Design & Implementation of a Comment Spam Filtering System on a Web Blog

Sau S.P.[1]., Roy S.[2] ,Mandal K.[3]
[1&2]Computer Science & Engineering & [3]Mechanical Engineering Department
NITTTR
Kolkata, India
E-mail: onlinesankar@gmail.com

Kunar S.[4]
[4]Production Engineering Department
Jadavpur University
Kolkata, India
E-mail: Sandip.sandip.kunar@gmail.com
(Corresponding author)

Abstract—**Spam can be defined as unwanted messages based on the polices of webmasters. Spam is no longer limited to email and WebPages. In the recent years one type of spam is most common, that is spam in blog sphere, is called comment spam. The increasing penetration of spam in blogs and social networks has given frustration to the webmasters. This work takes challenges posed by this type of spam in the blog sphere. The characteristic of comment spam investigates in the blog sphere based on their content analysis. To identify comment spam using an algorithm with taking help from some machine learning techniques such as Support Vector Machine and Naïve Bayes Classifier etc.**

Keywords-spam, content similarity, number of links, text similarity, comments.

I. INTRODUCTION

Spam is a problem that in recent years has caused a tremendous impact on the internet. It has grown to a significant level, and polluting the cyberspace in the different ways like emails, WebPages, blog posts, blog comments, instant messaging and social network comments. Spammers mainly target social network like Face book, MySpace etc. This type of spam mainly introduced to unethically advertise products, distribute different malwares, spread viruses and stealing personal information. Another's contributing factor to increase page rank that is used in every search engine optimization like Google and yahoo etc. One important criteria used to measure the relevance of page by search engine is link weighting. Link weighting is based on the concept that if a page is referred by many other pages, the relevance of this target page is increases. An important aspect of link weighting is that the rank of page will be higher if gets a reference from a high rank page. For this reasons spammers try to increase link weighting of his WebPages. Blogs are gaining constant popularity in recent years and become most important part of web [1]. A blog is a type of periodic articles with user comments and opinions. User can link different article in particular blog as well as in different blog. For this reason spammer gets advantage and places their web pages link in blog to increase their web page rank [2]. Various initiatives have been taken to reduce these kinds of comment spam. For example, may blogs now require users to register before they post any comment. However this restriction does not seem to hamper spammers completely. Major search engines like Google, yahoo, msn etc, came up with an alternative solution to add an attribute "rel = nofollow" to the hyperlinks that are automatically generated in the pages. But this solution does not stop spammers from posting link in comment as a legitimate comment. Another important characteristic of blog comments that make it fundamentally different from emails is that blog comments have a strong cohesion with the post and has a gradual build up on a topic while emails are independent writings in themselves. This work concentrates specifically in comment spam identification and comment spam filtering for blog comments using content analysis. Three main features including links, content repetitiveness and text similarity are based for comment spam identification. Content repetitiveness is determined by the length and frequency of the longest common substring, text similarity is calculated using vector space model.

II. CHARACTERISTIC OF SPAM

An electronic message needs to fulfill certain requirements to be classified as spam. The four most importance ones are:

(1) Spam is unsolicited
(2) Spam is send in large quantities
(3) Spam is not individualized
(4) Spam is cost-intensive for recipient

The two main characteristics, one and two are necessary for any type of message to be considered spam. The first main characteristic is that it is unsolicited. All messages considered spam are unsolicited; nevertheless not all unsolicited messages are considered spam, the third characteristic is a direct result of the second main characteristic. Since it is possible to send messages in large quantities at virtually no cost per message it has become unnecessary to individualize each message for the end user. This is true for most spam. The fourth characteristic

 ISBN: 978-81-923777-9-7

is that the costs of spam are born by the person receiving the spam message completely opposite to all other forms of communication where the sender pays a fee for the delivery of a message (post, telephone, and fax). In the case of spam the sender also seems to get a disproportional gain [3].

III. WEB SPAMMING

Apart from E-Mail spam, Web log (BLOG) spamming is probably the most well known example for unsolicited bulk advertisements lately. A Web log is a publicly accessible Web page that lists news, Web links or just comments on recent or not so recent events. There are public and private BLOGs. In public BLOGs it is possible for each registered user to post articles, comments.

In the fourth quarter of that year spam started to appear more frequently in Comment sections of BLOGs and news sites. Bloggers had to adapt to better spam Protection techniques than just blocking IP addresses. Besides drawing the blog readers' attention to the spammers' Web page BLOG spam also improves the search engine rating of the advertised Web site, which is the main reason why it becomes so popular [4].

IV. IDENTIFY THE FEATURES OF THE SPAM

This section analyze different features of a spam legitimate comment, all of them based on the content of the comments and the post, which may prove helpful for the detection of comment spams. Here spam comment defines as any unsolicited comment sent in response to a blog post which is not a spam, and a legitimate comment is a non-spam comment which is often also termed as harmful comment . Most of the time these spam comments are generated automatically with a computer program. Legitimate comments often have a specific pattern whereas Spam comments exhibit slightly random characteristic. A system is developed to analyze the comments based on the features mentioned below.

(A) Dataset
(B) Post Comment Similarity
(C) Word Duplication
(D) Noun Concentration
(E) Stop Words Ratio
(F) Spam Similarity

(A) Dataset

To analyze the characteristics and evaluate the detection algorithm based on features, the corpus is created by Mishne and Carmel. The corpus contains around seven random blog posts with fourteen comments posted to them. All the posts contain a mixture of spam and legitimate comments. Comments in the corpus were experimentally classified, leading to eleven legitimate comments and remaining being spam.

(B) Post comment similarity

The goal is to try to analyze the coherence of the comments compared to the post for which the comments are written. Legitimate comments are expected to have more coherent phrases compared to spam.

Hi, I just wanted to say thank you guys! I really like
Your site and I hope you'll continue to improving it.

Example of a comment spam which is not related to the subject matter.

$$\text{Similarity } (P_j, C_k) = \frac{P_j . C_k}{|P_j| . |C_k|} = \frac{\sum_{i=1}^{n} W_{i,j} . W_{i,k}}{\sqrt{\sum_{i=1}^{n} W_{i,j}^2} . \sqrt{\sum_{i=1}^{n} W_{i,k}^2}}$$

Here, $W_{i,j}$ is the frequency of the word occurring in the blog post.

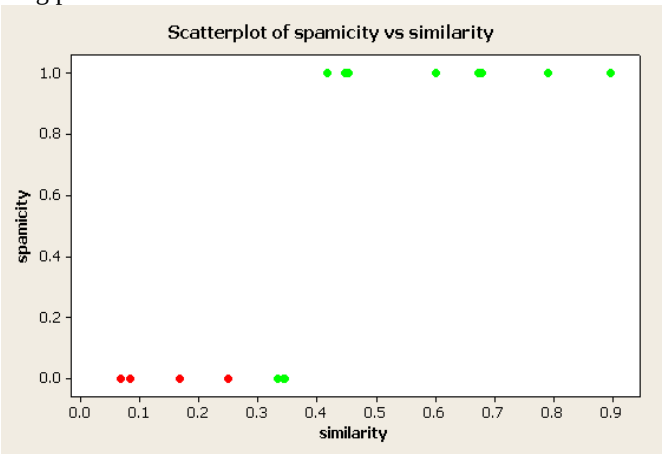

Figure 1. Distribution of spam and legitimate comments based on post-comment.

Here, the front horizontal axis represents the similarity of a comment to a post with scale 0 to 1. Spam city value of 0 represents a spam comment whereas 1 represents a legitimate comment. The distribution indicates that most of the comment spam's have similarity of less than 0.0833 to the posts, excluding some outliers whereas legitimate comments have higher similarity values.

(C) Word duplication

To analyze the behavior of blog comments based on their word repetition pattern. Word redundancy in context is defined mathematically as follows:

Word Redundancy Ratio=

$$1 - \frac{\textit{Number of unique words in the comment}}{\textit{Total number of words in the comment}}$$

Example: I am Sankar Prasad Sau
I am Animesh Patra

We Calculate, There are 5 unique words and total words are 9
Word Redundancy Ratio will be 1- 5/9=1- 0.5556=0.4444.

ISBN: 978-81-923777-9-7

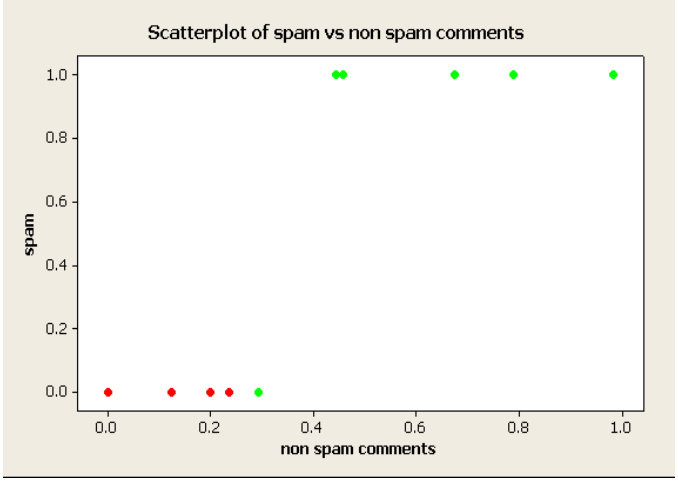

Figure 2. Distribution of spam and non spam comments based on Words Redundancy Ratio.

The distribution indicates that legitimate comments have fairly low word-redundancy compared to spam comments which have a redundancy ratio as high as 0.9.

(D) Noun concentration

Most auto-generated spam comments are populated either by some keywords in the form of noun-phrase chunks without the formation of a complete sentence or some links to keep the crawler going.

Noun-Concentration in our work was calculated using the following formulae:

$$\text{Noun Concentration} = \frac{\text{no of noun phases present in the commen}}{\text{total no.of words in the comments}}$$

Example:

Thanks to you for the wishes to all the student of M.Tech in this (institute (NITTTR KOLKATA)).

We Calculate, There are 6 noun phases in the comment & total words 17.

Noun Concentration = 6/17 = 0.3529

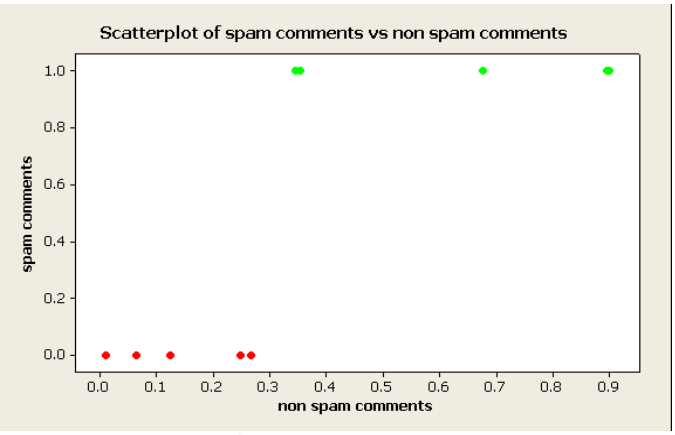

Figure 3. Distribution of spam and non-spam comments based on noun concentration

A group of spam comments is filtered which have high noun-concentration. The average noun-concentration for legitimate comments was found to be 0.3245 whereas for spam comments it was around 0.2511.

(E) Stop words ratio

Using a similar explanation as noun-phrase concentration, to hypothesize that legitimate sentences tend to have a fairly balanced stop words ratio compared to spam comments.

$$\text{Stop words Ratio} = \frac{\text{Number of stop words present in the comment}}{\text{Total number of words in the comment}}$$

Example:

Football, betting, football, betting, line online football betting college, football betting football, betting odds, football betting football betting, NCAA, football.

So, there are 10 stop words in the comment & total words 20.

$$\text{Stop words Ratio} = \frac{10}{20} = 0.5000$$

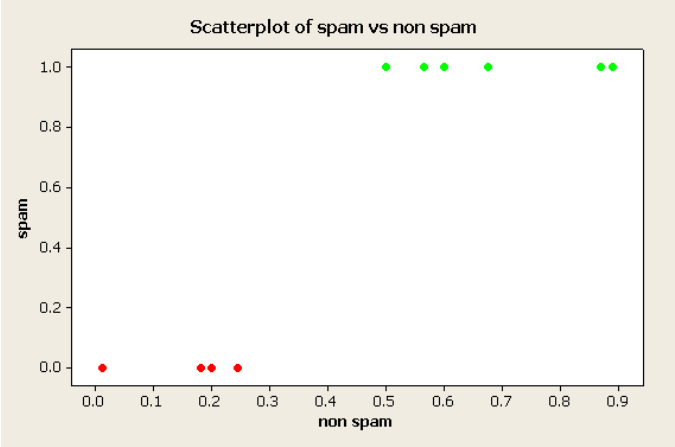

Figure 4. Distribution of spam and non-spam comments based on stop words ratio.

The distribution of spam/legitimate comments based on stop word ratio in figure 4 shows that legitimate comments almost always have a stop words ratio in the range 0.3 to 0.9 but spam comments have a wide variation in the stop words ratio. Clearly, as indicated by the graph, comments with less Stop words ratio are more likely to be spam than with high stop words ratio.

(F) Spam similarity

Spam similarity of any comment is expressed as

$$\text{Spam Similarity } (c_k) = \frac{\sum_{i=1}^{n} w_{k,i}}{i}$$

Here the distribution of the comments summarizes their spam-similarity.

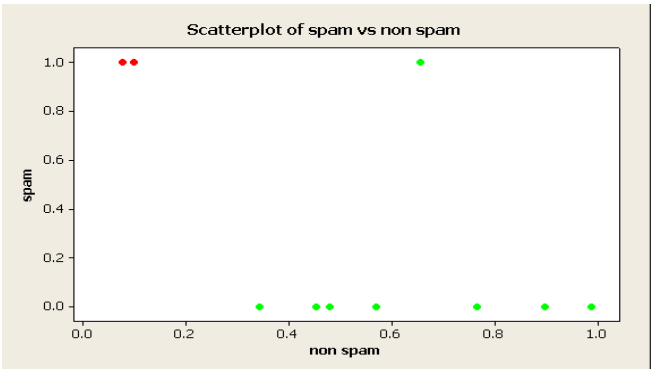

Figure 5. Distribution of spam and non-spam comments based on spam-similarity of the comment

the spam similarity of a legitimate page is almost 1 in most cases whereas spam comments have their similarity as less as 0.25, thus indicating that this feature could be a potential deterministic feature for the detection of spam

ISBN: 978-81-923777-9-7

comments[5].

V. EXPLAINATION OF THE STATEMENT

There are three main features for spam identification as well as preventing comment spam.

Text similarity: The content similarity between the comment spam and blog post is low. Valid comments can be seen as conversations related to the topic of the blog post. While comment spam is nothing but artificial trick for higher ranking scores, so they are not similar in content to the posts. We use support vector machine techniques to find out similarity between blog post and comment message.

TFIDF (Term frequency Inverse Document Frequency) algorithm is applied for text similarity calculation. Each document *d* is represented as a vector $(d^{(1)} \ldots d^{(|F|)})$ and documents with similar content have similar vectors. Each element $d^{(i)}$ represents a distinct word w_i. $d^{(i)}$ for a document *d* is calculated as combination of the statistics **TF (w_i , d)** and **DF(w_i)**. The term frequency **TF (w_i , d)** is the number of times word w_i occurs in document *d* and the document frequency **DF(w_i)** is the number of documents in which word w_i occurs at least once. The Inverse Document Frequency **IDF (w_i)** can be calculated from the document frequency.

$$\text{IDF}(w_i) = \text{LOG}(|D|/\text{IDF}(w_i)) \quad (1)$$

Where $|D|$ is the total number of documents, and the so-called weight $d^{(i)}$ of word w_i in document *d* is calculated by the bellow Equation.

$$d^{(i)} = \text{TF}(w_i, d)\text{IDF}(w_i) \quad (2)$$

The cosine similarity of the comment *C* with corresponding post *P* is calculated by the bellow Equation.

$$\text{Sim}(P_j, C_k) = \cos\theta = P_{j*} C_k / |P_j || C_k| \quad (3)$$

$$\text{Cos}\theta = (\text{Summation of } (w_{i,j} * w_{i,k})) / (\text{Square root of } (\text{sum of } w_{i,j}^2 * \text{sum of } w_{i,k}^2)) \quad (4)$$

Where $w_{i,j}$ is frequency of word occurring in blog post and w_{ik} **is** the frequency of word occurring in comment message. Here frequency means no of occurrences of word in the document content which is called as weight [6].

Links: There might be some out links pointing to irrespective pages in comment spam. The motivation of spammers is to artificially increase page rank by using link dumps that contribute to a link farm. For this purpose, spammers always put comments which include links pointing to their own pages on blogs [7]. Our motivation is to parse that WebPages content from the link and take content of that website and matched with blog post content to find that link is good or bad.

Content repetitiveness: Comment spam may partially or totally occur repeatedly. Spammers always copy the same comment spam to different pages to increase page rank in short time. Actually spammer used auto generated tool that is created comment message. So, probability of comment spam message may be similar. So, new comment messages can be found as spam if it is similar to the prior comment message posted by same user.

So, a comment message with high content repetitiveness and low similarity with the post is more likely to be a spam [8].

One is special character set that should be predefined. Special character set contain set (, , . , ?, “,”, etc). Second is stop word set which is also predefined? Stop word set contains like (am, is, are, I, we, etc). And create an empty set keyword set for each comment messages as well as blog post. Keyword set contains words after removal of stop words and ignore characters.

VI. ARCHITECTURE OF THE PROPOSED SYSTEM

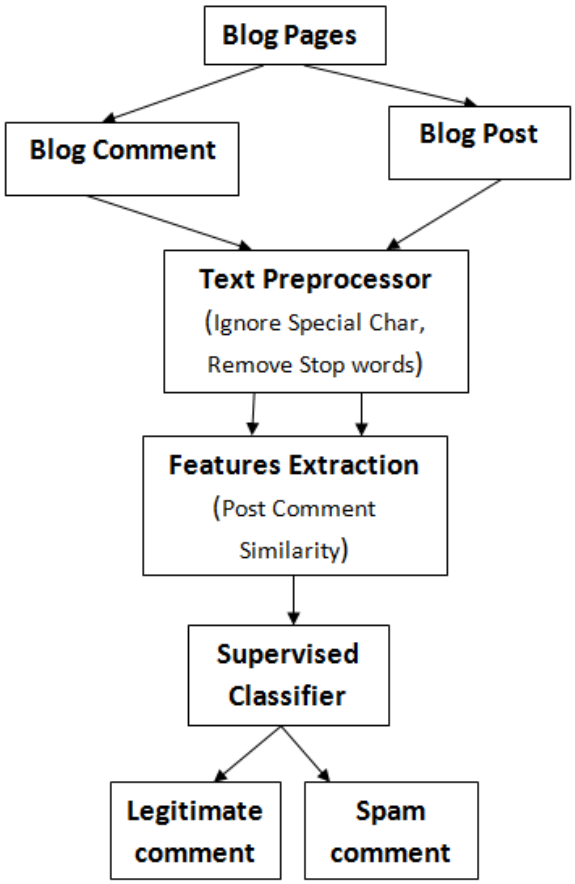

Figure 6. Architecture of the Proposed System

A. Explanation of the architecture of the proposed system

To write something on the blog page and post on the blog comment. Then according to blog pages, post some comment on the blog post.

- *The function of text preprocessor*

It is to find keyword from document content. For a given blog post, first we considered post as a content document. Then ignore special character list from the document’s content using string matching with the special character set. After that, remove all stop word from blog post document with string matching with the stop word set. Algorithm for text preprocessing is given below.

- *The function of features extraction*

It is to find the similarity between blog post keywords and comment message keywords. Vector space model is used to

ISBN: 978-81-923777-9-7

find out similarity.
From the algorithm, the set of keyword is found from blog post and comment message then create a matrix whose number of rows will be number of keyword in the document. And column will be the particular document. The weight or frequency of a keyword will place in the corresponding cell according to the document and particular keyword. The procedure should be same for both post and comment message. Then find the similarity between two documents using the above describe similarity function using weight of keywords. Two documents will be similar if result tends to one [9].

- *Algorithm of the proposed system*

Algorithm (Keyword extraction, Calculate weight, Create Count Matrix):-
Input: blog content, comment message content, stop word set, special word set
Step 1:- Find all special characters in the document and Remove specials character from the document.
Step 2:- Convert all words to lowercase.
Step 3:- For each word in document
If each word is not stop word set
Add token to keyword set.
End If.
End For.
Step 4:- For each keyword in keyword set
For each keyword in Count Matrix first column
If keyword not found
Add a row for keyword and its weight set to 1.
Else
Weight of the keyword increased by
End For.

VII. IMPLEMENTATION AND DESIGN

Proper implementation makes an application efficient. In approach there are some specific spaces where efficient implementation makes the application more accurate and give more desirable output. Some implementation snaps shorts are given below.

Figure 7. New Article Post:

Figure 8. Showing content matching similarity percentage

- *Experimental result and analysis*

Collect seven random English blog posts, along with the each two comments posted to them; all pages contain a mix of spam and non-spam comments. It is experimentally classified the comments: 11 (78.57%) were found to be "legitimate" comments (non-spam); the other 3 comments (21.42%) were

ISBN: 978-81-923777-9-7

spam comments. The Experimental Calculations are given bellow.

Example:

Blog Content: - I want to wish every student of this institute, a well wish for the examination.

Post Content**:** -Thanks to you for the wishes to all the student of M.Tech in this institute.

No of total keyword in the Blog Content = 08

No of keyword in the Post Content = 05

Content Matching Similarity =

$$: \frac{\textit{No of keyword in the Post Content}}{\textit{No of total keyword in the Blog Content}}$$

=5/8=0.6250

TABLE I. FOR BLOG AND POST COMMENT SIMILARITY CALCULATION

Blog	Post	Blog Keyword	Post Keyword	Similarity	Percentage
1	1	08	05	0.625	62.50
1	2	08	06	0.750	75.00
2	1	22	07	0.318	31.81
2	2	22	08	0.363	36.36
3	1	14	02	0.142	14.28
3	2	14	05	0.357	35.71
4	1	15	07	0.466	46.67
4	2	15	08	0.533	53.33
5	1	28	20	0.714	71.43
5	2	28	02	0.071	07.14
6	1	19	14	0.736	73.68
6	2	19	09	0.473	47.36
7	1	29	07	0.241	24.14
7	2	29	16	0.551	55.17

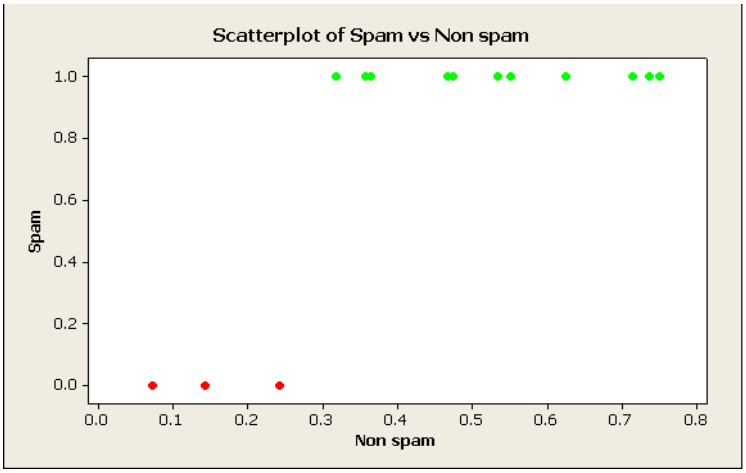

Figure 9. Scatter plot of Spam vs. non Spam

VIII. CONCLUSION

This work has introduced a architecture for content based spam detection without requiring any extra information, in future to include time factor for posting comment with respect to Blog post. Apart from the post-comment similarity, other methods can be equally applied to any form of short message spams such as instant messaging spams, social network spams, etc. Although system does not have a perfect detection mechanism, the undetected ones are more in the grey area even to a human analyst. The dataset used for analysis was not a large dataset. Thus, the future work would be to evaluate our methods on a larger dataset. The evaluation of the analyzed features on other type of short spams such as instant messaging spams, short messaging service spams and other comment spams are also a possible extension.

REFERENCES

[1] F. Zhou, L. Zhuang, B. Y. Zhao, L. Huang, A. D. Joseph, and J. Kubia-towicz. Approximate object location and spam filtering on peer-to-peer systems. In *Proceedings of Middleware*, 2003.

[2] W. Bachnik, S. Szymczyk, P. Leszczynski, R. Podsiadlo, E. Rymszewicz,L. Kurylo, D. Makowiec, and B. Bykowska. Quantitive and sociological analysis of blog networks, in preprint, 2005.

[3] T.M. Mitchell, Machine Learning, McGraw-Hill, New York,1997.

[4] J. Postel. On the junk mail problem. RFC706, 1975

[5] F. Assis, A text classification module for Lua – the importance of the Training method. In Fifteenth Text Retrieval conference, Gaithersburg, MD, 2006.

[6] A. Bratko, G. V. Cormack, B. Filipi˘c, T. R. Lynam, and B. Zupan Spam filtering using statistical data compression models. Journal of Machine Learning Research , 2006.

[7] Hearst, M.A., Hurst, M. and Dumais, S. T., What Should Blog Search Look Like, SSM, Napa Valley, California, USA, 2008.

[8] S.Webb, J. Caverlee, and C. Pu, Characterizing Web Spam Using Content and http session Analysis, CEAS, Fourth Conference on Email and Anti-Spam, 2007.

[9] G.V.Cormack, J. M. Gomez, and E. P. Sanz, Spam Filtering for short messages, 2007.

ISBN: 978-81-923777-9-7