International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:1, 2010

Analysis of Classifications of Unsolicited Bulk Emails

Jatinderkumar R. Saini, Apurva A. Desai

***Abstract*—**In recent times, the problem of Unsolicited Bulk Email (UBE) or commonly known as Spam Email, has increased at a tremendous growth rate. We present an analysis of survey based on classifications of UBE in various research works. There are many research instances for classification between spam and non-spam emails but very few research instances are available for classification of spam emails, per se. This paper does not intend to assert some UBE classification to be better than the others nor does it propose any new classification but it bemoans the lack of harmony on number and definition of categories proposed by different researchers. The paper also elaborates on factors like intent of spammer, content of UBE and ambiguity in different categories as proposed in related research works of classifications of UBE.

***Keywords*—**E-mail, Scams, Spam Email, Unsolicited Bulk Email (UBE)

I. Introduction

TECHNICALLY defined, E-mail, short for electronic mail and often abbreviated to email or simply mail, is a store and forward method of composing, sending, receiving and storing messages over electronic communication systems [24]. Since E-mail is fast, cheap and easy to send, it has gained enormous popularity not simply as a means for letting friends and colleagues exchange messages, but also as a medium for conducting electronic commerce. But these same features responsible for growth of email are also accountable for proliferation of a special sort of email called Unsolicited Bulk Email (UBE).

A large part of email traffic consisting of non-personal and non time critical information that should be filtered is called UBE and is synonymously known by various names, including spam email, bulk email, junk email, unimportant email and Unsolicited Commercial Email (UCE). E-mail spam is a subset of spam that involves sending nearly identical messages in bulk, to numerous recipients by e-mail without the consent of the recipients. "Unsolicited" means that the sender lacks affirmative consent from the recipient. "Bulk" means that a substantively similar message is sent to more than 200 addresses a day [10].

The problem of UBE has been increasing at a tremendous rate. SC Magazine in its December – 2008 statistical report on UBE has stated that approx. 200 billion spam messages are being sent per day [13]. For the end of year 2009, it predicted that spam volumes were to rise higher than 95 per cent [14]. UBE, hence, has become an increasing threat to the viability of Internet E-mail and a danger to Internet commerce. In addition to various technical problems like increasing Total Cost of Ownership (TCO), choking network and flooding file servers, it also poses serious non-technical problems like victimization of innocent people in financial scandals. Spoofing, Phishing and other fraudulent messages have become an order of the day, through UBE.

Keeping in view the technical problems posed by UBE and also the societal dimension of the issue, we feel that there is a need to analyze the types of UBE. We argue that such a structured discussion of the subject is important in identifying the instances of spam as well as its easy management. Most of all, it helps in a better understanding of spam-mailing and we believe that the first step towards fighting spam is to understand it. We further believe that by publicizing the types of UBE as well as the analysis of various types of UBE, we will raise the awareness and interest of the research community. A survey of classifications of UBE, as available in the related literature works of various researchers, has been presented by Saini et al. [12]. Moving on this line, this paper presents an analysis of this survey of various classifications of UBE.

II. Statistical Summarization Of UBE Classifications

Most of the research works, in the field of classifications of emails, have focused on classification of emails into spam and non-spam categories. Some researchers have also focused on classification of spam-emails into different categories. But the number of such research instances is quite few. Saini et. al. [12] in their work has attempted to provide an exhaustive list of classifications of spam-emails or UBE into different categories. Here we present the gist and analysis of this survey of classifications of UBE. If some researcher has given different classification categories of UBE, at different locations, then they have been considered as different sets of classifications. The statistical data for this is summarized and presented in Table I.

J. R. Saini is with the Narmada Education and Scientific Research Society's Narmada College of Computer Application, Bharuch, Gujarat, India as Senior Assistant Professor. 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).

TABLE I
STATISTICAL SUMMARY OF NO. OF UBE CATEGORIES

Sr. No.	Statistical Measure	Reference of Research Instance	No. of Categories
1	Maximum no. of UBE categories	spamregister.com [2]	17
2	Minimum no. of UBE categories	a. Ma et al. [8] b. Sahami et al. [11]	2 each

In all there were 36 research works found and analyzed for classifications of UBE into different categories. The number of UBE categories in each research work was counted and then this category-count was plotted against the frequency-of-category-count. This data, in Table II, is presented in tabular format in columns (2) and (3) respectively. Also, for Table II, the total of column (4), which is product of columns (2) and (3), presents the total number of UBE categories found in all research works. This value of 252 UBE categories when divided by value number of research works, i.e. 36, provides us with the average number of UBE categories in each research work as 7.

TABLE II
STATISTICS ON 'FREQUENCY OF CATEGORY-COUNT' OF LITERATURE WORKS

Sr. No.	UBE category-count	Frequency of column (2)	Product of columns (2) and (3)
(1)	(2)	(3)	(4)
1	2	2	4
2	3	2	6
3	4	4	16
4	5	7	35
5	6	3	18
6	7	2	14
7	8	5	40
8	9	5	45
9	10	3	30
10	12	1	12
11	15	1	15
12	17	1	17
Total	98	36	252

The graphical representation of data presented in Table II, for better comprehension, is presented in Fig. 1. The major interpretations for Fig. 1 are as follows:

- Most of the researchers in past have classified spam emails in 5 categories. This value could be derived based on the highest frequency of seven corresponding to the UBE-category-count of 5.
- As the number of categories increases, there is a drastic decrease in the number of researchers who have done so. But we believe that in order to classify the UBE properly, there is a need of more number of categories. This helps in creating a separate class for each type of class of UBE.

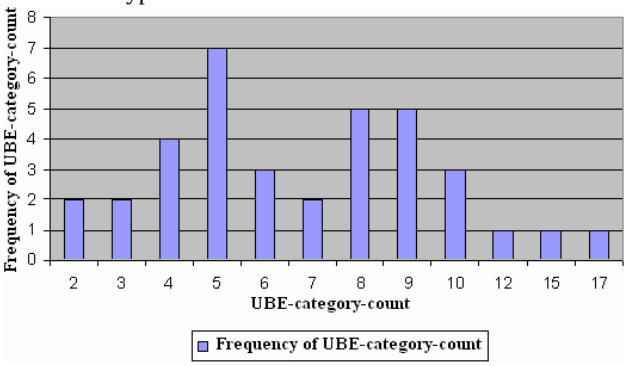

Fig. 1 UBE category-count and its Frequency

III. ANALYSIS OF UBE CLASSIFICATIONS

Previous section listed statistically summarized points for interpretation of Fig. 1 and summary of email classifications found in literature. Additionally, there is another set of important things that come out from analysis of various classifications proposed by work done in the past. These are delineated below.

1) No Solely Devoted Purpose of UBE Classification

The foremost thing to check for UBE classification is whether the researcher has classified the spam emails for the actual purpose of its classification or is this classification a side-product of some other process. This is important because it affects the way in which the emails are viewed and hence the way in which classification is done.

e.g. for many cases of literature study of our work we found that the classification has not been done for the sole purpose of classification of UBE; instead it is a part of discussion or work on some other related topic. The research works of Gajewski [4] and Sahami et. al. [11] are examples of this. There are also instances of research works who have explicitly classified the UBE. McAfee Inc. [9] and Sophos Inc. [18] are its examples.

2) No Hierarchical Classification of UBE

There is no classification of UBE which tries to categorize the UBE in a hierarchical taxonomy.

e.g. advertisement for printer-toner and advertisement for hair-growing medicine are both, ultimately the advertisements. Hence they both can be kept under a common head of Advertisements which in turn is divided into two sub-heads called I.T./Computer and Medicinal.

e.g. in absence of such a scenario an otherwise sub-category may be treated as another category. For instance, Stock Scams and Financial scams treated as separate categories is less suitable than treating Stock Scams as sub-category of Financial scams. Other sub-categories here may be those dealing with Lottery scams, Bank scams, etc.

3) UBE Fighting-Approach Used for UBE Classification

There is difference among the approaches used for fighting the different categories of UBE. So UBE fighting approach should not be used as a criterion for defining UBE categories. For instance anti-virus can identify UBE containing virus but

it will not identify the UBE containing a fake offer of genital enhancement medicines or a UBE from a person asking to transfer a few million United States Dollars (USD) to outside his country. This difference is also used by a few researchers as a criterion for classification of UBE.

e.g. Stephenson [20] in his research work has attempted to protect the enterprise from email-borne threats. He believes that the more threats that are managed appropriately, the better. Typically he sees anti-virus, anti-malware, anti-spam and anti-phishing. So from his perspective, spam emails and emails containing viruses are two separate groups but from our perspective they both are ultimately spam-emails, though anti-spam and anti-virus may work differently.

4) No Common Definition for Given UBE Category

There is no common definition of some category of spam email classification in corporate and research community. It should be standardized so that when classification done by one agency or researcher is communicated with the other one, a common element of concept without any ambivalence or ambiguity is interpreted by them.

e.g. FTC Division of Marketing Practices [3] categorizes Job offers under the category of Education while Threat Research and Content Engineering (TRACE) of Marshal Ltd. [22] categorizes it under the category of Other Offers.

5) Intent of Spammer Not Given Due Consideration

Classification of spam email needs to keep in focus the intent of spammer and not just the content of UBE.

e.g. The UBE for male-genital enhancement medicines contains some pictures of male-genitals for emphasizing the importance of their product. As a result, many classifiers will classify it as Adult or Porn UBE. The classifications proposed by Sen [17] and Kaspersky Labs [6] are examples of this. In the same classification Kaspersky Labs [6] have classified advertisements for weight loss, baldness, skin care, etc. in a separate category called Health and Medicine. It should be pointed out here that the intention of sender is to advertise his Medicinal product and not to promote Pornography; in spite of the fact that the picture is Pornographic.

6) Classification is Biased by Season and Concept Drift

For classification of spam emails the duration of research work should be large enough to get independent of seasonal changes. The short duration research suffers not only from seasonal bias but also from various sporadically occurring events in the world. But at the same time it should not be affected by concept drift [5, 7, 23].

e.g. during the time of elections in United States of America, many spam emails (particularly those containing Virus, Trojan or Porn material) had subject lines talking about the elections

7) UBE Classification Algorithm is not Hybrid

As in the case of various available programs (e.g. open source program called SpamAssassin [19]) for classification of emails, there is a need to develop a hybrid system for classification of spam emails. This will help in reducing the false positives. i.e. instead of designing systems based on a single test or a set of preliminary tests, a combination of tests should be employed for classification of UBE into various categories.

8) UBE Classification should Address Legality of Matter

Classification of spam is important from a legal perspective, because most spam legislation targets a specific category, such as commercial emails, or fraudulent spam [17].

9) UBE Classification should be Consistent

There have been classifications of UBE in past with respect to various factors. It should be borne in mind here that there has to be consistency in these classifications.

e.g. the report of Evett [21] while classifying UBE according to type, states one of the types as Adult while classifying the UBE according to percentage states data about Pornographic UBE. In this case it remains on the reader to assume that the two classifications namely Adult and Pornography refer to the same thing.

10) UBE Classification should be Language-wise

UBE needs to be classified on language-wise basis, i.e. instead of having two categories called Russian spam and Chinese spam along with Porn, Lottery, etc. spam categories as proposed by McAfee Inc. [9]; we need to work on spam in regional language separately, i.e. have Porn, Lottery, etc. UBE categories in UBE classification of each language.

11) Adult Kind of UBE needs Special Treatment

If the definition of adult mail states that it is a mail to be viewed or acted upon by a person above the age of 18 (or whatever age, depending on the legal, cultural or social conditions), than we can completely remove the category of adult mails. This is so because Pornographic mails will go in Porn category and UBE containing advertisements for male or female body-part enhancement medicines will go in category of Medicinal advertisements. In addition to this argument we can say that next-of-kin kind of mails, otherwise, could also be classified as adult emails because we can't expect a minor (as per laws of various countries) to enter a transaction of millions of dollars!

12) UBE Classification should Address All Types of UBE

As is evident from statistics of UBE classifications in literature, only ScamBusters Editors [15] have categorized Lottery as a UBE category. Others like Worcester Polytechnic Institute [25] have not included this as a separate UBE category and implicitly included it as part of some other category like Scams or Financial UBE. Still others have included this category in none of the other categories also. Sophos [18] and Zahren [26] are examples of this. Even for those who have considered Lottery as part of scam or financial spam, rare evidences could be found in support of their categorization. The instance of Lottery is just an example to highlight this point and the same is true for various other categories also.

13) UBE Classification is Not Descriptive

Many researchers have not given any detailed description of the categories they have created. This includes lack of details for type of emails to be classified in that category as well as the definition of each type of category and email type fitted therein. In absence of this type of information given by a researcher, it becomes very difficult to interpret the results of the researcher.

14) Derivation of UBE Categories should be Consistent

It is not suitable to propose spam classification consisting of spam categories derived on different basis.

e.g. The researchers at the Security Software Zone [16] have classified Dictionary Spam along with Virus Spam. These two, according to our opinion, are two different categories of UBE and are derivable on different basis. First category is classification of spam based on the approach used by spammers for spamming. Second category is classification of spam based on the actual contents of UBE. So classification of spam emails based on methods of spamming is different than the classification of spam emails based on the contents of spam emails.

15) UBE Classification is Ambiguous

Anderson et. al. [1] propose Illegal is a category of UBE. In the same work, the researchers have also identified other separate categories for spam emails, like Adult and Financial Data and Services. Actually any spam email dealing with Adult contents can be perceived as Illegal since the spammer does not know the age group, gender, social or cultural restrictions, etc. of the victim or reader. Further any financial transaction which is not reported to the Income Tax Department either in the form of earning or in the form of expenditure, is also illegal (in countries like India). Hence here is very narrow margin for classification of UBE.

16) UBE Classification Suffers from Fuzzy Behaviour

Spam classification is a fuzzy process. When to merge two or more identified categories into a single category or when to break a category into two or more categories, is difficult to be decided.

e.g. Coloncleanse and Viagra both are classified in category called Health & Medicines. But this can be further sub-classified into two different categories called Genital Enhancement/Sexual Enhancement Drugs and Bowel Clearing Drugs. Another way to create sub-categories may be using Allopathic Medicines and Homeopathic Medicines. So the solution here is to break a single category into more categories.

IV. Results And Findings

Through the analysis of the survey of the research literature related to classifications of UBE, we found that though much work has been done on the classification of emails; only a little work has been done on the classifications of UBE. We also found that there is vast difference between the numbers of UBE categories proposed by different researchers and a proper hierarchical classification of UBE has not been worked out. A list of UBE categories, comprising of detailed description of each category, is lacking. The approaches used by researchers for classification of UBE are different and there is no common definition of any kind of UBE category. Intent of spammer, content of UBE, language of UBE, algorithm used for UBE classification, duration of research, concept-drift and purpose of UBE classification are various factors based on which an analysis of the UBE classifications proposed in related research literature could be carried out. We presented a detailed analysis, for finding based on each of these factors, obtained by us from the existing UBE classifications in the preceding section of this paper.

V. Conclusions

The analysis of survey of various works related to classifications of UBE is presented here. We conclude that a lot of work needs to be done on UBE classification. The duration of research for UBE classification should be independent of seasonal-bias as well as concept-drift. UBE classification is a fuzzy process and there is very less differentiation among groups of UBE categories like Adult, Illegal, Medicinal, Porn, Dating, Romance, Matrimonial and others. Understanding spam of any kind is the first step towards fighting it and the best way to understand spam is to categorize it. We conclude that a hierarchical structure needs to be developed for a proper classification of UBE. The algorithm used for UBE classification should be hybrid, adaptive and scalable enough to accommodate various kinds of spam-emails roaming around. We conclude that an exhaustive list of UBE categories with a comprehensive description of each category is necessary for understanding UBE and fighting it.

References

[1] Anderson D. S., Fleizach C., Savage S. and Voelker G. M. “Spamscatter: Characterizing Internet Scam Hosting Infrastructure”, in *Proceedings of 16th USENIX Security Symposium*, Boston MA, Article 10, 2007. ISBN: 111-333-5555-77-9

[2] Evett D. “Spam Statistics 2006”, TopTenREVIEWS Inc. Available: http://spam-filter-review.toptenreviews.com/ spam-statistics.html

[3] Federal Trade Commission. “False Claims in Spam”, A report by the United States FTC Division of Marketing Practices, April 30, 2003

[4] Gajewski W. P. “Adaptive Naïve Bayesian Anti-spam Engine”, in *Proceedings of World Academy of Science, Engineering and Technology (PWASET 2005)*, Pages 45-50 Volume 7 August 2005 ISSN 1307-6884

[5] Hsiao W. F. and Chang T. M. (2008). “An incremental cluster-based approach to spam filtering”, in *International Journal of Expert Systems with Applications*, 34(3), April 2008, pp. 1599-1608

[6] Kaspersky Labs. “Types of Spam”, Available: http://www.viruslist.com/en/spam/info?chapter=153350533

[7] Klinkenberg R. and Joachims T. “Detecting concept drift with support vector machines”, in *Proceedings of 17th International Conference on Machine Learning (ICML-00)*, Stanford, CA, 2000, pp. 487–494

[8] Ma W., Tran D. and Sharma D. “Filtering Spam Email with Flexible Preprocessors”, *Advances in Communication Systems and Electrical Engineering, Lecture Notes in Electrical Engineering*, Volume 4, 2008, pp. 211-227. ISBN 978-0-387-74937-2

[9] McAfee Inc. “Current Spam Categories”. Available: http://www.mcafee.com/us/threat_center/ anti_spam/spam_categories.html

[10] Nelson R. “The Definition of Email Spam”. Available: http://www.crynwr.com/spam/definition.html

[11] Sahami M., Dumais S., Heckerman D. and Horvitz E. “A Bayesian Approach to Filtering Junk E-mail”, *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin: AAAI Workshop, TR AAAI WS-98-05, pp. 550-562

[12] Saini J. R., Desai A. A. “A Survey of Classifications of Unsolicited Bulk Emails”, in *National Journal of Computer Science and Technology (NJCST)*, 2010. ISSN 0975-2463 [to be published]

[13] SC Staff 1. “200 billion spam messages sent daily as spammers change tactics for 2009”. Available: http://www.scmagazineuk.com/200-billion-spam-messages-sent-daily-as-spammers-change-tactics-for-2009/emailArticle/122908/

[14] SC Staff 2. “Volumes of spam predicted to rise in 2009”. Available: http://www.scmagazineuk.com/Volumes-of-spam-predicted-to-rise-in-2009/emailArticle /122996/

[15] ScamBusters Editors. “Email Scam Analysis”, *Scamdex, Scambusters* [Online], Issue No. 292. Available: http://www.scamdex.com/MHON/E/msg08805.php

[16] Security Software Zone Inc. “Types of Spam”, August 2006. Available: http://www.securitysoftwarezone.com/reviews-spam-blocker-4.html
[17] Sen P. “Types of Spam”, *Interactive Advertising*, Fall 2004. Available: http://ciadvertising.org/sa/fall_04/adv391k/paroma/ spam/types_of_spam.htm
[18] Sophos Inc. “Sophos identifies the most prevalent spam categories of 2005”, August 3, 2005. Available: http://www.sophos.com/pressoffice/news/articles/2005/08/pr_uk_20050 803topfive-cats.html
[19] SpamAssassin. “SpamAssassin” Available: http://spamassassin.org
[20] Stephenson P. “Email Content Management”, in *SC Magazine for IT Security Professionals*, UK June 02, 2008. Available: http://www.scmagazineuk.com/Email-content-management-2008/GroupTest/129/
[21] The Spam Register. “Spam Email Directory: Categorized Spam Emails”. Available: http://www.spamreg.com/ directory.php
[22] Threat Research and Content Engineering (TRACE). “Spam Type Descriptions”, TRACE Blog. Available: http://www.marshal.com/TRACE/Spam_Types.asp
[23] Vipul. “The Fallacy of Corpus Anti-spam Evaluation”, *Balance Through Extremism, Archive for the Antispam category*, June 20, 2008. Available: http://blog.vipul.net/2008/06/20/the-fallacy-of-corpus-anti-spam-evalutation/
[24] Wikipedia, the free encyclopedia. “E-mail”, Wikimedia Foundation Inc. Available: http://en.wikipedia.org/wiki/Email
[25] Worcester Polytechnic Institute. “Junk (SPAM) Email Frequently Asked Questions (FAQ), Worcester Polytechnic Institute, Worcester, January 31, 2006. Available: http://www.wpi.edu/Academics/CCC/Help/Email/spamfaq.html
[26] Zahren B. “Blizzard of Spam”. Available: http://www.pcpitstop.com/news/blizzard.asp