Detection of spam using email signatures

E V Sharapova[1] and R V Sharapov[1]

[1]Vladimir State University, Orlovskaya street, 23, Murom, Russia, 602264

e-mail: info@vanta.ru

Abstract. Currently, unwanted emails are actively sent to the Internet. Millions copies of e-mails are sent simultaneously to various users. Often e-mails undergo minor modifications to complicate the detection of spam. The paper proposes options for determining the signature of e-mails that allow identify letters with the same content and structure. Content signature of the letter includes the basic phrases in the text of the e-mail with the exception of names, numeric codes, suspicious words that are not included in the dictionary. Structure signatures incorporate the same type of e-mails, such as paragraphs, tables, images. The paper shows the results of using signatures to detect e-mail spam.

1. Introduction

E-mail is one of the most popular services in the Internet. The ability to quickly communicate using electronic messages made e-mail used by billions people. However, users are faced with such a negative phenomenon as receiving unwanted e-mails. Currently, unwanted emails are actively sent to the Internet. These messages contain advertising of various goods and services, political advertising, are used for phishing and the spread of viruses. According to the Kaspersky Lab, at the beginning of 2019 the share of spam in e-mail traffic in Russia amounted to 54%. In other words, more than half of e-mail messages are spam.

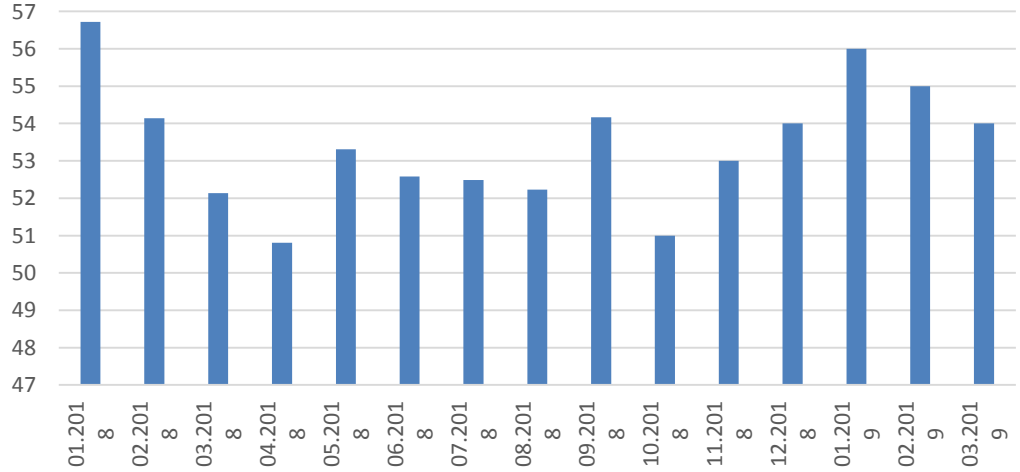


Figure 1. The share of spam in e-mail traffic in Russia.

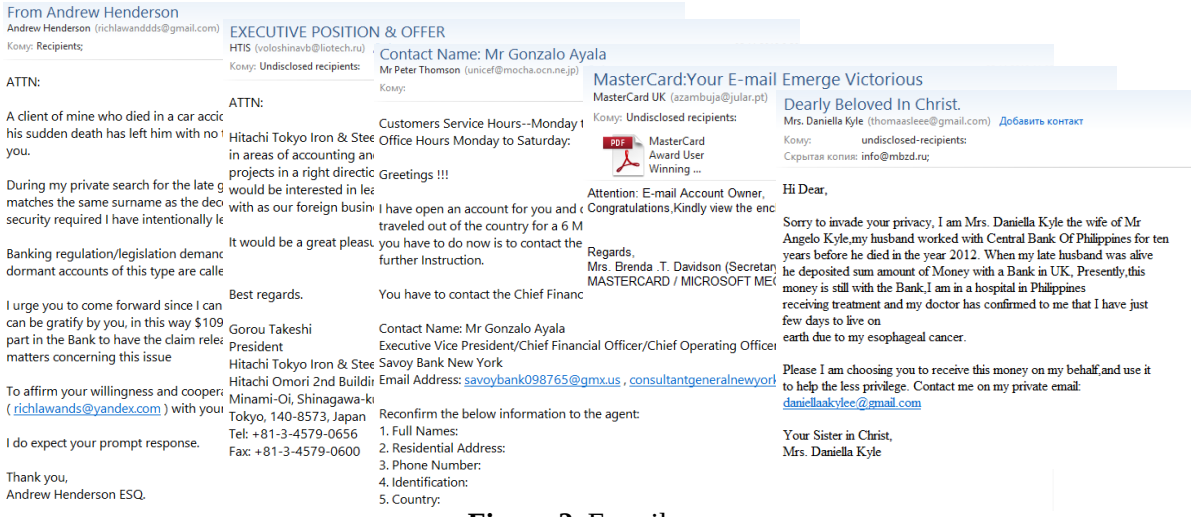


Figure 2. E-mail spam.

Spam is unsolicited mass mailing anonymous e-mail. Millions of e-mails copies are simultaneously sent to different users. Often copies differ from each other with a greeting (for example, an automatic indication of the sender's name from the dictionary - Leonty Lyudvigovich, Yadviga Svyatoslavovna) or a chain of characters (for example, 1c3790b4b8ad11e8aa21e41d2d101530).
The share of Russia in e-mail spam traffic is about 6% in 2019. Greater volume of spam is sent only in China (15%) and the USA (12%).

The uniqueness of messages is provided automatically, that is, random sequences of characters, greetings, etc. [1]. Thus, such messages can be considered as fuzzy duplicates [2], the detection of them is not a trivial task.

2. Analysis of the problem

Breaking e-mail spam has been known for a long time. More than 20 years, people are trying to stop receiving unwanted e-mail messages. This struggle is accompanied by varying success. E-mail filters are constantly improving. But to get around them, spammers come up with new ways.

Let's look at the main ways to combat a spam. Large mail services and information security companies use distributed anti-spam methods [3]. Companies collect information about the mail traffic passing through them and exchange this data between themselves. In this way, they get a full picture of the actions of spammers and can develop and select effective anti-spam defenses [4].

Another group of anti-spam methods is local. It does not use a data from external services, but works only with received messages. Local methods are used by both mail servers and final recipients. Often they are used to filter mail organizations [5, 6, 7].

Authentication of the sender and analysis of e-mail headers is carried out to spam detect. To do this, check information about the sending host, its IP address, server response codes, etc., are analyzed [8].

Often, trap addresses are used for checking - mailboxes intended solely for receiving mail spam and not used in normal life. Machine learning methods are successfully used in the fight against a spam. So, methods Bayesian filtering [9], decision trees [10], support vector machine [11], rule-based methods [12], etc., became popular.

Many works are devoted to the extraction and subsequent analysis of the distinctive properties and characteristics of e-mail items [13, 14, 15]. Various characteristics of messages are considered: visual, structural, system. In [16] it is proposed to use the dynamic property space of e-mail messages.

A number of papers related to the analysis of the text content of the e-mail [17, 18]. In [19] text information placed in images is analyzed. In [20] it is proposed to use social networks to combat with spam.

One of the ways to combat with spam is based on the use of various signatures. The way is based on counting e-mail checksums using various methods to detect duplicate e-mails.

Signatures are widely used in various tasks. They are widely used in information retrieval, in image processing. In [21] discusses methods for detect similar texts (fizzy duplicates).

Hash signature is easiest way to compare two messages. For this, the e-mail checksum is calculated using the MD5 or CRC32 algorithm. TF signature is based on counting the frequency of occurrence of words in a TF document. The signature is based on several most frequent words. The signature used is a CRC32 string checksum consisting of selected words arranged alphabetically. TF*IDF signature involves counting the weight of words not using the TF formula, but using the TF*IDF equation [22]. In this case, not only the word frequency in the document is taken into account, but also the total occurrence of words in all documents in the collection.

TF*RIDF signature based on the combination of the word frequency TF and the residual inverse frequency of the RIDF documents [23]. TF*IDF Optimal signature is a modified version of the TF*IDF signature. The modification consists in changing the principle of calculating the IDF value based on the so-called “optimal frequency”.

Long string signature built on the basis of the two longest sentences, makes it possible to find similar documents quite well. For this, the text contains the two longest sentences and concatenates into one line in alphabetical order. For the string, the control code CRC32 is calculated, which is the signature. Heavy string signature is based on a similar principle. Two sentences are selected from the text. However, sentences are selected on the basis of the sum of weights (calculated using the TF*IDF equation) of its words. The two sentences with the largest sum of weights are ordered alphabetically, concatenated into one line, for which the control code CRC32 is calculated.

I-Match signature is based on the calculation of the value of the I-Match function proposed in [24, 25]. A dictionary of words with an average IDF is compiled for the entire collection of documents (words with too large or small IDF values are not included in the list). For each document, a set of words is formed and its intersection with the dictionary is determined. When crossing some threshold, the hash function SHA1 (I-Match signature) is calculated for the set of words.

To calculate the Super Shingles signature for the entire set of shingles of the document, 84 different hash functions are calculated. Further, according to the criterion of the maximum or minimum of each function, 84 shingle are selected, which are divided into 6 groups, for each of which 6 super shingles are built [26].

MegaShingles signature is similar to no previous. 84 shingles are calculated. They are divided into 6 groups, for each of which 6 super shingles are built. The signature consists of 15 numbers (megashings), representing all possible pair combinations of the 6 super singles.

The signatures of Rabin [22] allow counting fuzzy checksums of letters. The signature of Winnowing [28] is local algorithms for document fingerprinting. The signature ensures that if there is at least one sufficiently long common substring in two files, then at least one label in their sets will match.

The Nilsimsa signatures [29] present the e-mail message by locality-sensitive hash. A Nilsimsa code is something like a hash, but unlike hashes, a small change in the message results in a small change in the Nilsimsa code.

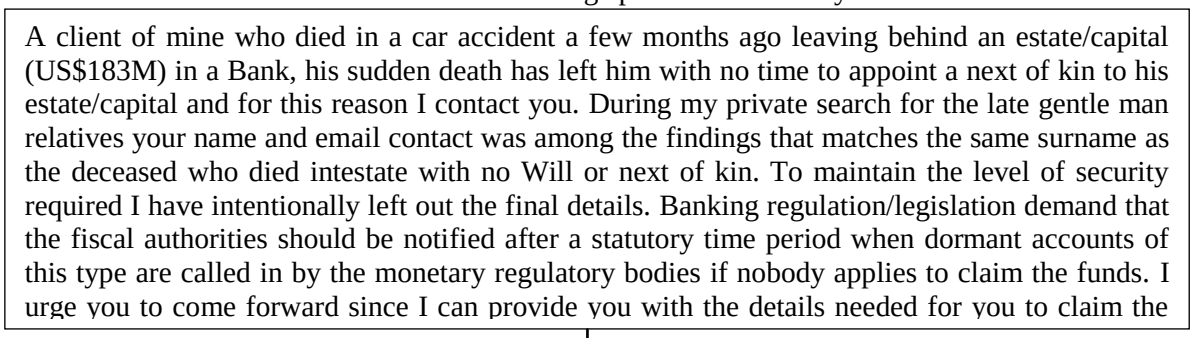
However, improved spamming techniques make existing signatures ineffective. Thus, it becomes necessary to modify the structure of signatures to more effectively detect duplicate letters.

3. Signatures of content and structure

Different e-mail signatures can be used to identify messages with the same content and structure. The signature of the contents of the letter SigData includes the main phrases in the text of the e-mail, with the exception of names, numeric codes, suspicious words that are not included in the dictionary. The difficulty lies in the degree of filtration content. With a weak filtering in the text may remain elements used to uniquely test the letter. With strong filtering (for example, taking into account only nouns or the most frequent words), different letters may be mistakenly recognized as identical.

According to the results of the experiments, it was decided to normalize the text and include in the signature word forms obtained after processing the AOT package by the LEMMATIZER module. At the same time, a package of candidate words for inclusion in the signature was programmatically generated from an e-mail and lemmatization was performed for each word using the AOT API

functions. In the absence of a candidate word in the dictionary, it was not included in the signature. The Russian Morphological Dictionary of A.A. Zaliznyak was used as a dictionary, including 161 thousand lemmas. Thus, it is possible to identify messages that have passed through the uniqueness (that is, fuzzy duplicates of letters). The signature of the message content SigData (see Fig. 3) is a hash code calculated for the text of the electronic message processed above by the indicated method.



SigData: 146bffd75c4c0a40cfb92df1a78395e0
Figure 3. Signature of content.

Massively distributed messages may have minor differences in content, but they do not differ in the design and arrangement of text elements. In other words, the structure of such messages is the same.

The signature of the structure SigStr includes structural elements of an e-mail type, such as paragraphs, tables, images. In this case, the content of the message is not taken into account. For the structure thus obtained, a hash code is calculated (see Fig. 4). Md5 algorithm is used to calculate the hash codes. Messages with the same internal structure will have the same hash codes.

SigStr: d1b37003288e83c5fdf5e34f0af0a252
Figure 4. Signature of the structure.

It should be noted that the signature of the structure may not always be applied. Many messages are plain text. Accordingly, the structure will not contain any markup tags. Similarly, some messages are very short and contain only a few tags, for example, new line breaks. The use of such structures as signatures will lead to incorrect accounting of various messages as identical. For this reason, messages that have more than 100 characters of markup tags are used to calculate the structure signature. For other messages, the value of the structure signature is taken equal to 0 and is not taken into account in comparison.

It should be noted that signatures for e-mail messages are calculated once. Further verification is carried out according to calculated signatures.

Although the structure and content signatures are similar to the receipt of the checksum of the message, there is a significant difference. Upon receipt of the checksum, the entire content is taken into account and even minor changes lead to different values of the checksum. The division of the

message into structure and content allows to take into account the individual characteristics of the messages, as well as to find mass mailings with varying content.

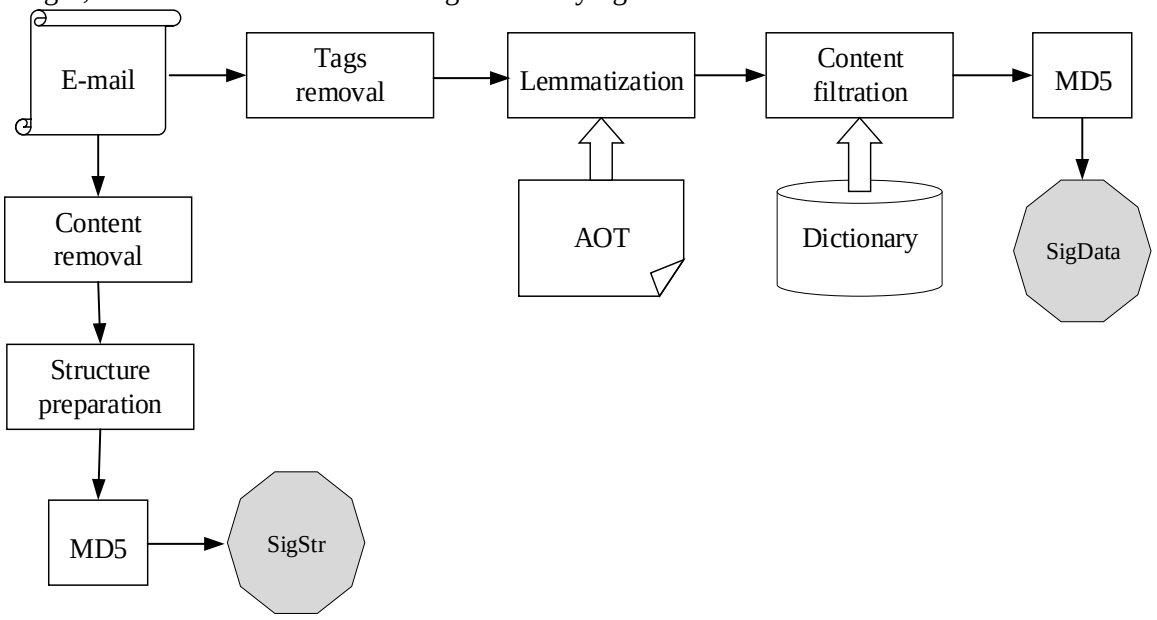


Figure 5. Calculation of signatures.

4. Using signatures to detect e-mail spam

The proposed signatures were used to detect e-mail spam arriving at the e-mail addresses of the Murom Institute of Vladimir State University mivlgu.ru and Internet resource addresses located on the commercial hosting Majordomo.ru (with the spam filter disabled). Mail messages that come to the addresses of popular mail services (gmail.com, yandex.ru, mail.ru, etc.) successfully pass spam filtering and cannot be used as a source of data for research.

A total of 30,000 e-mails were manually selected, which are e-mail spam. It should be noted that more than half of the messages (18638) were represented by several copies. The task was to detect such letters - letters that are fuzzy copies of other documents. In addition, 30,000 e-mails from real senders (that is, non-spam) were added to the message base.

At the beginning, an attempt was made to compare letters by body - content with the exception of a system header containing the sender, recipient, mail server address and other system information. Hash codes were calculated for each mail message. Messages with the same hash codes were recognized as duplicates. The number of identical messages turned out to be small - only 130 letters. The remaining letters have differences in structure and content.

When using the SigData content signature, 12237 similar messages were detected. In addition, due to the characteristics of content filtering when counting signatures (deleting non-informative elements) 42 messages were mistakenly counted as copies of other messages.

When using the signature of the SigStr structure, 14226 similar messages were detected. Due to the use of similar templates in the formation of e-mail messages, as well as messages in the form of unformatted text, 844 messages were mistakenly counted as copies of other messages.

When using the bundle of signatures content-structure SigData + SigStr, 15244 similar messages were found and 886 messages were mistakenly counted as copies of other messages.

The next metrics were used to assess the quality of work:

- Recall:

$$Recall = \frac{\text{Number of spam e-mails marked as spam}}{\text{Total number of spam e-mails}}$$

- Precision:

$$Precision = \frac{\text{Number of spam e-mails marked as spam}}{\text{Number of e-mails marked as spam}}$$

- F-measure:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The results of using signatures to detect spam are shown in table 1.

Table 1. Results of using signatures to detect spam.

Signature	Recall	Precision	Number of errors	F-measure
Content	0.007	1	0.993	0.014
SigData	0.656	0.996	0.332	0.791
SigStr	0.763	0.944	0.311	0.844
SigData+ SigStr	0.818	0.945	0.260	0.877

As can see, the greatest completeness is 0.818 and the smallest number of errors is 0.260 when using the content-structure signature bundle. The highest accuracy rates are achieved with a full comparison of the content of the letters, but fuzzy duplicates are not determined.

As a practical implementation, it was proposed to use the SigData and SigStr signatures in the spam filter of the mail server of the authors managed service hosted on a commercial hosting. For these purposes, signature counting scripts were implemented and new rules were added to the spam filter. The analysis showed that spam letters of the same content come to different recipients of the server with a frequency of several fractions of a second for several days. In addition, many mailings are repeated at intervals of several weeks to several months. For this reason, it was decided to store the signatures of each letter for three months and use them to decide on the spam membership of the newly received letters. It should be noted that the letters are marked as spam by the filter if at least one of the SigData and SigStr signatures match.

The results of practical use (see Fig. 6) showed the viability of the proposed method of combating postal spam. SigData and SigStr signatures began to be used from October (in September, another spam filter was used). As information accumulated and the system was adapted, it was possible to significantly reduce the number of not detected spam messages (from 42% in October 2018 to 18% in January 2019).

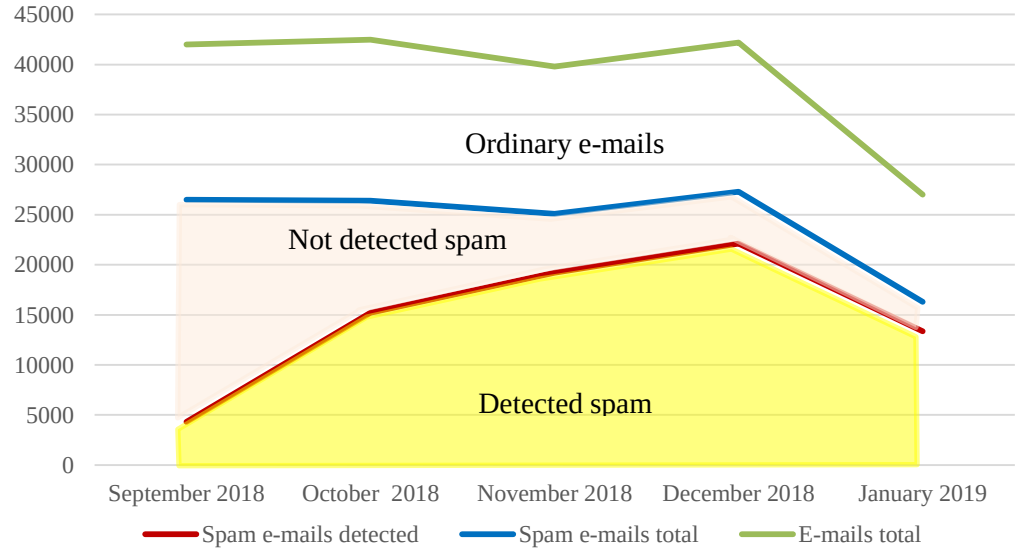


Figure 6. E-mail server spam filtering results.

During testing it was found that the fullness of spam detection increases with the number of pending mailboxes. When considering only the one mailbox address, number of detected spam is low,

because not often identical letters come several times. However, when considering the tens of mailboxes, the number of detected spam increases strongly. For this reason, the use of the proposed signature is justified when considering a group of mailboxes, such as a mail server.

5. Conclusion

The proposed content signatures and structures can be used to detect mass spam mailings, even if mailing is unique. Signatures can be used both individually and in pairs with each other. In the latter case, the best result is achieved in terms of completeness and the smallest number of errors.

To improve the quality of spam filtering, signatures can be used in conjunction with other methods for determining unwanted messages. The proposed signatures can also serve as separate message properties used as components in the application of machine learning methods.

6. References

[1] Lyapicheva N G 2018 Anti-spam issues: impact of cloud technologies *Bulletin of the Central Economics and mathematics Institute RAS* **1**
[2] Sharapov R and Sharapova E 2018 The problem of fuzzy duplicate detection of large texts *CEUR Workshop Proc.* **2212** 270-277
[3] Kovalev S S and Shishaev S S Modern methods of protection against unwanted mailings *Proc. of the Kola Scientific Center of the Russian Academy of Sciences* **7** 100-111
[4] Terentjev A M 2013 The corporate version of the implementation of Doctor Web antivirus packages in scientific institutions: implementation *National interests. Priorities and safety* **19** 40-45
[5] Baranchikova E A 2009 A method for filtering e-mail messages *Bulletin RGRTU* **2** 56-60
[6] Mironenko A N and Belim S B 2011 Multi-level spam filtering system *Information systems and technologies* **3** 125-128
[7] Mironenko A N and Belim S B 2011 Model filtering spam in email traffic *Bulletin of Computer and Information Technologies* **11** 34-36
[8] Subramaniam T, Jalab H A and Taqa A Y 2010 Overview of textual anti-spam filtering techniques *Int. J. Phys. Sci.* **5** 1869-1882
[9] Metsis V, Androutsopoulos I and Paliouras G 2006 Spam Filtering with Naive Bayes - Which Naive Bayes? *Proc.of 3 Conference on Email and Anti-Spam CEAS*
[10] Carreras X and Márquez L 2001 Boosting trees for anti-spam email filtering *Proc. of 4 international conference on recent advances in natural language processing* 1-8
[11] Drucker H, Wu D and Vapnik V 1999 Support vector machines for spam categorization *IEEE Transactions on Neural Networks* **10** 1048-1054
[12] Cohen W 1996 Learning rules that classify e-mail *Proc. of the AAAI spring symposium on machine learning in information access* 18-25
[13] Lee S M, Kim D S and Park J H 2010 Spam detection using feature selection and parameters optimization *Proc. of International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)* 883-888
[14] Wu C T and Cheng K T 2005 Using visual features for anti-spam filtering *Proc. of IEEE International Conference on Image Processing* 509-512
[15] Beiranvand A and Shadgar B 2012 Spam Filtering By Using a Compound Method of Feature Selection *Journal of Academic and Applied Studies* **2** 25-31
[16] Zhou Y, Mulekar M S and Nerellapalli P 2005 Adaptive spam filtering using dynamic feature space *Proc. of 17th IEEE international conference on tools with artificial intelligence* 302-309
[17] Sasaki M and Shinnou H 2005 Spam detection using text clustering *Proc. of international conference on cyberworlds* 316-319
[18] Chirita P A, Diederich J and Nejdl W 2005 Mailrank:using ranking for spam detection *Proc. of the 14th ACM international conference on information and knowledge management* 373-380
[19] Fumera G, Pillai I and Roli F 2006 Spam filtering based on the analysis of text information embedded into images *Journal of Machine Learning Research* **7** 2699-2720

[20] Boykin P and Roychowdhury V 2005 Leveraging social networks to fight spam *Computer* **38** 61-68
[21] Zelenkov Y and Segalovich I 2007 Comparative analysis of methods for fuzzy duplicate detection for Web-documents *Proc. of 9-th Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections»* 166-174
[22] Salton G and Buckley C 1988 Term-weighting approaches in automatic text retrieval *Information Processing & Management* **24** 513-523
[23] Church K and Gale W 1995 Poisson mixtures *Natural Language Engineering* **1** 163-190
[24] Chowdhury A, Frieder O, Grossman D and McCabe M 2002 Collection statistics for fast duplicate document detection *ACM Transactions on Information Systems (TOIS)* **20** 171-191
[25] Kolcz A, Chowdhury A and Alspector J 2004 Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization *Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 605-610
[26] Fetterly D, Manasse M and Najor M 2003 A Large-Scale Study of the Evolution of Web Pages *ACM* 669-678
[27] Rabin M 1978 Digitalized signature as intractable as factorization. Technical Report MIT/LCS/TR212 *MIT Laboratory for Computer Science*
[28] Schleimer S, Wilkerson D S and Aiken A 2003 Winnowing: Local Algorithms for Document Fingerprinting *Proc. of ACM SIGMOD International Conference on Management of Data*
[29] Damiani E, De Capitani di Vimercati S, Paraboschi S and Samarati P 2004 An open digest-based technique for spam detection *Proc. of the International Workshop on Security in Parallel and Distributed Systems* (San Francisco, CA USA)

Acknowledgments
The reported study was funded by RFBR according to the research project № 19-07-00692.