# Combined Classifier for Website Messages Filtration

*Veniamin Tarasov< tarasov-vn@psuti.ru>,*
*Ekaterina Mezenceva <katya-mem@mail.ru>,*
*Danila Karbaev <danila@karbaev.com>*
*Volga Region State University of Telecommunications and Informatics, 77 Moskovskoe sh., Samara, 443090, Russian Federation*

**Abstract.** The paper describes a new approach to website messages filtration using combined classifier. Information security standards for the internet resources require user data protection however the increasing volume of spam messages in interactive sections of websites poses a special problem. Spam messages vary significantly in content, however the common feature of these messages is that they are usually of little interest to the majority of the recipients. Many filtering approaches are based on the Naive Bayesian classifier - an effective method to construct automatically anti-spam filters with high performance. Unlike many email filtering solutions the proposed approach is based on the effective combination of Bayes and Fisher methods, which allows us to build accurate and stable spam filter. In this paper we consider the organization of combined classifier according to determined optimization criteria based on statistical methods, probability calculations and decision rules. We consider the optimization criteria for grading messages basing on statistical methods. The classifiers normally admit the compromise between the acceptable level of false-positive and false-negative errors, and use the threshold values for decision-making, which may vary. In order to receive more valid results of spam detection we need to analyze multitudes of results of various filters and a subset of their overlaps. The approach we suggest is to construct classifier organization, which presumes the combined use of Bayes and Fischer methods for improved the filtration quality based on the analysis of subsets and set overlaps identified by both methods (spam, non-spam, false triggering and spam leaks).

**Keywords:** combined classifier; spam filter; optimization criterion.

**DOI:** 10.15514/ISPRAS-2015-27(3)-20

**For citation:** Tarasov V., Mezenceva E., Karbaev D. Combined Classifier for Website Messages Filtration. Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 3, 2015, pp. 291-302. DOI: 10.15514/ISPRAS-2015-27(3)-20.

## 1. Introduction

The constantly growing volumes of data, number of uses as well as groups devoted to various subjects significantly decrease the effectiveness and the authenticity of communicated information. In this regard the task of increasing the efficiency of statistical data filtration and authentication algorithms becomes undoubtedly topical. The history of this subject in computer science accounts for more than 20-30 years and the trend is becoming more urgent. We can say that right now the antispam features of interactive sections of websites rest in the very initial stage of development.

The subject of message filtration in emails is widely developing, manual antispam methods are being used, and the issue of automated antispam protection of corporate websites becomes a priority on the agenda (including comments, forums and other interactive sections). In practice there are no universal software solutions to protect all types of interactive website sections from spam. There are only small number of specialized tools which prevent automatic messages posting. Some of them are designed for a particular content management system, such as WordPress in form of plugins: Akismet, Quiz, Spam Karma etc. These modules have some disadvantages: the distribution model "as is" do not include the statistical base, most of online services do not provide multilingual filtration and are limited only by the support of the English language. The other blog comment hosting services such as IntenseDebate, Disqus, Livefyre do not provide self-hosted option, except Discourse.

Thereby the spam filtering software solution should have the following properties: the use of multiple filtering methods, both formal and linguistic, united by a common intellectual decision making core; high speed and precision of the method; easy installation and use.

This work describes a new approach to spam filtration involving the combined use of Bayes and Fischer methods, allowing to significantly reduce the number of false triggering and increase spam detection.

## 2. Calculation of combined probabilities of conditions

The main idea of message classification is based on selection of all conditions, calculation of probabilities of select conditions, and further combination of all calculated probabilities into one value for the studied message. Messages with a large number of spam attributes and little non-spam attributes will have a value close to 1, and the messages with a large number of non-spam attributes and little number of spam attributes will gain a value close to 0.

We will build a classifier of messages received by the website to grade the incoming messages into three categories (spam, non-spam, unidentified). In this respect, we need to identify all conditions (words and word combinations) in the message to be analyzed, calculate statistical probabilities for some select conditions and combine all probabilities into one value for the whole message. In most cases the probability

of assigning a message to a certain category is a lot higher than to others, which results in further grading of such message.

Before calculating the combined probabilities of conditions, we need to calculate the probability of assigning a certain condition to a specific category. For this we can divide the identified number of messages with condition *i* in this category by the total number of messages in the same category, but we would rather use another method described below.

Let's assume:

$F_{ai}$ is the number of messages with condition $i$ in the spam group;

$F_{bi}$ is the number of messages with condition $i$ in non-spam group.

Then the statistical probability of appearance of $i$ in a spam message can be calculated as follows:

$$p_{ai} = \frac{F_{ai}}{F_{ai} + F_{bi}} \quad (1)$$

and the probability of appearance of $i$ condition in a non-spam message, as follows:

$$p_{bi} = \frac{F_{bi}}{F_{ai} + F_{bi}} \quad (2)$$

Thus, the number of messages with condition $i$ in one category will be divided by the total number of messages featuring this condition $i$.

The use of (1) and (2) takes into account the fact that with time the number of messages in both categories may be equal, i.e. these formulas do not depend on the number of messages in a specific category.

Note that formulas above give accurate result only to those conditions, which filter is used in both categories. As the result the spam filter becomes too sensitive on early stages of learning applying to rare words. To solve this problem we need to calculate new probability with expected a priori probability ($P_{ex}$) and applied weight ($w$), then according to (1) and (2) add calculated probabilities.

If the probability $P_{ex}$ = 0.5 and the weight of expected probability equals to one word ($w$ = 1), we estimate weighted probabilities using (1) and (2):

$$\overline{p_{ai}} = \frac{(w * P_{\text{ex}}) + p_{ai} * (F_{ai} + F_{bi})}{w + F_{ai} + F_{bi}},$$

$$\overline{p_{bi}} = \frac{(w * P_{\text{ex}}) + p_{bi} * (F_{ai} + F_{bi})}{w + F_{ai} + F_{bi}}.$$

This approach allows to avoid division by zero in the following formulas and to take into account rare words.

To obtain combined probabilities of the whole document (message) we will use the dictionary, which is built on the step of filter learning. We introduce the following

events: $A$ – document is spam, $B$ – document is non-spam. We assume that the probabilities are independent, thus the multiplication is allowed:

$$P(A) = \overline{p_{a1}} \times \overline{p_{a2}} \times \ldots \times \overline{p_{aM}} \quad (3)$$

- for the probability of words co-occurrence in spam;

$$P(B) = \overline{p_{b1}} \times \overline{p_{b2}} \times \ldots \times \overline{p_{bM}} \quad (4)$$

- for the probability of words co-occurrence in non-spam [[1]].

## *3. Decision rules based on bayes theorem*

To estimate the probability that word belongs to one of three categories (spam, non-spam, unidentified messages) we consider the two methods of classification. In this case we apply Bayes formulas using a priori knowledge [[1]].

We introduce two hypotheses for any given message:

$H_A$ if the message is a spam,

$H_B$ if the message is a non-spam.

Further, we introduce the following notation:

$F_a$ is the total quantity of spam messages;

$F_b$ is the total quantity of non-spam messages;

$p_a = \dfrac{F_a}{F_a + F_b}$ is a priori probability that a message is a spam;

$p_b = \dfrac{F_b}{F_a + F_b}$ is a priori probability that a message is not a spam;

$O_a = \dfrac{P_a}{1 - P_a}$ is a priori expectations that a message will be a spam;

$O_b = \dfrac{P_b}{1 - P_b}$ is a priori expectations that a message will be a non-spam.

Then basing on Bayes theorem using a priori knowledge we obtain:

$P(H_A) = \dfrac{P(A) \times O_a}{P(A) \times O_a + P(B) \times O_b}$ - a posteriori probability that a message is a spam;

$P(H_B) = \dfrac{P(B) \times O_b}{P(A) \times O_a + P(B) \times O_b}$ - a posteriori probability that a message is non-spam.

The probabilities $P(A)$ and $P(B)$ are estimated according to (3) and (4).

Given algorithm is implemented in spam detection and filtering system for websites. [[2]].

## 4. Decision rules based on fisher's method

According to Fisher method all probabilities are multiplied together in a similar manner to Bayes method, then the natural logarithm is taken of the product and the result is multiplied by -2. To do this we introduce variable *hisqv*, which is estimated by the following expressions:

$hisqv = -2 * \ln(P(A))$ or $hisqv = -2 * \ln(P(B))$,

where probabilities $P(A)$ and $P(B)$ are calculated according to (3) and (4).

Fisher proved that if the set of independent and random probabilities (3) and (4) is given, the value $-2 * \ln(P(A))$ follows the distribution of $\chi^2$ with *2n* degrees of freedom (n – the number of words in the document):

$$F(x) = \int_0^x \frac{t^{n-1} e^{-t/2}}{2^n \Gamma(n)} dt \qquad (5)$$

where $\Gamma(n)$ is the gamma function.

In view of foregoing using a representation of the gamma function of even argument (5) can be written as:

$$F(x) = \frac{1}{2^n (n-1)!} \int_0^x x^{n-1} e^{-x/2} dx \;\Big|\; x = hisqv \qquad (6)$$

The calculation of the factorial and the integrand in (6) could cause the overflow error due to floating point numbers range in PHP programming language. Thus the recurrence formula is used in the calculation algorithm. Calculation the probability of (6) is implemented by Gaussian quadrature formula with 15 nodes:

$$\int_a^b f(t)dt \approx \frac{b-a}{2} \sum_{i=1}^{n} A_i f(t_i),$$

where $t_i = (b+a)/2 + (b-a)x_i/2$, and $x_i$ are the nodes of Gaussian quadrature formula;

$A_i$ are the Gaussian coefficients, ($i = 1, 2, \ldots, 15$)[[3]]. In our case $a = 0$, $b = hisqv$ .

The value returned by the function $F(hisqv)$ is low if a text contains many spam conditions. We need the opposite result to rate the message correctly. For this purpose we subtract the value from 1. The use of this subtraction for a large number of non-spam conditions allows us to get the probability that message is not spam.

However the Fisher method is not symmetrical. We need to combine the probabilities of spam and non-spam into a single value in the range between 0 and 1. For this we use the Fisher index:

$I = \frac{1 + P(H'_A) - P(H'_B)}{2}$, where:

$P(H'_A) = 1 - F(-2\ln(P(A))$ is the probability that a document belongs to spam;

$P(H'_B) = 1 - F(-2\ln(P(B))$ is the probability that a document belongs to non-spam [[4]].

## 5. Optimization criteria for grading messages based on statistical methods

Let's assume that all set of conditions is divided into classes A and B, where A – class of spam messages, and B – class of non-spam messages. The task of assigning a message to any of these classes is not directly connected to the statistical verification of the following hypotheses: simple hypothesis HA: X A against the alternative HB: X B, where X is the message qualifying condition. As we know from the math statistics, if a message appertains to class A and it was qualified as class B, it will result in 1st type error with the conditional probability of - level of importance. It will be an error of the alternative hypothesis selection HB instead of the correct HA. If HB hypothesis is fair but, nevertheless, HA was selected, the 2nd type error will occur with the conditional probability of.

The 1st type error or false-negative error occurs if the spam filter erroneously leaks an undesired message through identifying it as non-spam (spam leakage or insufficient method completeness). Whilst the spam filter is capable of identifying a large share of undesired messages, the task of minimizing the number of faulty filtering of desired (non-spam) messages may become a higher priority, i.e. the task of 2nd type of error minimization.

The 2nd type error or false-negative error occurs if the spam filter erroneously classifies a legitimate message as spam (faulty triggering or method accuracy). The spam filter will be efficient with a lower number of such errors, i.e. with minimal 2nd type error level. However currently all antispam systems demonstrate correlation between 1st and 2nd type errors.

The classifiers normally admit the compromise between the acceptable level of 1st and 2nd type errors, and use the threshold values for decision-making, which may vary. This results in the "strictness" or "softness" of the classifier. The level of significance set during the statistical hypothesis verification is taken as the threshold value. Whereas, the increase of the filter sensitivity leads to the increased occurrence of 1st type errors (spam leaks), and decrease of sensitivity – to increased occurrence of 2st type of error (false triggering).

## 6. Bayes optimization criterion

We need to consider the losses related to 1st and 2nd type errors for evaluating the classification quality. For this we need to split the space of condition $X$ into two semispaces $X_A$ and $X_B$ with point $x_0$. Let's define $c_1$ as the conditional price of 1st

type error and $c_2$ – conditional price of 2nd type error, $P(A)$ – a priori probability of $A$ class, $P(B)$ – a priori probability of class $B$, $P(A) + P(B) = 1$. The values $c_1$ and $c_2$ depend on the price matrix coefficients $\mathbf{C_{2x2}}=\{c_{ij}\}$ and on the 1st and 2nd type errors:

$$c_1 = c_{12}\,\alpha + c_{11}\,(1 - \alpha) \tag{7}$$

$$c_2 = c_{21}\,\beta + c_{22}\,(1 - \beta) \tag{8}$$

These values are also called conditional risks with proven fairness of hypotheses $H_A$ and $H_B$, respectively.

According to the decision making theory, we introduce the decision rule of classification, which minimizes the function of losses (risk) [[3]]:

$$R = c_1 P(\mathrm{A}) + c_2 P(B) \tag{9}$$

where $c_1$ and $c_2$ are determined by (7) and (8).

Function (9) represents the average risk, which depends on the threshold value $x_0$, because the values $c_1$ and $c_2$ depend on the $x_0$ value through type I and type II errors, therefore these errors are correlated.

Minimum value $R_{min}$ of risk function (9) at the point $x_0$ is called Bayes risk.

$$\frac{f_1(X)}{f_2(X)} = \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)} \tag{10}$$

where $f_1(X)$ and $f_2(X)$ are the probability density distributions of $X$ condition on $A$ and $B$ classes respectively.

The right part in (10)

$\frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}$ is called likelihood ratio, which is constant for the selection of $c_{ij}$. Thus, if the inequality $\frac{f_1(X)}{f_2(X)} > \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}$ is true, the observable vector $X$ is related to $A$ class; if the inequality $\frac{f_1(X)}{f_2(X)} < \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}$ is true, then observable vector $X$ is related to $B$ class. If the equality $\frac{f_1(X)}{f_2(X)} = \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}$ is true, the observed vector $X$ is related to one of the classes A or B. The latter expression is the equation for the boundaries of A and B classes. This decision rule is related to Bayes rules [[5]].

The technique can be applied to many practical problems formulated in terms of statistical decision making theory with assumption that probability densities $f_1(X)$ and $f_2(X)$ are known. In most practical cases functions $f_1(X)$ and $f_2(X)$ are not

known, and we need to determine estimations $\tilde{f}_1(X), \tilde{f}_2(X)$ on training sets using approximation method [[5]], which can cause the classifier to slow down. Considering this fact we use the following approach: on the stage of filter learning the estimations $\tilde{f}_1(X), \tilde{f}_2(X)$ are determined on small training sets of 100-200 elements, and the optimality criterion to get such estimations can be excluded excluded from the program flow.

Results of numerous tests on training selections allowed identifying optimal threshold values for decision-making:

$x_{\mathrm{H}} = 0{,}95$ for higher threshold and $x_{\mathrm{L}} = 0{,}4$ for lower threshold.

Thereby we set strict limits for spam and regular for non-spam messages. Such threshold values provide minimum leakage of desired messaged into spam, i.e. minimum false triggering. However, it's notable that any system administrator will be able to easily set more convenient threshold values to suit his needs.

## *7. Combined filter*

In order to receive more valid results of spam detection we need to analyze multitudes of results of various filters and a subset of their overlaps.

We suggest exactly this kind of approach to classifier organization, which presumes the combined use of Bayes and Fischer methods for improved the filtration quality based on the analysis of subsets and set overlaps identified by both methods (spam, non-spam, false triggering and spam leaks).

Let's assume $\mathbf{S}=\{s_i\}$ ($i=1\div M$) – multitude of documents (messages), including both desired and spam messages; $\mathbf{S}_B \subset \mathbf{S}$ and $\mathbf{S}_F \subset \mathbf{S}$ – multitude of documents, identified by Bayes and Fischer classifiers, respectively. Then the subset resulting from the overlap $\mathbf{S}_B \cap \mathbf{S}_F$ against all indicated categories may be used for evaluating the quality of the combined filter operation (see Fig. 1).

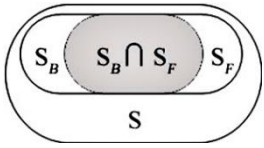


*Fig. 1. Illustration of overlap degree of two subsets SB and SF.*

The completeness of such overlap $\mathbf{S}_B \cap \mathbf{S}_F$ will also grade the subsets $\mathbf{S}_B \backslash \mathbf{S}_F$ and $\mathbf{S}_F \backslash \mathbf{S}_B$. As a measure of overlap degree of two sets $\mathbf{S}_B$ and $\mathbf{S}_F$ we suggest to use the absolute measure $N(\mathbf{S}_B \cap \mathbf{S}_F)$ – number of shared documents in these subsets. Thus, the maximum value of measure of $l$ category (spam, non-spam, false triggering and spam leaks) will be used as the optimality criterion for spam filter self-teaching evaluation:

$$N_l(\mathbf{S}_B^l \bigcap \mathbf{S}_F^l) \to \max.$$

Once the best values of sets $\mathbf{S}_B$ and $\mathbf{S}_F$ overlap are reached across all categories, the administrator will be able to choose a filter for further application (see Fig. 2).

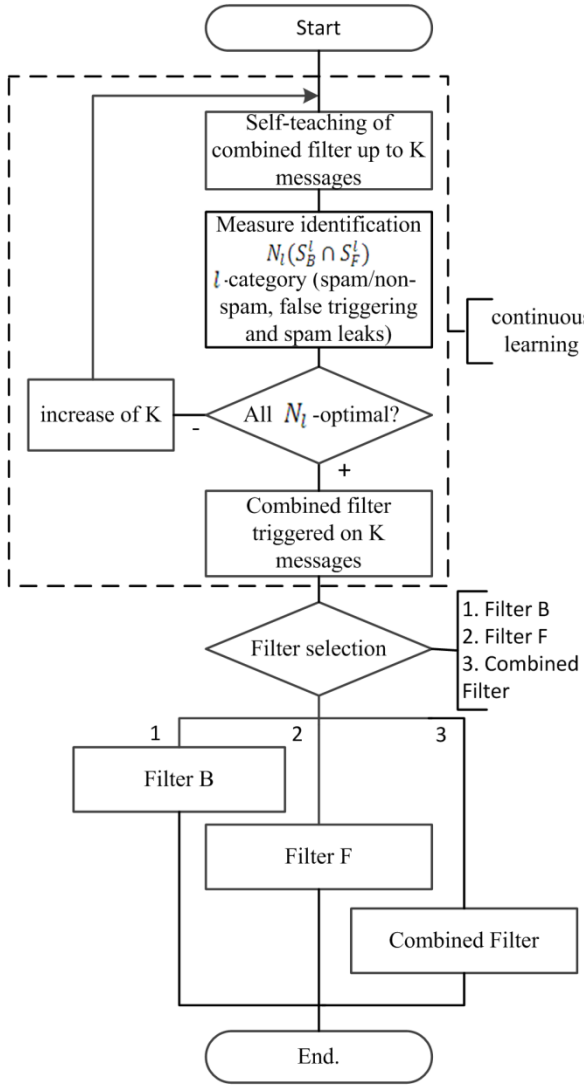


*Fig. 2. The algorithm of combined filter accuracy evaluation.*

As a benefit of the combined filter implementation the evaluation of all components of the overall picture became possible:

- spam messages caught by both filters;

- spam filters caught only by Bayes or only Fischer filters;
- simultaneous false triggering of both filters;
- false triggering of each individual filter;
- simultaneous spam leaks by both filters;
- spam leaks of each individual filter.

Before testing filter was trained on 1100 messages (400 spam and 500 non-spam). The tests were run on the flow of 1223 messages. The Bayes method showed 2.9 percent of the false triggering, 9.8 percent of spam omission. The Fisher method showed 1.5 and 4.5 percent accordingly. The combined filter showed the best result with 1.0 and 4.5 percent.

The experimental results confirmed the feasibility of using the selected filtering algorithms. Only having a whole picture, we will be able to make a reasonable comparison of the combined filter self-teaching quality.

## References

[1]. E. Mezenceva, V. Tarasov. “Securing computer networks. The method of multi-module spam filtering on websites,” Information Technologies, 2012. vol. 6, P. 18-22 (in Russian).
[2]. E. Mezenceva. “The software system of recognition and spam filtering on the sites,” Certificate of state registration of the computer program №2011619160, [Registered in the Computer Program Registry, Moscow, on November 25th, 2011] (in Russian).
[3]. S. Nikolskiy. Quadrature Formulas. “Nauka”, Moscow, 1974. 224 p. (in Russian).
[4]. E. Mezenceva, V. Tarasov. “Computer networks security. Web programming of the multi-module spam filter,” Software Engineering, 2012. vol. 4, P. 27-32 (in Russian).
[5]. E. Mezenceva, V. Tarasov. “An optimal filter construction based on combining statistical classifiers,” Information and communications technologies, book 1, 2013. vol. 4, P. 53-57 (in Russian).

# Совмещенный классификатор для фильтрации сообщений на веб сайтах

*Вениамин Тарасов< tarasov-vn@psuti.ru>,*
*Екатерина Мезенцева <katya-mem@mail.ru> ,*
*Данила Карбаев <danila@karbaev.com>*
*ФГОБУ ВПО Поволжский государственный университет телекоммуникаций и информатики, 443090, Россия, Самара, Московское шоссе д. 77.*

**Аннотация.** В работе рассмотрен новый подход к фильтрации сообщений на сайтах с использованием совмещенного классификатора. Уровень защиты пользовательских данных определен стандартами информационной безопасности для Интернет-ресурсов, кроме того постоянно растет число спам-сообщений в интерактивных разделах сайтов.

Предлагаемый подход, в отличие от распространенных решений для электронной почты, основан на совместном использовании методов Байеса и Фишера, что позволило разработать эффективное программное решение фильтрации спама. Основная идея классификации сообщений заключается в выделении всех признаков, вычисления вероятностей для отдельных признаков, и затем объединения всех вычисленных вероятностей в значение для всего сообщения. Рассмотрены критерии оптимальности при классификации сообщений на основе статистических моделей. В качестве примера были установлены пороговые значения, обеспечивающие минимум пропуска в спам нужных сообщений, т.е. минимум ложных срабатываний. Для получения более достоверных результатов выявления спама необходимо проводить анализ множеств результатов работы отдельных фильтров и подмножества их пересечений. В работе рассмотрен подход к построению совмещенного классификатора, удовлетворяющего критериям оптимальности и обеспечивающего принятие решений при классификации сообщений на основе статистических методов. Нами предлагается именно такой подход к организации классификатора, который заключается в совместном использовании методов Байеса и Фишера для повышения качества фильтрации на основе анализа подмножеств пересечения множеств, распознанных обоими методами (спам\не спам, ложные срабатывания и пропуск спама). Благодаря реализации совмещенного фильтра можно обоснованно сравнивать качество обученности совмещенного фильтра.

**Ключевые слова:** совмещенный классификатор, спам фильтр, критерий оптимизации.

**DOI:** 10.15514/ISPRAS-2015-27(3)-20

**Для цитирования:** Тарасов В., Мезенцева Е., Карбаев Д. Совмещенный классификатор для фильтрации сообщений на веб сайтах. Труды ИСП РАН, том 27, вып. 3, 2015 г., стр. 291-302 (на английском языке). DOI: 10.15514/ISPRAS-2015-27(3)-20.

## Список литературы

[1]. Е.М. Мезенцева, В.Н. Тарасов. “Организация защиты компьютерных сетей. Метод многомодульной фильтрации спама на web-сайтах,” Информационные технологии, 2012 г., № 6, с.18-22.

[2]. Е.М. Мезенцева. “Программная система распознавания и фильтрации спама на сайтах,” Свидетельство о государственной регистрации программы для ЭВМ № 2011619160, [Роспатент, Москва, 25.11.2011].

[3]. С. М. Никольский. Квадратурные формулы. “Наука”, Москва, 1974. 224 с.

[4]. Е.М. Мезенцева, В.Н. Тарасов. “Защита компьютерных сетей. Веб программирование многомодульного спам фильтра,” Программная инженерия, 2012 г., № 4, с. 27-32.

[5]. Е.М. Мезенцева, В.Н. Тарасов. “Построение оптимального спам фильтра на основе совмещения статистических классификаторов,” Инфокоммуникационные технологии, том 1, 2013г., № 4, с.53-57.