SPAM AND PHISHING DETECTION IN VARIOUS LANGUAGES

Liana Ermakova

Abstract*: The majority of existing spam filtering techniques suffers from several serious disadvantages. Some of them provide many false positives. The others are suitable only for email filtering and may not be used in IM and social networks. Therefore content methods seem to be more efficient. One of them is based on signature retrieval. However it is not change resistant. There are enhancements (e.g. checksums) but they are extremely time and resource consuming. That is why the main objective of this research is to develop a transforming message detection method. To this end we have compared spam in various languages, namely English, French, Russian and Italian. For each language the number of examined messages including spam and notspam was about 1000. 135 quantitative features have been retrieved. Almost all these features do not depend on the language. They underlie the first step of the algorithm based on support vector machine. The next stage is to test the obtained results applying trigram approach. Proposed phishing detection technique is also based on SVM. Quantitative characteristics, message structure and key words are used as features. The obtaining results indicate the efficiency of the suggested approach.*

Keywords*: spam, corpus linguistics, phishing, filtering, text categorization.*

ACM Classification Keywords*: I.2.7 Text analysis*

Introduction

Kaspersky Lab defines spam in the following way:

Spam is unsolicited anonymous mass email [Kaspersky Lab, 2010].

According to Kaspersky lab, in the last quarter of 2010 spam made up 77.1% of total email traffic [Наместникова, 2011]. It should also be mentioned that Russian spam became more carefully designed: more spam messages have the HTML format [Kaspersky Lab, 2010]. Nowadays spam concerns not only email, but also social networks, instant messaging (IM) and other systems. Traditional approaches such as blacklisting and message header analysis are efficient enough for email filtering. Though, they fail to deal with spam in social networks, IM and forums. In this case content and link analyses seem to be more effective. Moreover the last two ones may be applied to identify phishing.

State-of-the-arte

Spam appeared in the nineties of the XX century. Firstly, spam was sent from proper spammers' addresses. The earliest messages were similar. That spam is easy to filter. Content analysis development forced spam to evolve. All messages became different. One of the ways to do it is to add an address to the beginning of a letter (e.g. adding «Hello, joe!» to the message to joe@user.com). The trick may be detected by applying fuzzy signature or statistical learning filters (like Bayesian filtering). A message may begin or end with an extract from classical literature or a sequence of random words. HTML message may contain an unreadable text (e.g. printed in very small font or the same color as the background). These additions provide obstacles to fuzzy signature and statistical filters. In response new techniques appear such as quotation searching and detailed HTML parsing. Usually it is possible to detect spammer's trick it-self and classify a message as spam without detailed content analysis. An advertisement may be sent as a picture. Therefore image analysis techniques which enable to retrieve a text from a picture are used.

Transforming messages are messages which have the same meaning but different forms. Every message looks like a connected text. Only if one has a number of these letters it is possible to establish a paraphrasing fact.

Nowadays the major part of junk emails is delivered from compromised user machines. The most widely used tricks are transforming messages, spam sent as a graphic attachment and unreadable text addition. And not all spam filters can deal with them [Kaspersky Lab, 2009].

Yandex divides spam detections methods into two categories:

- Techniques based on text samples (it is difficult to make them and to keep them up to date);
- Manual analysis and email monitoring (e.g. signature approach) [Yandex, 2010].

Yandex uses, inter alia, white listing [Yandex, 2010]. This approach suffers from some serious disadvantages. In the systems with authorization mechanism it is not so easy to send a message to a user for the first time. Moreover, the practice indicates that white listing is not efficient in IM (e.g. qip, icq) and social networks (e.g. ВКонтакте, Facebook) as far as there the lager half of spam is distributed from the accounts of authorized people. Some researchers believe that spam may be filtered only by end user [Сегалович, 2010]. According to another survey conducted by Yandex, 40% of the respondents have difficulties in distinguishing spam from legal mail [Kaspersky Lab, 2009].

Today the improvement of signature methods seems to be crucial. There are two basic approaches:

- Syntactical (i.e. operating with word chains);
- Lexical (i.e. operating with dictionary) (e.g. key words) [Yandex, 2010].

In current syntactical methods based on shingles (i.e. contiguous subsequences of tokens in a document) [Broder, 2003] [Manber, 1994], for each shingle a check sum is computed and then a random sample is constructed from this set. Shingles make it possible to find similar texts rather reliably. However, real-world

problems, such as spam filtering, require too many shingles and consequently too many resources in order to cluster messages [Yandex, 2010].

The major drawback of every lexical method is that it may be applied only to a single language.

Peculiarities of Spam in Various Languages

Spam classification may be made in terms of two criteria: by structure and by subject. Spam may be divided by structure into three types:

- Spam disguising as legal mass mailing;
- Spam disguising as a personal message;
- Advertising spam.

Regardless of the language, advertisement is spam dominating subject, especially medicine, tourism and education offers. English courses are very popular in non-English-speaking countries. Other subjects such as cheap software and pornography are common for various countries.

Advertising spam disguises as legal mass mailing and contains many links (especially French spam) and words related to a commerce sector. It often begins with an exclamatory or interrogative sentence. Bulleted and numbered lists are also common features of spam in various languages. Nevertheless these features may not be used for spam filtering since they occur in legal mass mails.

Another popular subject is easy money (Internet casino, lottery and so on). Sometimes it is related to phishing and identity theft as well as Nigerian scam. The latter resembles personal mail and is difficult to be filtered. Nigerian scam in French is designed according to the rules of business correspondence. However official letters usually contain an expression «à l'attention de» with a position and/or a name, while in spam one can see «à votre attention». There are a lot of email addresses in business correspondence as well as in phishing. The fraud is that a user may respond to a spam message. In this case the spammer will know that the email is active. The share of spam disguising personal messages is comparatively small. However it is necessary to take them seriously because legal messages can be lost.

French spam is more carefully designed than English and especially Russian ones. Usually it has HTML format therefore there are phrases like "Si ce mailling ne s'affiche pas correctement". Sometimes spammers suggest unsubscribing ("Cliquez ici pour ne plus recevoir nos emails"). If a person clicks on this link the spammer will know that this e-mail address is active and as a result the person will receive more spam or even download a virus. Sometimes spammers "explain" why people receive spam ("Vous êtes inscrit sur", "You are receiving this message because"). Due to perception peculiarity verb forms such as imperative, future simple and present are widespread in spam unlike solicited messages. Direct Impératif is not enough polite. Spammers try to control

readers and that is why imperative usually occurs in the aim of a junk mail («push the button now», «achetez maintenant»). An action in indicative mood is thought as a real one.

Many Anglicisms can be found in French and Russian spam. French spam contains less pronouns and possessive determinative than legal messages. There is not such a tendency in the Russian and English languages.

All types of spam appeal to feelings (curiosity, covetousness, laziness, credulity, boredom etc.). Spam features may appear according to subject, structure or aim of a message.

Methods of Message Transforming

Transliteration is often used in Russian spam. Besides, there are a lot of deliberate word distortions (e.g. unnecessary symbols, deliberate misprints, Latin letters in Cyrillic text etc.). However these features do not definitely indicates spam. Sometimes transliteration is used by emigrants and travellers for lack of Russian keyboard layout. Encoding problems may appear. People often apply different transliteration rules. In this case a human being may easily read a message but it is difficult to perform an automated analysis.

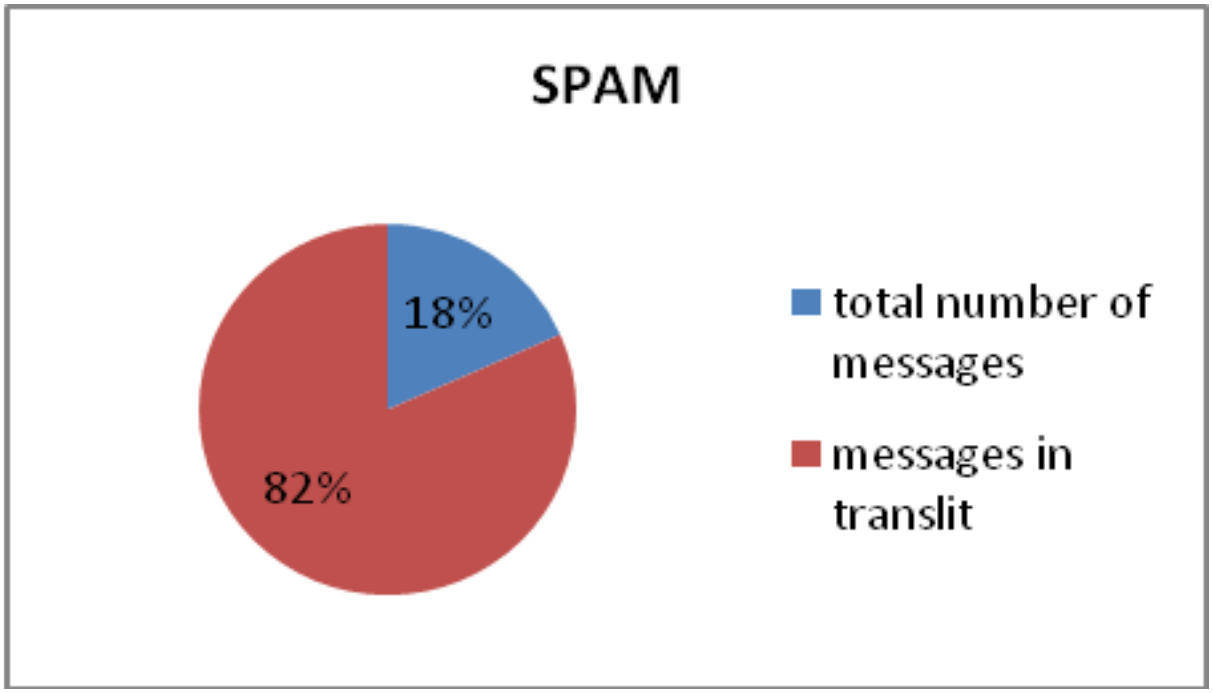

Fig. 1. Share of letter written in transliteration in spam

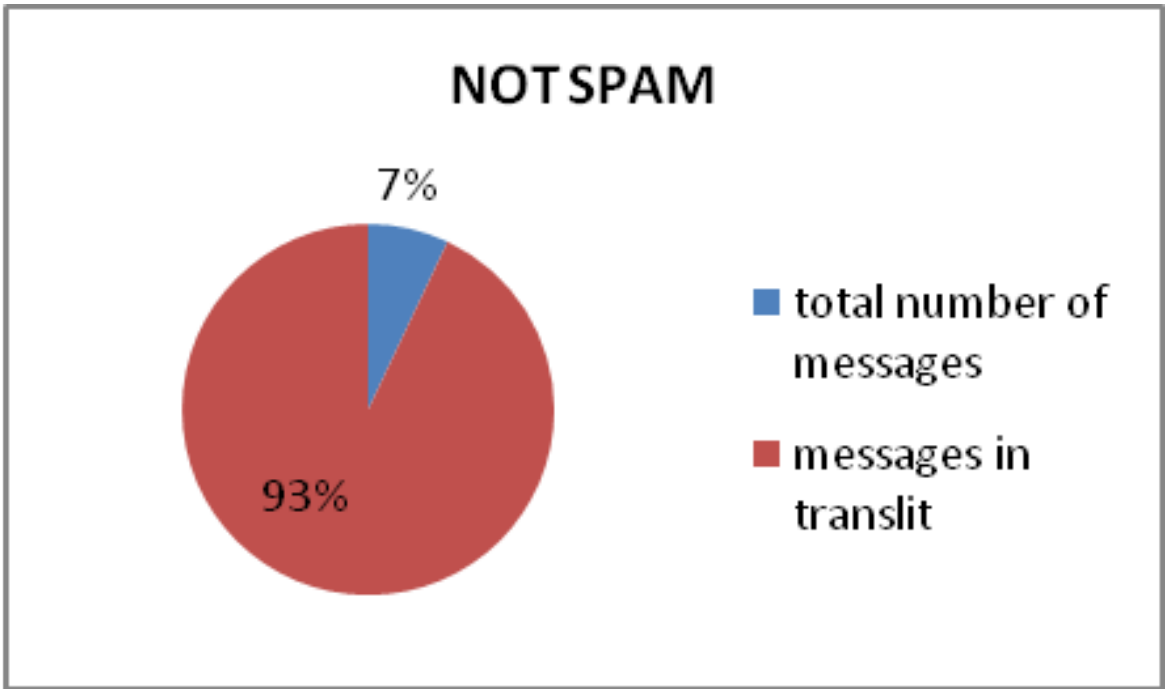

Fig. 2. Share of letter written in transliteration in legal messages

Spam	Not spam
pRODAVA email BAZ pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) eSLI wY OBLADAETE SOBSTWENNYMI INSTRUMENTAMI PROWEDENIQ email RASSYLOK, TO DLQ wAS MY MOVEM PREDLOVITX BAZY DANNYH SOBSTWENNOGO SBORA. <…> cENA ZA 1 MLN. - 50$ cENA ZA WS@ BAZU - 500$ <…>PO L@BYM WOPROSAM: tELEFON:	Privet , zolotze. Nakonez-to posylayu tebe fotki. Ya vybrala nemnozhko bolshe , chto-by ty vybrala kakie hochesh i posmeyalas nemnozhko. Ya kogda smotrela, u menya srazu podnyalos nastroenie. Vse- taki my klassno s toboj syezdili v Ust- Kachku. Esli hochesh, ya tebe vse ostalnye tozhe pereshlu. Pishu tebe iz doma pervyj raz. Ladno, pobezhala delat chto- nibud. A - to zeloe utro za kompiuterom sizhu. Lublu, zeluyu. Mame i koshkam privet!

Here are some examples of transforming messages written in transliteration.

sWEVIE email BAZY pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает потребность в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Предлагаем Вам: получение выписки ЕГРЮЛ за 1,2ОО рублей справки Госкомстата за 1 200 руб. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2.0ОО рублей Доставка курьером, оплата по факту. Контактная информация + 7495 222+07.68
sWEVIE email BAZY pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает необходимость в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Мы предлагаем Вам: получение выписки ЕГРЮЛ за 1 2ОО рублей справки Госкомстата за 1 тыс. 200 р. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2 тыс. 0ОО руб-й. Доставка курьером, оплата по факту. Телефон: + 7495 222_07;68
aDRESA DLQ email RASSYLOK pRODAVA BAZ email ADRESOW (ADRESA DLQ email RASSYLOK) <...>	В начале года всегда возникает потребность в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Мы предлагаем Вам: получение выписки ЕГРЮЛ за 1 тыс. 200 руб-й справки Госкомстата за 1 200 рублей. заказ выписки ЕГРЮЛ + справки Госкомстата составит всего 2,000 р. Доставка курьером, оплата по факту. Контакты + 7(495) 222-07-68

In Russian spam one can find a lot of "spammers' tricks":

- Substitution of letters by digits and vice versa (4-ч, 0-о, 3-з, 1-l);
- Substitution of Cyrillic symbols by similar Latin letters (k-к, a-а, H –Н и т.д.);
- Unnecessary symbols and blanks («Вы хотите ве рнуть вашего любимо го челове ка навсегда и полность ю избавиться от измен?»);
- Interchanging of different symbols (e.g., in telephone number).It is important to mention another transformation technique, namely synonymous expressions (sWEVIE email BAZY = sWEVIE email BAZY = aDRESA DLQ email RASSYLOK, Предлагаем Вам = Мы предлагаем Вам, необходимость= потребность).

It happens that only an address or a link transforms:

<...> La preghiamo di rispondere solo alla mia personale e-mail:khhaykanush@yahoo.com Tua amica Haykanush.
<...>La preghiamo di rispondere solo alla mia personale e-mail:haykanusharm@yahoo.com Tua amica Haykanush.
<...>La preghiamo di rispondere solo alla mia personale e-mail:khaykanush@yahoo.com Tua amica Haykanush.

Medicine advertisement is the most changeable. Both a subject and a text transform. They may even substitute each other. Usually all links are different (they are automatically created in free hosts). Meanwhile sense is the same.

Subject	Text
Desire to impress and please your lover tonight	The only bluepill you need to get bigger python. http://wanzulkifli.com/c6ave6lc.html
Gain in size and win your wife's addiction	Desire to act like a pornstar? Bang a magicpilule! http://bpyasociados.com.ar/9vh6w3lf.html
Wish to act like a porn-director Nail a blu colored med!	0% amorous failure risk http://mikloswowmobile.com/uaagzeib.html
Dream to act like a porn-director Bang a blu colored pil!	Long manliness is great http://antalyagunlugu.com/d4zz8qan.html

The same can be said about casino. It should be noticed that French and English spam is more intricate than Russian and Italian one; especially it concerns such areas as casino, medicine, stock market games, porno and software. In Spanish there are almost no transformations.

Subject	Text
Comme Faire _200 de _20 - nous APPRENDRONS	Bonne journee Jessikaparsons, {http://yxaqih983.o-f.com/kerizev.html} Accueillez la fortune dans votre vie avec de grandes opportunitйs de gagner, avec l'assurance que vos informations personnelles sont protйgйes et vos gains seront payйs rapidement. Une demi-heure et Ђ200 dans ta poche
Gagner _100 pour une demi-heure c'est rйel	Du jour reussi Shirley_patel, {http://gamingworldshop.ru} Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de faзons de gagner. Faire Ђ100 pour une demi-heure - Apprendre?
Faire -100 pour une demi-heure - Apprendre	Bonne journee Nvshamshik, {http://beluwulod.maddsites.com/abimogek.html} Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de faГons de gagner. Gagner -100 pour une demi-heure c'est rИel
Jouer ici, c'est le bonheur ! Telechargez maintenant	{http://opakypiwel.dreamstation.com/jededila.html} On ne peut pas faire plus simple, il suffit de vous inscrire, de faire un versement et vous recevez un fantastique bonus de bienvenue - alors foncez et gagnez ! La meilleure selection de jeu sur internet ! Jouez ici
Jouez plus, gagnez plus	Salut Shea.swan Des options bancaires sыres qui conviendront a tous sont disponibles. Relaxez-vous et soyez certains que vos informations confidentielles sont sйcurisйes et ne seront p& #97;s divulguйes. {http://durl.me/554k6}Comment aimeriez-vous commencer au mieux dans le jeu en ligne avec 1,200 Gratuits? Ils sont dйja a vous, rйclamez-les, jouez et gagnez!

Trigrams in Transforming Message Detection

There are many approaches to find the distance between two documents (e.g. Jaccard coefficient, Hamming distance, edit distance) [Chakrabarti, 2003]. In this research we have used trigram distance.

Traditionally trigrams are used in problems of plagiarism detection [Coulthard, 2004] [Halteren, 2004] and language and encoding identification [Sotnik, 2006] [Cavnar, 1994]. Another group of affiliation methods is based on quantitative text characteristics [Mesheryakov R., 2005] [В.П.Фоменко, 1983] [Рахимова, 2005]. Firstly quantitative features were used in Flesch index and Flesch-Kincaid Index [Галяшина, 2003]. Within the bound of this work these two approaches have been combined. We have used 135 quantitative text features such as share

of content and function words, share of sentences, paragraphs and words of specified length, share of various parts of speech (POS), punctuation marks, co-occurrence of POS etc. Trigram method was modified. Firstly, we have considered as a gram a word and not a symbol. We have examined the sequence of 3 elements in order to determine POS using Zalizniak's grammar dictionary.

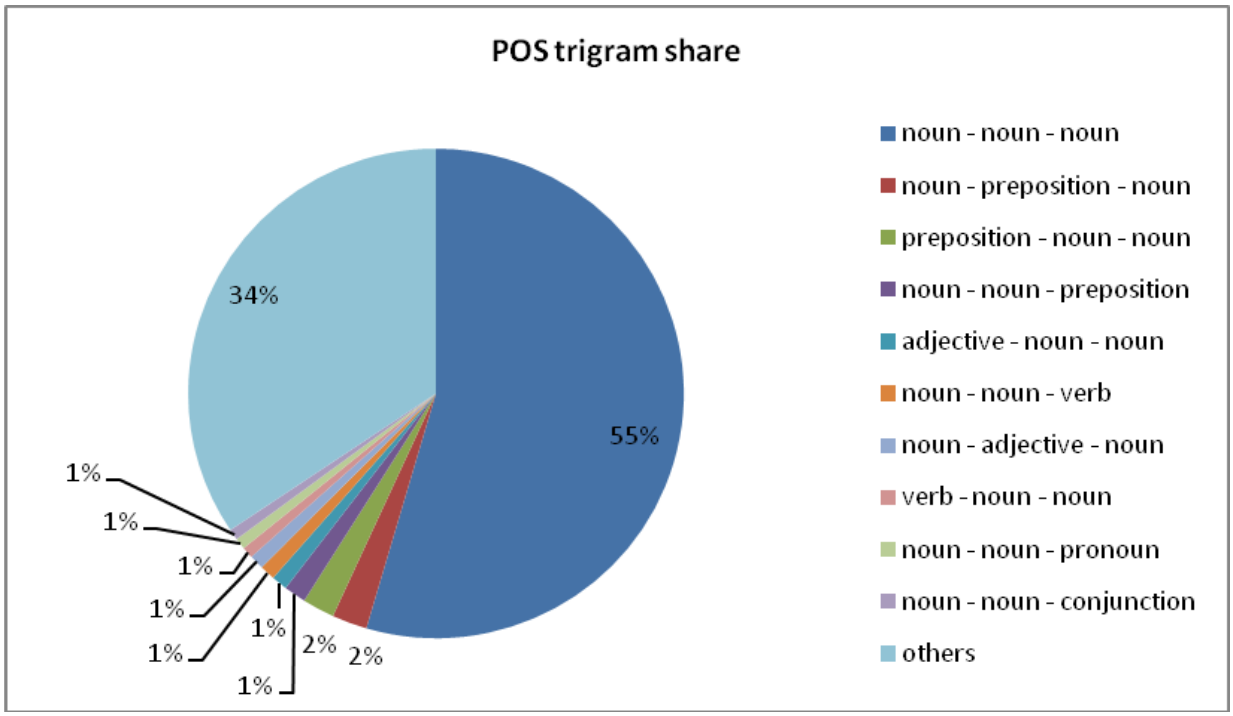

Fig. 3. Share of various POS sequences

Secondly, we have computed similarity of two messages:

similarity=2*NumberOfMatches/(NumberOfTrigramsIn_1_text+NumberOfTrigramsIn_2_text)

This quantity is not normalized. Similarity of Russian and Italian transforming messages is extremely high. Moreover, it varies slightly. Similarity of English and French letters is much smaller and has a large scatter (Fig. 4 - Fig. 9).

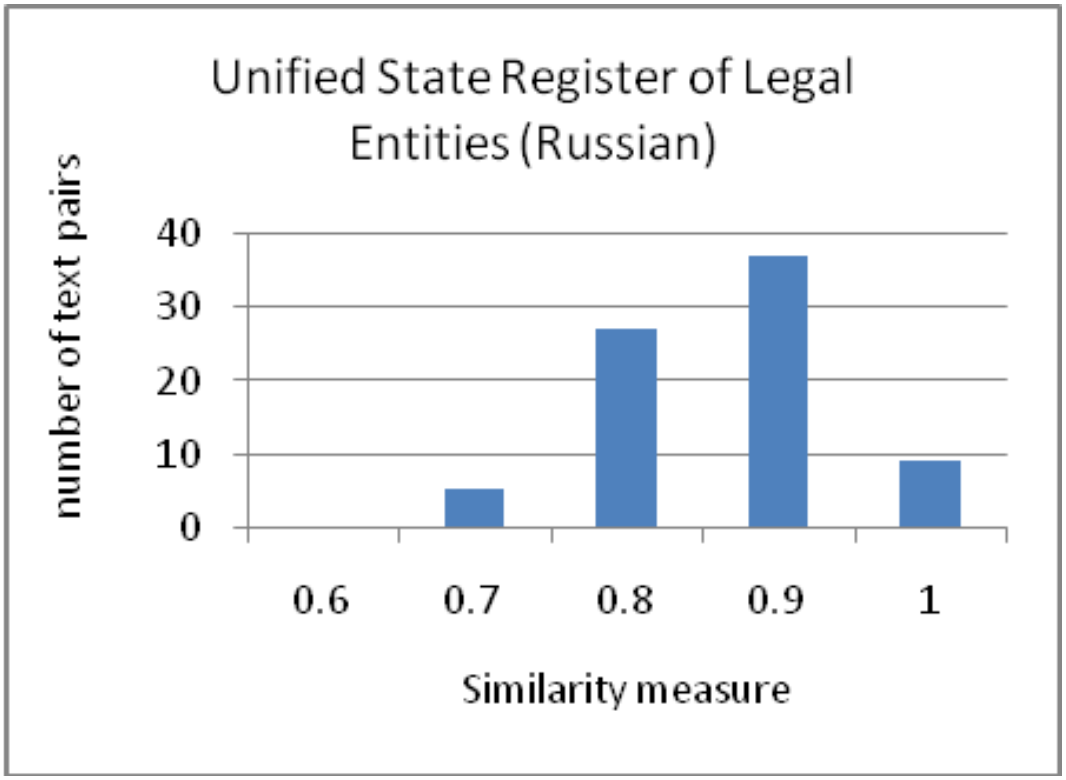

Fig. 3. Trigram similarity measure of "ЕГРЮЛ" mails

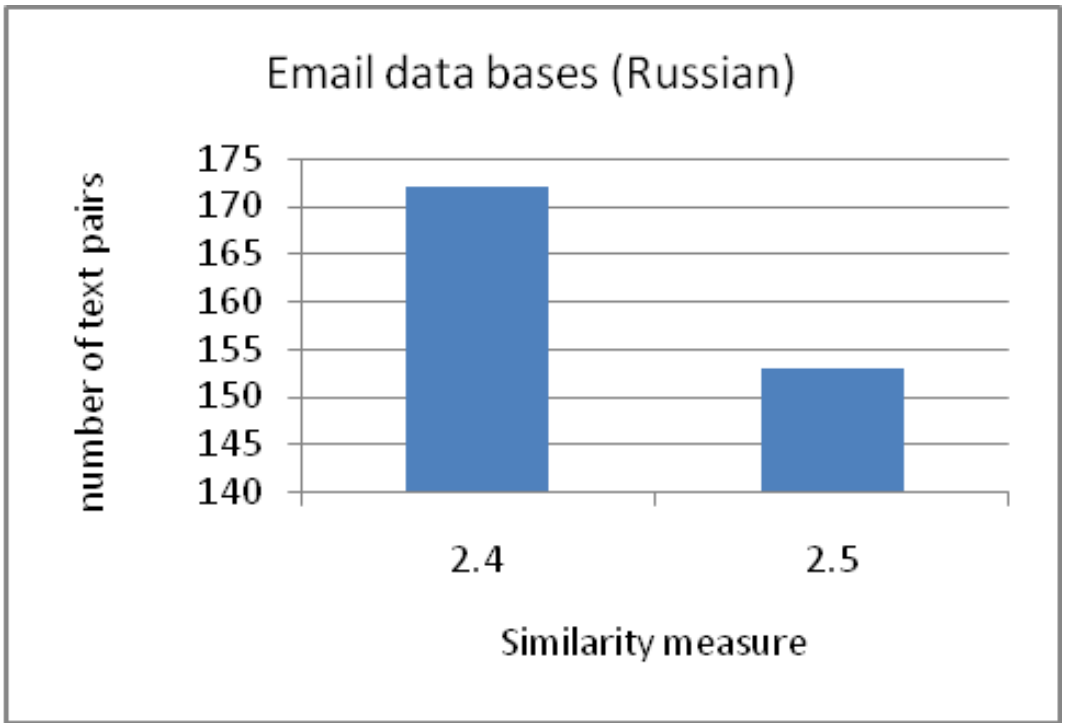

Fig. 4. Trigram similarity measure of "Email базы" mails

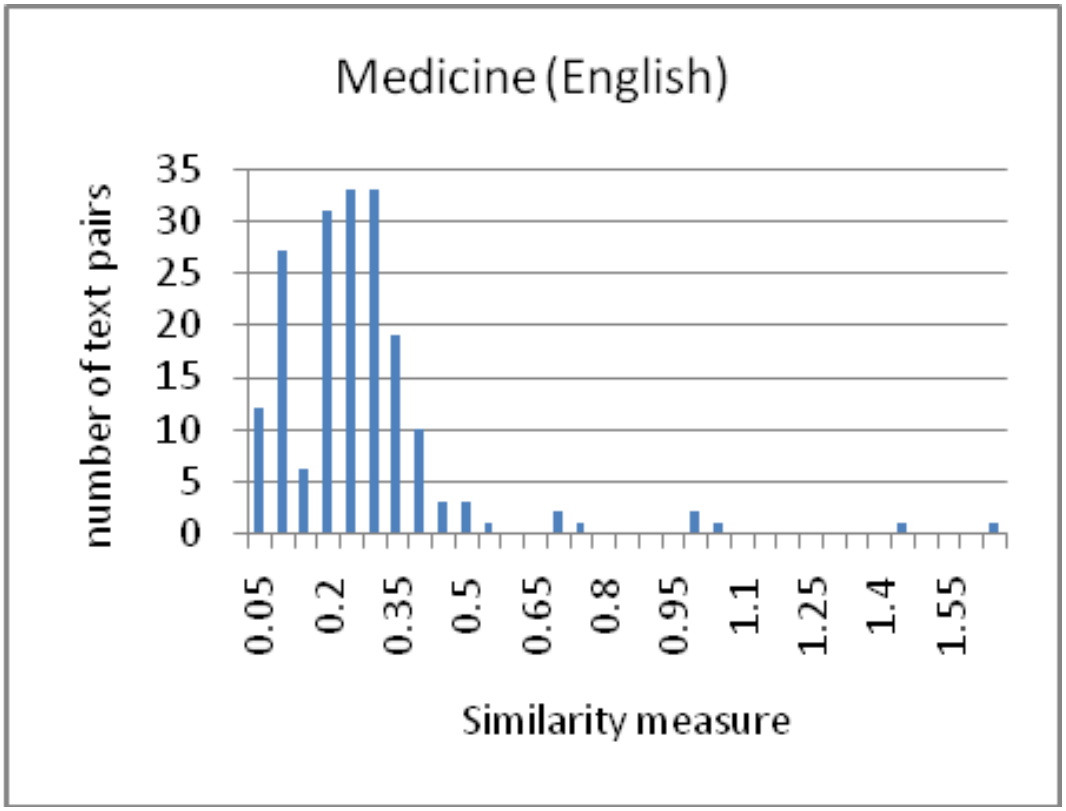

Fig. 5. Trigram similarity measure of "Medicine" mails

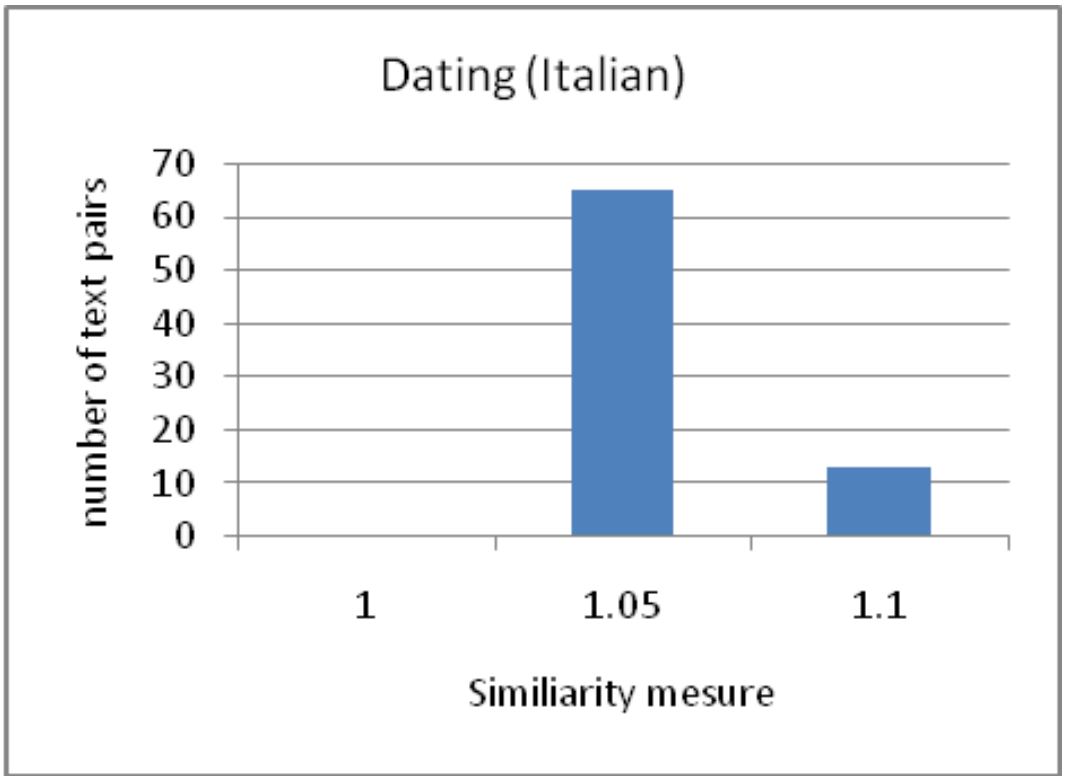

Fig. 6. Trigram similarity measure of "Dating" mails

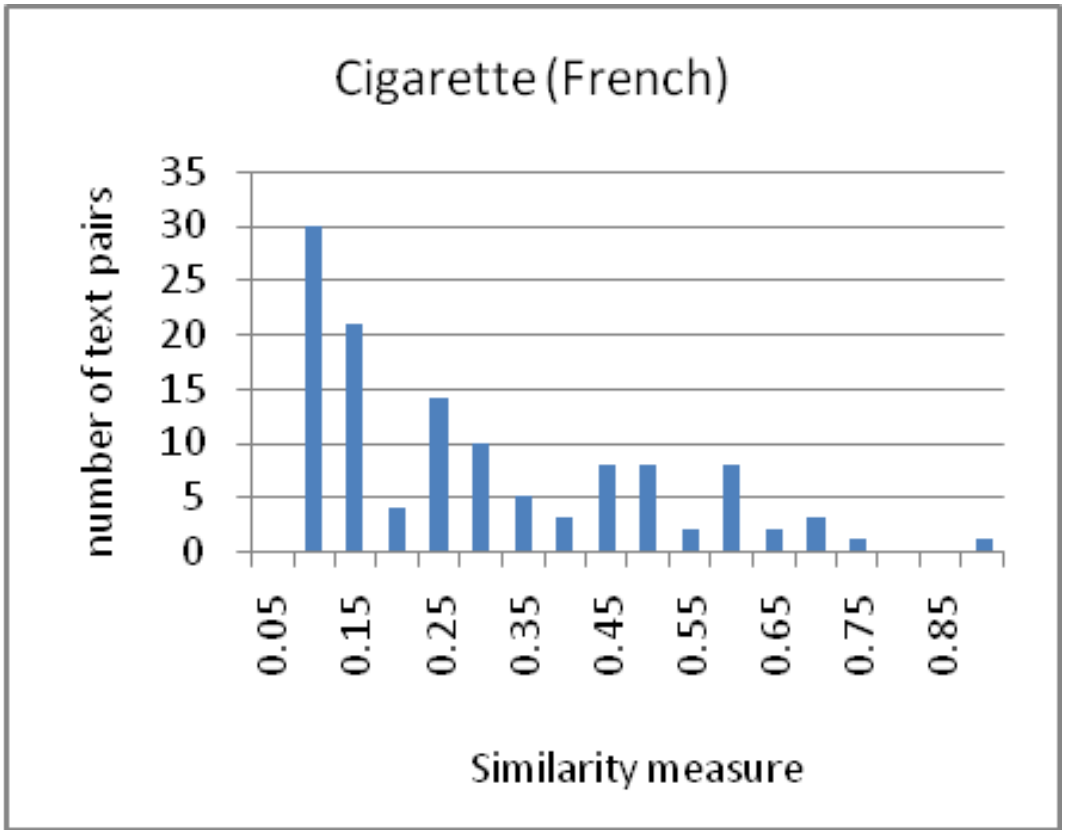

Fig. 7. Trigram similarity measure of "Cigarettes" mails

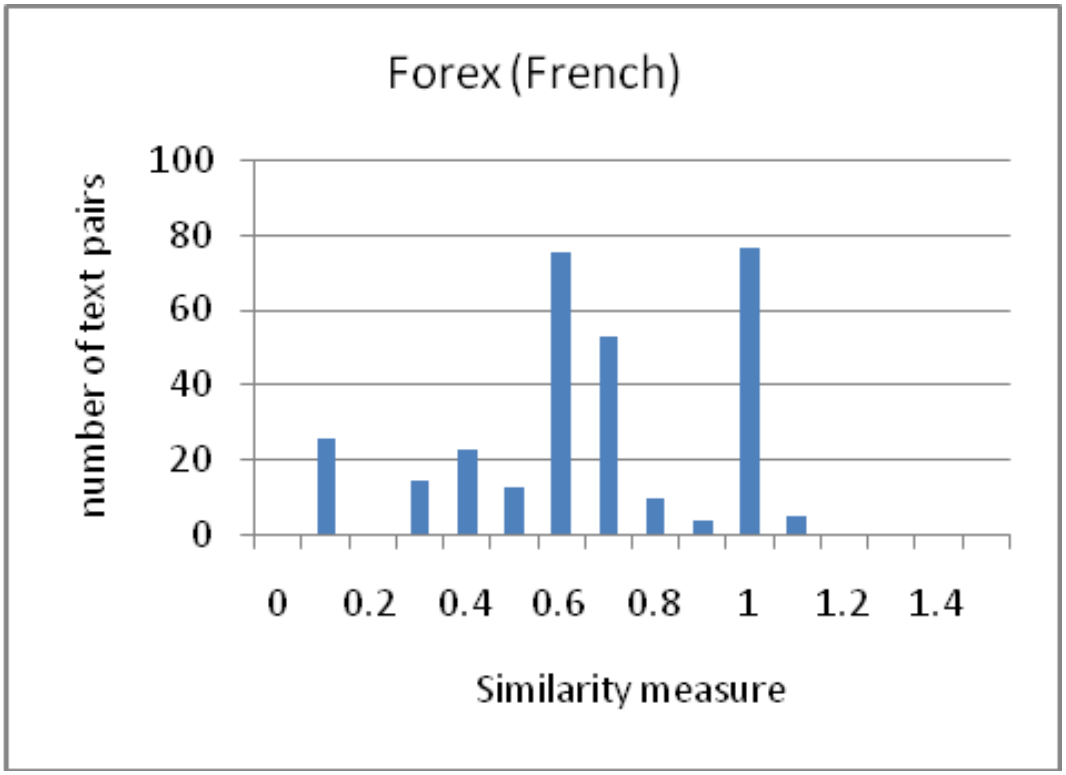

Fig. 8. Trigram similarity measure of "Forex" mails

It seems that trigram approach is not efficient because words may be rearranged. However, even in natural languages with flexible word order (e.g. Russian) there are syntagmatic regularities. Deviations from these regularities perform an emphatic function or make a text difficult to understand. Perception difficulties are not desired in spam since they reduce the response.

As a classifier a support vector machine (SVM) has been chosen. We have used STATISTICA 8.0. Quantitative features of Russian messages enable to identify transforming messages with high accuracy. SVM parameters are given in the Table 1. As we can see SVM detects transformers with extremely high accuracy. However, the results obtained by SVM may be checked by trigram method. It is possible to use other classifiers (e.g. neural networks are quite efficient).

Table 1. SVM parameters for the identification of Russian transforming messages

Sample size = 707 (Train), 236 (Test), 943 (Overall)
Support Vector machine results:
SVM type: Classification type 1 (capacity=10,000)
Kernel type: Radial Basis Function (gamma=0,007)
Number of support vectors = 118 (0 bounded)
Support vectors per class: 94 (0), 16 (1), 8 (2)
Class. accuracy (%) = 100,000(Train), 100,000(Test), 100,000(Overall)

Thus, there are three main steps of transforming messages detection:

1. Quantitative features retrieval;
2. Classification using SVM;
3. Trigram verification.

Phishing detection

According to definition given by Kaspersky Lab, "phishing is a type of Internet fraud that seeks to acquire a user's credentials by deception" [Kaspersky Lab, 2009]. It made up 0,57% in the first quarter of 2010 [Kaspersky Lab, 2010].

We have carried out a survey in order to establish what spam is considered to be, whether people think it is dangerous and how they protect themselves against it. 200 people have taken part in the survey. The respondents have been divided into several categories:

- According to profession/education: related to IT and non-related to IT
- According to age: 0-25, 25-40, more than 40
- According to sex.

Respondents should provide a spam definition and reveal its features. Moreover they have been asked about social engineering and identity theft.

59% of respondents consider spam as a kind of advertisement. 90% think that this phenomenon is not related only to email. 1% believes that any distributed advertisement (e.g. flyer) is spam. People have revealed the following spam indicators:

- 16% - links;
- 16% - unknown sender;
- 2% think that spam is the same thing as phishing;
- 45% consider spam as useless information;
- 11% define spam as nonsense messages.

The majority of respondents are not acquainted with concepts identity theft and social engineering. 50% do not consider social networks and IM to be dangerous.

A special phishing detecting software has been implemented. The algorithm is based on quantitative features and message structure. Moreover, we considered specific vocabulary related to phishing.

Here are the most frequent words occurring in phishing and other messages.

Table 2. Words occurred in phishing as well as in other messages

Word	Occurrence in phishing	Occurrence in notphishing
compte	55	592
paypal	53	3
carte	31	13
informations	29	52
images	24	224
cliquez	19	810
free	19	84
banque	16	38
visa	15	2
cher	13	71
client	13	23

Table 3. Words occurred only in phishing

Word	Occurrence in phishing
verified	10
activer	9
en_us	9
freebox	8
facturation	7
free.fr	7
desjardins	6
caisse	6
xxxx	6
accesd	5
suspendue	5

Table 4. Words not occurred in phishing

Word	Occurrence in notphishing
cliquez	810
compte	592
argent	583
dйmonstration	514
gagner	497
commencez	434
montres	384
йvиnements	372
prix	317
bonus	292
experts	273

securite	13	19
jour	12	106
service	11	63
lien	10	45
mettre	10	4
html	9	1594
passe	9	27
dessous	8	10
postale	8	2
ligne	7	84
information	7	9
confirmer	7	1
connexion	7	1
dernier	6	19
page	6	15
sites	6	8
limite	6	7
login	6	2

suspension	4
temporairement	4
suspendre	4
rappel	4
mesures	4
pixel	4
connecter	4
curitr	4
faveve	3
frauduleux	3
btn_org_arrow	3
contraints	3
bloquйe	3
populaire	3
protegez	3
kunstgeschichte	3
labanquepostale	3
inhabituelles	3
retablir	3

gratuitement	268
annonces	266
formation	263
assistance	256
grвce	251
intйgrale	245
arrкter	245
entraonez	244
toujours	232
images	224
ouvrir	222
atteindre	221
travailler	212
temps	210
sacs	208
bijourama	207
faire	199
forex	199
maintenant	184

As a classifier we also have used SVM algorithm built in STATISTICA 8.0.

Table 5. SVM parameters for phishing identification

Sample size = 994 (Train), 333 (Test), 1327 (Overall)
Support Vector machine results:
SVM type: Classification type 1 (capacity=10,000)
Kernel type: Radial Basis Function (gamma=0,009)
Number of support vectors = 57 (18 bounded)
Support vectors per class: 43 (0), 14 (1)
Class. accuracy (%) = 98,592(Train), 98,498(Test), 98,568(Overall)

Conclusion

Nowadays there are quite a lot of spam filters. Nevertheless, they are not efficient enough or they are very time and resource consuming. The majority of techniques are suitable only for email filtering. In contrast to them content methods may be applied to spam filtering in various message systems (IM, social networks etc.). The improvement of signature methods seems to be topical. The proposed techniques enable to identify transforming messages in a very efficient way. It is not extremely resource consuming as shingle approach and at the same time may be applied for various languages.

The performed survey allows drawing a conclusion that people underestimate threat related to IM, social networks and email. The majority of users are not familiar with the term "phishing". Proposed phishing detection technique is based on SVM. Quantitative characteristics, message structure and key words are used as features. Classification accuracy is above 98%. This approach may be improved by link analysis.

Bibliography

[Kaspersky Lab,2010] Kaspersky Lab, What spam is // Securelist, 2010, http://www.securelist.com/ru/encyclopedia/spam?chapter=151

[Namestnikova M.,2011] Namestnikova M. Spam v dekabre 2010 goda // Securelist, 2011, http://www.securelist.com/ru/analysis/208050676/Spam_v_dekabre_2010_goda

[Kaspersky Lab, 2010] Kaspersky Lab. Spam v pervom kvartale 2010 goda // Kaspersky Lab, 2010, http://www.kaspersky.ru/news?id=207733226

[Kaspersky Lab, 2009] Kaspersky Lab, Spam evolution // Securelist, 2009, http://www.securelist.com/ru/encyclopedia/spam?chapter=155

[Kaspersky Lab,2009] Kaspersky Lab, What is phishing? // Securelist, 2009, http://www.securelist.com/ru/encyclopedia/spam?chapter=155

[Yandex, 2010] Yandex, Nekotorye avtomaticheskie metody detektirovaniya spama dostupnye bolshim pochtovym sistemam // Yandex Company, 2010, http://company.yandex.ru/public/articles/antispam.xml

[Segalovich I., 2010] Segalovich I., Teyblum D., Dilevsky A. Principy i tekhnicheskye metody raboty s nezaprashivaemoy korrespondenciey // Yandex, 2010, http://download.yandex.ru/company/spamooborona-latest.pdf

[Kaspersky Lab,2009] Kaspersky Lab , Spamtest, 2009, http://www.kaspersky.ru/news?id=143937135

[Broder A, 2003] Broder A. On the resemblance and containment of documents // Digital Systems Research Center, 2003, http://ftp.digital.com/pub/Digital/SRC/publications/broder/positano-final-wpnums.pdf

[Manber U,1994] Manber U. Finding similar files in a large file system // USENIX Conference, 1994

[Chakrabarti S.,2003] Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data, 2003

[Coulthard M. 2004] Coulthard M. Author Identification, Idiolect and Linguistic Uniqueness. 2004

[Halteren H.,2004] Halteren H. Linguistic Profiling for Author Recognition and Verification// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004

[Sotnik S.,2006] Sotnik S. Identifikacya yazyka UNICODE teksta po N-grammam dlinoy do 4 vkluchitelno // Matematicheskoye modelirivanie, 2006, p. 111-114

[Cavnar W. B.,1994] Cavnar W. B., Trenkle J. M. N-Gram-Based Text Categorization // Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, 1994, стр. 161-175

[Mesheryakov R.,2005] Mesheryakov R., Vasukov N. Identifikacya avtora metodami iskustvennogo intellekta, 2005

[Fomenko V.,1983] Fomenko V., Fomenko T. Avtorsky invariant russkikh literaturnykh tekstov // Metody kachestvennogo analisa tekstov, 1983

[Rakhimova A.,2005] Rakhimova A. Lingvisticheskaya ekspertisa // Vestnik KASU, 2005

[Galyashina E.,2003] Galyashina E. Osnovy sudebnogo rechevedeniya, 2003

Author' Information

Liana Ermakova *– e-mail: liana87@mail.ru*

Major Fields of Scientific Research: text classification, filtering, information retrieval