THE SPAM FILTRATION BY MEANS OF PROBABLE METHOD

Volodymyr Mesyura[1], Vyacheslav Sedletskyy[2], Bohdan Vlasyuk[3]

[1]Vinnytsia National Technical University
Khmelnytske shosse, 95, Vinnytsia, 21021, Ukraine, Tel.: 8 (0432) 43-78-80,
E-mail: vimes@vstu.vinnica.ua[1], slawik@bodq.org.ua[2], bohdan@bodq.org.ua[3]

Abstract

The distribution of advertising postals and spam is enough widespread phenomenon in the World network Internet? That results in charges as time on the treatment of evry correspondence sheet as money? Which an user spends on connecting with Internet. There are enough methods and measures for fighting with such unpleasant phenomenon. The offered method is based on the appropriation of spam belonging probability to evry element the title and the body of sheet and by means of full probability calculation determine sheet belonging to spam.

Introduction

With development of scientific and technical progress and newest technologies, which has caused introduction in all orbs of human activity intellectual, programm and simple мультиагентних of systems, which considerably has facilitated work of the user, having made it it is easier and more pleasantly [1]. For this reason impossible began to use the computer, not using for want of it in work multiagent systems. One from many examples of use is the dialogue by telephone with the help of Internet (IP- telephony) or programmed agents, which work with the computer mail and other.

The majority of problems, to which the use of the computer mail (spam, distribution of network viruses, diverse attacks on confidentiality of the letters etc.) is connected, is connected to a unsufficient guard of modern mail systems. It is necessary to deal with these problems and users of readily available public systems and organizations. The practice shows, that the instantaneous solution of a problem of a guard of the computer mail is impossible [2].

The premises of some problems connected immediately to confidentiality of the mail messages, were mortgaged for want of origin of the computer mail of three decades back and consists in the following [1,2]:

- Any from the standard mail protocols (SMTP, POP3, IMAP4) does not include mechanisms of a guard, which would guarantee confidentiality of a rewriting;
- Absence of a reliable guard of the protocols allows to create the letters with false addresses. Probably to not be sure on 100 % in the one who is of an appropriate real by the author of the letter;
- Electronic letters are easy for changing. The standard letter does not contain means of check of an own wholeness and for want of to transfer through a huge set of servers, can be read and is changed; the electronic letter similar today on a card;
- It is usual in work of the computer mail there are no warranties of delivery of the letter.

The solution of a part of problems connected to use of the computer mail, bases on application of reliable and high-power methods both means rather отслежавания and blocking of mail streams, which contain спам, viruses and commercial dispatch.

The offered method is based on assignment to each element of the electronic letter of a parameter θ, that characterizes probability of a membership of an element to one from elements спаму. By an example of such membership the words such as commercial, busness, significance of colour of the text (#ffff00) ect. The parameter θ is considered as magnitude casual with a known a priori density function f (θ). Thus, the set of frequency functions is set not simply which are supposed, from which it is necessary to select one, and are set their apriory probable scales [3]. Aposteriori probable scales θ is calculated on selection $x_1,\ldots, x_N$, obtained in the correspondence with a denseness f (x).

The estimation of density function is determined as continuous mixture of densities with aposteriori probable scales

$$f[x(x_1,\ldots,x_N)] = \int_{A\theta} f(x/\theta) f[\theta/(x_1,\ldots,x_N)]d\theta$$

where Aθ - area of possible significances θ [3].

Aposteriori density function θ is determined by the Buyes's rule [3]:
Using full probability formula [4] for all electronic sheet elements:

$$f[\theta / x_1,...,x_N)] = \prod_{i=1}^{N} f(x_i / \theta) f(\theta) \Big/ \left(\int_{A\theta} \prod_{i=1}^{N} f(x_i / \theta) f(\theta) d\theta \right)$$

$$\mathbb{P}_{(A)} = \sum_{k=1}^{n} \mathbb{P}_{(A|B^k)} \cdot \mathbb{P}_{(B^k)}.$$

and comparing them with estimated we conclude a type of the letter. If the a posteriori full probability exceeds the a priori probability is given, then the letter is noticed as спам and on the contrary, if smaller, the letter is considered normal.

The defect of such approach consists in vast mathematical accounts and the time of processing of one letter is considerably increased.

The advantage of the given method is more objective analysis of the electronic letter, which is stipulated by the analysis and testing of each element of the letter, allows to increase percent(interest) of a correct filtration of the letters.

References:

[1] Materials of the electronic library InfoCity. (www.infocity.ru)
[2] Materials of the information serever www.citforum.ru
[3] Горелик А.Л. Методы распознавания. М., "Высшая школа", 1986 г.
[4] Теория вероятностей - проф. Топчий В.А., Дворкин П.Л., проф. Ватутин В.А., Леонов И.В., Печурин А.В., Нелин Д.А., ОФИМ СО РАН, 1999