# MECHANISM OF CLASSIFICATION OF TEXT SPAM MESSAGES COLLECTED IN SPAM PATTERN BASES

**Saadat Nazirova**

Institute of Information Technology of ANAS, Baku, Azerbaijan
*saadatn@mail.ru*

**Introduction.** The final set of words, figures and symbols combined with lexical, grammatical, semantic, frequency relations and forming the informative message is a text message. E-mail spam, known as unsolicited bulk Email (UBE), junk mail, or unsolicited commercial email (UCE), is the practice of sending unwanted e-mail messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients.

There are some approaches to information representation in databases for maintenance of the subsequent analysis of this information. We will consider the most popular approaches to representation of the text information dynamically arriving in databases of information systems. The first approach is based on the theory of the sets, the second on vector algebra, and the third on probability theory. All these approaches are based on the assumption, that document content, its basic maintenance is defined by set of keywords - terms and concepts which are in it. Such representation of text messages is called as not structured model «bag of words». Each text message consider as a bag of words, and each word – an independent random variable out of a context and communication with in other words in text [1]. In real systems this simplification is overcome, for example, expanded boolean model considers contextual affinity (operators near, adj). Later in Salton's works [3,6] it has been offered the vector model as alternative to free indexing of the lexical context. In the elementary case the vector model assumes comparison to each document of a frequency spectrum of words and accordingly a vector in lexical space [2]. In the given article as such text messages are considered, unwanted e-mail messages (spam).

**Representation model of text email spam**. *In vector model* any text email describe in form of points in N-dimensional space, where N number of different terms in the set of spam patterns $S = \{s_1,...,s_M\}$ [3]:

$$S_i = \begin{bmatrix} w_{1j} \\ \vdots \\ w_{ij} \\ \vdots \\ w_{Mj} \end{bmatrix}, \qquad (1)$$

where $w_{ij}$ is a weight of term $j$ in spam message $i$ ($i = 1,...,M$), $M$ is a number of spam messages in sample, $j = 1,...,N$. The main advantage of vector model is the possibility of ranking of messages on similarity that is their proximity in the vector space.

Sample of text spam messages can be represented *in the form of a matrix*:

$$S = \begin{bmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1N} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{iN} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{M1} & \cdots & w_{Mj} & \cdots & w_{MN} \end{bmatrix}. \qquad (2)$$

Such a matrix can be called a "message-term" matrix as its rows are spam messages, and columns are terms.

**Classification of text spam messages.** Classification is a reference of each spam pattern to a certain class with known characteristics obtained during system learning. The number of classes in the classification is advanced known. Clustering spam patterns collected in the spam pattern bases means automatically grouping of thematically similar spam messages. In the case of information flows, this task is complicated by the need to carry out this process in real time. There are some complications associated with multiple-choice algorithms of clustering of spam patterns. Different methodologies use different algorithms for proximity of email in the presence of a large number of features. Once the classes are determined by the clustering method, there is a necessity of their support as the spam constantly varies. In this case, classification comes to help. Thus the classification mechanism at first studies on the basis of identifing of spam patterns' characteristics corresponding to certain themes. At this stage training of the classification mechanism is carried out - correlations between separate characteristics are defined, and the system is capable to classify new messages. Classification and clustering represent two extremes concerning human participation in the process of grouping spam patterns. The classification mechanism is trained on selected spam patterns only after the training of automatic detection of classes (clusters) comes to an end. The number of clusters can be arbitrary or fixed. If the classification allows attribution of spam emails with certain known features, then clustering is a more complex process, which allows not only to attribute spam emails of some peculiarities, but also the identification of these very traits - the classes. However, in contrast to classification, thematic orientation of these groups not known in advance. The purpose of clustering of some spam pattern set is to separate these subsets (clusters) that all emails falling into one cluster, is similar to each other. Cluster can be considered as a group of patterns with similar characteristics. The aim of all clustering methods is similarity of patterns falling into the cluster should be maximum and semantically justified.

Numerical clustering techniques are based on these definitions of the cluster as a set of patterns, the importance of semantic proximity between any two elements of which not less than a certain threshold or the value of proximity between any pattern of set and the centroid is also not less than a certain threshold. In this case centroid of the cluster is the vector calculated, for example, as an arithmetic average of vectors of all spam patterns in the cluster. Non-numerical semantic methods, corresponds to the cosine of the angle between the vectors – images of spam messages $s_1$ and $s_2$. Obviously, $sim(s_1, s_2)$ belongs to a range $[0,1]$. The larger $sim(s_1, s_2)$ means the messages $s_1$ and $s_2$ more close to each other. For any pattern $s_i$ we have $sim(s_i, s_i) = 1$ [2].

Above methods of text document representation have common flaw related to high dimensions. Due to this, the task of reducing the dimensions of terms spaces used for classification of spam messages becomes important. In more advanced vector models the space dimension is reduced by discarding the most common or uncommon terms, thus increasing the percentage of core importance of terms.

The procedure of dimension reducing consists of selection from initial $T$ terms of the most informative $N$ terms, possessing the best dividing properties. For revealing informative terms in classification of text documents there are effectively used some theoretical approaches:

- term weighting;
- transition to new system of signs (the factorial and componential analysis);
- the statistical approach ( $\chi^2$ - statistics);
- the theoretical-informative approach;
- feedback with user;
- application of genetic algorithms [4,5].

**Term weighting methods.** The given method is based on the assumption that the semantic component of any message can be represented in the form of set of terms meeting different frequency in the text. Thus

- the more often a term occurs in the email, the more it reflects theme of the email;
- the more often a term occurs in the whole sample of emails, the smaller secretary (discriminating) ability it has.

Thus, for classification carrying out it is desirable to select mid-frequency terms which is better describe the message of the set subjects.

Let $f_{ij}$ is a frequency of term $j$ in a spam message $i$, $N$ is a number of terms in the sample after removing of syntactic words and allocation of a root of a word, $M_j$ is a total number of the messages containing term $j$. Further the most common methods of determining of weight of a term $w_{ij}$ are considered.

1. *Logical weighting.* The simplest approach is to assign a weight value "1" for the term $j$ if it occurs in the message, and "0" otherwise

$$w_{ij} = \begin{cases} 1, f_{ji} > 0; \\ 0, f_{ij} = 0. \end{cases} \tag{3}$$

The main advantage of this method is the simplicity of implementation, and lack of is ignoring frequency of occurrence of a term in different messages.

2. *Weight-frequency of term.* Another simple method for determining the weight $w_{ij}$ is to calculate the frequency of occurrence of term $j$ (Term frequency) in the message $i$:

$$w_{ij} = f_{ij}. \tag{4}$$

Using term frequency gives about 25% increase in the efficiency of the classification than the logical weighting (3).

3. *Tf-idf - weighting (term frequencies – inverse document frequencies).* The previous two methods do not consider the term frequency in all messages of sample, its discriminating ability. For elimination of it, it is offered to use so-called tf-idf - weighting [3,6] which appropriates weight to a term $j$ in the message $i$ proportionally to the number of messages in sample into which the term at least, once enters:

$$w_{ij} = f_{ij} \log\left(\frac{M}{M_j}\right). \tag{5}$$

4. *tfc - weighting*. In tf-idf – weighting the fact is not taken into consideration that the messages can be various length therefore weight of terms in "short" and "long" messages can essentially be differ. In tfc-weighting the formula (5) is modified by carrying out of normalization of lengths of messages [7]:

$$w_{ij} = \frac{f_{ij} \log\left(\frac{M}{M_j}\right)}{\sqrt{\sum_{j=1}^{N}\left[f_{ij} \log\left(\frac{M}{M_j}\right)\right]^2}}. \tag{6}$$

The summation in denominator of fraction is taken over all terms of message $i$.

5. ltc- weighting. The given approach The given approach consists in use of the logarithm of term frequency instead of $f_{ij}$. It allows to essentially reduce disorder in

frequencies of different terms which is the majority of text spam messages. The formula ltc-weighting is in the form [7]:

$$w_{ij} = \frac{\log(f_{ij}+1)\log\left(\frac{M}{M_j}\right)}{\sqrt{\sum_{j=1}^{N}\left[\log(f_{ij}+1)\log\left(\frac{M}{M_j}\right)\right]^2}}. \quad (7)$$

6. atc - weighting. In such weighting the weight will change from 0,5 to 1 that in some cases leads to improvement of quality of classification, allowing to consider significant terms, which rarely occur in concrete sample [6]:

$$w_{ij} = \frac{\left(0{,}5+0{,}5\frac{f_{ij}}{\max\limits_i f_{ij}}\right)\log\left(\frac{M}{M_j}\right)}{\sqrt{\sum_{j=1}^{N}\left[\left(0{,}5+0{,}5\frac{f_{ij}}{\max\limits_i f_{ij}}\right)\log\left(\frac{M}{M_j}\right)\right]^2}}. \quad (8)$$

here $\max\limits_i f_{ij}$ is the frequency of most accruing term in message $i$.

**Conclusion.** Using above methods we can get different results in classification of text spam messages. Continuously analyzing the spam-patterns collected in bases of an anti-spam system [8], with the methods of Text Mining it is possible to receive an information portrait of each class of spam messages. Information portrait can represent such characteristics of spam messages, as:

- theme;
- language;
- size of messages (small, average, big);
- source country;
- name of source;
- keywords for the given sample of spam messages.

Classifying and parametring spam patterns will also establish a thematic relation the geographical (e.g., what topics are prevalent in spam messages sent from certain countries). Thus, the system will be capable to reveal purposeful information attacks if these things occur. It is possible to determine the organized spam groupings through analyzing the sources of spam messages from spam pattern base [8].

## References

1. S. Scott, S. Matwin, Text classification using wordnet hypernums // Proceedings of the ACL Workshop: Usage of WordNet in natural language processing systems. 1998, pp. 45-51. http://acl.ldc.upenn.edu/W/W98/W98-0706.pdf
2. D.V. Lande, The fundamentals of integration of information flow, Kiev, 2006, 237 p (in Russian)
3. G. Salton, Dynamical library-information system. Moscow, 1979. 557 p. (in russian)
4. D.R. Tauritz, J.N. Kok, I.G. Sprinkhuizen-Kuyper, Adaptive information filtering using evolutionary computation // Information Sciences. 2000. N 2-4, pp. 121-140. http://portal.acm.org/citation.cfm?id=338012
5. R. Alguliyev, R. Aliguliyev, Effective summarization method of text documents/ Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2005, pp. 264 – 271.

http://portal.acm.org/citation.cfm?id=1092481
6. G. Salton, C. Buckley, Term Weighting Approaches in Automatic Text Retrieval // Information Processing and management, 1988. Vol. 24. N 5, pp. 513-523 http://portal.acm.org/citation.cfm?id=866292
7. K. Aas, L. Eikvil, Text Categorization: A Survey // Technical report. N 941. Norwegian Computer Center. Oslo. June 1999, pp. 11-21.
8. R.M. Alguliyev, S.A. Nazirova, Mechanism of forming and realization of anti-spam policy // Telecommunications, 2009. №12, pp. 38-43 (in Russian)