International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:4, 2012

# Identification of Most Frequently Occurring Lexis in Body-enhancement Medicinal Unsolicited Bulk e-mails

Jatinderkumar R. Saini, Apurva A. Desai

***Abstract*****—**e-mail has become an important means of electronic communication but the viability of its usage is marred by Un-solicited Bulk e-mail (UBE) messages. UBE consists of many types like pornographic, virus infected and 'cry-for-help' messages as well as fake and fraudulent offers for jobs, winnings and medicines. UBE poses technical and socio-economic challenges to usage of e-mails. To meet this challenge and combat this menace, we need to understand UBE. Towards this end, the current paper presents a content-based textual analysis of more than 2700 body enhancement medicinal UBE. Technically, this is an application of Text Parsing and Tokenization for an un-structured textual document and we approach it using Bag Of Words (BOW) and Vector Space Document Model techniques. We have attempted to identify the most frequently occurring lexis in the UBE documents that advertise various products for body enhancement. The analysis of such top 100 lexis is also presented. We exhibit the relationship between occurrence of a word from the identified lexis-set in the given UBE and the probability that the given UBE will be the one advertising for fake medicinal product. To the best of our knowledge and survey of related literature, this is the first formal attempt for identification of most frequently occurring lexis in such UBE by its textual analysis. Finally, this is a sincere attempt to bring about alertness against and mitigate the threat of such luring but fake UBE.

***Keywords*****—**Body Enhancement, Lexis, Medicinal, Unsolicited Bulk e-mail (UBE), Vector Space Document Model, Viagra

## I. Introduction

WITH the increase in usage and availability of Internet, there has been a tremendous increase in usage of e-mail. It has proved to be an important medium of cheap and fast electronic communication. But the same thing that has increased its popularity as a communication medium has also proved to be a source of non-personal, non-time critical, multiple, similar and un-solicited messages received in bulk. This type of message is called Unsolicited Bulk e-mail (UBE) and is known by various other names like Spam e-mail, Junk e-mail and Unsolicited Commercial e-mail (UCE). The spread of UBE has posed not only technical problems but also posed major socio-economic threats. Also, the definition of spam e-mail is 'relative' [5], [10], [14]. This means to say that all e-mails going to spam folder may not be spam for a person – same as all e-mails going to inbox may not be ham (i.e. non-spam) e-mails. Further, all spam e-mail is not harmful, some is just annoying [3], [7], [12].

J. R. Saini is with the Narmada College of Computer Application, Bharuch, Gujarat, India as Associate Professor and I/C Director. He is PhD from Veer Narmad South Gujarat University, Surat, Gujarat, India. (phone: +91-9426861815; e-mail: saini_expert@yahoo.com).

A. A. Desai is with the Veer Narmad South Gujarat University, Surat, Gujarat, India as Professor and Head of Department of Computer Science. He is PhD from Veer Narmad South Gujarat University, Surat, Gujarat, India. (e-mail: desai_apu@hotmail.com).

UBE incidences range from fake job offers and viruses to pornography. Another area of concern is of spam e-mails that advertise the body enhancement medicinal products. The target areas of these products range from enhancement of male and female genitals to loose or gain weight, improve hair growth and reduce blood-sugar.

The dangerous thing about these emails is that they demand a handsome amount of money for delivery of the product, which is never delivered or in worst case a fake product is delivered. But due to the fear of society and feeling of embarrassment, the victim rarely comes out to declare of the way he/she was cheated through non-delivery or delivery of a fake product against a heavy payment of a so-called body enhancement medicine. Further, these kind of UBE mostly target medicines or drugs like Viagra, Xanax and Phentrimine for the genitals and many times the advertising pharmacies include pictures and textual statements in the emails which are largely pornographic. Even though there are many target areas of such medicinal products as advertised and offered in the UBE, in general this paper refers to this kind of UBE as body enhancement medicinal UBE.

In past, researchers have worked in direction of understanding the spam for combating it [1], [9], [19]. We also believe that first step in combating spam is to understand spam. A novel idea proposed in this paper is that the best way of understanding spam is to analyze it. Most importantly, spam can be differentiated by content [16] and in this paper we target content-based analysis of un-structured UBE documents which advertise fake medicines for body enhancements. The present work aims towards identification of lexis occurring in such UBE. The basic structure of spam e-mail message is same as of ham e-mail, consisting of 'header' and 'body' parts. In this paper, we have treated spam e-mail as un-structured because in addition to consideration of contents of structured 'header' part, we propose content analysis of 'body' part also. The structure of 'body' part is not fixed with respect to number of words, lines, format, etc. and hence we treat UBE as an un-structured document. From a technical perspective, identification of most frequently occurring lexis in UBE documents is a Text Parsing and Tokenization task and we propose to solve it using Bag of Words (BOW) and Vector Space Document Model approach.

## II. Related Works

As far as, our study of past and contemporary literature for this field is concerned, this is the first formal attempt for identification of lexis occurring in body enhancement medicinal UBE. The survey of related work shows that the researchers have made many attempts to classify e-mails into ham and spam groups. The numbers of attempts targeted

towards classification of spam e-mails are also there, but very scarce, per se. Kiritchenko et al. have treated e-mail classification as a special case of text classification [11]. Martin et al. have laid emphasis on individual user's behaviour for identifying spam messages [13]. Fette et al. have presented and evaluated a classification method for spam and non-spam e-mails based on Training-data Set approach [6]. Gajewski has discussed the use of a naïve Bayesian classifier based on a BOW representation of an e-mail. He has defined Spam as an unsolicited commercial mail and something strictly related to 'commerce' and 'consumption goods' or some very well defined sectors of activity as 'Pharmacy' or 'Pornographic' industry [8].

There are many other instances too where the researchers have identified as well as classified body enhancement medicinal UBE as a category of spam emails. This categorization, though, differs a lot as far as the terminology and rationale of classification is concerned. Threat Research and Content Engineering (TRACE) group of Marshal Ltd. [18] has broadly classified spam into nine categories. Under the title of 'Health', amongst the nine categories, they have classified primarily spam of a pharmaceutical nature, advertising all manner of drugs, pills, potions and herbal remedies. They are of the opinion that this spam often promises better skin, weight loss, sexual enhancement, lengthening, invigoration and energy. Lambert [12], in his report on 'Analysis of Spam', has classified this kind of UBE into three categories viz., 'Body Enhancement', 'Viagra' and 'Weight Loss'. The researchers working with website mahalo.com [2] have listed 'Viagra' as an explicit category for body enhancement medicinal UBE.

Zahren B. [20], in an attempt to classify spam emails, has provided the categories 'Erectile Dysfunction', 'Online Pharmacy' and 'A larger penis' for the classification of body enhancement medicinal UBE. The Spamregister collection [17] has provided an elaborate categorization of spam emails into various categories. They have classified body enhancement medicinal UBE into two categories for 'Pharmacy' and 'Weight Loss'. The researchers at knujon.com [1] website have classified such UBE as a kind of 'product-driven' spam that includes 'Prescription Drugs' and offers sales of controlled substances. The related literature review, hence, provides us with sufficient evidence of classification of body enhancement medicinal UBE into different categories. In this paper, we attempt to combine such similar categories together and identify the most frequently occurring lexis in body enhancement medicinal UBE. At the same time, the textual analysis of such UBE and identification of most frequently occurring lexis therein is a vent through which we express our concern for the gravity of problem of UBE in general and body enhancement medicinal UBE in specific.

## III. Methodology

In this section, we describe the detailed methodology followed by us for the identification of most frequently occurring lexis in body enhancement medicinal spam e-mails. For the sake of simplicity and better understanding, the entire section is divided into three major sub-sections for Data Collection & Clustering, Data Pre-processing and Feature Extraction & Feature Selection.

### *A. Data Collection & Clustering*

We first collected various UBE documents of all types together. We used 40 e-mail addresses for collecting the required data. Another 18 websites providing online archives of UBE were also used for data collection. This formed a text corpus amounting to approximately 1.5 GB of data-size and consisted of 30074 UBE documents. To prevent the data from 'contributor bias' [4], it was sourced from different locations and at different times from e-mail addresses owned by different persons.

As a next step, we identified the data clusters. For this, we used hierarchical divisive clustering approach in which initially all the UBE documents formed one text corpus of a single cluster. The process of clustering was based on the analysis of the contents of UBE documents in the text corpus. This text corpus was processed to yield 2 clusters in such a way that one cluster contained the body enhancement medicinal UBE whereas the other cluster did not contain such UBE. The cluster comprising body enhancement UBE was the cluster of interest and the number of instances in it was 2711, which amounted to nearly 178 MB of data size. Given the inherent in-secure nature of UBE documents, a noteworthy thing here is that the collection of such UBE is a difficult process. Our intention was to create a corpus of UBE which advertise the body enhancement medicines or medicinal products like for genitals, hair, fat and weight. Our task of data collection was eased by the fact that many of this kind of UBE have an explicit subject line which makes it easy to identify the category of UBE under question. Besides our naïve approach for categorization of UBE, the spam filters provided by the e-mail providers also helped us confirm the categorization by actually classifying the UBE under the spam folder.

### *B. Data Pre-processing & Cleaning*

The main motive of this phase was to clean the data. At this stage, we pre-processed the collected text-files in the UBE corpora by removing 'obvious noise' from them and converting them in a common format. By 'obvious noise', we mean the location and site specific data slipped into the UBE documents when sourced from different locations, e.g. website name. This data-cleaning is also required for making the data ready for further processing – specifically, easing the subsequent phase of feature extraction.

### *C. Feature Extraction & Feature Selection*

This is the most important and bulkiest phase of data-processing. The types of operations done during this phase are often referred to as 'Feature Extraction' and 'Feature Selection' by the research literature of text analysis and text mining. Here, we picked the corpus of body enhancement medicinal UBE. The corpus under consideration is actually formed of UBE which are eventually text documents. For each text document, we performed sentence-splitting in order to treat it as a Bag Of Words (BOW). In BOW representation of a text document, lexis or terms or tokens in the document are identified with words in the document. Hence this

representation is also called Set of Words (SOW) [15]. We then performed Syntactic Text Analysis by Parsing the UBE document, for extraction of Tokens.

In English language the tokens are words [21] and the act of breaking the text into tokens is called Tokenization. A noteworthy thing here is that our tokenization is not case-sensitive. This means that a word appearing in any combination of lower-case or upper-case letters is treated as the same word. As a next step we counted the number of unique tokens in each UBE. This resulted in each document being represented as sub-set of Vector Space Document Model (VSDM). A vector corresponding to each UBE in this model is 2-dimensional, consisting of unique tokens and their frequency and is sorted on frequency column in descending order. This resulted in a total number of 2711 vectors, one each for the 2711 UBE in the cluster of interest.

Further, the UBE vectors are designed not to include stop-words. Special kinds of stop-words considered by us are Domain stop-words. These are the words which are statistically irrelevant in the context of current research work because of their presence in both clusters, i.e. cluster formed of body enhancement medicinal UBE and the cluster formed of non body enhancement medicinal UBE. Hence, the entire stop-list considered by us, consists of following four types of stop-words:

a. HTML stop-words e.g. html, body, img
b. Generic stop-words e.g. his, thus, hence
c. Noise stop-words e.g. isdfalj, asdfwg
d. Domain stop-words e.g. salary, academy, phone

As a final step towards simplification of data processing, we created a single vector from the 2711 vectors of UBE documents. This 2-dimensional vector consisted of 16879 unique tokens and was sorted on the frequency count of tokens in descending manner. The number of tokens in this single vector was naturally less than the sum of number of tokens in each of 2711 vectors. The frequency count for a given token in this vector is the aggregate sum of the frequency count of the token in the 2711 vectors. This means to say that those vectors which do not contain the given token, contribute a value of zero towards the aggregate sum. Next, our motive was to keep only the desired lexis in this vector of extracted lexis. As the stop-words were already removed, this was a second level of refinement of the vector. For this we removed all lexis of length greater than 30, as we did not deemed them to be of statistical relevance. The frequency of 1 in the aggregated vector is an indication that the token has appeared only 1 time in 2711 documents. As a result we also removed all those tokens with a frequency of 1. The number of lexis with length greater than 30 and with frequency of 1 was 8 and 8366 respectively. The removal of such words resulted in the highly refined selected lexis set of 8505 lexis.

## IV. Results And Findings

Based on the processing of more than 2700 body enhancement medicinal UBE, we obtained a vector containing 8505 lexis. This vector is a set of words contained in the body enhancement medicinal UBE. On the analysis of this vector, we were able to identify the most frequently occurring lexis in such UBE. The identification of lexis with highest frequency is possible from this vector as it is sorted in descending manner on the frequency count of the lexis. A snap-shot of listing of such top 100 lexis is given in Table I. The third column of Table I depicts the frequency of the word in the set of 2711 UBE whereas the fourth column is the ratio of frequency of the word to the number of UBE. In Table I, this is expressed in terms of percentage of the value and is called 'Percentage of Presence'.

The first record of Table I can be interpreted to say that the word 'PENIS' appears for 2203 times in 2711 UBE with a presence percentage of {(2203 / 2711) x 100 =} 81.26%. The other records of the Table I can be interpreted similarly. If a word appears for more than 2711 times in a set of 2711 UBE, evidently the word registers a presence of more than 100% in the UBE set under consideration. We were not able to find even a single word that could register a presence of more than 100% in the UBE set. We believe that the reason behind this is the 'word-mutation" technique used by the spammers. For instance, instead of always using the word 'VIAGRA' directly, the spammer would use its mutated formation like 'VAIGRA' and 'VIARGA'.

The single most frequently occurring word in the body enhancement medicinal UBE was found to be 'PENIS' with a presence percentage of 81.26%. This was followed by six words, 'VIAGRA', 'MEGADIK', 'PILLS', 'PHARMACY', 'SEXUAL' and 'MEDS', in the presence percentage range of 25 to 49. An important interpretation of this result is that the presence of these words is a clear indication of a high probability of the UBE under consideration to be one containing advertisement of body enhancement medicinal product. There were 93 words in the presence percentage range of 0 to 24.

TABLE I
Frequency And Percentage Of Presence Of Most Frequently Occurring Top 100 Lexis

| Sr. No. | Lexis | Freq-uency | Percentage of Presence |
|---|---|---|---|
| 1 | PENIS | 2203 | 81.26 |
| 2 | VIAGRA | 1311 | 48.36 |
| 3 | MEGADIK | 1262 | 46.55 |
| 4 | PILLS | 1142 | 42.12 |
| 5 | PHARMACY | 1003 | 37.00 |
| 6 | SEXUAL | 752 | 27.74 |
| 7 | MEDS | 717 | 26.45 |
| 8 | DRUGS | 587 | 21.65 |
| 9 | INCHES | 581 | 21.43 |
| 10 | LOGON | 498 | 18.37 |
| 11 | MEDICATIONS | 414 | 15.27 |
| 12 | XUAL | 404 | 14.90 |
| 13 | MALE | 396 | 14.61 |
| 14 | ENLARGEMENT | 378 | 13.94 |
| 15 | CIALIS | 366 | 13.50 |
| 16 | PILL | 362 | 13.35 |
| 17 | DRUG | 300 | 11.07 |
| 18 | ERECTION | 295 | 10.88 |
| 19 | ENLARGE | 276 | 10.18 |
| 20 | PLEASURE | 245 | 9.04 |
| 21 | CANADIANPHARMACY | 244 | 9.00 |
| 22 | ERECTILE | 216 | 7.97 |
| 23 | PRESCRIPTION | 215 | 7.93 |

| | | | |
|---|---|---|---|
| 24 | COCK | 214 | 7.89 |
| 25 | DRUGSTORE | 201 | 7.41 |
| 26 | AIRPLANE | 197 | 7.27 |
| 27 | TOOL | 192 | 7.08 |
| 28 | HAM | 186 | 6.86 |
| 29 | SURGERY | 181 | 6.68 |
| 30 | VPXL | 180 | 6.64 |
| 31 | HERBAL | 173 | 6.38 |
| 32 | CODEINE | 168 | 6.20 |
| 33 | MANSTER | 166 | 6.12 |
| 34 | DYSFUNCTION | 159 | 5.86 |
| 35 | PUMPS | 151 | 5.57 |
| 36 | CIA | 145 | 5.35 |
| 37 | VISITS | 144 | 5.31 |
| 38 | MANHOOD | 144 | 5.31 |
| 39 | EXERCISES | 140 | 5.16 |
| 40 | BOYFRIEND | 140 | 5.16 |
| 41 | FDA | 135 | 4.98 |
| 42 | STIMULATION | 130 | 4.80 |
| 43 | CHECKOUT | 128 | 4.72 |
| 44 | INSTRUMENT | 121 | 4.46 |
| 45 | PEER | 120 | 4.43 |
| 46 | YOURPENIS | 116 | 4.28 |
| 47 | ERECTIONS | 113 | 4.17 |
| 48 | JEALOUS | 112 | 4.13 |
| 49 | GUYS | 112 | 4.13 |
| 50 | EJACULATION | 111 | 4.09 |
| 51 | PHENTERMIN | 111 | 4.09 |
| 52 | GIRTH | 108 | 3.98 |
| 53 | LAUGH | 108 | 3.98 |
| 54 | CELEB | 108 | 3.98 |
| 55 | DIC | 107 | 3.95 |
| 56 | ANATRIM | 106 | 3.91 |
| 57 | TRAM | 106 | 3.91 |
| 58 | CAILIS | 105 | 3.87 |
| 59 | DOSE | 104 | 3.84 |
| 60 | INTERCOURSE | 103 | 3.80 |
| 61 | CIALIX | 99 | 3.65 |
| 62 | NIS | 98 | 3.61 |
| 63 | VIAGRACAILIS | 98 | 3.61 |
| 64 | MIRACLE | 98 | 3.61 |
| 65 | LEVITRA | 97 | 3.58 |
| 66 | GRA | 97 | 3.58 |
| 67 | AMBIEN | 97 | 3.58 |
| 68 | INCH | 96 | 3.54 |
| 69 | LUCAS | 96 | 3.54 |
| 70 | VIAGRAB | 93 | 3.43 |
| 71 | VIAGRAPILLS | 91 | 3.36 |
| 72 | ORGASMS | 90 | 3.32 |
| 73 | MEDICATION | 90 | 3.32 |
| 74 | LOSER | 90 | 3.32 |
| 75 | MUSCLE | 88 | 3.25 |
| 76 | ORDERING | 85 | 3.14 |
| 77 | JUMPCUT | 84 | 3.10 |
| 78 | ENHANCEMENT | 83 | 3.06 |
| 79 | DIK | 82 | 3.02 |
| 80 | GIRLFRIEND | 80 | 2.95 |
| 81 | BEDROOM | 80 | 2.95 |
| 82 | HES | 80 | 2.95 |
| 83 | PHARMACEUTICALS | 80 | 2.95 |
| 84 | PRACTICALLY | 78 | 2.88 |
| 85 | DISCRETE | 77 | 2.84 |
| 86 | PHALLUS | 76 | 2.80 |
| 87 | ORGAN | 74 | 2.73 |
| 88 | IIS | 74 | 2.73 |
| 89 | PHENTREMINE | 74 | 2.73 |
| 90 | WEAPON | 74 | 2.73 |
| 91 | LEVITR | 74 | 2.73 |
| 92 | PHARMACEUTICAL | 73 | 2.69 |
| 93 | PFIZER | 73 | 2.69 |
| 94 | GAGGING | 72 | 2.66 |
| 95 | CHOKED | 72 | 2.66 |
| 96 | GODNESS | 72 | 2.66 |
| 97 | TABLET | 71 | 2.62 |
| 98 | REMEDY | 70 | 2.58 |
| 99 | PHENTRIMINE | 70 | 2.58 |
| 100 | SEXUALLY | 70 | 2.58 |

Moving on these lines, we divided the entire presence percentage data of Table I into 5 ranges. The pertinent data is presented in Table II. The third column of Table II depicts the frequency of the values of percentage of presence of the fourth column of Table I.

TABLE II
RANGE OF PERCENTAGE OF PRESENCE OF LEXIS AND FREQUENCY OF LEXIS IN THAT RANGE

| Sr. No. | Range of Percentage of Presence of Lexis | Frequency in the Range |
|---|---|---|
| 1 | 0-24 | 93 |
| 2 | 25-49 | 6 |
| 3 | 50-74 | 0 |
| 4 | 75-99 | 1 |
| 5 | >=100 | 0 |
| Total | 5 | 100 |

The most important interpretation of Table II is that as we move from first record to the fifth record in this table the probability of a given UBE containing advertisement of body enhancement product, increases. This increase is proportional to the occurrence of word(s) from the number of words listed corresponding to the range. For instance, the probability of a given UBE containing such advertisement is more if a word 'PILLS' occurs in it, compared to occurrence of word 'MANSTER' in the UBE. This is so because the word 'PILLS' forms an element of the set with range '25-49' whereas the word 'MANSTER' forms an element of the range set '0-24'.

TABLE III
FREQUENCY AND PERCENTAGE OF PRESENCE OF LEXIS FROM UBE ADVERTISING HAIR-GROWTH MEDICINAL PRODUCTS

| Sr. No. | Lexis | Freq-uency | Percentage of Presence |
|---|---|---|---|
| 1 | MEDICALHAIRRESTORATION | 42 | 1.55 |
| 2 | MEDICALHAIRRESTORATION OFFER | 12 | 0.44 |
| 3 | CLOUDNINEHAIRDESIGN | 10 | 0.37 |
| 4 | HAIRDO | 9 | 0.33 |
| 5 | SUAVEHAIR | 8 | 0.30 |
| 6 | QCHAIRMEN | 4 | 0.15 |

In addition to the snap-shot of top 100 lexis out of the vector containing a total of 8505 lexis, occurring in the body enhancement medicinal UBE, Table III presents the other specially selected lexis from this vector. The analysis of lexis of Table I show that their main area of focus is genital enhancement whereas the lexis of Table III have a focus on the medicinal products focusing on hair-growth. The trivial percentage of presence of lexis of Table III leads us to say that the main area of focus of body enhancement medicinal UBE is advertisement of genital enhancement products.

## V. CONCLUSION

Based on the textual analysis of more than 2700 UBE containing advertisement of body enhancement medicinal products, we conclude that it is possible to identify the lexis

which are occurring in these UBE. We identified such lexis based on the criteria of their frequency of occurrence in the data set under consideration. We also attempted to analyse the identified lexis of UBE of interest. Based on the analysis of such top 100 most frequently occurring lexis in the body enhancement medicinal product advertising UBE, we conclude that the presence of words, 'PENIS', 'VIAGRA', 'MEGADIK', 'PILLS', 'PHARMACY', 'SEXUAL' and 'MEDS' is an indication of a high probability that the given UBE will be containing an advertisement of body enhancement medicinal product. Also, it is concluded that genital enhancement is the main focus area of advertising medicinal products in UBE advertising body enhancement products. We advocate that our results could be put to use for text-based identification of such UBE. It is further concluded that the identification of presence of combination of lexis presented in this paper can be put to use for further research work concerning the UBE containing advertisements of body enhancement medicinal products.

We believe that the best way to fight spam is to understand it. The current paper is an attempt to understand the UBE which claim to provide cheap medicinal products for enhancement of genitals, hair, height, weight, etc. These medicinal products are never delivered and if at all delivered never work as expected. The current work can be extended to implement a naïve anti-UBE fighter for such fake medicinal product announcing UBE.

Our results are best reported on the dataset used. We do not promote or discourage either the use of specific word or of lexis in the designing of body enhancement medicinal product announcing UBE. We just present the identification of lexis which occur most frequently in such UBE. The current work is having a wide range of general applicability to other text domains including the other categories of UBE. On the sidelines of the current study, we advocate that it has also provided an insight into behaviour of spammer' preference for selection of lexis for designing fake body enhancement medicinal product announcing UBE. Finally, we sincerely believe that only awareness and alertness can help protect the general masses against the fake and sometimes lethal-consequences bearing net of greedy persons. Such persons are always looking for victimizing the innocent persons through their luring offers targeting the psychologically vulnerable points like enhancement of genitals.

## REFERENCES

[1] Berry R. "The 100 Most Annoying Things of 2003", January 18, 2004, http://www.retrocrush.buzznet.com/archive2004/annoying2003/

[2] Castillo C., Donato D., Becchetti L. et al. "A Reference Collection for Web Spam", *ACM SIGIR Forum*, December 2006. 40(2). 11-24p. ISSN: 0163-5840

[3] Crucial Web Hosting Ltd. "How Consumers Define Spam", March 06, 2007, http://www.crucialwebost.com/blog/how-consumers-define-spam/

[4] Fette I., Sadeh N. and Tomasic A. "Learning to Detect Phishing Emails", *Institute for Software Research International School of Computer Science (ISRI)*, Carnegie Mellon University (CMU), CMU-ISRI-06-112, June 2006

[5] Frederic E. "Text Mining Applied to Spam Detection", *Presentation given at University of Geneva* on January 24, 2007, http://cui.unige.ch/~ehrler/presentation/ Spam%20Filtering.pdf

[6] Gajewski W. P. "Adaptive Naïve Bayesian Anti-spam Engine", *Proceedings of World Academy of Science, Engineering and Technology (PWASET 2005)*, August 2005. 7. 45-50p. ISSN: 1307-6884

[7] Gyongyi Z. and Garcia-Molina H. "Web Spam Taxonomy", First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb, 2005), Chiba, Japan, April 2005

[8] Infinite Monkeys & Co. "Spam Defined", http://www.monkeys.com/spam-defined/definition.shtml

[9] Kiritchenko S. and Matwin S. "Email Classification with Co-Training", *Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research*, Toronto, Canada, 2001. 8p.

[10] Knujon.com "Categorizing junk eMail", http://www.knujon.com/categories.html

[11] Lambert A. "Analysis of Spam", *Dissertation for Degree of Master of Science in Computer Science*, Department of Computer Science, University of Dublin, Trinity College, September 2003

[12] Mahalo.com "How to stop spam email", http://www.mahalo.com/How_to_Stop_Spam_Email

[13] Martin S., Sewani A., Nelson B., et al. "Analyzing Behavioral Features for Email Classification", *Proceedings of the Second Conference on Email and Anti-Spam (CEAS, 2005)*, Stanford University, California, U.S.A. July 21-22, 2005

[14] Roth W. "Spam? Its All Relative", published online on December 19, 2005, http://www.imediaconnection.com/content/7581.asp

[15] Sebastiani F. "Machine Learning in Automated Text Categorization", in *ACM Computing Surveys*, March 2002. 32(1), 1-47p. ISSN: 0360-0300

[16] Sen P. "Types of Spam", *Interactive Advertising, Fall 2004*, http://ciadvertising.org/sa/fall_04/adv391k/paroma/spam/types_of_spam.htm

[17] The Spam Register "Spam Email Directory: Categorized Spam Emails", December 17, 2008, http://www.spamreg.com/directory.php

[18] Threat Research and Content Engineering (TRACE) "Spam Type Descriptions", http://www.marshal.com/TRACE/Spam_Types.asp

[19] Youn S. and McLeod D. "Spam Email Classification Using an Adaptive Ontology", *Institute of Electrical and Electronics Engineers (IEEE) Journal of Software*, April 2007

[20] Zahren B. "Blizzard of Spam", http://www.pcpitstop.com/news/blizzard.asp

[21] Zhang T. "Predictive Methods for Text Mining", *Machine Learning Summer School - 2006*, Taipei, http://videolectures.net/mlss06tw_zhang_pmtm