УДК 004.896

И.С. Созинова, А.С. Романов, Р.В. Мещеряков

ОПРЕДЕЛЕНИЕ ПОИСКОВОГО СПАМА С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

Созинова И.С., Романов А.С., Мещеряков Р.В. **Определение поискового спама с использованием метода опорных векторов**.
Аннотация. В работе рассматривается классификация поискового спама. Выделяются признаки контента веб-страниц, характерные для спама. Предлагается использование метода опорных векторов для определения принадлежности веб-страницы к спаму. Приводятся результаты проведенных экспериментов.
Ключевые слова: поисковый спам, поисковый запрос, машина опорных векторов.

Sozinova I.S., Romanov A.S., Meshcheryakov R.V. **Search Spam Identification Using Support Vector Machine.**
Abstract. The paper deals with the classification of web spam. It is marked out the features of web pages context, which are typical for spam. It is offered to use the method of support vector machines to figure out whether the web page is the spam. The results of the experiments are given here.
Keywords: search spam, search query, support vector machine.

1. Введение. В настоящее время все большую актуальность приобретают методы и средства обработки текстовой информации в окружающем интеллектуальном пространстве [1] и прикладные аспекты генерации различных искусственных текстов. Одним из таких направлений, активно использующимся в сети Интернет, стали методы генерации поискового спама.

Поисковый спам (далее спам) - это попытки обмана поисковой системы и манипулирования результатами поиска в целях изменения позиции того или иного веб-сайта [2].

В правилах поисковых систем оговорены пункты, согласно которым использование спама запрещено, а веб-сайты, использующие спам могут быть понижены при ранжировании или вовсе исключены из базы поисковой системы.

Негативное последствие спама заключается в существенном ухудшении качества результатов поисковых запросов. Например, по результатам такого запроса в числе первых выдается веб-сайт, в котором содержится минимум полезной для пользователя информации (в лучшем случае плагиат с оригинального веб-сайта) с максимальным количеством страниц, повторяющих друг друга. Такие веб-сайты могут использоваться для переадресации пользователя на другой ресурс, в продвижении которого заинтересован заказчик, или обычного

мошенничества. Кроме того зачастую такие веб-сайты занимают более высокие позиции в поисковой выдаче, чем более информативные для пользователя веб-сайты, продвижение которых в рейтинге осуществляется их владельцами легальными методами поисковой оптимизации без нарушения правил и влияния непосредственно на алгоритм поиска. Последнее оборачивается для владельцев веб-сайтов дополнительными затратами на их продвижение и возможными финансовыми потерями от упущенных клиентов.

Решение данной проблемы лежит на стыке нескольких областей знаний и носит междисциплинарный характер, при этом имеет большое значение в сфере информационной безопасности, защищая интересы как конечных пользователей поисковой системы, так и владельцев веб-сайтов, не относящихся к поисковому спаму.

2. Классификация и методы выявления поискового спама. Общепринятая классификация подразумевает деление спама на два вида: контентный и ссылочный [3]. Ссылочный спам связан с манипулированием внутренними и внешними ссылками на веб-странице и использованием их для перенаправления пользователя на другие веб-сайты. Контентный спам связан с манипулированием содержимым веб-сайта с целью привлечения пользователей.

В соответствии с этой классификацией можно выделить два базовых подхода для проверки результатов поисковой выдачи. Методы выявления ссылочного спама [4] лежат на поверхности, являются хорошо изученными и устоявшимися, в то время как исследование контентного спама требует более глубокого анализа текста веб-страницы и включает несколько аспектов, затрагивающих такие области знания, как статистика, теория вероятностей, лингвистика, морфология, семиотика и пр., а также предполагают наличие знаний об инструментах и методах создания веб-сайтов, их продвижения в сети Интернет [3,5-6]. С точки зрения повышения качества обнаружения спама контентные методы являются более перспективным направлением для исследований.

Существует несколько способов выявления поискового спама, которые подразделяются на его автоматическое детектирование, ручной анализ, а также совместное использование данных методов. Добиться наилучшего результата можно при изучении экспертом веб-страницы на предмет ее принадлежности к поисковому спаму. Но поисковые машины используют автоматическую проверку результатов поисковой выдачи.

Важными направлениями в борьбе с поисковым спамом являются методы обнаружения дубликатов текстов, автоматически сгенерированных и неестественных текстов.

Обзор методов обнаружения дубликатов приведен в работе [7]. В их основе лежит эффективное обнаружение фрагментов скопированных текстов на основе алгоритмов шинглирования.

В основе многих методов обнаружения неестественных текстов лежит подход, предложенный в работе [8]. Этот подход основывается на анализе статистических характеристик текстов и применении машинного обучения для построения автоматического классификатора поискового спама. Развитием данного подхода является работа [9]. В ней предлагается использовать метод скрытого распределения Дирихле для определения спам-текстов.

В работе [10] предлагается подход, основанный на анализе сочетаемости пар слов для обнаружения неестественных текстов. В основе подхода лежит предположение, что неестественные тексты с большей вероятностью содержат редкие пары слов. Авторы предлагают алгоритм для подсчета доли редких пар слов и показывают, что эта характеристика улучшает качество определения поискового спама.

В работе [11] предлагается подход к определению неестественных текстов, в основе которого лежит гипотеза, что такие тексты не могут одновременно удовлетворять всем ограничениям, свойственным естественным текстам. При обучении алгоритма выделяется большое количество статистических признаков, связанных с читаемостью, единством стиля и жанровыми особенностями, которые впоследствии используются в автоматическом классификаторе.

Подход, описанный в работе [12], существенно развивает исследование [11] за счет учета свойств рассматриваемой модели тематической структуры текста для определения неестественных текстов.

Несмотря на существование большого числа методов противодействия спаму, поисковые системы продолжают стабильно выдавать спам-страницы на выходе в определенном проценте случаев (от 1% до 3,5%).

3. Методика идентификации контентного поискового спама с помощью машины опорных векторов. В данной работе разработан подход, связанный с анализом характеристик спам-контента, отличающих его от «легальных» веб-страниц. На основе данных характеристик, аналогично работам [11, 12], связанным с машинным обучением, будет производиться классификация с использованием метода опорных векторов (SVM). Стоит отметить, что данный классификатор показывал отличные результаты при решении авторами смежных задач, связанных с обработкой текстов [13-16].

На рисунках 1-2 представлена методика идентификации контентного поискового спама в виде IDEF0 диаграммы (2 уровня).

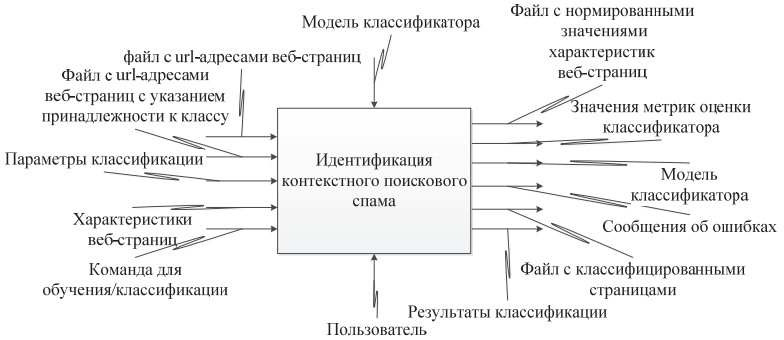

Рис. 1. Схема IDEF0 методики идентификации поискового спама (1 уровень)

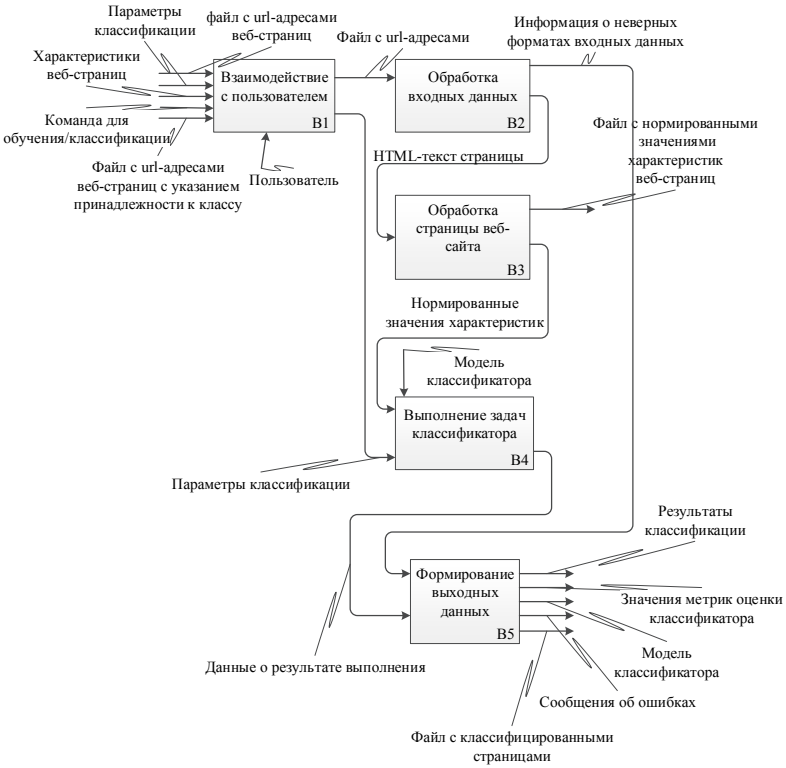

Рис. 2. Схема IDEF0 методики идентификации поискового спама (2 уровень)

Отобранные характеристики с их описаниями приведены в таблице 1.

Таблица 1. Отличительные характеристики контента веб-страницы

№ пп	Характеристика	Описание
	Специфика веб-страницы	
1	Доля текста на странице	Заметно снижается для определенного процента спам-сайтов (берем все, что между тегов)
2	Мета - тег «keywords» (50-80 символов) Количество слов, Плотность слов	Одна из наиболее распространенных характеристик, не показательна для части классического спама.
3	Тег «title» (50-80 символов) Количество слов, Плотность слов	Перенасыщение ключевыми словами заголовка является устойчивой характеристикой для спама
4	Мета-тег «description»: Количество слов, Плотность слов.	Может быть сгенерирован автоматически и перенасыщен ключевыми словами.
5	Анкор: Число ссылок, Без анкоров, Количество внешних, Внешние без анкоров.	Текст анкора может быть перенасыщен ключевыми словами, вследствие чего являться отличительной особенностью для спама
6	Доля анкорного текста	Увеличивается для спама
7	Доля видимого текста	Увеличивается для спама в 50% случаев
	Текстовые характеристики (автоматические/искусственные тексты)	
8	Среднее количество знаков пунктуации на предложение	Требуются дополнительные эксперименты
9	Число слов	Заметно увеличивается для спама
10	Средняя длина слова	Увеличивается для спама
11	Количество длинных слов	Увеличивается для спама
12	Количество знаков экспрессивной пунктуации («!», «?», «…»)	Требуются дополнительные эксперименты
13	Количество слов, начинающихся с заглавной буквы	Требуются дополнительные эксперименты

Метод опорных векторов сводится к задаче оптимизации следующего вида:

$$\begin{cases} \arg\min_{w,b} \|w\|^2, \\ y_i(\langle w, x_i \rangle + b) \geq 1, i = 1, \dots, m \end{cases},$$

которая является стандартной задачей квадратичного программирования и решается с помощью множителей Лагранжа.

Классифицирующая функция F принимает вид:

$$F(x) = sign(\langle w, \varphi(x)\rangle + b).$$

Выражение:

$$k(x, x') = \langle \varphi(x), \varphi(x)' \rangle$$

называется ядром классификатора.

Выбор подходящей ядерной функции – ключевая задача при получении качественной модели регрессии.

Классический SVM классификатор не всегда на практике справляется с задачей классификации:

– когда невозможно провести плоскость, разделяющую классы;

– когда выборка из другого класса оказывается слишком мала.

Для того чтобы избежать данных проблем вводится коэффициент регуляризации C:

$$\min_{\vec{\omega}, \vec{\xi}, b} \frac{1}{2} \vec{\omega} \cdot \vec{\omega} + C \cdot \sum_{i=1}^{n} \xi_i.$$

4. Результаты экспериментов. Последовательность шагов проведения экспериментов для оценки точности классификации приведена ниже:

– выбор характеристик веб-страницы на основе ее особенностей как HTML-объекта, а также наиболее удачных текстовых характеристик;

– создание тестового, а также обучающего наборов url адресов релевантных сайтов и не релевантных веб-страниц;

– генерация спам-страниц для составления обучающей и тестовой спам-базы;

– подсчет интересующих параметров в выборках;

– нормирование параметров выборок в диапазон [0...1];

– обучение модели SVM на данных обучающей выборки;

– подача на вход обученной модели SVM данных тестовых выборок, работа классификатора, считывание результатов;

– подбор параметров C *и* γ методом перебора возможных значений из трех областей: *{C <= 1, 1 < C <= 100, C > 100}, {* γ *<= 1, 1 <* γ *<= 100,* γ *> 100}*. Повтор с шага 5 для каждого выбранного значения. Выбор значений параметров с наилучшим показателем F-меры. Повтор с шага 5 для каждого;

– переопределение набора характеристик веб-страницы; повтор с шага 4 для нового набора характеристик;

– анализ полученных результатов.

На рисунке 3 представлен график зависимости величины f-меры от объема выборки текстов. Вычисление значений f-меры проводилось на основе результатов скользящей проверки с разбиением массива данных на 10 частей. Как видно из графика, классификация для заданных тестовой и обучающей выборок имеет высокие показатели. Полученный результат можно объяснить схожестью групп файлов спама между собой. Небольшие отклонения кривой f-меры укладываются в построенные для точек кривой доверительных интервалов.

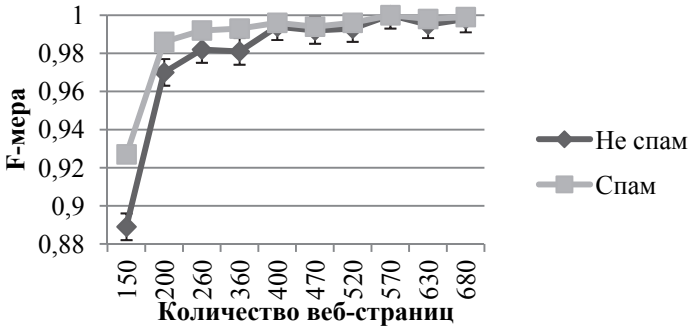

Рис. 3. График зависимости f-мера = f(n выборки)

Были проведены эксперименты по подбору типа ядра классификатора, дающего наиболее высокие показатели. Как видно из гистограммы (рисунок 4), лучше всего с задачей классификации справляется классификатор с заданной функцией сигмоида (SIGMOID), а также радиальной базисной функцией ядра (RBF). Менее значительными являются результаты работы классификатора с полиномиальным (POLY) и линейным (LINEAR) ядром.

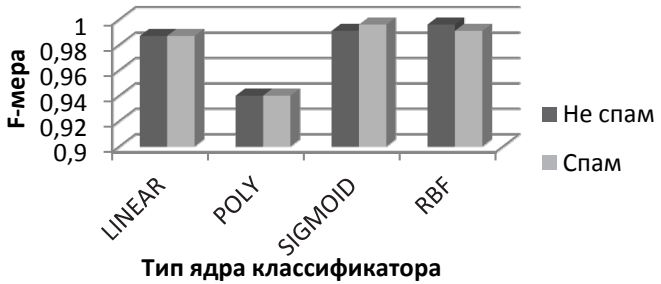

Рис. 4. Гистограмма значений f-меры в зависимости от типа ядра классификатора

В таблице 2 представлен перечень всех рассчитанных метрик для функций ядра: сигмоид и RBF.

Таблица 2. Метрики классификации для двух типов ядер классификатора

	Аккуратность		Точность		Полнота		F-мера	
Ядро	Не спам	Спам	Не спам	Спам	Не спам	Спам	Не спам	Спам
SIGMOID	0,383	0,612	1	0,991	0,986	1	0,993	0,996
RBF	0,357	0,598	1	0,983	0,973	1	0,986	0,991

В ходе экспериментов были подсчитаны метрики классификации, для двух разных групп наборов характеристик веб-страницы (таблица 3). Как видно из гистограммы (рисунок 5) наименее показательными являются характеристики анкоров веб-страницы.

Таблица 3. Группы характеристик веб-страницы

Main	Anchor	Text
Доля текста на странице	Анкор: число ссылок, без анкоров, количество внешних, внешние без анкоров.	Среднее количество знаков пунктуации на предложение
Мета - тег «keywords» (50-80 символов): количество слов, плотность слов.	Доля анкорного текста	Число слов
Тег «title» (50-80 символов) количество слов, плотность слов.	-	Средняя длина слова
Мета-тег «description»: количество слов, плотность слов;	-	Количество длинных слов
Доля видимого текста	-	Количество знаков экспрессивной пунктуации («!», «?», «…»)
-	-	Количество слов, начинающихся с заглавной буквы

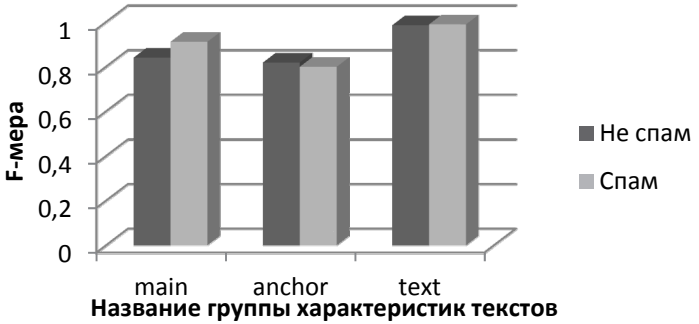

Рис. 5. Гистограмма значений величины f-меры в зависимости от группы исследуемых характеристик

Сравнение результативности методов, основанных на схожем подходе к определению поискового спама, приведено в таблице 4.

Таблица 4. Результаты классификации спама разными методами

	Метрика	Метод обнаружения поискового спама, порожденного с помощью цепей Маркова	Метод контентного анализа характеристик веб-страницы как HTML-объекта	Метод данной работы
Спам	Точность	92,8%	97,8%	98,3%
	Полнота	59,4%	98,7%	100%
	F-мера	72,5%	98,3%	99,1%

Величины метрик классификации в данной работе были взяты для выборки, состоящей из 244 страниц не спама и 120 страниц спама. Полученные результаты в большинстве носят закономерный характер и позволяют сделать вывод о работоспособности предложенного подхода для оценки принадлежности веб-страницы к поисковому спаму. Более того, данная система позволила добиться лучших результатов при использовании аналогичных методов по сравнению с предыдущими исследованиями.

5. Заключение. В ходе данной работы были решены следующие задачи:

– анализ и систематизация знаний о предметной области;

– исследование в рамках методов выявления поискового спама, связанных с анализом текстовой составляющей;

– разработка методики выявления поискового спама. В ходе экспериментов были отобраны 18 характеристик, отличающих

поисковый спам от не спама. Была сформирована обучающая база, состоящая из 244 примеров веб-страниц, не относящихся к спаму, и 120 примеров поискового спама;

– анализ результатов автоматизированной классификации поискового спама.

Показатели точности и полноты работы метода для отобранных методом экспертной оценки данных являются довольно высокими. При объеме выборки, составляющей 244 отобранных веб-страниц в качестве представителей класса «Не спам» и 120 страниц-представителей «Поискового спама», F-мера классификации «Не спама» составила 98,6%, «Спама» - 99,1%.

Для классификации веб-страниц из других наборов данных требуются дополнительные эксперименты с увеличением доли поискового спама в выборках для лучшего обучения классификатора.

В дальнейшем планируется провести модификацию разработанной методики и программной системы с целью улучшения показателей классификации. В частности будут проанализированы дополнительные характеристики веб-страницы, обработка которых пока не реализована в существующей версии программы, а также морфологические и синтаксические характеристики самого текста [17].

Литература

1. *Ronzhin A.L., Karpov A.A.* Russian voice interface // Pattern Recognition and Image Analysis. 2007. vol. 17. no. 2. pp. 321-336.
2. Лицензия на использование поисковой системы Яндекса. URL: http://legal.yandex.ru/termsofuse/ (дата обращения: 13.01.14).
3. *Gyongyi Z., Garcia-Molina H.* Web Spam Taxonomy // Chiba: First International Workshop on Adversarial Information Retrieval on the Web. 2005. URL: http://infolab.stanford.edu/ (дата обращения: 26.01.2014).
4. *Зеленков Ю.Г., Сегалович И.В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Тр. IX Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). Переславль. 2007. Т. 1. С. 166–174.
5. *Золтан Д., Гарсиа-Молина Г.* Таксономия веб-спама // Стенфорд: Кафедра информационных технологий Стенфордского университета. URL: http://wseob.ru/seo/web-spam-taxonomy (дата обращения: 20.02.2014).
6. Дорвей // свободная статья из Википедиа. URL: http://www.webeffector.ru/wiki/Дорвей (дата обращения: 20.11.2013).
7. *Abernethy J., Chapelle O., Castillo C.* WITCH: A new approach to Web spam detection // Proc. Of the 4th Int.Workshop on Adversarial Information Retrieval on the web. Beijing: ACM. 2008. pp. 61–62.
8. *Ntoulas A., Najork M., Manasse M., Fetterly D.* Detecting spam Web pages through content analysis // Proc. Of the 15th Int. Conference on World Wide Web. Edinburgh: ACM. 2006. pp. 83–92.

9. *Biro I., Siklosi D., Szabo J., Benczur A.A.* Linked latent Dirichlet allocation in Web spam filtering // Proc. of the 5th Int. Workshop on Adversarial Information Retrieval on the Web. Madrid: ACM. 2009. pp. 37–40.
10. *Гречников Е.А., Гусев Г., Кустарев А.А., Райгородский А.М.* Поиск неестественных текстов // Тр. XI Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронныеколлекции»: Петрозаводск. 2009. С. 306–308.
11. *Павлов А.С., Добров Б.В.* Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Тр. XI Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск. 2009. С. 311–317.
12. *Павлов А.С., Добров Б.В.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // Тр. XII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск. 2010. С. 210–218.
13. *Романов А.С., Мещеряков Р.В.* Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). М.: РГГУ 2009. Вып. 8 (15). С. 432–437.
14. *Романов А.С., Мещеряков Р.В.* Идентификация авторства коротких текстов методами машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). М.: Изд-во РГГУ 2010. Вып. 9 (16). С. 407–413.
15. *Романов А.С., Мещеряков Р.В.* Определение пола автора короткого электронного сообщения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25 - 29 мая 2011 г.). М.: Изд-во РГГУ. 2011. Вып. 10 (17). С. 620–626.
16. *Романов А.С., Резанова З.И., Мещеряков Р.В.* Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции // Доклады томского государственного университета систем управления и радиоэлектроники. Томск: Издательство Томского государственного университета систем. 2014. № 2(32). С. 264-269.
17. *Karpov A., Kipyatkova I., Ronzhin A.* Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis // In Proc. 12th International Conference INTERSPEECH-2011. ISCA Association. Florence. Italy. 2011. pp. 3161-3164.

References

1. Ronzhin A.L., Karpov A.A. Russian voice interface. Pattern Recognition and Image Analysis. 2007. vol. 17. no. 2. pp. 321-336.
2. Licenzija na ispol'zovanie poiskovoj sistemy Yandeksa [License to use the search engine Yandex]. Available at: http://legal.yandex.ru/termsofuse/ (accessed 26.10.2014) (In Russ.).
3. Gyongyi Z., Garcia-Molina H. Web Spam Taxonomy Chiba: First International Workshop on Adversarial Information Retrieval on the Web. 2005. Available at: http://infolab.stanford.edu/ (accessed 26.10.2014).
4. Zelenkov Ju.G., Segalovich I.V. [Comparative analysis of methods of near-duplicate detection for Web-documents] *Tr. IX conference "Jelektronnye biblioteki:*

perspektivnye metody i tehnologii, jelektronnye kollekcii" [Digital Libraries: Advanced Methods and Technologies] (RCDL'2007). Pereslavl. 2007. vol. 1. pp. 166–174. (In Russ.).
5. Zoltan D., Garsia-Molina G. [Taxonomy of web-SPAM]. Stanford: Dept. information technology Stanford university. Available at: http://wseob.ru/seo/web-spam-taxonomy (accessed 26.10.2014).
6. Dorvej [Doorway Page]. Available at: http://www.webeffector.ru/wiki/Dorvej. (accessed 26.10.2014) (In Rus).
7. Abernethy J., Chapelle O., Castillo C. WITCH: A new approach to Web spam detection. Proc. Of the 4th Int.Workshop on Adversarial Information Retrieval on the web. Beijing: ACM. 2008. P. 61–62.
8. Ntoulas A. M., Najork M., Manasse D., Fetterly D. Detecting spam Web pages through content analysis. Proc. Of the 15th Int. Conference on World Wide Web. Edinburgh: ACM. 2006. pp. 83–92.
9. Biro I., Siklosi D., Szabo J., Benczur A.A. Linked latent Dirichlet allocation in Web spam filtering. Proc. of the 5th Int. Workshop on Adversarial Information Retrieval on the Web. Madrid: ACM. 2009. pp. 37–40.
10. Grechnikov E.A., Gusev G.G., Kustarev A.A., Rajgorodskij A.M. [Search unnatural texts]. *Tr. XI conference "Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii*" [Digital Libraries: Advanced Methods and Technologies]. Petrozavodsk. 2009. pp. 306–308. (In Russ.).
11. Pavlov A.S., Dobrov B.V. [Detecting web spam created with Markov Chains]. *Tr. XI conference "Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii*" [Digital Libraries: Advanced Methods and Technologies]. Petrozavodsk. 2009. pp. 311–317. (In Russ.).
12. Pavlov A.S., Dobrov B.V. [Detection method massively generated unnatural texts based on an analysis of the thematic structure] *Tr. XII conference "Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii*" [Digital Libraries: Advanced Methods and Technologies]. Petrozavodsk. 2010. pp. 210–218. (In Russ.).
13. Romanov A.S., Meshcheryakov R.V. [Identification of the author of text by using support vector machine]. *Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog»* [Computational Linguistics and Intellectual Technologies]». M.: RGGU. 2009. vol. 8(15). pp. 432–437. (In Russ.).
14. Romanov A.S., Meshcheryakov R.V. [The identification of authorship short texts machine learning methods]. *Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog»* [Computational Linguistics and Intellectual Technologies]. M.: Izd-vo RGGU. 2010. vol. 9 (16). pp. 407–413. (In Russ.).
15. Romanov A.S., Meshcheryakov R.V. *Opredelenie pola avtora korotkogo jelektronnogo soobshhenija* [Sexing the author of a short e-mail message]. *Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog»* [Computational Linguistics and Intellectual Technologies]». M.: Izd-vo RGGU. 2011. vol. 10 (17). pp. 620–626. (In Russ.).
16. Romanov A.S., Rezanova Z.I., Meshcheryakov R.V. [Methodology for testing homogeneity of the text and plagiarism detection method based on support vector machines and fast correlation filter]. *Doklady TUSUR – Reports of TUSUR*. Tomsk: Izdatel'stvo Tomskogo gosudarstvennogo universiteta sistem. 2014. vol. 2(32). pp. 264-269. (In Russ.).

17. Karpov A., Kipyatkova I., Ronzhin A. Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis. In Proc. 12th International Conference INTERSPEECH-2011. ISCA Association. Florence. Italy. 2011. pp. 3161-3164.

Созинова Ирина Сергеевна — инженер кафедры Комплексной информационной безопасности электронно-вычислительных систем ТУСУР. Область научных интересов: информационная безопасность. Число научных публикаций — 5. irishechka7371@gmail.com@gmail.com; 634045, г. Томск, ул. Красноармейская 146, ауд. 509; р.т. +7 (3822) 900-111, факс +7 (3822) 900-111.

Sozinova Irina Sergeevna — engineer, Dept. of Complex Security of Electronic-computing Systems of Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: information security. The number of publications — 5. alexx.romanov@gmail.com; KIBEVS Dept. TUSUR, 146, Red Army street Tomsk, 634045, Russia; office phone +7(3822)900-111, fax +7(3822)900-111.

Романов Александр Сергеевич — к-т техн. наук, доцент кафедры Комплексной информационной безопасности электронно-вычислительных систем ТУСУР. Область научных интересов: информационная безопасность, интеллектуальный анализ данных, искусственный интеллект, обработка текста. Число научных публикаций — 41. alexx.romanov@gmail.com; 634045, г. Томск, ул. Красноармейская 146, ауд. 509; р.т. +7(3822) 900-111, факс +7 (3822) 900-111.

Romanov Aleksandr Sergeevich — Ph.D., associate professor, Dept. of Complex Security of Electronic-computing Systems of Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: information security, data mining, AI, word processing. The number of publications — 41. alexx.romanov@gmail.com; KIBEVS Dept. TUSUR, 146, Red Army street Tomsk, 634045, Russia; office phone +7(3822)900-111, fax +7(3822)900-111.

Мещеряков Роман Валерьевич — д-р техн. наук, профессор кафедры Комплексной информационной безопасности электронно-вычислительных систем ТУСУР. Область научных интересов: системный анализ, информационная безопасность, вопросы обработки информации в интеллектуальных системах, особое внимание уделяется вопросам создания информационно-безопасных систем. Число научных публикаций — 247. mrv@security.tomsk.ru; 634050, г. Томск, пр. Ленина, 40, ауд. 210; р.т. +7(3822)900111, факс +7 (3822) 900-111.

Meshcheriakov Roman Valerievich — Ph.D., Dr. Sci., professor, Dept. of Complex Security of Electronic-computing Systems of Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: speech analysis, speech recognition, medical technology, information security. The number of publications — 247. mrv@security.tomsk.ru; KIBEVS Dept. TUSUR, 40, Lenin-avenue Tomsk, 634050, Russia; office phone +7(3822)413426, fax +7(3822)900-111.

РЕФЕРАТ

Созинова И.С., Романов А.С., Мещеряков Р.В. **Определение поискового спама с использованием метода опорных векторов.**

Поисковый спам – это попытки обмана поисковой системы и манипулирования результатами поиска в целях изменения позиции того или иного веб-сайта. Решение данной проблемы лежит на стыке нескольких областей знаний и носит междисциплинарный характер, при этом имеет большое значение в сфере информационной безопасности, защищая интересы как конечных пользователей поисковой системы, так и владельцев веб-сайтов, не относящихся к поисковому спаму.

Общепринятая классификация подразумевает деление спама на два вида: контентный и ссылочный. Важными направлениями в борьбе с поисковым спамом являются методы обнаружения дубликатов текстов, автоматически сгенерированных и неестественных текстов.

В данной работе используется подход, связанный с углублением в характеристики спам-контента, отличающих его «легальных» веб-страниц. В ходе экспериментов были отобраны 18 характеристик, отличающих поисковый спам от не спама. На основе данных характеристик производится классификация с использованием метода опорных векторов (SVM). Показатели точности и полноты работы метода для отобранных методом экспертной оценки данных являются довольно высокими. При объеме выборки, составляющей 244 отобранных веб-страниц в качестве представителей класса «Не спам» и 120 страниц-представителей «Поискового спама», F-мера классификации «Не спама» составила 98,6%, «Спама» – 99,1%.

SUMMARY

Sozinova I.S., Romanov A.S., Meshcheryakov R.V. **Search engine spam detection using support vector machine.**

Search engine spam is attempts to deceive search engine and manipulate search results in order to change the position of a website. The solution of this problem lies at the intersection of several disciplines and interdisciplinary character, and has a great importance in the field of information security, protecting the interests of the search engine end-user and website owners, not related to search engine spam.

Common classification implies the division of spam into two types: a content and reference. Important areas in the fight against search spam are detection methods of duplicate texts, automatically generated and unnatural texts.

In this paper we use the approach of a detailed examination of the characteristics of spam content, distinguishing it from "legal" web pages. During the experiments were selected 18 characteristics that distinguish search engine spam from not spam. Based on these characteristics, we classify web pages using support vector machine (SVM). Accuracy and precision of the method are quite high. When the sample size is 244 selected web pages as representatives of a class "Not Spam" and 120 pages, representatives of the "search spam», F-measure of classification "Not Spam" amounted to 98.6%, "spam" – 99.1%.