THE IMPLEMENTATION OF ROCCHIO'S CLASSIFICATION TO CATEGORIZE SPAM COMMENTS IN BLOG

Andre Prilly Kurniawan[1], Ednawati Rainarli[2]

Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer. Universitas Komputer Indonesia
Jl. Dipati Ukur No. 112-116 Bandung
E-mail : dreprilly@gmail.com[1], irene_edna@yahoo.com[2]

ABSTRACT

Spamming refers to the unwanted and irrelevant information for the users. This phenomenon is widely spread and often seen on emails, short messages, blogs, and forums. This study aimed to examine the spam problem in blog. The comment system in blog, which was provided by the owner to facilitate the interaction with the readers, was being targeted by the spammers. The present actions by the blog owners such as monitoring and managing comments manually and using CAPTCHA could not prevent and solve this problem. Therefore, this study offered the Rocchio Classification method to minimize the occurrence of spam comment attacks. The features used in this study were the using of anchor text, referring usernames and calculating the words ratio in comments, and measuring the similarity level and time difference between blog posts and comments. By testing 400 data sources, the results showed that Rocchio Classification was able to classify spam or organic comments with an average accuracy of 95% of various test scenarios.
Keywords: Spam, Blog, Spam Comments, Rocchio Classification, Classification

1. INTRODUCTION

Over the last decade, the blog became very popular on the internet. According to WordPress, one of the blog publishing service providers stated that its users make an average of 69.8 million new postings and 42 million new comments every month [1]. Unfortunately, with the amount of traffic there is a large gap management is not good so the blog could become a target for spammers. Right now, in fact the blog owners are already using some of the techniques to reduce comment spam. Some blog owners choose the do monitoring and managing comments manually. Another technique that is used to distinguish the blog owner comments are done automatically by a bot with the original comment made by a user are using CAPTCHA [2]. CAPTCHA typically-shaped image that contains letters and numbers which are difficult to be recognized automatically by the bot. However, research has proved that this method is very easy to botched [3].

In the year 2005 Mishne et al. [4] uses the language modeling approach to detect spam comments with the Kullback-Leibler divergence method get 83% accuracy rate. In the year 2011 Bhattarai et al. [5] using content analysis to identify spam with words duplications, stopwords ratio etc., with best results using the method of Support Vector Machine (SVM). From the approach to the highest degree of accuracy obtained by 86%. In the year 2012 Ashwin et al. [6] using the analysis of the relationship between the posting comments and blogs with comments by using several methods of classification of the highest degree of accuracy is obtained using the method of decision tree with an accuracy of 92%.

In other studies in the year 2013 Pausta dkk. [7] compares the SVM with Rocchio library catalogue to search result Rocchio have smaller processing time 57.2% and 37.8% precision level greater than the SVM. Rocchio Classification taken from the concept of Relevance Feedback Rocchio has design concepts for classifying only two classes that is relevant and irrelevant [8]. Based on the concept on this research will be very suitable due on this research will be classified into the category of comment spam or not spam. Therefore, this research will be carried out on the implementation of the Rocchio Classification in identifying spam comments in hopes of getting a better degree of precision.

Based on the explanation that has been described above, then the expected method used is the solution to minimize the spamming actions happen on comments on the blog.

The expected goal will be achieved in this research are:

1. Categorize comments spam with method Rocchio Classification.
2. Test the level of accuracy in Classification categorizes Rocchio spam comments.

2. THEORETICAL BASIS

2.1. Spam in Blog

Spam in blogs (also called simply blog spam, comment spam, or spam social) is a form of utilizing spamdexing. This is done by post (usually automatically) random comments, copying material from another place that is not genuine, or promoting commercial services to blogs, wikis, guestbooks, or other online discussion forums are publicly accessible. Every web application

accepts and displays hyperlinks submitted by visitors may be a target. Adding links that lead to web site artificially boosts/spammer sites on search engines where the popularity of URL contribute to the value of tersiratnya, an example is the PageRank algorithm as used by Google Search. Doing so will enhance the commercial sites listed spammers ahead of other sites for certain searches, increasing the number of potential visitors and customers who pay [9].

2.2. Classification

Classification is an employment rate data objects to integrate it into a particular class from a number of classes that are available. In the classification there are two main work is done. First, the construction of the model as a prototype to be stored as memory. Second, the use of models to do the recognizing/classification/prediction on an object to another, so that known in the classroom where the data objects in the model that is already about them [10].

2.3. Rocchio Classification Algorithm

Rocchio classifiers is one method of learning supervised document classification. Classification method of comparing the similarity of content between rocchio data training and test data with merepresentsaikan all the data into a vector. In using vector space model required the boundaries between classes to find out the appropriate classification. Rocchio technique applying those limits in the form of centroid to give such restrictions. Centroid of a c grade is the average of all the vector class c. to calculate centroid value can be seen in equation (1).

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d) \quad (1)$$

With:

$\vec{\mu}(c)$ = centroid class c

= total of document class c

= vector of document that has been normalized

To determine the similarity of two vector space model by measuring the distance. In determining the distance between two vector space model of euclidean distance is used which can be seen in equation (2).

$$d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (2)$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

With:

p and q = a vector that has been normalized

3. RESEARCH METHOD

In this study the research method used is experimental research methods. Experimental method is a method that has the purpose to explain the causal relationship between one variable and another [14]. The flow of research that will be carried out in this study can be seen in Figure 1 as follows:

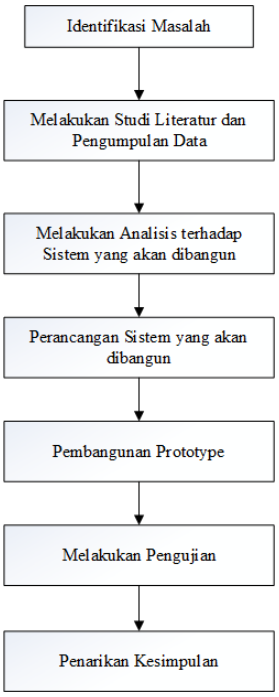

Figure 1. Research Flow

3.1. Method of Collecting Data

Data collection methods using data in research conducted by Mishne et al. on the other hand used in this study are as follows:

Literature study is carried out by studying, researching and analyzing various literatures from libraries sourced from books, scientific journals, internet sites, and readings that are related to the research topic.

3.2. Software Development Method

The software development method used in this study is the Prototype model. The following stages are carried out in this study:

1. Analysis

Problem analysis is done to understand the problems that arise and find solutions to solve problems in generating spam commentary classifications..

2. Data Requirements

At this stage, researchers will collect data data input system for comments..

3. Prototype Development

At this stage it will be implemented from the analysis process and system requirements that have been obtained and researchers try to implement the Rocchio Classification method into the program logic.

4. Prototype Evaluation

The program will be tested where trials are conducted to find out the shortcomings in the program. If there are still deficiencies, then the prototype is revised with the steps previously carried out.

The prototype stages carried out in this study will be explained in Figure 2.

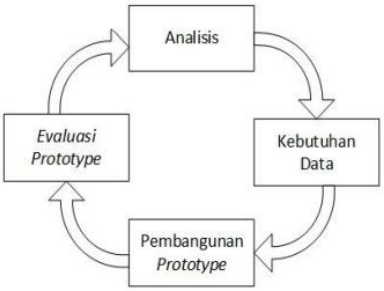

Figure 2. Prototype Model

4. RESULT AND DISCUSSION

4.1. System Analysis

Classification of comment spam using the Rocchio Classification problems which will be solved and discussed on this research, the implementation of classification used in Classification Rocchio spam comments, a system of classification comment spam based on the features that will be a reference in the process of classification. System designed can be run on the device of a computer (PC) with the Javascript programming language, which can be used to view the results of the classification of spam comments.

In building a spam commentary classification system several stages of analysis are carried out. The stages of the system created are shown in Figure 3. The stages carried out by the system are divided into two stages. The first stage is training and the second stage is testing.

In the first stage begins by extracting the training data to get features such as the number of anchor text, the difference in the date of the blog post with the comments and the appearance of the user name in the comment column. Then proceed to the preprocessing stage. The results of preprocessing are extracted again to get other features such as word repetition and post-commeny simmilarity ratios.

In the system process flow described above, there are two stages in the collection of features used, this is due to features such as the number of anchor text, the difference in the date of the blog post with the comments and the appearance of the user name in the comments column can only be obtained before the preprocessing stage. Then after the five features in the form of vectors are obtained, the training phase is carried out using the Rocchio Classification method. In the second stage, testing, the process stage is not much different from the training stage, which distinguishes the testing stage and the classification process is carried out to determine whether a comment is considered spam or not spam.

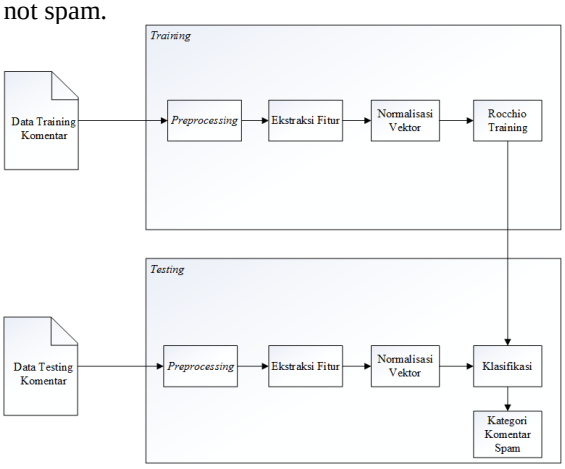

Figure 3. System Architecture

4.2. Input Data Analysis

Input data analysis needed in this study is to have blog post variables, blog post dates, author comments, comments, comment dates and comment class categories formatted in json form taken from Mishne and Carmel research totaling 400 data from an estimated 20 blog posts .

4.3. Feature Extraction

Spam comments usually have certain characteristics which will be used as a feature to distinguish non-spam comments. This research will use five features: number of anchor text, word repetition ratio, post-comment similarity, appearance of user name in the comments column and difference in the date of posting with comments.

4.3.1. Number of Anchor Text

The text that appears in HTML between the <a ...> and tags is referred to as anchor text. These texts basically create links to other pages that can be connected by clicking on the hyperlink. Web crawlers usually follow this link iteratively to browse web pages on the internet. Spam comments try to include a lot of link text that leads to spammer sites to increase their page rank on search engines. The following are examples of spam comments with some anchor text shown in Figure 4 below.

Spy Software http://spy-software.seo.se.com
Internet Spy Software http://internet-spy-software.seo.se.com

Figure 4. Spam Comment Example which contains many URLs

4.3.2. Time Difference Data Post with Comments

An article is usually written because many people talk about this topic, such as the topic of elections when entering the election period, or the topic of the soccer world cup when the world cup was rolling. Non-spam comments are usually done when the topic is still in the warmest area, while spam comments are not time dependent. The following is the formula for calculating the difference in the date of posting time with comments shown in Figure 5 below.

Selisih Waktu = Tanggal Posting Blog – Tanggal Komentar

Figure 5. Time Difference Formula Date Post with Comments

4.3.3. Appearance of Author in the Comment

The comment system always provides a column of names which is used to enter the name of the commentator. Non-spam comments generally won't enter their names in their comments while spammers use keywords as their name and include them in comments, it aims to increase keywords in search engines. The following is an example of the appearance of a user name in the comment column shown in the Figure 6 below.

Get a Generic Viagra alternative at Cheap Generic ViagraGet a Cheap Generic Viagra alternative at Cheap Generic Viagra

Comments posted by: Cheap Generic Viagra at March 11, 2005 09:27 PM

Figure 6. Examples of Spam Comments Containing User Names in their Comments

4.3.4. Word Duplication Ratio

Spam comments use repetition of words to attract search engines while organic comments flow more frequently in the context of related articles. Because most blog comments are short, the same word is rarely repeated in organic comments. The following formula for calculating the word repetition ratio is shown in figure 7 below.

$$\text{Ratio Pengulangan Kata} = 1 - \frac{\textit{Jumlah kata unik di komentar}}{\textit{Jumlah total kata di komentar}}$$

Figure 7. Word Duplication Ratio Formula

4.3.5. Post-Comment Similarity

Spammers use computer-generated scripts to generate millions of spam comments that are ready to be sent. However, in many cases, this automatic spam comment is not related to the context of blog articles. The following examples of spam comments that are not related to the context of the blog article are shown in the figure 8 below.

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

Figure 8. Examples of Spam Comments that Are Not Related to the Context of Blog Articles

4.4. First Scenario Testing

The first scenario testing is done by testing the comments included in the training data, this test aims to determine the level of recognition of the comment data that has been trained. Comment data used amounted to 400 data consisting of 2 classes with 200 classes each.

Table 1. Confusion Matrix First Scenario Testing

Class		Prediction		Accuracy
		Spam	Non Spam	
Target	Spam	195	5	97,5%
	Non Spam	12	188	94%
Average				95,7%

4.5. Second Scenario Testing

The second scenario testing is done by testing different comments with training data, this test aims to determine the level of recognition of comment data outside the training data. Training data used amounted to 300 data consisting of 2 classes with each class there were 150 data and test data used amounted to 100 data consisting of 2 classes with each class there were 50 data.

Table 2. Confusion Matrix Second Scenario Testing

Class		Prediction		Accuracy
		Spam	Non Spam	
Target	Spam	47	3	94%
	Non Spam	50	0	100%
Average				97%

4.6. Third Scenario Testing

4.6.1. Testing with the 5-Fold

Testing with k-fold 5 is testing 5 rounds, meaning that the dataset is divided into 5 equals. In this study, the data used is 400 data and will be divided into 5 namely data A1 = 80, data A2 = 80, data A3 = 80, data A4 = 80 and data A5 = 80. In the first round, A1 data is used as test data while A2 to A5 are used as training data. In the second round, A2 data is used as test data while A1, A3, A4 and A5 data are used as training data. Likewise in the third round and so each data group will get a turn into test data and training data.

Testing	Spam	Non Spam	Correct Prediction	Accuracy
A1	58	22	78	97,5%
A2	43	37	79	98,7%

Testing	Spam	Non Spam	Correct Prediction	Accuracy
A3	62	18	75	93,7%
A4	37	43	73	91,2%
A5	0	80	78	97,5%
Average				95,72%

4.6.2. Testing with the 8-Fold

Testing with k-fold 8 is testing 8 rounds, meaning that the dataset is divided into 50 equals. In this study, the data used is 400 data and will be divided into 8 namely data A1 = 50, data A2 = 500, data A3 = 50, data A4 = 50, data A5 = 50, data A6 = 50, data A7 = 50 and data A8 = 50. In the first round, A1 data is used as test data while A2 to A8 are used as training data. In the second round, A2 data is used as test data while A1, A3, A4, A5, A6, A7 and A8 data are used as training data. Likewise in the third round and so each data group will get a turn into test data and training data.

Testing	Spam	Non Spam	Correct Prediction	Accuracy
A1	36	14	50	100%
A2	25	25	48	96%
A3	30	20	49	98%
A4	46	4	47	94%
A5	30	20	47	94%
A6	28	22	47	94%
A7	5	45	45	90%
A8	0	50	50	100%
Average				95,75%

4.7. Testing Conclusion

Based on the results of the first test scenario that is testing the training data similar to the test data, it can be concluded that the Rocco classification method can classify with an accuracy of 95.7%. Then based on the results of the second test scenario that is testing the test data is not included in the training data, the Rocchio classification method can classify with an accuracy of 97%.

Based on the results of the third testing scenario, that is testing using the k-fold cross validation method, the Rocchio classification method can classify with an average accuracy of 95.72% with a value of k is 5 and 95.75% with a value of k is 8.

From the test results, the non-spam comment category is more difficult to identify than the spam comment category. The accuracy of the Rocco classification method has a fairly good level in categorizing comments.

5. CONCLUSION

Based on the discussion of the testing phase, it can be concluded that the Rocchio Classification method has good accuracy in classifying comments from several test scenarios and in some cases spam comments that tend to be similar to organic comments, Rocchio Classification is quite difficult to predict correctly.

The suggestions that can be given for further development are handling non-standard words and adding new features to learn specific characteristics of spam and organic comments.

REFERENCES

[1] WordPress, “Stats – WordPress.com”, [Daring]. Tersedia pada: https://wordpress.com/activity/. [Diakses 27 Agustus 2016].

[2] Saini, B.S, Bala, A.: “Bot Protection using CAPTCHA: Gurmukhi Script”, Vol. 2, pp. 267, May 2013.

[3] Mori, G., Malik, J.: “Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA” In IEEE, 2003.

[4] Mishne et al.: “Blocking Blog Spam with Language Model Disagreement”, 2005.

[5] Bhattarai et al.: “A Self-supervised Approach to Comment Spam Detection based on Content Analysis”, 2011.

[6] Ashwin et al.: “Comment Spam Classification in Blogs through Comment Analysis and Comment Blog Post Relationships”, 2012.

[7] Yugianus dkk.: “Pengembangan Sistem Penelusuran Katalog Perpustakaan Dengan Metode Rocchio Relevance Feedback”, 2013.

[8] Manning et al.: “Introduction to Information Retrieval”, Chapter 14, 2009, hal. 292-295.

[9] Saleh, Rachmad. Spam dan Hijacking Email. Jakarta : Andi Publisher, 2008, hal. 06 – 46.

[10] Prasetyo, Eko. 2014. Data Mining. Yogyakarta: Andi Offset.