Filtering Spam Mail in Non-Segmented Languages Using Hybrid Approach: the Integration of Stopword Removal, N-gram Extraction and Classification Techniques

Ployphailin Khumsong, Todsanai Chumwatana, and Supanit Augsirikul

Rangsit University, Thailand, {ploypailin.k54; todsanai.c; supanit.a}@rsu.ac.th

ABSTRACT

Junk mail or spam mail has been regarded as a major problem in today's world. The spam mail can lead to cybercrime that impacts all individuals and organization. Many people and businesses seek for spam mail prevention technique in order to protect their own data and computer system. The spam mails normally contain advertise products or services contents and also conveys viruses, malwares, spywares and so forth. Many people thought spam mails do not cause any damage. In fact, the spam mails made a management cost increased and resources will be used ineffectively. Therefore, verifying and filtering spam mails need to be taken into consideration. The objective of this paper is to introduce the hybrid approach, which combines three techniques including stop-word removal, n-gram extraction and data classification, for filtering spam emails and simplifies system development. The proposed hybrid approach can be widely applied for all different languages due to being language independent technique. To examine the approach, CSDMC2010 spam mail corpus comprising of 198 common emails, 202 spam mails, and 10 selective emails were used in experimental study. The results showed that the proposed technique enabled to monitor whether the email is spam with 93.2% accuracy. Hence, this hybrid approach could provide benefits for all users and organization to decrease the computer risk.

Keywords: Spam mail, N-gram extraction, Classification, Stop-word removal, Non-segmented languages

I INTRODUCTION

In Thailand, the Internet recently plays a significant role which makes the communication easier for people in all areas. Electronic mails or e-mail serve as the popular communication channel which widely used due to cost and time saving. More importantly, the e-mail is an online service and unlimited that can be sent to multi people in different places at a time.

Regarding the advantages, the email has become a primary way for online communication among the Internet users (Statistics and facts of email marketing, 2013) With the rapid growth and various benefits of the email, some groups of people make an effort to earn their income by using email containing soliciting message to advertise, promote or selling their products and services. This potential threat is known as "Spam Mail". The spam mail or unsolicited message email causes the interruption to any receiver in which they are offended their privacy and waste their time to eliminate those spam mails. The spam mails steadily increase on the Internet because it is instant and very cost-effective medium (Puckdeewongs, 2003). Consequently, the spam mail is used for money purpose. Furthermore, the large growth of spam mail also impacts to all people and organization in the aspect of computer risk since spam mail embeds not only persuasive texts but also viruses, malwares, spywares, and so forth. These are harmful to all users including individuals, organizations and a group of agencies.

In general, the spam mail are basically prevented or detected by examining words which are comprised in any sections of emails such as the title or the content (Tubsee, 2010). Words used in the email content are separated and brought to investigate the characteristic of spam mail. Nevertheless, procedures in separating words of each language are different, for example, English words are separated by white space. In contrast to English language, many Asian languages (e.g. Chinese, Japanese, Thai and Korean) are written without spaces. Thus, the traditional technique can be only applied with English languages. Regarding to the limitation, n-gram has become more important technique for separating words from email content because it can be used with all different languages without considering the meaning of words. Moreover, the study of Stop-word removal for more accuracy in filtering the spam mails should be concerned. Based on the survey, the email both normal and spam mails mainly composed of Stop-word approximately 30 to 50% of the content in many languages. The set of Stop-word is normally shared by most emails which make the accuracy of distinguishing and filtering the spam mails is lower. Hence, this research aims to propose the hybrid approach integrating three techniques: Stop-word Removal, N-gram and Data Classification, for better categorization and filtering spam mails.

II RELATED WORKS

Spam mail is a term of unsolicited commercial message which has been sent to numerous recipients without their permission or reception (Yusuk, 2013). Those spam mails are unwanted mail which encompasses with product, services, advertisement, online gamble contents (Rojkangsadan, 2016). In traditional way, screening the spam mails can be done through data classification technique, which previous data are collected and brought to train the system and then to establish a new model for testing. The classification technique consists of three main procedures: 1) creating a new model by analyzing the previous data with a variety of data classification techniques, 2) testing the efficiency of the build model, and 3) applying the model for screening new upcoming data (Pacharawongsakda, 2014).

Before classifying the data, information extraction (IE) process is primarily used for classifying the proper information. Many algorithms may be adopted in the information extraction in order to enhance computer system for greater comprehension on any document, texts, and data (Mansamut, 2010). Another technique is to employ Natural Language Processing (NLP), the artificial intelligence that enables a computer program to understand intended meaning of any human languages. Nevertheless, the diversity of human speech reflects various kinds of language difficulties such as ambiguity, pronouns, reference and ellipsis. As a result, the classification of text documents can be divided into two main techniques: language dependent and language independent (Todsanai, 2014).

In the language dependent approach, there are several techniques proposed to split non-segmented texts such as Chinese (Kwok), Japanese (Croft, 1993), Korea (Ahn, 1996) and Thai (Smith, 2001) into term tokens. A word segmentation technique is usually required to extract a bag of words before they can be used in classification process. This technique usually rely on language analysis or on the use of dictionary. The preparation of such method is very time consuming. Therefore, word segmentation has become a challenging task in Natural Language Processing (NLP) for many languages.

Apart from language dependent technique, there is other technique, called n-gram technique, which is language-independent (Adams, 1991). This technique is a language-independent approach which has been adapted in word delimitation process. n-gram is a sequence of items which typically collected from either spoken and written texts. The n-gram model word type provides more efficient to establish "Language Statistic Model" and to retrieve any information without language dependency (Majumder, 2002). For instance, the analysis of a word 'university' by bi-grams resulted as : un, ni, iv, ve, er, rs, si, it, ty, y_, using '_' for space.

Both language dependent and independent approaches have been applied in several previous studies in the area of document classification in multi-languages. The study of Majumder investigated five groups of Indian language adhering to Algorithm theory proposed by Cavnar (Majumder, 2002). The researchers selected a set of 100 documents from five different Indian languages for creating n-gram profiles and examining the shortest space. Their findings showed that using alphabets and words building n-gram profiles offer positive results in both cases.

Another study is done by Mansur (Mansur, 2006) employing n-Gram Algorithm to categorize Bangla in Prothom-Alo newspapers which was published for a year. The overall results revealed that 2-gram or 3-gram could perform as the most proper methods for data classification. Furthermore, Zhihua adopted text representative and data selection techniques to classify data in Chinese language (Zhihua, 2009). They examined n-grams for three main purposes: 1) to compare the effectiveness on performance of data selection features, 2) to compare the sparseness, and 3) to compare the ability among three weight methods.

The data from Chinese corpus TanCorpV1.0 (over 14,000 texts divided into 12 classes) were used as the source of data in their study. Their research findings are detailed as follows: 1) by using less than 3000 features, the n-gram frequency basically provides better outcomes rather than those selected by text frequency (absolute or relative). Relative frequency results reflect more negative consequences than absolute frequency. In the cause of using greater than 3000 features, both absolute and relative frequency results contained much similarity, 2) the selective method using n-gram frequency caused a dense of "text feature" matrix, which was more than the relative frequency, 3) the n-gram frequency selection generated the feature which carried less correlation than those selected in the text frequency, and 4) text representation at TF weight contained similar performance to those text using tf*idf , in which both methods had greater capability than 0/1 logical weight.

III FILTERING SPAM MAIL IN NON-SEGMENTED LANGUAGE USING HYBRID APPROACH: THE INTREGATION OF STOP-WORD REMOVAL, N-GRAM EXTRACTION AND CLASSIFICATION TECHNIQUES

This section describes five procedures of spam mail filtration by applying the integration of three techniques: Stop-word Removal, N-gram extraction and classification techniques. Workflow of the proposed technique is illustrated in Figure 1.

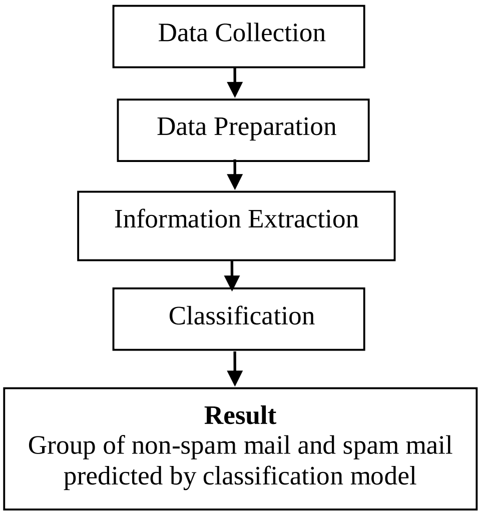


Figure 1. Workflow of the proposed technique.

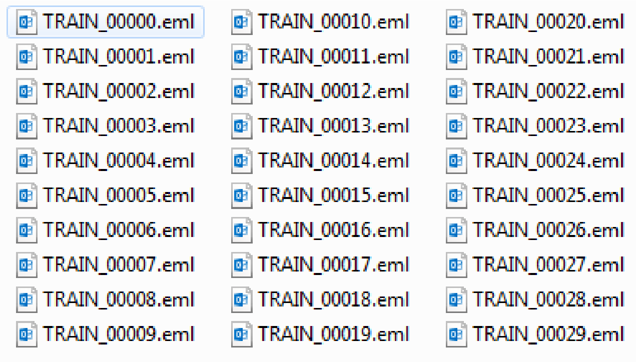
From Figure 1, the processes in screening or filtering spam mails are divided into four steps including data collection, data preparation, information extraction and classification. These four steps comprise of distinct function as given below.

A. Data Collection

For data collection (Spam email datasets, 2010), all data was gathered from CSDMC2010 SPAM corpus by selecting the content of emails and used to experiment spam mail filtering system. 410 sets of information were brought and grouped into three folders: 1) 198 normal electronic mails, 2) 202 junk mails or spam mails, and 3) 10 selected mails opened with outlook program for sighting all content in the mails.

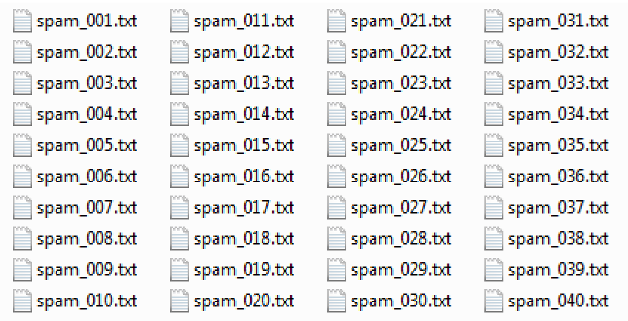
B. Data Preparation

The data preparation is the second procedure which aims to select the information from email content and to convert the data. Before information extraction, '.eml' files were converted into '.txt' or text document files for being used in the other steps. The data preparation stage is presented in Figure 2 and 3.



TRAIN_00000.eml
TRAIN_00001.eml
TRAIN_00002.eml
TRAIN_00003.eml
TRAIN_00004.eml
TRAIN_00005.eml
TRAIN_00006.eml
TRAIN_00007.eml
TRAIN_00008.eml
TRAIN_00009.eml
TRAIN_00010.eml
TRAIN_00011.eml
TRAIN_00012.eml
TRAIN_00013.eml
TRAIN_00014.eml
TRAIN_00015.eml
TRAIN_00016.eml
TRAIN_00017.eml
TRAIN_00018.eml
TRAIN_00019.eml
TRAIN_00020.eml
TRAIN_00021.eml
TRAIN_00022.eml
TRAIN_00023.eml
TRAIN_00024.eml
TRAIN_00025.eml
TRAIN_00026.eml
TRAIN_00027.eml
TRAIN_00028.eml
TRAIN_00029.eml

Figure 2. Collection of Files Used in the Study.

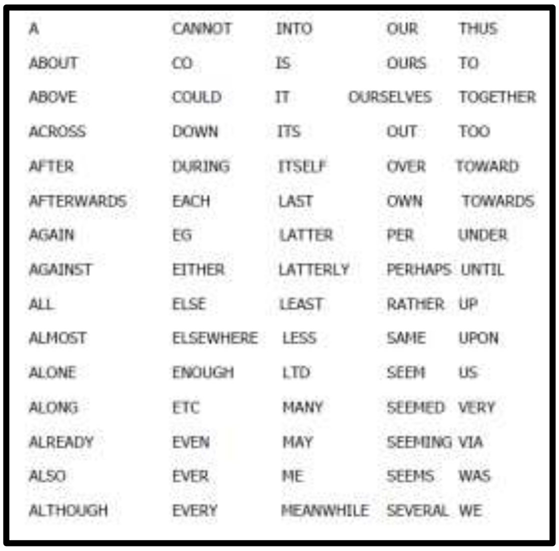


spam_001.txt
spam_002.txt
spam_003.txt
spam_004.txt
spam_005.txt
spam_006.txt
spam_007.txt
spam_008.txt
spam_009.txt
spam_010.txt
spam_011.txt
spam_012.txt
spam_013.txt
spam_014.txt
spam_015.txt
spam_016.txt
spam_017.txt
spam_018.txt
spam_019.txt
spam_020.txt
spam_021.txt
spam_022.txt
spam_023.txt
spam_024.txt
spam_025.txt
spam_026.txt
spam_027.txt
spam_028.txt
spam_029.txt
spam_030.txt
spam_031.txt
spam_032.txt
spam_033.txt
spam_034.txt
spam_035.txt
spam_036.txt
spam_037.txt
spam_038.txt
spam_039.txt
spam_040.txt

Figure 3. Collection of Converted Files.

C. Information Extraction and Natural Language Processing

At this recent stage, two methods including Stop-word Removal and n-gram Extraction were then performed. Stop-word Removal method refers to words or vocabularies which frequently appear in all types of written documents, but those words are unable to represent the focal meaning of the message such as a, an, the, and, about, is, am, are in English language as shown in figure 4. Typically, many languages in the world have their own stop word list which appears approximately 30 to 50% of the content. The frequent occurrence of those stop words would make most email are similar to each other which makes it difficult in distinguishing the spam mail from non-spam mails. This is because most emails share the same stop words in emails. Consequently, these stop words should be excluded from the email before information extraction can be performed.



A	CANNOT	INTO	OUR	THUS
ABOUT	CO	IS	OURS	TO
ABOVE	COULD	IT	OURSELVES	TOGETHER
ACROSS	DOWN	ITS	OUT	TOO
AFTER	DURING	ITSELF	OVER	TOWARD
AFTERWARDS	EACH	LAST	OWN	TOWARDS
AGAIN	EG	LATTER	PER	UNDER
AGAINST	EITHER	LATTERLY	PERHAPS	UNTIL
ALL	ELSE	LEAST	RATHER	UP
ALMOST	ELSEWHERE	LESS	SAME	UPON
ALONE	ENOUGH	LTD	SEEM	US
ALONG	ETC	MANY	SEEMED	VERY
ALREADY	EVEN	MAY	SEEMING	VIA
ALSO	EVER	ME	SEEMS	WAS
ALTHOUGH	EVERY	MEANWHILE	SEVERAL	WE

Figure 4. Example of Stop Words in English.

At this step, stop words are eliminated by creating Hash Table for all stop words and compare with email content in the collection. After that, stop words shown in the email were then deleted; only important content remains in the document. The next step is extracting n-gram from the emails. All emails without stop words were processed in the information extraction by using n-gram technique. In this paper, 2-gram, 3-gram, 4-gram and 5-gram were used as a dimension of gram in information extraction process which will be described in the following.

Assume that email *d* consists of a string *s* of characters a_1, a_2, ..., a_N. An n-gram term is a substring of *n* overlap or non-overlap successive characters extracted from the string. Extracting a set of n-gram terms from the email *d* can be done by using the 1-sliding technique. Therefore, the *i*th n-gram term extracted from email *d* is the substring a_i, a_{i+1}, ..., a_{i+n}. Table 1 shows 2-gram, 3-gram, 4-gram, 5-gram overlap sequence of the email *d* containing the string *s* 'you won'

Table 1. Samples of 2-gram, 3-gram, 4-gram, 5-gram of string s 'you won'.

n-gram	List of n-gram terms
2-gram terms	yo, ou, u_, _w, wo, on
3-gram terms	you, ou_, u_w, _wo, won
4-gram terms	you_, ou_w, u_wo, _won
5-gram terms	you_w, ou_wo, u_won

After dividing terms with *n* difference, the researcher produced the model for information processing and predicting types of emails in the next procedures.

The different size of *n* (2-5) was used to assess the performace and to investigate the propotion of *n* which reflects the most effective results.

D. Data Classification

The research was conducted using data mining techniques which includes Decistion Tree, Naïve Bayes, K-NN and SVM with the same set of data. The results of SVM technique is appropriate for the data sets and the highest performance. Therefore, SVM technique is used as a model applied in this research. In this step, the SVM model were constructed from training data method. After extracting n-gram terms from emails, the SVM model was tested with a set of testing data for assessing the accuracy of the model. The researcher emperimented the model three rounds with the specific different random seed staring from 500, 1000, and 1500. In addition, the comparison of model performance with differenct size of *n* from 2 to 5 was assessed. The average or mean of various *n*-gram performance are presented in Table 2.

Table 2. Percentage of accuracy for diffent grams.

Random Seed	2-gram	3-gram	4-gram	5-gram
500	75.06	90.93	91.18	93.2
1000	74.81	90.93	91.18	93.2
1500	76.07	92.19	91.18	93.2
Average	**75.31**	**91.35**	**91.18**	**93.2**

From table 2, result shows the percentage of accuracy of predicting or categorizing whether the sample emails are normal or spam mails. The accuracy of the analysis would be monitored. The experimental study and result will described in detail in next section.

IV EXPERIMENTATL STUDY

Set of input emails used in this paper were collected from CSDMC2010 SPAM Corpus (from: http://csmining.org/index.php/spam-email-datasets-.html). The researcher gathered 410 data which were separated into three sets: 198 normal mails, 202 spam mails, and 10 testing emails. These email will be sent to Stop-word removal and Information Extraction processes. Then the SVM model were created and learn from the data set by using 10-Fold Cross Validation method. The operator used in training process is shown in Figure 5. Moreover, the performance of the SVM model and the build SVM model are presented in Table 3 and Figure 6.

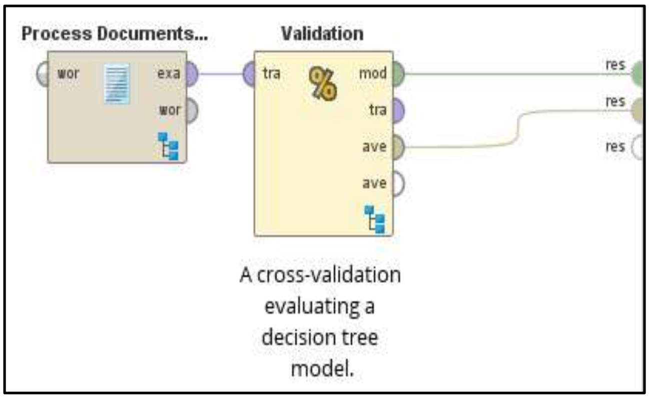
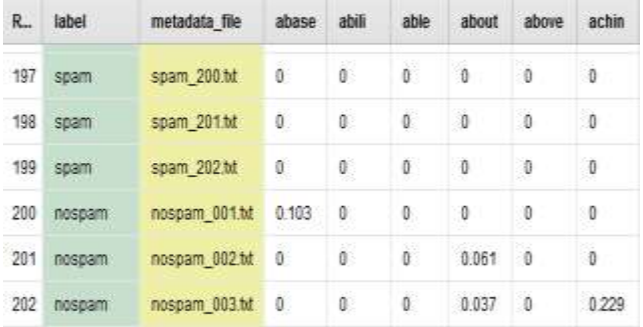


Figure 5. Operator Used in Training Data Method.

Table 3. Performance of the SVM Model.

accuracy: 93.20% +/- 2.28% (mikro: 93.20%)

	True spam	True nospam	Class precision
Pred.spam	190	18	91.35%
Pred. nospam	9	180	95.24%
Class recall	95.48%	90.91%	



R..	label	metadata_file	abase	abili	able	about	above	achin
197	spam	spam_200.txt	0	0	0	0	0	0
198	spam	spam_201.txt	0	0	0	0	0	0
199	spam	spam_202.txt	0	0	0	0	0	0
200	nospam	nospam_001.txt	0.103	0	0	0	0	0
201	nospam	nospam_002.txt	0	0	0	0.061	0	0
202	nospam	nospam_003.txt	0	0	0	0.037	0	0.229

Figure 6. Sample of the Build SVM Model.

Once building and evaluating the efficiency of the SVM model is completed, it is then used to test with testing email as shown in Figure 7. The results of this experiment are presented in Table 4.

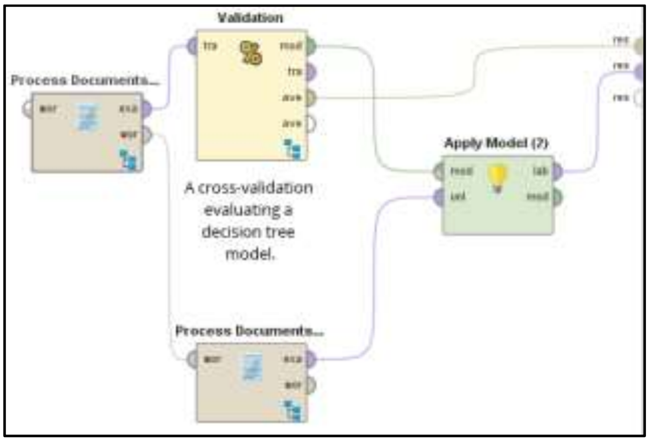


Figure 7. Operator Used in Testing Data Process.

Table 4. The Results from Prediction.

	Label	Pred. (label)	Confidence (spam)	Confidence (nospam)	Meta datafile
1	Don't know	nospam	0.262	0.738	dn_ns08 .txt
2	Don't know	nospam	0.357	0.643	dn_ns09 .txt
3	Don't know	nospam	0.342	0.658	dn_ns10 .txt
4	Don't know	spam	0.831	0.169	dn_ns01 .txt
5	Don't know	spam	0.775	0.225	dn_ns02 .txt
6	Don't know	spam	0.657	0.343	dn_ns03 .txt
7	Don't know	spam	0.782	0.218	dn_ns04 .txt
8	Don't know	spam	0.788	0.212	dn_ns05 .txt
9	Don't know	spam	0.778	0.222	dn_ns06 .txt
10	Don't know	spam	0.828	0.172	dn_ns07 .txt

After testing the SVM model with the testing data process, the researcher then conducted three consecutive experiments and computed the average score of a particular *n* dimension: 2-gram, 3-gram, 4-gram, 5-gram for investigating the highest accuracy of prediction. The results are shown in Figure 8.

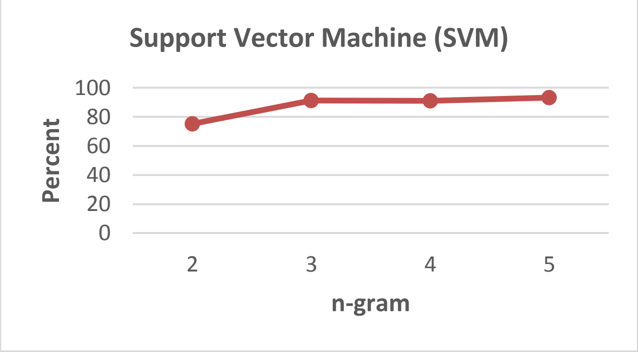


Figure 8. Means of the Model Performance in Different n-Grams.

Regarding the experiment result, it found that 5-gram terms indicated the most accurate prediction which accounts for 93.2%, followed by 3-gram (91.35%), 4-gram (91.18%) and 2-gram (75.31%) respectively.

V CONCLUSION

Spam mail remains one of the dilemmas in electronic mail services which steadily contribute many disadvantageous for the users. The spam mails are unwanted or unsolicited message that not only sent for commercial purpose, but also transmit viruses, malwares, and spywares to users' computer systems. To reduce computer risk, this paper proposed the integrated approach combing three techniques: Stop-word Removal, N-gram extraction and Classification for screening or filtering spam mail precisely and effectively. This hybrid approach is available to all language used in the world. The experimental result shows that the proposed technique is able to identify spam mail with performance 93.2% of accuracy. Furthermore, the researcher discovered that the number of *n* dimension which conveys the most precise prediction was 5-grams, following by 3-grams (91.35%), 4-gram (91.18%), and 2-gram (75.31%). Based on the results, it can be concluded that the proposed technique can serve as the effective model to categorize spam mail from other emails. Significantly, the model may be helpful for users either individuals or organizations to avoid and computer risk and cybercrimes. It also can be applied to any other method or theory. To achieve a more accurate prediction

REFERENCES

Adams, E. (1991). A Study of Trigrams and Their Feasibility as Index Terms in a Full Text Information Retrieval System. *PhD thesis, George Washington University*, USA, 1991.

Ahn, J. (1996). Using n-Grams for Korean Text Retrieval, *In Proc. Int'l Conf. on Information Retrieval, ACM SIGIR*, Zurich, Switzerland, pp. 216-224.

Croft, W. (1993). A Comparison of Indexing Techniques for Japanese Text Retrieval, *In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 237-246.

Kwok, L. (1997). Comparing Representations in Chinese Information Retrieval," in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information on 27-31 July. Philadelphia, PA: ACM Publication , 34-41.

Majumder, M. M Mitra and B.B. Chaudhuri. (2002). N-gram a language independent approach to IR and NLP, Kolkata.

Mansamut, K. (2010). Information Extraction (IE). Retrieved 2015 April 21, From : https://ejeepss.wordpress.com/2007/10/01/information-extraction-ie/.

Mansur, M. Naushad UzZaman and Mumit Khan. (2006). Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus (Bacherlor's Thesis). Department of Computer Science and Engineering, BRAC University, Bangladesh.

Pacharawongsakda, E. (2014). An Introduction to Data Mining Techniques. (2nd ed.). Bangkok : Asia Digital Printing Limited.

Puckdeewongs, A. (2003). Anti-Spam Mail System. *Master thesis, M.S., King Mongkut's Institute of Technology North*, Bangkok

Rojkangsadan, T. (2016). Cognizant of spam. Retrieved 2015 April 22, From : http:// www.dailynews.co.th/it/376076.

Spam email datasets. (2010). Retrieved 2016 April 26, From : http://csmining.org/index.php/spam-email-datasets-.html.

Statistics and facts of email marketing. (2013). Retrieved 2015 April 22, From : http:// thumbsup.in.th/2013/06/email-marketing-2013-and-beyond.

Todsanai, C. (2014). Using Clustering Techniques for Non-segmented Language Document Management: A Comparison of K-mean and Self Organizing Map Techniques. In the Knowledge Management International Conference (KMICe), 12-15 August. Langkawi, Malaysia.

Tubsee, H. (2010). Thai Spam Mail Filter Using Neural Network. *Master thesis, M.S., King Mongkut's Institute of Technology North, Bangkok.*

Yusuk, S (2013). junk mail. Retrieved 2015 April 22 From : http://www.cdoae.doae.go.th/56/ked-it/15.pdf.

Zhihua, W., Jean-Hugues, C., Rui, Z., and Wen, L. (2009). N-grams based feature selection and text representation for Chinese Text Classifiacation. International Journal of Computational Intelligence Systems, 2(4), pp. 365-374.