. An algorithm is written that 4 different types of time window in order to SVM training is selected (100,200,500 and all the preset data or open window). The criteria for assessing include accuracy rate, recall, and precision rate. The results show that the techniques that some specialists have some criticisms to it.

Key Words

Spam Detection, Email Classification, Support Vector Machine (SVM).

I. INTRODUCTION

Almost all Web users make use of postal mail to communicate electronically. These people count on the actual postal mail process to produce their particular mails to the beneficiary. Spam mail has built the actual postal mail process more hard to rely on simply because postal mail can get falsely captured simply by unsolicited mail filtration system in route to the beneficiary or perhaps postal mail could die amid unsolicited mail within the recipient's inbox. The goal of the internet community ought to be to operate toward a far more workable World-wide-web with less spam mail. Achievable methods to do this are generally over the legislation and also the legitimate process, techie remedies and also user awareness.

Not long ago, the huge number of electronic mail spams provides caused significant issues within important email communication. Conventional spam mail filters goal at examining electronic mail written content to help define your capabilities which have been normally incorporated into spams. Nonetheless, it's discovered which crafty tricks built to stay clear of content-based filters are going to be limitless on account of the fiscal advantages of sending spams. Because of this scenario, there has been a lot study attempt toward carrying out junk mail prognosis in line with the trustworthiness of senders in lieu of what is from electronic mails. Enthusiastic through the fact those spammers are prone to get strange habits as well as particular styles of electronic mail conversation, checking out electronic mail support systems to help detect spams provides obtained a lot interest [19].

Unwelcome email messages also referred to as spams, is more important.in a research by Message- Labs in 2006 the statistics on the number of spams shows that there was 40% (12.4 billion from 31 billion every day) associated with email messages were being considered as spams. The serious problem is that the spam rate persistently remains high. Traditional filter systems, for example Naive Bayes classifiers, happen to be commonly handed down by means of imprecise rules along with hit-or-miss part gain access to. Together with essential information that will spammers wish to disclose, there are numerous unimportant contents included into spams. Due to the fact awesome methods working at mail written content tend to be perpetual, studies in this area have focused on show who is sender of the mail as an alternative to what exactly is contained in the mail.

Researches has an interesting on data mining .more interesting of them is on systems for analyzing this things. Data mining play a basic role in real world such as network traffic scanning, intrusion identification, web click-stream interpret and credit card cheat detection [18]. Nowadays Most of researches studies on projects on fast mining algorithms ,it is a huge data streams in which can be sense with real-time answers [10, 8, 9, and 7].basically , a lot of studies have devoted to producing the information streams constructed from these purposes [11, 12, and 13]. Experiments with promoting exploration algorithms are lower [14]. This situation seems big difference; it needs exploration technique with regard to static information exploration, has been

immediately tolerable [17] and possesses one on one to systems such as, Weka [15] and also OLE DB with regard to DM [16]. Additionally, static exploration algorithms made inside procedural language work with a cache exploration approach which makes tiny use of DBMS principles. Nonetheless, online exploration tasks can't be arranged while set-alone algorithms, simply because they call for a lot of DSMS standard, For instance I/O buffering, windows, synopses, load shedding, and so forth. Evidently, KDD research workers and also practitioners prefer to devote to the down sides involving information mining tasks and also pruning the complications involving controlling data streams, through allowing the exploration technique cope with. In brief, even though exploration systems are using with regard to protecting, they are necessary for data streams.

A method of ordering linear and nonlinear data is Support vector machines (SVMs). In a case, SVM is an algorithm and the function of it is as follows. To change the original training data into a higher dimension, it applies a nonlinear mapping. It seeks for the linear ideal separating hyper plane through this new dimension. A hyper plane can always separate the data into two classes with a suitable nonlinear mapping to an appropriately high dimension. The SVM discovers this hyper plane utilizing support vectors that is “essential” training tuples and margins which is explained by the support vectors. Vladimir Vapnik and colleagues Isabelle Guyon and Bernhard Boser (1992) have done the first research on support vector machines since the groundwork for SVMs has been around. Even though the training time of SVMs is very extremely slow, they are very precise and can to model compound nonlinear decision limitations. In compare to other methods, they are much less predisposed to over fitting. The provision vectors also are a compressed explanation of the trained model. SVMs also are able to utilize for numeric calculation along with classification. They have been used for many areas such spam email detection. [6].

There are More strategies on anti-spam.in this article we try to explain a research on spam protective strategies namely Support Vector Machines (SVMs). It is a good theoretical foundation, mount with large data, and present itself to the text classification problem. In this study, we perform a function of SVMs. For doing this study, first the received emails would be preprocessed then stream data in order to learning the classification would be given to the proposed data miner system. The number of training data set with window based solution will be selected with default, W=100, the first 100 data would be used as training set. Each received email input to SVM to be classified in to 2 predefined classes named: Non spam, and Spam. A programming that 4 different kinds of time window in order to SVM training are selected (100,200,500 and all the preset data or open window). There are some criteria for assessing such as rate, recall, and precision rate. But some researchers are disagreeing with these assessing criteria.

II. RELATED WORKS

Nowadays email spams are more problematic. There are some approaches to relieve this problem. Based on characteristics of email there are three factors for this problem: (1) content-based methods, (2) non-content-based methods, and (3) integrated methods. To start with; researchers consider written content involving e-mail and also form this specific side effect like a binary text categorization activity. Proponents in this classification are generally Naive Bayes [20, 21] and also Support Vector Machines (SVMs) [22, 23] methods. In general, Naive Bayes methods train some sort of possibility style employing grouped email messages, and also every single word inside email messages will probably be offered some sort of possibility to be some sort of dubious spam mail search phrase. As for SVMs, this is a supervised learning technique, which in turn boasts outstanding efficiency about text classification tasks. Conventional SVMs [23] and also improved SVMs [22] are actually looked at. Since above arbitrary machine learning methods are in effect with fixed data categories, one important favorable is that it is expense restrict to regularly engaging these techniques with current data to conform to quick developing characteristic of spams. In addition, artful ideas ambiguous cheating has always been followed to debase the performance of these techniques. In contrast, selected unique attributes including Urls [24] in addition to images [25] are taken into account for junk e-mail recognition.

The other team attempts to help task non-content info including e mail header, e mail traffic [26], along with e mail social network [27, 28] to reduce spams. At first, Gathering disreputable and simplicity delivery addresses (or IP addresses) via e mail header to generate monochrome list is surely a normal used techniques. In [26], researchers try to consider methods for analyzing email traffic streams to preserve irregular machines and anomaly sending and receiving email. In addition, an authentic distinction system is created to put in a huge webmail service. This system is performed by the previous behavior of each sender with SPF and Domain Key authentication.

Moreover, some researchers study in combining the advantage of some methods [30, 31, and 32]. Since the using of category incorporates looks important, there is still no result on what is the best category. Furthermore, how to mastery update the whole included arrangements is another unanswered problem.

In [33], particular network based on characteristics are developed to identify each user. An improved k-Nearest Neighbor (k-NN) model is then applied to deal with the particular junk e-mail categorization. Throughout [34], graph theoretical research of systems is conferred to find the very best distinguish involving tolerable electronic mails in addition to spams. Throughout [35], the particular authors propose an email rating way which interprets respected reviews involving persons in networks. Throughout [27], the particular writers take advantage of the particular characteristic of clustering coefficient in systems in order to formulate a new recognition system. At the least, these kinds of activities are generally then a couple troubles.

First, they are not powerful in different conditions. The other difficulty is that the refresh design, which is all-important for developing networks, has been neglected in these works.

III. SUPPORT VECTOR MACHINES

We study spam detection as a text categorization problem. There are two categories for email content: y_i€ {-1, +l} where -1 represent no spam and +1 spam. An aspect is a word in email content and an aspect vector x_i describe an email in the peculiar place. Given n labeled instruction examples: (x_1,y_1)... (x_n,y_n), the activity is to understand from the instruction examples a hypothesis that can be applied for categorizing hidden email contents.

Support vector machines can be a member in learning tactics [l]. Linear hard-margin SVMs can be a distinct design within SVMs and they also named the actual maximum perimeter groups designed to use simply for info linearly a part in the specific circumstance. The particular linear hard-margin SVMs detachment many facets of vectors in the couple of categorize by means of thinking about a new hyper plane with maximum perimeter. The suitable vectors closest thing towards hyper jet named service vectors. The particular maximum perimeter hyper plane related the overall glitches on the linear machines presented a new training approach S, and can always be gotten by means of capitalizing on the actual coaching.

$$W(\alpha) = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j \boldsymbol{x}_i . \boldsymbol{x}_j$$

Soft-margin SVMs [1] can be employed with regard to non-linearly separable facts. Soft-margin SVMs let training mistakes. The optimization trouble at this point gets making the most of the following:

$$W(\alpha) = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j \boldsymbol{x}_i . \boldsymbol{x}_j$$

The C could be the parameter in which we have to track to make the design match towards the non-linearly separable facts. The soft perimeter SVMs act similar to hard-margin SVMs if the parameter C is large plenty of. Observe [2, 3, 4, and 5] with regard to details.

IV. RESEARCH DESIGN AND PROPOSED FRAMEWORK

According to Figure 1, first the received emails would be preprocessed then stream data in order to learning the classification would be given to the proposed data miner system. The

number of training data set with window based solution will be selected with default, W=100, the first 100 data would be used as training set.

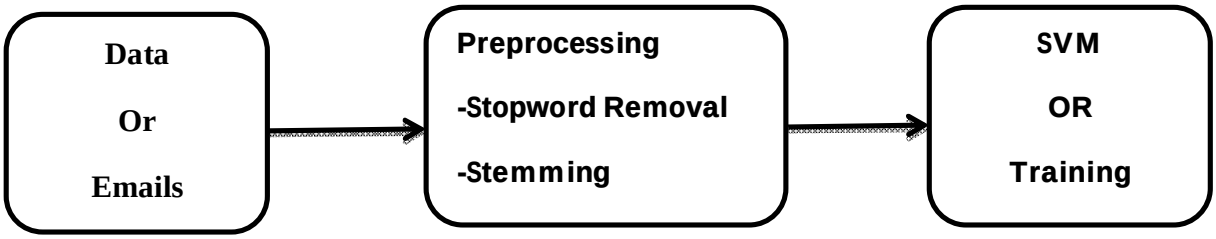

FIGURE 1: PROPOSED FRAMEWORK

Before sending email or data there should be some preprocessing on them. This action named stop word .pruning Stop words are the words that frequently use in the internet and facilitate forming sentences but do not show any content of the documents. Content, prepositions and conjunctions and pronouns are samples of this kind of stop words. In several dialects, any concept offers various syntactical varieties dependent on which in turn wording it truly is employed. For example, within English, nouns can be utilized within plural varieties, verbs have also within gerund varieties (by including “ing”), and kind of verbs in the past anxious are wide and varied from the current anxious. Most of these variations recognize seeing that syntactic adjustments in the exact same main kind. Such adjustments complete a way of identification for search engines like google since a suitable spam mail electronic might include a difference of an query concept although not the precise concept itself. This issue might be partially deemed simply by stemming. Figure 2 illustrate that each received email input to SVM to be classified in to 2 predefined classes named: Non spam, and Spam. Data set Usnet1 and Usenet 2 are applied in order to training and proposed data miner learning.

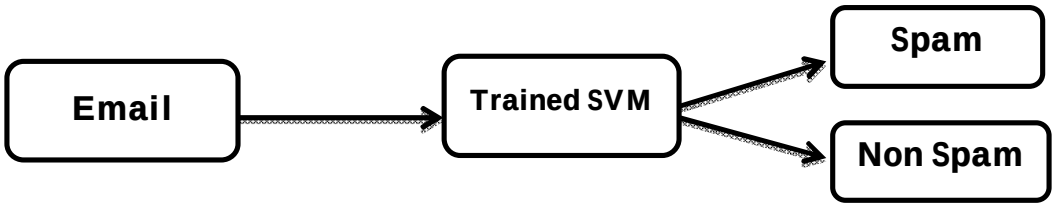

FIGURE 2: CLASSIFICATION OF RECEIVED EMAILS TO SVM

V. EXPERIMENTAL RESULT

Table 1 shows the stream data classificiation exprimental results through using SVM in Spam data set. In this table, 4 different kinds of time window in order to SVM training are selected (100,200,500 and all the preset data or open window) that 3 evaluations criteria’s

including precision , recall and accuracy are evaluated that is shown in Table 1. Figures 3 to Figure 6 show the mean of vector precision for 1000 experimental samples.

TABLE 1: STREAM DATA CLASSIFICIATION EXPRIMENTAL RESULTS

Label	Precision	Recall	Accuracy
Simple Incremental	0.9987	0.9172	0.9320
Time Windows(W=100)	0.9600	0.9165	0.9140
Time Windows(W=200)	0.9660	0.8954	0.9050
Time Windows(W=500)	0.9937	0.8275	0.8990

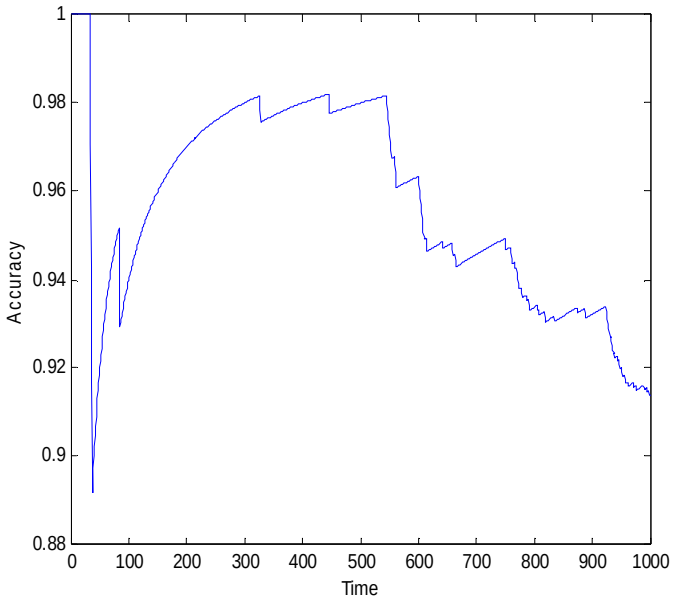

FIGURE 3: THE MEAN OF VECTOR PRECISION FOR 1000 EXPERIMENTAL SAMPLES

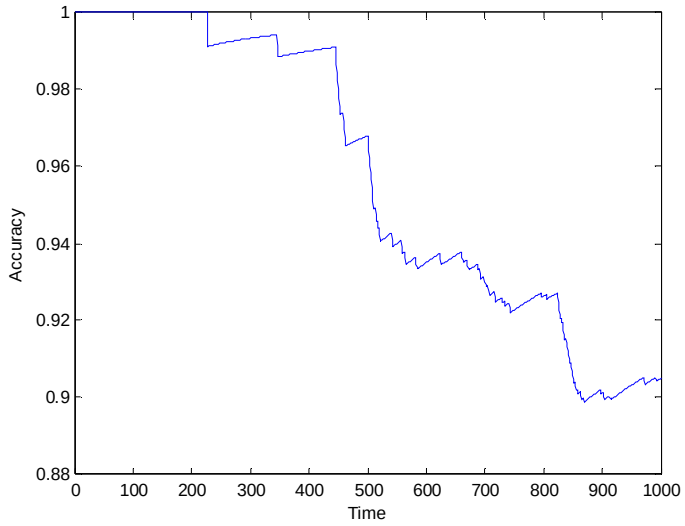

FIGURE 4: THE MEAN OF VECTOR PRECISION FOR 1000 EXPERIMENTAL SAMPLES

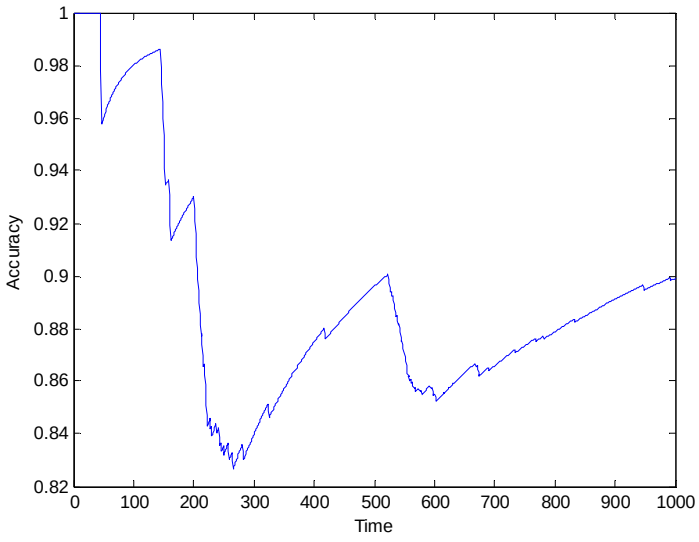

FIGURE 5: THE MEAN OF VECTOR PRECISION FOR 1000 EXPERIMENTAL SAMPLES

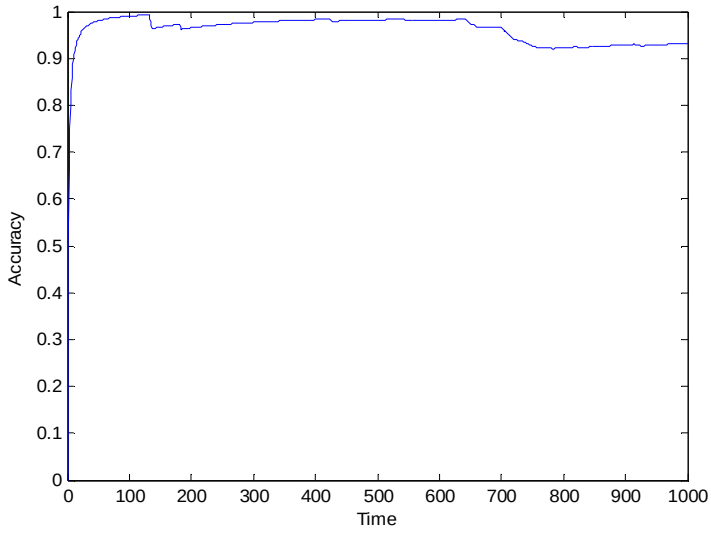

FIGURE 6: THE MEAN OF VECTOR PRECISION FOR 1000 EXPERIMENTAL SAMPLES

VI. CONCLUSION

Many researchers have an interest throughout issue connected with On-line Data mining, nevertheless analysis in systems with regard to on-line exploration is incredibly low simply because engaged using troubles. On-line data stream explorations have a very simple function throughout raising number of authentic software, for example community traffic monitoring, intrusion detection, World Wide Web click-stream analysis, along with bank card fraud detection. Many initiatives get recently manufactured a shot with regard to building rapidly exploration algorithms simply because a lot of data streams are usually attend using actual-time reaction.

In this research, the particular linear SVMs possess some strength. One of these simple strengths is actually of which SVMs usually are a smaller amount impacted by how many training instances inside the a pair of classes since they are not really participate in reducing the particular error rate, but instead make an effort to separate the ones with professional space. Since, the particular action is the powerful concern. In the event that there are large numbers of training point, learning process of action could make long time. Execution could make a slower rate regarding nonlinear SVMs.

For the result, such research needs to be done with a larger number of data. This study also concluded that that the performance of categorization methods really in affection together with training cases, i.e. the particular characteristic vectors taken on the traditional mail contents.

Within these kinds of researches, simply words in the content of information were being applied since the sign on the attributes. A lot more vital symbolism might be loosed in the feature extraction actions. For example, the issue subject of e mail is most likely the beneficial sign on the attributes. In addition, latest unsolicited mail articles are usually coded together with html so it could be the best actions to be able to involve the html limitations in the attributes. The preprocessing before training would be the crucial methods to the learners to execute much better categorization. For the additional model, the spam e-mail intended for training and studying need to be engaged while using the multiple categories in line with the types of spam emails like factor and trips. This specific makes the outcomes of researches much more valuable and accurate.

REFERENCES

[1] Matsumoto, R., Zhang, D., & Lu, M. (2004, November). Some empirical results on two spam detection methods. In Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on (pp. 198-203). IEEE.

[2] Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory and algorithms (p. 205). Kluwer Academic Publishers.

[3] Osuna, E., Freund, R., & Girosi, F. (1997, September). An improved training algorithm for support vector machines. In Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop (pp. 276-285). IEEE.

[4] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

[5] Vapnik, V., & Kotz, S. (2006). Estimation of dependences based on empirical data. Springer.

[6] Moradi Koupaie, Hossein, Suhaimi Ibrahim, and Javad Hosseinkhani. "Outlier Detection in Stream Data by Machine Learning and Feature Selection Methods." International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol 2 (2014): 17-24.

[7] Mozafari, B., Thakkar, H., & Zaniolo, C. (2008, April). Verifying and mining frequent patterns from large windows over data streams. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on (pp. 179-188). IEEE.

[8] Chu, F., & Zaniolo, C. (2004). Fast and light boosting for adaptive mining of data streams. In Advances in Knowledge Discovery and Data Mining (pp. 282-292). Springer Berlin Heidelberg.

[9] Wang, H., Fan, W., Yu, P. S., & Han, J. (2003, August). Mining concept-drifting data streams using ensemble classifiers. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 226-235). ACM.

[10] Forman, G. (2006, August). Tackling concept drift by temporal inductive transfer. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 252-259). ACM.

[11] Law, Y. N., Wang, H., & Zaniolo, C. (2004, August). Query languages and data models for database sequences and data streams. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 492-503). VLDB Endowment.

[12]Arasu, A., Babu, S., & Widom, J. (2004, January). CQL: A language for continuous queries over streams and relations. In Database Programming Languages (pp. 1-19). Springer Berlin Heidelberg.

[13]Abadi, D. J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., ... & Zdonik, S. (2003). Aurora: a new model and architecture for data stream management. The VLDB Journal—The International Journal on Very Large Data Bases, 12(2), 120-139.

[14]Thakkar, H., Mozafari, B., & Zaniolo, C. (2008, March). Designing an inductive data stream management system: the stream mill experience. In Proceedings of the 2nd international workshop on Scalable stream processing system (pp. 79-88). ACM.

[15]Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[16]Tang, Z., Maclennan, J., & Kim, P. P. (2005). Building data mining solutions with ole db for dm and xml for analysis. ACM SIGMOD Record, 34(2), 80-85.

[17]Imielinski, T., & Mannila, H. (1996). A database perspective on knowledge discovery. Communications of the ACM, 39(11), 58-64.

[18]Thakkar, H., Mozafari, B., & Zaniolo, C. (2008, December). A data stream mining system. In Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on (pp. 987-990). IEEE.

[19]Tseng, C. Y., & Chen, M. S. (2009, August). Incremental SVM model for spam detection on dynamic email social networks. In Computational Science and Engineering, 2009. CSE'09. International Conference on (Vol. 4, pp. 128-135). IEEE.

[20]Hovold, J. (2005, July). Naive Bayes Spam Filtering Using Word-Position-Based Attributes. In CEAS.

[21]Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes?. In CEAS (pp. 27-28).

[22]Blanzieri, E., & Bryl, A. (2007). Evaluation of the highest probability SVM nearest neighbor classifier with variable relative error cost.

[23]Li, S., Kwok, J. T., Zhu, H., & Wang, Y. (2003). Texture classification using the support vector machines. Pattern recognition, 36(12), 2883-2893

[24]Schneider, K. (2004). Brightmail url filtering. In Spam Conference

[25]Dredze, M., Gevaryahu, R., & Elias-Bachrach, A. (2007, August). Learning Fast Classifiers for Image Spam. In CEAS.

[26]Clayton, R. (2007, August). Email traffic: a quantitative snapshot. In CEAS.

[27]Oscar, P., & Roychowdbury, V. P. (2005). Leveraging social networks to fight spam. IEEE Computer, 38(4), 61-68.

[28]Chirita, P. A., Diederich, J., & Nejdl, W. (2005, October). MailRank: using ranking for spam detection. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 373-380). ACM.

[29]Taylor, B. (2006, July). Sender Reputation in a Large Webmail Service. InCEAS.

[30]Hershkop, S., & Stolfo, S. J. (2005, August). Combining email models for false positive reduction. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 98-107). ACM.

[31]Cormack, G. V., Smucker, M. D., & Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval, 14(5), 441-465.

[32] Segal, R. (2007, August). Combining Global and Personal Anti-Spam Filtering. In CEAS.

[33]Castillo, C., & Davison, B. D. (2011). Adversarial web search. Foundations and trends in Information Retrieval, 4(5), 377-486.

[34]Moradi, F., Olovsson, T., & Tsigas, P. (2012, April). Towards modeling legitimate and unsolicited email traffic using social network properties. InProceedings of the Fifth Workshop on Social Network Systems (p. 9). ACM.

[35]Golbeck, J., & Hendler, J. A. (2004, July). Reputation Network Analysis for Email Filtering. In CEAS.