КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 681.513.7

ПРОГРАММНАЯ СИСТЕМА ОБНАРУЖЕНИЯ ТЕКСТОВОГО СПАМА

А. С. Павлов

МГУ им. М. В. Ломоносова

Поступила в редакцию 14.11.2011

Аннотация. В данной работе предлагается новый метод определения текстового спама, основанный на анализе разнообразия тематической структуры текстов и применении методов машинного обучения. На основе разработанного метода строится эффективная система обнаружения поискового спама. Качество предложенного решения подтверждается экспериментально.
Ключевые слова: поисковый спам, неестественный тексты, машинное обучение.
Abstract. This article is dedicated to a new method for content spam detection. The method is based on topical diversity analysis and machine learning. An efficient web spam detection system is built based on the developed algorithms. The quality of the system is proved empirically.
Keywords: web spam, unnatural texts, machine learning.

1 ВВЕДЕНИЕ

В настоящее время поисковые машины стали одним из основных источников информации в сети Интернет. Задача поисковой машины – по каждому пользовательскому запросу отранжировать страницы, находящиеся в ее индексе по релевантности. Релевантность – это мера соответствия страницы запросу. Манипуляции, направленные на незаслуженное повышение оценки релевантности страницы в поисковой системе, называются поисковым спамом.

Текстовый спам – это разновидность поискового спама, связанная с манипуляциями с текстами страниц и массовым порождением текстов. Поисковый спам был признан одной из основных угроз для современных поисковых систем [1]. По некоторым оценкам до 20% всего содержимого сети Интернет является поисковым спамом [2].

Существует два основных подхода к массовому порождению текстов:

• Копирование существующих естественных текстов;

• Синтез текстов на основе естественных документов-образцов.

В настоящее время существует целый ряд эффективных методов обнаружения дубликатов, которые позволяют обнаруживать скопированные тексты в масштабах сети Интернет [3]. В связи с этим большое распространение получили алгоритмы автоматического порождения текстов.

Данная работа посвящена алгоритмам обнаружения неестественных текстов и построению эффективной системы обнаружения текстового спама.

2 ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

В основе многих методов обнаружения неестественных текстов лежит подход, предложенный в работе [4]. В этой работе предлагается 10 эвристик для анализа статистических характеристик текстов. Признаки, полученные на основе эвристик, объединяются в автоматический классификатор поискового спама с помощью методов машинного обучения.

Развитием данного подхода является работа [5]. В данной работе предлагается использовать метод скрытого распределение Дирихле, для определения спамерских текстов. Данный метод ориентирован на обнаружение текстов определенных тематик, свойственных поисковому спаму. В этой работе также рассматривается алгоритм на основе объединения текстовых характеристик и характеристик ссылочной структуры.

В работе [6] предлагается подход к определению неестественных текстов, в основе которого лежит гипотеза, что неестественные тексты не могут одновременно удовлетворять всем ограничениям, свойственным естественным тек-

стам. При обучении алгоритма выделяется большое количество статистических признаков, связанных с читаемостью, единством стиля и жанровыми особенностями, которые впоследствии объединяются в автоматический классификатор.

В работе [7] предлагается оценивать тематическое разнообразие документов для определения текстового спама. В основе метода лежит наблюдение, что неестественные тексты обладают более размытой тематикой. В работе предлагают характеристики оценки тематического разнообразия полезные при обнаружении текстового спама.

3 МЕТОД ОБНАРУЖЕНИЯ ТЕКСТОВОГО СПАМА

Решение, предлагаемое в данной работе, сочетает метод оценки тематического разнообразия, предложенный в работе [7], и большое количество трудноконтролируемых характеристик, предложенных в работе [6]. Данный раздел посвящен краткому изложению этих методов.

3.1 МЕТОД ОЦЕНКИ ТЕМАТИЧЕСКОГО РАЗНООБРАЗИЯ

Одной из важных характеристик естественных текстов является глобальная тематическая связность текстов. У текстов чаще всего есть одна основная тема, и несколько второстепенных. Изучение неестественных текстов показывает, что они зачастую бессмысленны и лишены единой тематики.

Предлагаемый подход к формализации понятия тематики аналогичен лингвистической теории, сформулированной авторами [8]. В рамках данной теории утверждается, что любой осмысленный документ – это некоторое высказывание над несколькими макроконцептами. При этом разные концепты в разной степени участвуют в формировании текста, что соответствует основным и второстепенным тематикам в документе.

Также в модели предполагается, что тематик конечное число. Пусть $\Theta = \{\Theta_1,..,\Theta_K\}$ – множество всех тематик, тогда вектор $\theta^d = (\theta_1^{\,d},..,\theta_K^{\,d})$ называется тематической структурой документа d, если доля слов принадлежащих тематике i равно $\theta_i^{\,d}$:

$$\theta_i^{\,d} = \frac{|\{w : w \in d, w \in \Theta_i\}|}{|d|} \qquad (1)$$

Теоретические исследования методов порождения неестественных текстов [7] позволяют формально доказать нарушение тематической структуры в порожденных текстах.

Теорема 1. Пусть $D_{\text{образцы}}$ – набор документов-образцов для генератора текстов, и задана тематическая структура каждого документа. Пусть с помощью генератора порождается документ $d_{\text{порожд}}$ длины l. Тогда с ростом l тематическая структура порожденного документа $\theta^{\text{порожд}}$ сходится по вероятности к усредненной тематической структуре документов-образцов:

$$\theta^{\text{порожд}} \xrightarrow{\text{P}} \frac{\sum_{d \in D_{\text{образцы}}} |d|\,\theta^d}{\sum_{d \in D_{\text{образцы}}} |d|} \qquad (2)$$

Важным следствием данной теоремы является то, что распределение тематик более разнообразно в порожденных текстах, чем в документах-образцах. Если в естественных текстах зачастую присутствует одна ярко выраженная тематика, то в порожденных документах тематика более расплывчатая.

В настоящее время существует несколько подходов к моделированию тематик текстов. В данной работе использовалась статистическая модель для текстов скрытое распределение Дирихле (СРД), также известная как Latent Dirichlet Allocation (LDA) [9].

На основе теоремы 1 вводятся две характеристики текстов, пригодные для обнаружения текстового спама с неестественной тематической структурой:

- Критерий Пирсона для проверки гипотезы о равномерном распределении тематик;
- Параметры закона Ципфа [10] для распределения тематик документа.

3.2 КОМБИНИРОВАННЫЙ МЕТОД ОБНАРУЖЕНИЯ ТЕКСТОВОГО СПАМА

Автоматическая генерация текстов в настоящее время является нерешенной задачей. Естественным текстам свойственно большое количество закономерностей, которые сложно воспроизвести автоматическими методами. Исследования автоматических методов построения аннотаций показывают, что даже специализированные алгоритмы плохо воспроизводят все характеристики естественных текстов [11].

Фундаментальная идея предлагаемого метода учитывает следующие обстоятельства:

- Нормальный связный текст является значительно информационно избыточен с формальной точки зрения;

• В настоящий момент не существует успешных реализаций генерации связного текста.

Тогда можно попытаться выделить некоторые характеристики текста, поведение которых будет отлично для естественных и искусственных текстов.

Предлагаемый метод обнаружения поискового спама основывается на выделении большого числа трудно контролируемых автором статистических характеристик текстов. Затем выделенные характеристики используются для построения автоматического классификатора неестественных текстов с помощью методологии машинного обучения.

Признаки, рассматриваемые в данном разделе можно разделить на следующие группы:

• Признаки, связанные с читаемостью текста;

• Особенности жанра и авторского стиля;

• Глобальные статистические закономерности текстов;

• Характеристики тематической структуры, описанные в разделе 3.1;

Всего применяется 75 различных признаков текстов. Для объединения всех признаков в единый классификатор применяется метод машинного обучения, основанный на деревьях решений C4.5 и подробно описанный в работе [6].

4 ПРОГРАММНАЯ СИСТЕМА ОБНАРУЖЕНИЯ ПОИСКОВОГО СПАМА

В данном разделе описывается система обнаружения текстового спама, реализующая метод обнаружения, описанный в предыдущем разделе.

4.1 СЦЕНАРИИ ИСПОЛЬЗОВАНИЯ СИСТЕМЫ

Система построена по модульной схеме, различные модули задействуются в зависимости от сценария использования системы. Программная система поддерживает три основных сценария использования.

Построение модели скрытого распределения Дирихле (СРД). В данном сценарии система получает на вход набор HTML-документов. Из документов извлекаются тексты, тексты разбиваются на слова. Специальный модуль по набору текстов формирует модель СРД. На данном этапе не требуется наличие разметки спам-неспам для тренировочного набора СРД, так как СРД – это алгоритм обучения без учителя.

Обучение классификатора поискового спама. В данном сценарии используется ранее построенная модель СРД. На вход системе поступает обучающий размеченный набор документов. Для каждого документа с помощью нескольких специализированных модулей выделяется набор признаков, описанный в разделе 3.2. Затем получившиеся размеченные вектора признаков поступают на вход модулю обучения, который строит деревья решений согласно алгоритму, описанному в работе [6].

Классификация веб-документов. В данном сценарии на вход системе поступает поток HTML-документов, для каждого из которых требуется определить, является ли он поисковым спамом. На данном этапе используется ранее построенная модель СРД и дерево решений, полученное в предыдущем сценарии. Данный режим работы системы предусматривает постоянную обработку поступающих документов по мере их скачивания роботом поисковой системы.

4.2 ОСНОВНЫЕ МОДУЛИ СИСТЕМЫ

Модули разработанной системы решают основные задачи классификации HTML-документов, начиная с предобработки, заканчивая вынесением вердикта. Далее приводятся описания основных модулей системы классификации спама.

Модуль предобработки документа. Данный модуль является общим во всех сценариях использования системы. Его основная задача обработать поступающий на вход HTML-документ и выделить из него текст документа. В данном модуле производится анализ дерева HTML-тегов. Текст документа, выделенный на данном этапе, поступает на вход различным модулям выделения признаков и также модулю обучения модели СРД.

Модуль обучения модели СРД. Основная задача данного модуля – обучение модели СРД на основе коллекции текстов, поступивших на вход. Вначале поступающие тексты приводятся к нижнему регистру и разбиваются на слова, затем полученные последовательности слов используются как обучающий набор для алгоритма обучения СРД. Данный модуль реализует алгоритм обучения СРД на основе итерационного процесса известного как семплирование Гиббса [12]. Данный процесс позволяет приближенно вычислить параметры модели СРД, которые затем сохраняются для дальнейшего использования.

Модуль анализа тематического разнообразия. Данный модуль сходен по строению с модулем обучения модели СРД. Он также представляет документы в виде последовательности слов, а затем применяет семплирование Гиббса чтобы по известной модели СРД восстановить веса тематик для нового документа. В данном модуле также вычисляются характеристики тематического разнообразия.

Модуль выделения глобальных характеристик. Данный модуль вычисляет частоту встречаемости всех слов в документе и на основе этих частот оценивает параметры закона Ципфа для документа. В данном модуле также вычисляется степень сжатия документа алгоритмами gzip и bz2.

Модуль выделения характеристик читаемости. Данный модуль производит разбиение текста документа на слова и предложения. На основе разбиения в нем вычисляются признаки читаемости текстов, такие как средняя длина предложений и слов.

Модуль выделение стилистических характеристик. Данный модуль получает на вход текст документа и вычисляет по нему признаки авторства и стиля. Для морфологического анализа применялись существующие реализации морфологических анализаторов: парсер mystem[1] для русского языка; анализатор частей речи Stanford Part of Speech Tagger[2] для английского языка.

При этом система поддерживает добавление других морфологических анализаторов для других языков. После этапа морфологического анализа в данном модуле происходит подсчет признаков, связанных с особенностями жанра и авторского стиля.

Модуль обучения классификатора. Данный модуль работает с документами в виде векторов признаков, которые он получает из модулей выделения характеристик. В этом модуле реализуется алгоритм построения ансамбля деревьев решений, описанный в работе [6]. Данный модуль на выходе формирует дерево решений, обученное для классификации поискового спама.

Модуль классификации. Этот модуль также работает с документами в виде векторов признаков. Он загружает ранее сформированное дерево решений и использует его для классификации документов, поступающих к нему из модулей выделения характеристик. На выходе выдается вещественное число в интервале от -1 до 1, которое характеризует спамность документа с точки зрения системы.

Система спроектирована с учетом возможности добавления поддержки других языков и морфологических анализаторов, а также с учетом возможности добавления новых модулей выделения признаков. На рис. 1 приведена схема работы всех описанных модулей системы для сценария классифкиации.

5 ЭКСПЕРИМЕНТЫ

Важной частью работы является экспериментальное подтверждение применимости предлагаемого алгоритма обнаружения текстового спама.

Чтобы сравнить предлагаемые алгоритмы обнаружения поискового спама с существующими аналогами был проведен ряд экспериментов на наборе данных WebspamUK-2007 [13]. Этот набор представляет собой набор всех страниц из доменной зоны .uk, собранный за 2007 год. 6000 сайтов из данного набора размечены вручную авторами набора на предмет принадлежности поисковому спаму. Набор размеченных сайтов разделен на обучающую и тестовую выборки.

В рамках эксперимента на обучающей выборке обучались две версии алгоритма – базовый алгоритм без характеристик тематической структуры и комбинированный, содержащий характеристики, описанные в работе [6]. Версии алгоритма сравнивались между собой, а также с лучшими результатами других исследователей на данном наборе. Сравнение проводилось с алгоритмом победителя соревнований по обнаружению поискового спама Web Spam Challenge 2008 [14], а также с лучшим результатом на данном наборе, показанным алгоритмом Linked LDA [5].

Общепринятой метрикой для измерения качества классификаторов поискового спама на данном наборе является площадь под ROC-кривой (Area Under ROC-Curve, AUC). ROC-кривая – это кривая, которую описывает алгоритм классификации на графике, осями которого являются верно-положительные и ложно-положительные срабатывания. Чем больше площадь, тем лучше качество алгоритма.

[1] Данный парсер доступен по адресу: http://company.yandex.ru/technology/mystem

[2] Данный морфологический анализатор доступен по адресу: http://nlp.stanford.edu/software/tagger.shtml

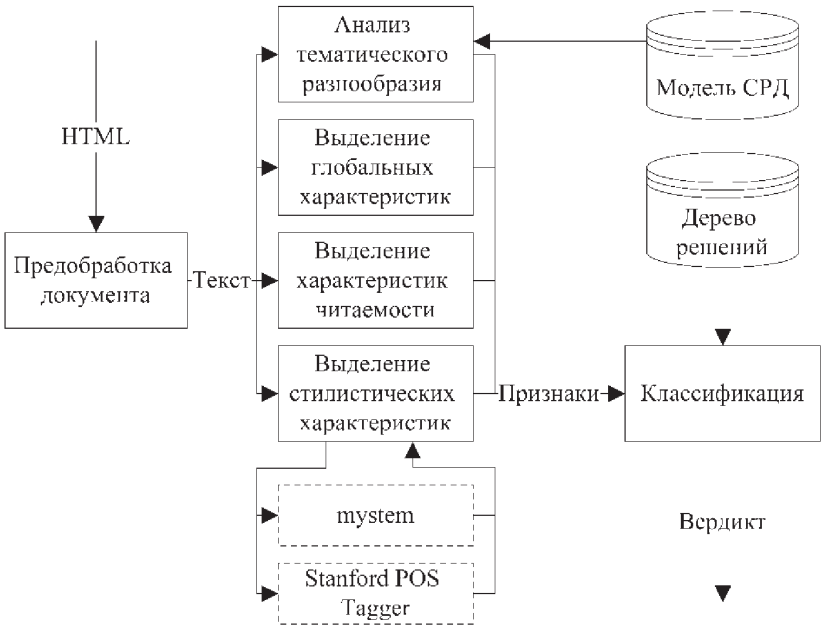

Рис. 1. Схема взаимодействия модулей разработанной системы обнаружения поискового спама

Результаты эксперимента и сравнение с существующими аналогами приведено в таблице 1. Как видно, комбинированный алгоритм существенно превосходит, как базовую версию, так и лучшие на текущий момент алгоритмы классификации поискового спама.

6 ЗАКЛЮЧЕНИЕ

Для решения задачи определения текстового спама разработан новый алгоритм на основе оценки разнообразия тематик документа и большого числа трудноконтролируемых характеристик текстов.

На основе предложенного алгоритма реализована программная система определения поискового спама. Разработанная система апробирована на стандартном наборе данных реальных сайтов WebspamUK-2007. Получены более высокие характеристики классификации вебспама, по сравнению с известными методами.

Таблица 1

Результаты эксперимента на наборе WebspamUK-2007

Алгоритм	AUC
Победитель WSC-2008 [14]	0.850
Linked LDA [5]	0.854
Базовый алгоритм [6]	0.847
Комбинированный алгоритм	0.870

СПИСОК ЛИТЕРАТУРЫ

1. *Henzinger M., Motwani R., Silverstein C.* Challenges in web search engines // ACM SIGIR Forum. No. 2. ACM, 2002. P. 22.

2. *Henzinger M., Motwani R., Silverstein C.* Challenges in web search engines // ACM SIGIR Forum. No. 2. ACM, 2002. P. 22.

3. *Зеленков Ю.Г., Сегалович И.В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль, Россия, 2007. – Том 1, С. 166–174.

4. *Ntoulas A., Najork M., Manasse M., Fetterly D.* Detecting spam web pages through content analysis // Proceedings of the 15th international conference on World Wide Web – WWW '06. 2006. P. 83.

5. *I. Bıry, D. Siklysi, J. Szaby, A. A. Benczır* Linked latent Dirichlet allocation in web spam filtering, // Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web – AIRWeb '09. 2009. P. 37.

6. *Павлов А.С., Добров Б.В.* Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Труды 11й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL'2009, Петрозаводск: 2009.

7. *Павлов А.С., Добров Б.В.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // Вычислительные методы и программирование:

новые вычислительные технологии. 2011. Т. 12. С. 58–72.

8. *Дейк Т.А., Кинч В.* Стратегии понимания связного текста // Новое в зарубежной лингвистике. Вып.23. М.: Прогресс. 1988. С.153–211

9. *Blei D., Ng A., Jordan M.* Latent Dirichlet allocation // Journal of Machine Learning Research, 3(5):993-1022, 2003.

10. *Li W.* Random texts exhibit Zipf's-law-like word frequency distribution // Information Theory, IEEE Transactions on. 2002. Vol. 38, no. 6. Pp. 1842–1845.

11. *Dang H.* Overview of DUC 2006 // Proceedings of the Document Understanding Conference. Vol. 2005. 2006.

12. *Heinrich G.* Parameter estimation for text analysis // Bernoulli. 2005. no. 1. P. 1–31.

13. Yahoo! Research: "Web Spam Collections". http://barcelona.research.yahoo.net/webspam/datasets/ Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.dsi.unimi.it/. URLs retrieved May 2007.

14. *Geng G., Jin X., Wang. C.-H.* CASIA at Web Spam Challenge 2008 Track III // Proceedings of the 4th international workshop on Adversarial information retrieval on the web, Beijing, China. ACM, 2008. 32–33.

Павлов А. С. – МГУ им. М.В. Ломоносова, факультет Вычислительной Математики и Кибернетики. E-mail: pavvloff@yandex.ru

Pavlov A. S. – Moscow State University МГУ им. М. В. Ломоносова, факультет Вычислительной Математики и Кибернетики. E-mail: pavvloff@yandex.ru