Proceedings of the 10th International Conference "Reliability and Statistics in Transportation and Communication" (RelStat'10), 20–23 October 2010, Riga, Latvia, p. 243-249. ISBN 978-9984-818-34-4
Transport and Telecommunication Institute, Lomonosova 1, LV-1019, Riga, Latvia

THE PROBLEM OF E-MAIL GATEWAY PROTECTION SYSTEM

Aleksey Latkov, Nikita Skabcov, Yury Svirchenkov

Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019, Latvia
Ph.: +371 26598191. E-mail: latkov@tsi.lv

Currently, the problem of unwanted mailings is topical around the world. 90% of the total number of e-mails is SPAM. Due to the SPAM problem, companies have multimillion-dollar damages. Also through SPAM and phishing attacks carried out the spread of viruses. At the moment, a sufficiently effective solution for blocking SPAM does not exist.

Keywords: phishing, shingle, "white" and "grey" lists, SPAM filtration

1. Introduction

Nowadays, the problem of SPAM is one of the main problems of electronic Internet mail. More than 90% of total amount of electronic mail is SPAM. SPAM is the main reason of huge economic losses for the states and companies, as well as for individuals. Due to the SPAM, providers should use automatic systems of e-mail filtration. SPAM filters can't divide normal and harmful e-mail traffic with 100% accuracy, therefore providers have large problem with filter false decision. Every year, approximately losses from SPAM can reach hundreds of billions dollars worldwide.

In the article authors compare the effectiveness of SPAM filtering systems of leading world companies and open-source communities. The real flow of incoming mail was duplicated to the experimental model. A filter statistical data collection the quality of filtering system was valued by the following indices: central processor unit usage, random access memory usage, amount of filter false decisions.

Further, new methods to increase effectiveness of SPAM filtering are offered. The offered filter method is realized in software product. Next – the research of the offered filter is done on the scale model. Then a comparison between indices of offered method and method of Bayesian filtration is made. Finally, a conclusion concerning the common usage Bayesian filtering and the offered filter is done.

2. Description of Testing

To estimate the effectiveness of e-mail gateway protection system scale model is made (see Figure 1).

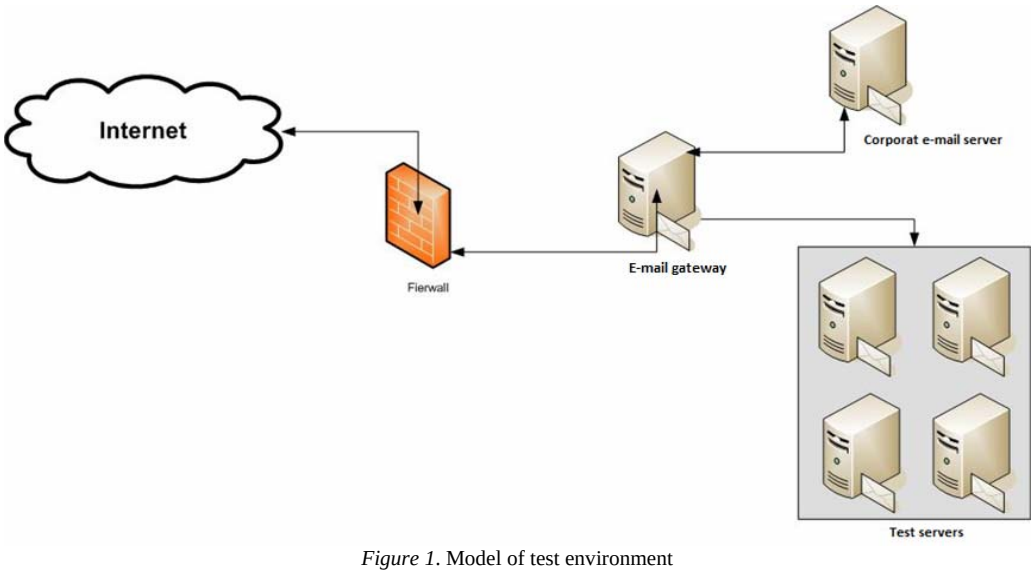

Figure 1. Model of test environment

To make the research of the existing models of e-mail gateway protection system the following products of leading world companies for Windows and Unix like systems have been chosen:

- McAfee Email and Web Security VMTrial v5.5;
- TrendMicro InterScan Messaging Security Virtual Appliance 7;
- Kaspersky Mail Gateway 5.6.28;
- MailCleaner 3.0;
- GFI MailEssentials for Exchange/SMTP x86 14.1;
- Symantec Mail Security For MSE 6.5.0.67;
- E-mail Security Suit 1.5

2.1. Description of research model

To test e-mail gateway protection system server is installed in provider's datacenter. Server has a role of virtual server that hosts several virtual machines.

Table 1. Configuration of physical server

Server model	HP ProLiant DL380 G6
Amount of CPU	2
CPU model	Intel® Xeon® 5600 series
CPU speed	3.7 GHz
Amount of RAM	8 GB
HDD total size	2 TB
RAID level	5

Table 2. Configuration of virtual server

Amount of CPU	1
Amount of RAM	1Gb
HDD total size	80Gb
OS	CentOS 5.3 / Windows Server 2008 R2

Protection system testing has been carried out by means of real post flow that the following parameters have been measured:

- CPU usage;
- RAM usage;
- The number of regular e-mails marked as SPAM – „false negative";
- The number of SPAM e-mails that pass through filter – „false positive".

Load on server was approximately 30000 e-mails per day. The testing took two days. To calculate „false positive" and „false negative" all incoming post archive was filtered using regular expressions. Several categories of letters did not run trough filtering. Namely, the letters witch where filtered out by MTA due to such parameters as: "the letter not meeting RFC821" [1], sender's address was found in "black list", impossibility to identify sender's address, etc. The letters that pass filtering where checked manually because there is no other method to distinguish with 100% accuracy a regular letter from SPAM.

Table 3. Test results

Protection system	**CPU %**	**RAM %**	**False positive %**	**False negative %**
McAfee	77	63	11	1.3
TrendMicro	73	89	15	1.4
Kaspersky	69	49	8.5	1.9
MailCleaner	63	52	12	1.1
GFI	50	49	9	1.22
Symantec	54	56	8.2	1.5
EMSS	46	90	9	1.7

After analyses of test results the following conclusions can be made:

- Researched products utilize too many system resources, especially processor time, which will lead to the failure of a part of protection system or even of the whole e-mail server – if the load increases.
- All protection systems have rather high level of "false negative", i.e. a letter, no being SPAM, was marked by filter as SPAM. It is a weary undesirable situation which can appear due to polymorphous technologies, or when is used wrong self-studied filter.
- All protection systems have a high level of "false positive", when SPAM in not recognized. It is caused by the imperfection of technologies of undesirable letter blocking.
- Spammers can avoid filter applying some technologies such as: using template alteration, polymorphous technologies, social engineering and methods of letter distribution.
- Even then 2-3 undesirable letters a day can contain harmful code or phishing, which can causes problems and is very dangerous for business and corporate network.

3. Description of Suggested Method

Taking into consideration everything mentioned above, the authors offer a filtering method, which helps to improve SPAM filtering indices. The offered method is based on the following postulates:

- E-mail filtration algorithm modification;
- Modified shingles algorithm; [1]
- The use of the notions of "white", "grey" and "black" lists;

The block-scheme of offered algorithm of letter testing is given on Figure 2. This solution lets decrease the system recourses usage. As well gives a possibility to test new rules for SPAM blocking.

For more effective filtration of SPAM it is suggested to use queues. When e-mail comes to server it will be queued while one of events will not happen – either the determined amount of letters is collected or it is passed the sufficient amount of time. Combining of these terms is not eliminated also.

For the analysis of letters, which are in a queue, it is very convenient to use the shingle algorithm. This algorithm will help to expose similar letters and find a SPAM and even polymorphic SPAM.

Realization of algorithm implies a few stages:

- text canonization;
- breaking up the text on shingles;
- checksums finding;
- search of identical sub lines;

In the algorithm of shingles the comparison of the texts checksums is realized. As it is generally known, the checksums of the static functions are very sensitive to the changes. Therefore, at a search almost of doublet, all exclamation marks, points, commas and other will be deleted from text. Only words will be left.

Now it is necessary to break up each of the received texts on sub lines – shingles. Further, checksums can be found from marked out shingles. Secondly, must be determinate how they must go – lapped (that is how exactly "shingles" are obtained), or butt [1].

To create a sample – the classical algorithm written by Broder suggests choosing either a fixed number of shingles with a minimal value or all shingles whose value can be divided into some small number (10-30). In the first case we get fixed size sample and large-sized set of shingles, but we cannot judge about the subdocument. In the second case the number of shingles is proportional to the size of the document, but it is possible to evaluate by the sample of shingles the attachment of documents to each other or the percentage of their intersection.

Finally, the last algorithm generates a fixed size sample which size is determined by a given number (e.g. 85) of different independent random functions, each of which store exactly one shingle, which is the minimal by the value of the checksum. This approach combines the advantages of the previous two [2].

There is a problem with short documents, for which the algorithm of the shingle selection may not select a single suitable, or to choose too little. There are two solutions: one of them: to circle text of the document, that is virtually continuing its beginning after end. The second approach is to use a sample, size of which has a logarithmic dependence on the size of the document.

If, for each letter to select more than one shingle, we are confronted with the task of identifying documents that have only a few matching shingles. If we reduced the number of shingles, it still would be a nontrivial amount of work. Therefore, over a set of shingles of the document, it is possible to calculate another checksum, the so-called "super shingle" [2].

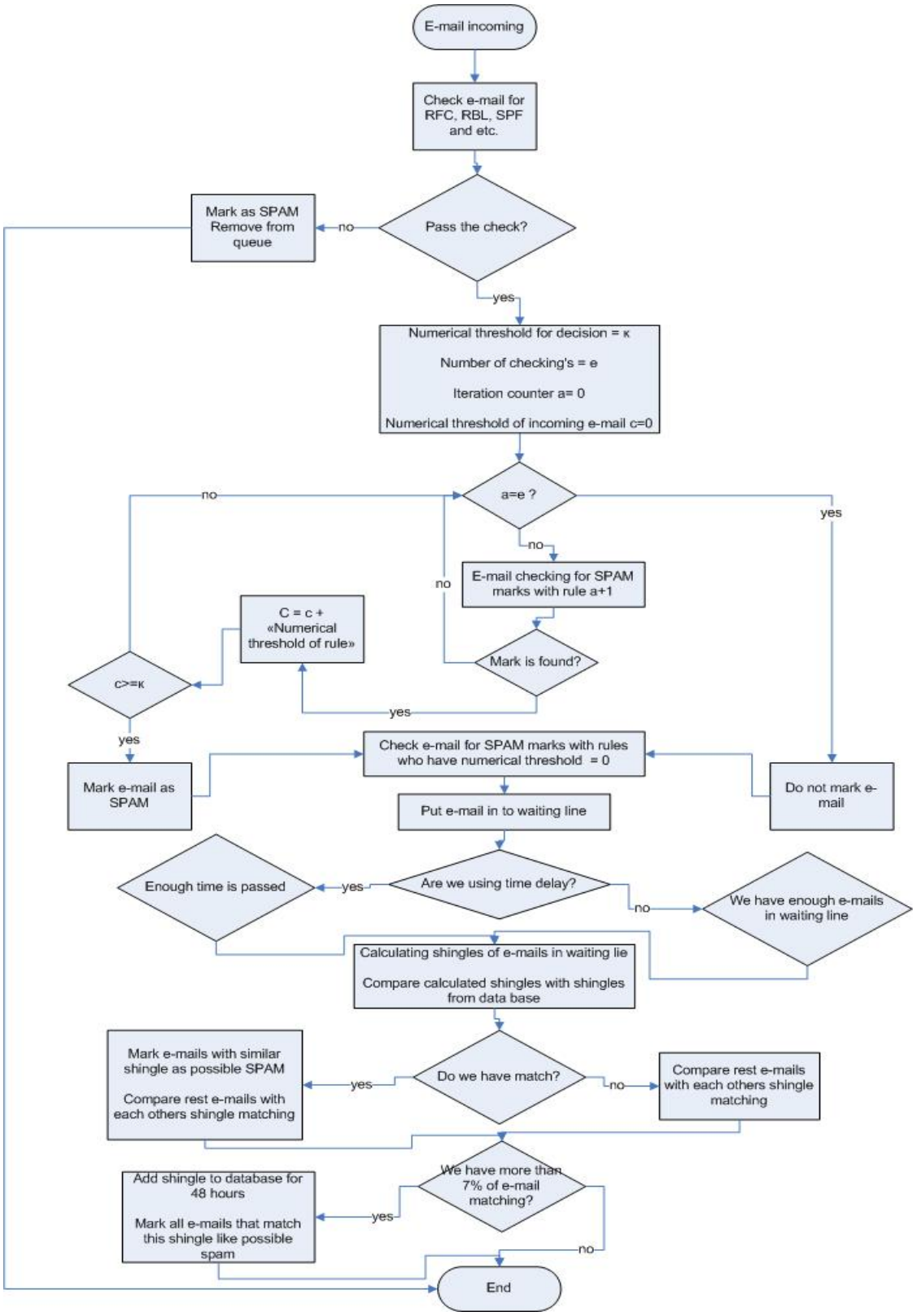

Figure 2. Offered solution block-scheme

To improve performance, it is possible to compare not all of the obtained checksums, but only those who, for example, can be divided by 25. It makes a significant gain in speed and not greatly reduces the accuracy, but only when processing large volumes.

When using the algorithm of shingles, a situation may arise where the normal mail delivery will be adopted as SPAM. To solve this problem, we propose to use "white" lists.

This algorithm was realized on the python language as software modules functioning in windows and UNIX environment. On the scale model (Figure 1) the testing of realized filter has been carried out. The results of the testing are seen on the graph (Figure 3). The graph shows that though the offered generalized algorithm of filtering enlarges the load on server by 7-12%, it substantially improves "false negative" and "false positive" indicates by 8.5%.

Yet, the offered generalization method of filtering has few drawbacks:

- If there is only one letter or there are several of them (on condition of big except of letters) from the started distribution, they won't be recognized as SAPM;
- The given method can create delay of letters up to 5 min;
- The load on server processor increases;

To assess filter false activation the following formula has been used:

$$P_{err} = P(A)P(B/A)+P(B)P(A/B), \quad (1)$$

where:
P(A) and P(B) – possibilities of event (B) occurrence on filtering entrance of both regular letters and SPAM;
P(B/A), P(A/B) – conditional possibilities of false filter activation. The results obtained in formula (1) and on scale model have divergence of on more than 5-7%;

Applying the notion of "white", "grey" and "black" lists lets solve a number of problems while filtering. For example, creation of dispersed database of "white" lists simplifies the usage of "grey" lists. Decreases the load on server and improves "false negative" indices. Block-scheme of the offered generalized SPAM filtering algorithm is given on Figure 2.

4. Offered Method Testing

After implementation of new method we have tested system performance.

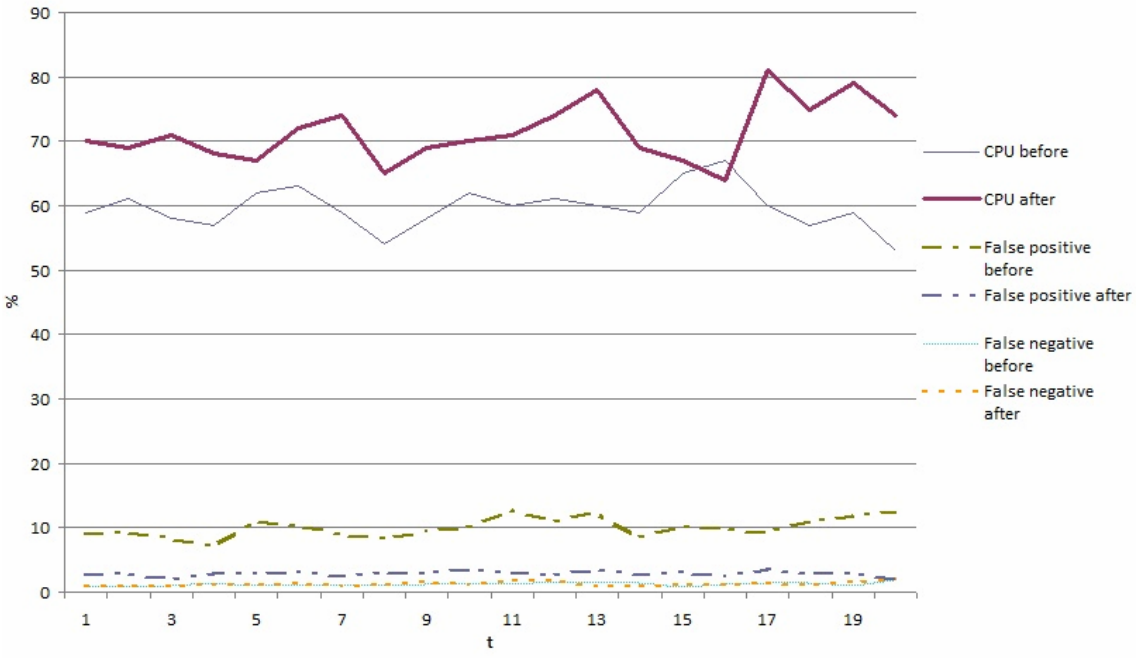

Figure 3. Improved algorithm testing results

Further, to estimate the effectiveness of the offered generalized filtering method operation, the authors have compared it with the well-known Bayesian filtering on the scale model [3]. The test was carried out on different amount of SPAM in total mail flow.

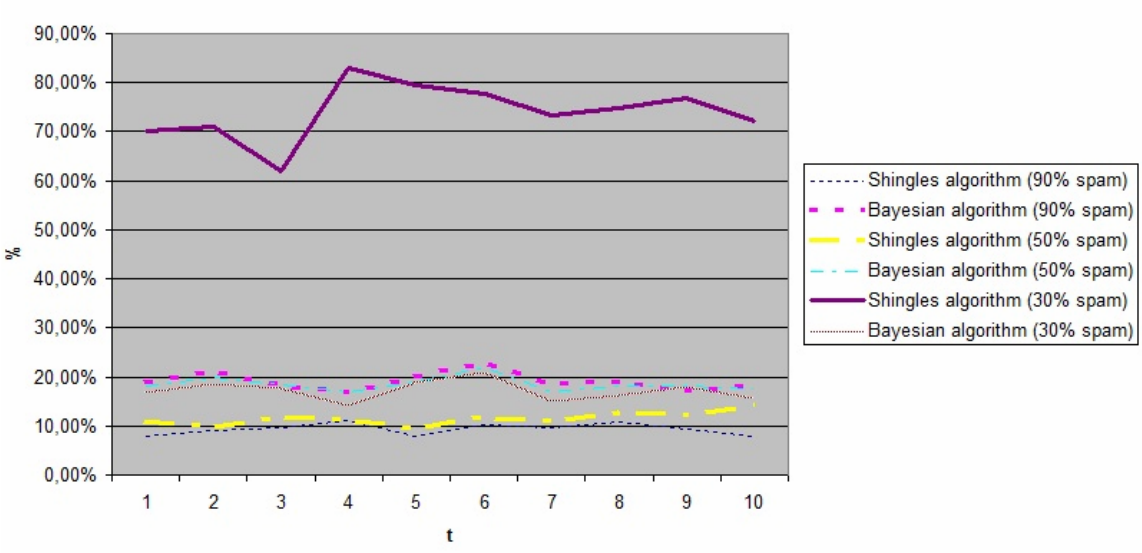

Figure 4. Compare Shingles and Bayesian algorithm

Figure 4 shows the results of the testing. As we can see, it is evident from the graph, that "false positive" index for the generalized filtering algorithm substantially worse than for Bayesian filter in case when total e-mail traffic contains small (under 30%) part of SPAM.

Therefore it seems expedient that to solve the problem of SPAM filtering both methods should be used in combination: Bayesian filtering and generalized filtering algorithms.

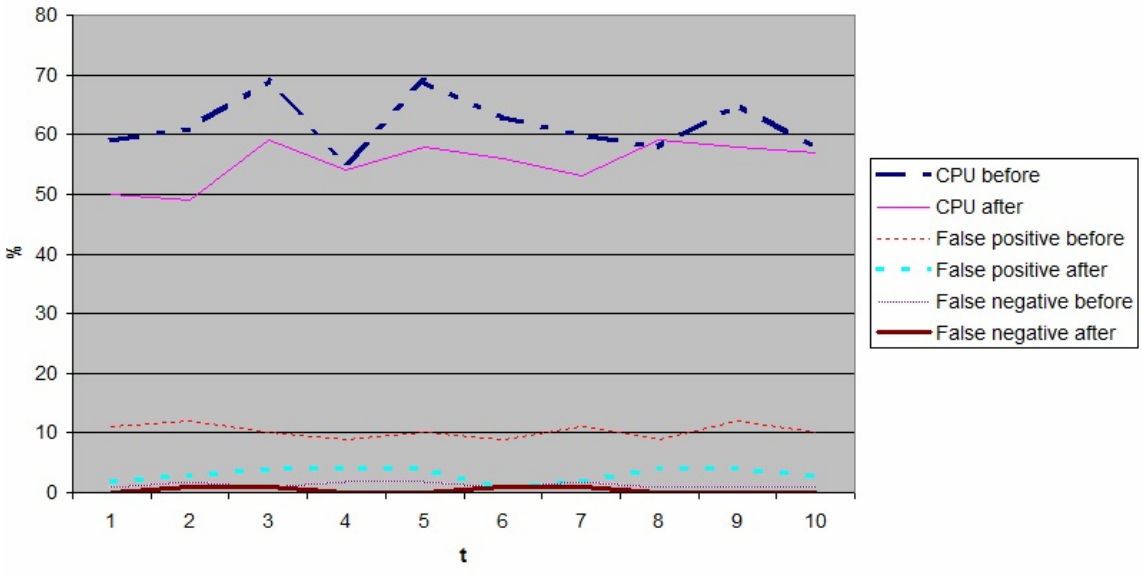

Figure 5. Testing result after new methods implementation

5. Conclusions

The above mentioned combined filtering method will allow to significantly improving SPAM filtering indices:

- "False positive" by 5.1 %;
- "False negative" by 0.2%;
- CPU usage by 5%.

Based on these results we can conclude that the problem posed at the beginning of the work was successfully carried out.

References

1. Jonathan, B. Postel *SIMPLE MAIL TRANSFER PROTOCOL*. Marina del Rey: Information Sciences Institute, University of Southern California, 1982. 71 p.
2. Broder, A. *Algorithms for duplicate documents* – http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/Princeton.pdf
3. *Bayesian Spam filtering* – http://en.wikipedia.org/wiki/Bayesian_spam_filtering
4. Ilyinsky, S. *An efficient method to detect duplicates of Web documents with the use of inverted index* – http://www2002.org/CDROM/poster/187/ (2002)
5. Chowdhury, A. *Collection statistics for fast duplicate document detection. ACM Transactions on Information Systems (TOIS)* – http://ir.iit.edu/~dagr/2002collectionstatisticsfor.pdf (April 2002)