Filtering Spam by Hybrid Approach

Mohamed Tabris, Youssif B. Alnashif, and Salim Hariri
Autonomic Computing Lab
the University of Arizona
m.tabris@gmail.com, *alnashif@ece.arizona.edu*, and *hariri@ece.arizona.edu*

***Abstract*— Spam has become a main cause of financial loss in most of the organization. It was seen that 81.6% of the email Traffic in 2006 was spam [1]. The loss incurred by the companies is growing exponentially and so is the number of spam emails. This makes spam detection and spam filters are critically important. There are various techniques used in order to filter spam, two of most prominent techniques available are IP blacklisting/white-listing and content-based filtering. An email is identified as spam based on the reputation of the source done by grey listing the source from their previous history [2]. In this paper we introduce a method for improving the spam filters by using a hybrid technique (Content Based & Anomaly based detection approach). In this work, we show how to identify whether the email is spam or not by implementing models that capture the nature of emails' headers and patterns found in the emails' content. The general behavior of spam and legitimate emails for each of these models is obtained and assigned a score; the value of this score is used to differentiate between a legitimate emails and spam. By using this hybrid approach, we were able to detect spam with a false positive rate of .54% and a false negative rate of 1.34%. We also discuss the relation between phishing and spam and how some anti-phishing techniques can be used in spam filters.**

***Index Terms*— Hybrid Spam Filtering, Spam, Anomaly based, Spam Detection, Email Protection**

I. INTRODUCTION

Spam is defined as unsolicited, unwanted email that endangers the very existence of the e-mail system with massive and uncontrollable amounts of messages [3]. Spam is used by bots to overload the email inboxes and increase the traffic on a server which eventually leads to DoS. A number of approaches have been taken in order to defend against DoS attacks [4]. Around 80% of the received emails today are spam. According to surveys, companies spent around $600 million in 2003 in order to prevent Spam and according to a recent survey by university of Maryland the cost has gone up to 22 Billion $/year [5]. Even after such precautions spam is estimated to cost a total of $130 billion worldwide, of which $42 billion is in the U.S. alone [6].

The rest of the paper is organized as follows. In section 2, we present an overview of related work is. Section 3 presents an overview of the framework and presents insights into our approach to the modeling methodology. Section 3.4 describes the characteristics of spam headers and the models developed to evaluate the authenticity of SMTP/email headers. Section 3.5 deals with describing the content reliability model and different Anti-Phishing techniques we incorporate in this model. Section 4 the preliminary results of the proposed approach. Section 5 presents the conclusion and outlines the future research activities.

II. RELATED WORK

There are several approaches applied in to block spam and phishing emails like blacklisting, whitelisting, signature/content based techniques, and rule-based anomaly methods. But these approaches have failed because spammers have been successful in finding some loop holes in these techniques. A brief description of these approaches is given below.

A. Blacklisting

This method uses the history of the sender's IP address as the criteria to filter email. According to this approach, if the IP address of the sender is involved with spamming in the past, then it is added to a list of spammers and the emails is blocked from reaching the targeted inbox. Any email received from an IP which is blacklisted is considered to be a spam. Spammers have come up with many ways to evade Blacklisting; sending less amount of spam over a certain amount of time and keep changing the IP address by using dynamic IP addresses so that they do not get blacklisted. Stealing IP blocks using BGP Route Hijacking [7], stealing IP from local networks is another common approach. One of the most common methods the Spammers use is by hiring Bots(Spam Campaign) and sending huge number of mails in a short period of time and by the time the IP is blacklisted thousands of spam have already been sent which result in heavy traffic or DoS attack. This method assumes IP address of the sender to be static which is not true nowadays as the machines generally get the IP address from a pool of available IP's (Dynamic IP allocation) and hence it is not very effective. This method also requires the blacklists to be updated very often which makes it not very attractive.

B. White-listing

This method is very similar to Blacklisting; it also uses reputation of the IP as the criteria to differentiate between spam and legitimate mail. Instead of maintaining a list of Blacklisted IP addresses, it maintains a list of allowable IP addresses which do not have any reputation of sending spam mails. It shares the same disadvantages as Blacklisting, the list has to be updated very often as IP addresses [8] can start sending spam mail suddenly and IP addresses do not remain

The research presented in this paper is supported in part by National Science Foundation via grants numbers IIP-0758579, CNS- 0855087, and IIP-1032048, and its conducted as part of the NSF center for

COLLABORATECOM 2011, October 15-18, Orlando, United States

DOI 10.4108/icst.collaboratecom.2011.247188

static so the chances of a legitimate IP address being removed from the white list are high.

C. Content Based Filtering

There are many filters available which use content of the mail as the criteria for filtering spam [9]. One of the most popular and effective example is the Bayesian approach [10, 11]. A spam dictionary is set up which consists of the probability values of various well known signatures. In this method the probability of few unique signatures occurring in a Spam is calculated from spam word's dictionary and each mail is checked for these signatures and the total probability (Bayesian Probability) is calculated which helps to decide whether it is a spam or not. In order to degrade the performance of Bayesian filters spammers have come up with a method called Bayesian poisoning [12, 13]. In this method, Bayesian Probability is degraded by using certain signatures which frequently occur in legitimate mails or by modifying a word by adding some special characters along with it. Another disadvantage of this method is that it has to update spam word's dictionary at regular intervals [14].Other disadvantages of this approach are given in [15].

D. Anomaly Based Approach

Anomaly Detection involves identifying observations that deviate from the normal behavior of a system [16]. In this method we characterize a system's normal behavior and any anomaly in the system's behavior is detected. In email system we study the characteristics of normal and spam mails and define the region of normal operation and anything deviating from this is considered to be a spam. This method overcomes the disadvantages of the Content-Based filters as the Anomaly Detection is adaptive and changes with time in accordance to changes in system behavior. The main challenges of this technique are defining the normal region, the used features, and the training period.

III. Hybrid Approach

A. Introduction

We have discussed various approaches taken in order to filter out spam, out of which content- based filtering is very popular and widely used. In this paper, we have developed a Hybrid Approach which utilizes the advantages of anomaly based techniques and content based filtering. In this approach we have applied techniques similar to that of Anti-Phishing for spam detection. We have divided the spam filter into three models i) Anomaly Based Model, ii) Content Based Model, and iii) Reputation Based Model. Anomaly based model has a number of weak estimators which help us categorize the authenticity of the email headers. The Content Reliability Model checks for some basic patterns in the body of the Mail and uses anti-phishing techniques when any hyperlink appears in the body. Our proposed system integrates the desirable aspects of both categories of spam detection system (Anomaly & Content Based).

B. Hybrid SPAM Filter

The main design principle of our filter is to extract headers and type of content in email body such as text, HTML, Images. In this framework, we create models for dealing with header and Text/HTML content. Anomaly based models takes the SMTP/email headers as the input and the body of email is given to the content based models.

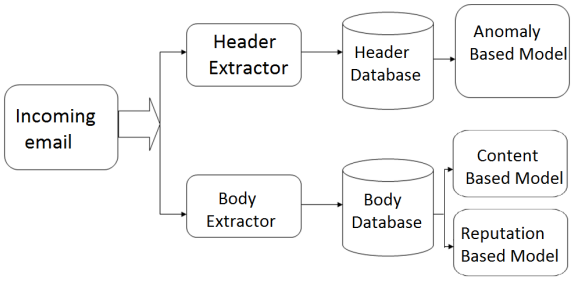

Figure 1. Block Diagram of Hybrid SPAM filter.

The block diagram of our framework is given in Figure 1. The internal Block diagram of each of the models will be discussed later in the following sections.

C. Anomaly Based Model

The basic block diagram of the anomaly based model is shown in Figure 2. The system is basically divided into three phases i) Data Collection phase, ii) Training, and iii) Decision Making module. Data Collection Module is used to collect the emails and pass them to the extractor. The extractor is used to extract the header and store the data in the Database.

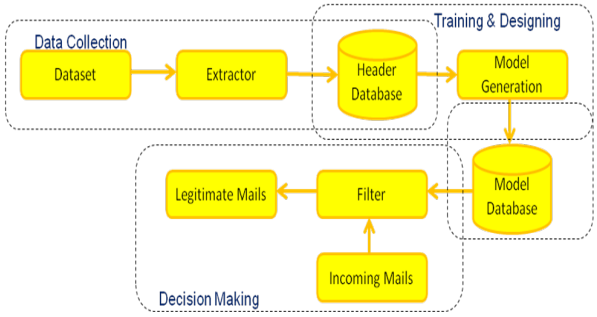

Figure 2. Basic Block Diagram of Anomaly Based Model.

The training module uses the data from the database in order to analyze various characteristics of SPAM and legitimate mails and do appropriate feature selection. Based on the features selected a number of models have been designed in order to distinguish between SPAM and legitimate emails. This model uses the characteristic of normal email to define a region of normal operation.

1) Model Based on Characteristics of Spam Headers

It is often noted that the Headers of the mails are being manipulated in order to avoid spam detection by IP blacklisting or Signature based method. Headers in spam mail can be used to detect the source or origin of the mail and hence it necessitates the spammers to forge them in order to conceal their Identity. In order to find out the characteristics of a SPAM mail and legitimate mail we studied a dataset of more than 5000 spams and 3000 normal emails. A set of characteristics of headers in SPAM are observed which can be used as indicators of forging. One of the most common ways used by the spammer is by changing the header such that it

looks like the email has been sent by the receiver itself ; this is known as *Identity forgery* [17]. An example of an email header that fells into our spam trap is given in figure 3.

Sample Email Header:-
From: VIAGRA ® Official Reseller
<mtabris@gmail.com>
To: mtabris@gmail.com
Subject: For mtabris! Discount ID99626
Content-Type: text/html; char set="ISO-8859-1"

Figure 3. Spam Email Header illustrating the Spam Header Characteristics

2) Degree of Randomness Model

One of most efficient ways of sending huge amount of spam in a short period of time deployed by spammers is by hiring bots. Bots are machines which send out huge number of spam mails in a small period of time. Bots are one of the major security threats [18, 19, 20]. The Botnets are rented for a small amount of time and they are able to send thousands of mails in a small amount of time. In order to identify a spam mail we observe characteristics of Bots and developed a model to identify emails sent by Bots. Spammers collect a huge number of email Id's either by using tools called **Harvesting Bots** [21] or they purchase them from other spammers. In either way the information regarding the target is limited (just the email ID, username etc). To make their spam look real they try to use this information (Username) in the subject field and they try to address them with their username rather than their real name. The characteristic of a mail sent by Bots are observed from the data collected and it is observed that the degree of randomness in the mail sent by Bot is very less when compared to mails sent by humans. An example of spam mail header collected by our spam trap is given in figure 4.

Sample Email Header:-
From: VIAGRA ® Official Reseller
<mtabris@gmail.com>
To: mtabris@gmail.com
Subject: For mtabris! Discount ID99626
Content-Type: text/html; char set="ISO-8859-1"

Figure 4. Spam Header - Bot Characteristics.

From the sample header we notice the following

1. *Domain name is same for all recipients*
2. *Usernames are quiet similar*

It is often seen that the domain name is the same since the probability of a Bot sending mail to different domains is small. We use features to exploit this property of Bots, by checking the randomness property in email Id's, Domain name, etc... We have selected three main features in order to find the proximity of usernames and domain name. The features are given below:

Avg_Pos : Average number of similar characters in username

Avg_Con : Average number of consecutive similar characters regardless of position.

Avg_Dom_Sim: Average Similarity in the Domain Name of the receivers.

The 2-Dimensional and 3- Dimensional plots obtained after applying these features on our database are given in figure 5. It is clear that almost all spams collected by our spam trap have been directed to receivers of the same domain. It is also clear from the 3-D plot that the Degree of Randomness in usernames receivers is more in normal email when compared to the Spam.

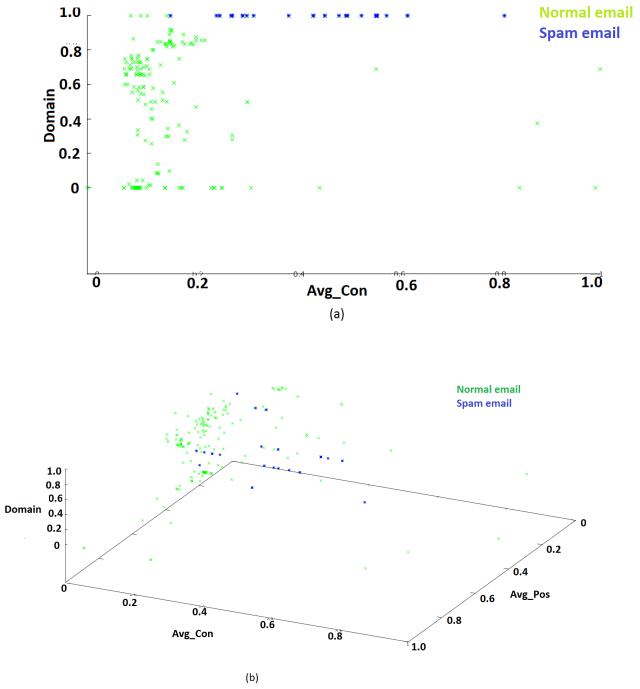

Figure 5. a) 2-D Plot demonstrating similarity of Domain Name b) 3-D plot- Demonstrating Degree of Randomness in usernames

D. Content Reliability Model

This Model is designed based on the characteristics of the body in spam emails collected by our Spam Trap. Most of the spams are sent with an intension of phishing or advertising products. Recent Internet trend analysis report by Commtouch (Figure 6) shows the distribution of spam sent/received during the first-quarter of 2010[22].

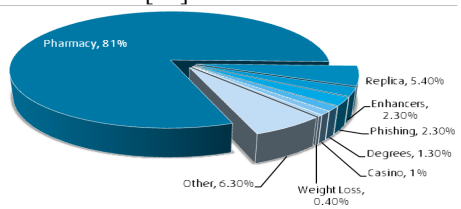

Figure 6. Spam Distribution in first-quarter of 2010.(Source:Commtouch)

It is clear that 93.7% of the spam mails sent/received were advertisements and phishing mails. It is obvious that every spam sent with the intention of advertisement or phishing will contain hyperlinks or domain names which they want the receiver to visit. In our spam trap around 95% of the spam collected had more than one hyperlink. The block diagram of

the Content Reliability Model & Reputation Based Model is given in figure 7.

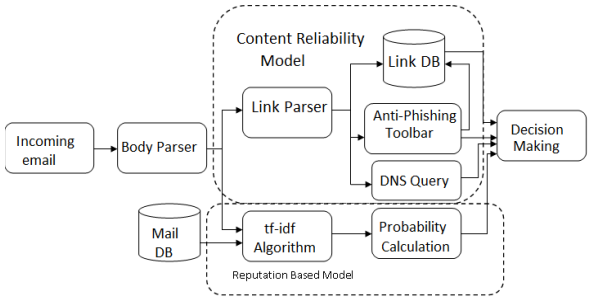

Figure 7: Block Diagram of Content Reliability Model & Reputation Based Model

Hence we evaluated the reputation of Domain Names used in the spam mails by using one of the leading Anti-Phishing Toolbar, Web of Trust (WOT) to see whether they will be able to classify these Domains as spam. We have used the mywot toolbar API which helps us evaluate the hyperlinks present in the email [23]. This toolbar evaluates a Domain Name based on four parameters Trust, Reliability, Privacy and Adult Content. The plot obtained by applying the Anti-Phishing toolbar technique on the database is given in figure 8. It is seen that 98% of the domain names present in spam emails of our database have parameter values less than 80.

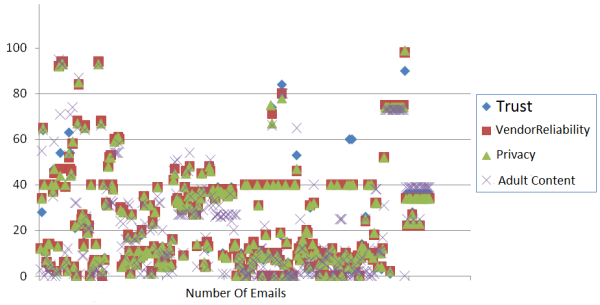

Figure 8. Anti-Phishing Toolbar Plot

Some of the domain names do not have enough history to build a reputation. In order to address those domain names we query the DNS registry with the domain name to get the information about the owner, time for which the domain is active. This information helps us to find out the where about of the Domain name by using *who* command and check whether it was recently created to spam the network. The older the domain name the more reliable it is.

E. Reputation Based Model

In this model, decision is made by skimming through the body of the mail. The method we use is similar to that which is use in CANTINA to do Anti-Phishing [24]. Information retrieval algorithm like *tf-idf* is implemented in order to search for all words in the body of the email and select top 'n' rare/unique words occurring in the Database. The *tf-idf* Algorithm is explained in [25]. *Term Frequency* (*tf*) is defined as the number of times a term t_i appears in the particular mail. *Inter-Document Frequency* (*idf*) is defined as the frequency at which the term t_i appears in the corpus (Database). We use the idf value to sort the numbers based on the rarity and choose 'n' words from it.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{j,k}}$$

$n_{i,j}$: *number of occurrences of the considered term* (t_i) *in the mail*(d_j)

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$|D|$: *total number of mails in the corpus.*

$|\{d : t_i \in d\}|$: *Number of mails where the term* t_i *appears.*

The probability of the word present in the mail being associated with the Spam is calculated. The formulae for calculating the Probability is given by:

$$p_i = \frac{n_{si}}{|n_{si} + n_{ni}|}$$

n_{si}: *number of mails in which a term* t_i *appears in Spam Database.*

n_{ni}: *number of mails in which a term* t_i *appears in Normal mail Database.*

$$AnomalyFactor = \frac{\sum_{i=1}^{n} pi}{n}$$

The Anomaly Factor is calculated based on the above mentioned formula. The decision of the model is based on the anomaly factor value; we can preset the threshold for this measure using experimental data. in our evaluation of our approach, we use this a value of 0.6.

IV. Experimental Results

In order to create a Corpus of mails both Spam and normal, we collected mails received by five users(from gmail domain) from December 2009 - July 2010. This corpus was used for filter learning it consisted of more than 5000 Spam mails and 10000 normal emails. It was seen that around 35% of the spam received were sent by forging the receiver's address “Self Spamming". Around 15% of the Spam mails were sent to more than one recipient and were detected by the Degree of Randomness Model. Around 95% of the spam emails had suspicious links in the body of the email which were detected by the content
reliability model. Out of the links present in spam emails 3% of the links had no parameter values from the Anti-phishing toolbar and we had to query DNS age to get the reputation of the domain. Based on these results we implemented the Anomaly Based, Reputation Based and Content Reliability Models. After implementing these models we evaluated the spam filter by checking each received mails and taking decisions based on the models. The results obtained by using our spam filter to filter out received mails for the duration of approximately 3 months from Jul 2010 - Oct 2010 is given in figure 9. The experimental results shown below are obtained by setting the threshold value of 0.6 for The anomaly factor obtained from the Reputation Based Model. The x axis reprints the period in weeks (1 time unit = 2 weeks) and y axis

represents the Number of mails received during that particular period.

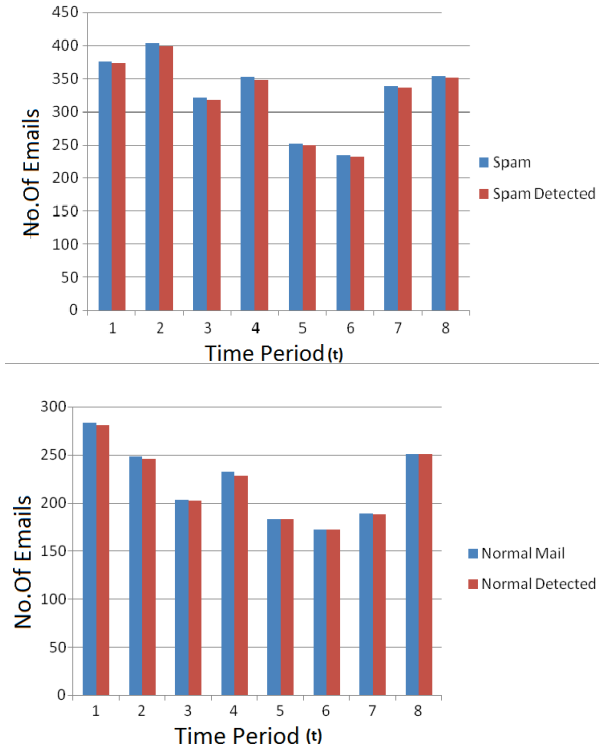

Figure 9. a) Total Spam received vs. Spam detected by Hybrid Filter. b) Total normal emails vs. Normal emails detected by Hybrid filter.

It is clear from the results above that the false Negative rate of Hybrid Spam filter is ~1.34% where as the False positive rate is ~.54%. The high False Negative rate is due to the fact that the parameter values we chose to distinguish Spam from Normal are quiet high. For example the decision made by Content Reliability Model to get the above graph is given below:

If ((Trust <80)||(Privacy < 80)) Then Spam = 1

Another factor contributing to the false negative rate is that few Spam's use Attachments and Images as a tool for sending viruses. We have not addressed in this paper the models for analyzing the attachments and images included into the body of the mail.

V. Conclusions and Future Work

In this paper we have designed hybrid behavior analysis models based on Text/Html in the body of the mail and email headers. We plan to develop models based on Attachments and Images that could be part of the body of emails in order to improve the detection rates of our models. The spam filter performance is tested at the user level and it can be extended to the server level, where more data can be analyzes and will significantly improve our methodology.

References

[1] Mikko Siponena,Carl Stuckeb," Effective Anti-spam Strategies in Companies: An International Study", Proceedings of the 39th Hawaii International Conference on System Sciences - 2006.
[2] Ramachandran, Santosh Vempala and N. Feamster, "Filtering Spam with Behavioral Blacklisting", CCS '07 Proceedings of the 14th ACM conference on Computer and communications security.
[3] Fulu Li, Mo-han Hsieh, "An empirical study of clustering behavior of spammers and Group based Anti-spam strategies", CEAS 2006, pp 21-28, 2006.
[4] Habib, A.; Roy, D.; "Steps to defend against DoS attacks", Computers and Information Technology, 2009. ICCIT '09. pp. 614 - 619.
[5] Available online: http://www.highbeam.com/doc/1G1-132679051.html
[6] Ferris Research.(2010).Industry Statistics [Online].Available:http://www.ferris.com/research-library/industry-statistics/
[7] Ramachandran and N. Feamster, Understanding the Network-Level Behavior of Spammers. In Proc. ACM SIGCOMM, Pisa, Italy, Aug. 2006.
[8] Eric Allman "Features: Spam, Spam, Spam, Spam, Spam, the FTC, and Spam" Queue- Vol. 1 Issue 6, pages 62 - 69, September 2003.
[9] Wong, Tak-Lam, Chow, Kai-On, Wong, Franz, "Incorporating Keyword-Based Filtering to Document Classification for Email Spamming", Vol.7,19-22 Aug. 2007, pp.3899 -3904, Machine Learning and Cybernetics
[10] Available online: http://bogofilter.sourceforge.net
[11] P.Graham, Better Bayesian Filtering, http://www.paulgraham.com/better.html
[12] Available online :http://en.wikipedia.org/wiki/Bayesian_poisoning
[13] Available online: http://blog.jgc.org/2006/04/bayesian-poisoning-paper-pointers.html
[14] Paul Graham (2003, Aug 1).Stopping Spam[Online].Available: http://www.paulgraham.com/stopspam.html
[15] Available online: http://math.uc.edu/~siva/mdbayes/chap2p.pdf
[16] Zhan, J.; Oommen, B.J.; Crisostomo, J., "Anomaly Detection in Dynamic Social Systems Using Weak Estimators" in Computational Science and Engineering, 2009, CSE '09. pp. 18 - 25.
[17] Robert L.Vaessen. (2009, Jan, 2). Forgery [Online].Available:http://www.robsworld.org/forgery.html
[18] M, Overton, Bots and Boetnets: Reisks, Issues, and Prevention, Virus Bulletin Conference, Dublin, Ireland, October 2005.
[19] B, Schneier, How Bot Those NEts? Wired Magazine, July 27, 2006.
[20] R, Narasine, Money Bots: Hackers Cash In on Hijacked PCs, eWeek, Sept. 2006.
[21] Govil,J.;, "Examining criminology of Bot zoo" in Information, Communications & Signal Processing, 2007 ICICS '07. pp. 1 – 6
[22] Commtouch.(2010,April). Internet Threat Trend Report[Online].Available:http://www.commtouch.com/sites/default/files/newsletter/April-10.html
[23] Mywot. My web of trust API [Online]. Available: http://www.mywot.com/wiki/API
[24] Sanglerdsinlapachai, N. ; Rungsawang, A. ; Knowledge Discovery and Data Mining, 2010. WKDD '10. pp. 187 - 190
[25] Available online: http://en.wikipedia.org/wiki/Tf%E2%80%93idf