МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА

УДК 681.513.7

МЕТОД ОПРЕДЕЛЕНИЯ НЕЕСТЕСТВЕННЫХ ТЕКСТОВ НА ОСНОВЕ ХАРАКТЕРИСТИК ТЕМАТИЧЕСКОГО РАЗНООБРАЗИЯ

Павлов А.С.
Московский государственный университет им. М.В. Ломоносова,
факультет вычислительной математики и кибернетики, г. Москва

Поступила в редакцию 31.03.2011, после переработки 20.05.2011.

В данной работе предлагается новый метод определение неестественных текстов, основанный на анализе разнообразия тематической структуры текстов и применении методов машинного обучения. Эффективность предложенного метода подтверждается экспериментально.

This article is dedicated to a new method for unnatural texts detection. The method is based on topical diversity analysis and machine learning. Efficiency of proposed method is proved empirically.

Ключевые слова: неестественные тексты, поисковый спам, тематическая структура текстов.
Keywords: unnatural texts, search spam, topical structure.

Введение

В настоящее время поисковые системы являются основным источником информации в сети Интернет. Высокие позиции в выдаче поисковых систем являются серьезным коммерческим преимуществом, так как обеспечивают постоянный приток посетителей на сайт. В связи с этим существует постоянное давление на алгоритмы поисковых систем с целью повлиять на результаты поиска. Это явление получило название поискового спама, под которым понимается [2] любое намеренное действие, направленное на незаслуженное повышение релевантности или важности страницы в поисковой системе по сравнению с ее истинной ценностью. Поисковый спам признан одной из основных угроз для современных поисковых систем [3], до четверти всего содержимого сети Интернет может быть отнесено к поисковому спаму [13]. В настоящее время во всех ведущих поисковых системах уровень спама в первой десятке результатов составляет около 5% (http://analyzethis.ru), то есть в среднем каждый второй запрос содержит один спамерский документ.

При создании некоторых видов поискового спама спамерам требуется породить большое количество текстов. Одним из таких видов являются дорвеи. Дорвеи — это специальные сайты и страницы, которые созданы только для перенаправления пользователей с поисковых систем на другие страницы. При создании дорвеев требуются тексты, чтобы создать страницы, релевантные как можно большему

количеству запросов, таким образом, спамеры максимизируют переходы пользователей на дорвеи.

При порождении текстов у спамеров возникает альтернатива - копировать существующие тексты, возможно с небольшими изменениями, или порождать новые. Существующие методы определения дубликатов позволяют эффективно обнаруживать копии текстов, поэтому спамеры зачастую используют системы автоматического порождения текстов. В настоящее время не существует автоматических генераторов текста, которые бы воспроизводили все закономерности, свойственные естественным текстам.

В данной статье рассматривается подход к обнаружению неестественных текстов, основанный на измерении тематического разнообразия.

1. Обзор существующих решений

Поисковый спам нацелен на различные алгоритмы в поисковой системе, и разделяется на несколько направлений. Ссылочный спам создается, чтобы для некоторой страницы улучшить место в поисковой выдаче. Существует обширный класс алгоритмов, нацеленных на борьбу со ссылочным спамом, например [9, 10].

Другим важным направлением в борьбе с поисковым спамом является обнаружение дубликатов текстов. Обзор методов обнаружения дубликатов приведен в работе [1].

В основе многих методов обнаружения неестественных текстов лежит подход, предложенный в работе [4]. Этот подход основывается на анализе статистических характеристик текстов и применении машинного обучения для построения автоматического классификатора поискового спама. Развитием данного подхода является работа [5]. В данной работе предлагается использовать метод скрытого распределение Дирихле, для определения спаммерских текстов.

В работе [8] 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. В работе предлагается алгоритм для подсчета доли редких пар слов и показывается, что эта характеристика улучшает качество определения поискового спама.

В работе [7] предлагается подход к определению неестественных текстов, основанный на большом количестве статистических факторов. В основе подхода лежит гипотеза, что неестественные тексты не могут одновременно удовлетворять всем ограничениям, свойственным естественным текстам. При обучении алгоритма выделяется большое количество статистических признаков, связанных с читаемостью, единством стиля и глобальными закономерностями, которые впоследствии объединяются в автоматический классификатор.

Подход, предлагаемый в данной работе, опирается на работу [7], но существенно расширяет его, используя метрики тематического разнообразия для определения неестественных текстов.

2. Генераторы текстов на основе цепей Маркова

В настоящий момент спамеры используют разнообразные методы порождения неестественных текстов. В основе всех этих методов лежит использование образцов естественных текстов. Одним из распространенных алгоритмов порождения синтетических текстов являются генераторы на основе цепей Маркова.

При порождении документов с помощью цепей Маркова вероятность порождения каждого последующего слова зависит от N предыдущих порожденных слов, где N - длина цепи Маркова. При этом эти вероятности предварительно собираются по коллекции естественных текстов. Таким образом, тексты, порожденные с помощью цепей Маркова, сохраняют локальную связность, что усложняет их обнаружение. На практике спамеры применяют цепи Маркова длины 2 или 3. Ниже приведен пример текста, порожденного цепью Маркова длины 2, в качестве образцов использовались документы из коллекции ROMIP.ByWeb (http://romip.ru):

> *Бесплатная и маркетингом развлекательных программ канала. А если так дальше будет затягиваться этот конфликт, тем самым дают возможность использовать внешний шаблон, например, уже успевших изголодаться по хоккею. О юридическом лице я и сообщил похитителю, что данные нужно дублировать и работать с соответствующими именами в виде регулярного выражения.*

На данном примере хорошо видно, что при порождении текстов сохраняется локальная связность текста, но при этом глобальные закономерности, такие как единство тематики, нарушаются, так как фрагменты предложений были взяты из различных исходных документов, посвященных разным тематикам (например, хоккей и регулярные выражения). Это свойство и легло в основу предлагаемого метода.

3 Тематическое разнообразие текстов

Предлагаемый подход является развитием метода, предложенного в работе [7]. Изучение порожденных текстов показывает, что они зачастую бессмыслены и лишены одной общей тематики. При этом в тексте встречаются слова из различных документов-образцов, посвященных разным тематикам. Таким образом, на интуитивном уровне тематика неестественных текстов более разнообразна и расплывчата, а характеристики разнообразия тематики текстов могут быть полезными при обнаружении спама. Для оценки разнообразия текста в данном аспекте необходимо формализовать понятие тематики.

Предлагаемый подход к формализации понятия тематики аналогичен лингвистической теории, сформулированный авторами [11]. 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 документа.

3.1 Скрытое распределение Дирихле

Для моделирования тематик текстов удобно воспользваться статистической моделью для текстов, известной как скрытое распределение Дирихле (СРД) или Latent Dirichlet Allocation [6]. В рамках данной модели считается, что тематика определяется вероятностью порождения слов из словаря. Считается, что существует ограниченное число тематик N. При этом одно и то же слово имеет ненулевую вероятность порождения в разных тематиках. В данной модели каждому документу ставится в соответствие вектор вероятностей тематик θ. При этом считается, что каждое слово в документе порождено строго одной тематикой.

В рамках модели СРД предлагается следующий порождающий процесс для набора документов:

- Задается матрица вероятностей порождения каждого слова в каждой тематике.
- Для каждого документа:
  - Выбирается набор весов тематик θ, исходя из распределения Дирихле с вектором параметров α.
  - При порождении каждого слова в документе:
    - Выбирается тематика t исходя из мультиномиального распределения с вектором параметров θ.
    - Выбирается слово, которое надо породить, исходя из распределения вероятностей порождения слов в выбранной тематике.

Модель СРД также позволяет по имеющемуся набору документов восстановить вероятности слов в тематиках и веса тематик θ для каждого документа. Тематики в модели СРД, востановленные по коллекции текстов, обладают рядом свойств, которые делают их похожими на тематики в интуитивном представлении:

- слова, которые часто встречаются вместе в одних и тех же текстах, получают высокий вес в одних и тех же тематиках;
- любое слово может порождаться разными тематиками с разной вероятностью;
- часто употребляемые слова, такие как предлоги и союзы, будут иметь высокую вероятность порождения в любой тематике.

На рис. 1 приведен пример тематической структуры текста, полученной с помощью модели СРД. Вначале модель была обучена на наборе из 10000 документов из коллекции ROMIP.ByWeb. Затем для одного из документов с помощью модели все слова были размечены по тематикам СРД. В приведенном тексте есть одна основная тематика, и две второстепенных, остальные слова распределяются по другим тематикам с меньшим весом.

Описанные свойства позволяют рассматривать веса тематик для документов, полученные в модели СРД, как некоторую модель интуитивного понятия о тематиках документа. Назовем распределение весов тематик θ для документа тематической структурой документа.

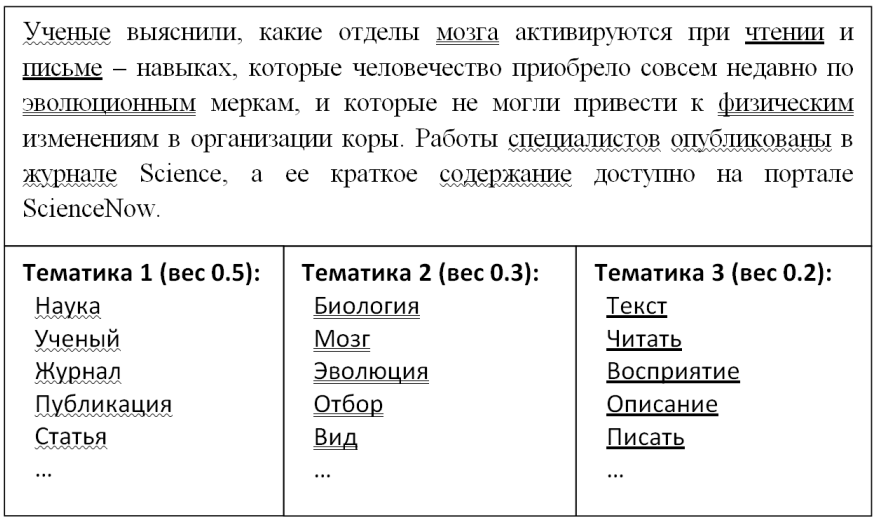

Рис. 1: Пример тематической структуры текста, полученной с помощью модели СРД. Для удобства приведены три темы с наибольшим весом.

3.2 Закон Ципфа для тематической структуры

Естественным текстам свойственен ряд статистических закономерностей, таких как закон Ципфа [12]. Закон Ципфа утверждает, что если упорядочить слова текста по частотности, то частота каждого слова будет обратно пропорциональна его порядковому номеру.

Предлагаемый подход опирается на гипотезу, что для весов тематик в модели СРД справедлива аналогичная закономерность - если упорядочить тематики по весу в документе, то вес тематики будет обратно пропорционален ее порядковому номеру. Вес тематики θ_k с порядковым номером k подчиняется следующему соотношению:

$$\theta_k(s, c) \approx \frac{c}{k^s}, \tag{1}$$

где s — параметр, характеризующий разнообразие тематик в тексте, c - константа.

Чем больше параметр s, тем больший вес будет у основных тематик, чем меньше s, тем более разнообразны тематики в документе. Чтобы определить, насколько разнообразие тематической структуры может использоваться для определения неестественных текстов, была собрана статистика по параметру закона Ципфа для естественных и неестественных текстов. Результаты приведены на рис. 2. На этом графике сплошной линией обозначено распределение параметра для 10000 произвольных документов из коллекции ROMIP.ByWeb, а пунктирной — для 10000 документов, порожденных генератором на основе цепей Маркова. На графиках видно, что естественные тексты в среднем менее разнообразны (значения параметра Ципфа для них больше). При этом неестественные тексты хорошо выделяются с помощью данной характеристики.

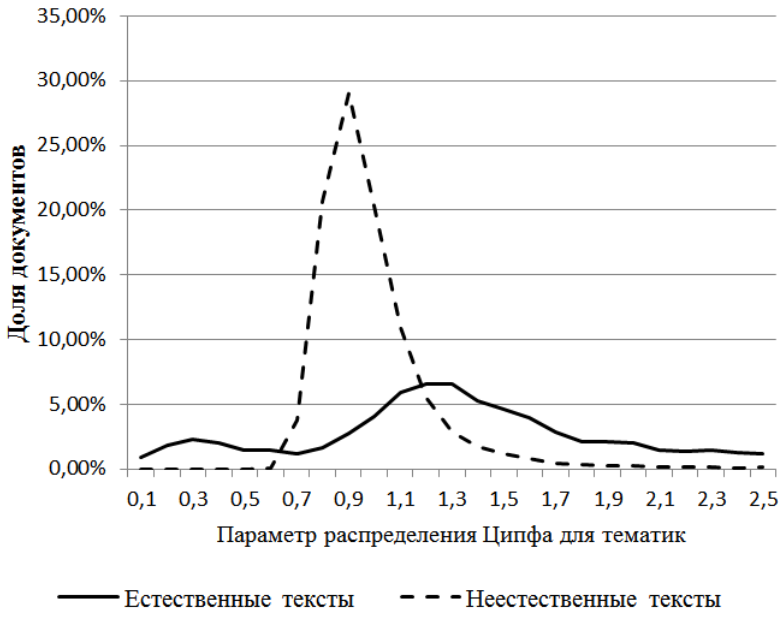

Рис. 2: Доли естественных и неестественных документов в зависимости от параметра распределения Ципфа

Таким образом, распределения Ципфа для тематик естественных и неестественных текстов существенно отличаются. Для оценки разнообразия тематик в тексте можно по частотам слов в тексте оценить параметры s и c. Для вычисления значения s формулу (1) удобно привести к логарифмической шкале:

$$\log(\theta_k(s,c)) \approx \log(c) - \log(k^s). \tag{2}$$

Чтобы из этого уравнения получить приближенное значение s для текста, воспользуемся методом наименьших квадратов:

$$f_k = \log(\theta_k(s,c)), r_k = \log(k),$$
$$s = -\frac{n\sum r_k f_k - \sum r_k \sum f_k}{n\sum (r_k)^2 - (\sum r_k)^2}. \tag{3}$$

Характеристика разнообразия тематик в тексте, вычисленная по формуле (3), может использоваться как один из факторов для определения неестественных текстов.

3.3 Критерий Пирсона для тематической структуры текстов

В рамках данной работы также исследовался вероятностный подход к оценке разнообразия тематической структуры текстов. Модель скрытого распределения Дирихле для каждого документа определяет наиболее вероятные веса тематик.

Для того чтобы оценить разнообразие тематической структуры можно применить критерий согласия Пирсона.

Рассмотрим случайную величину – тематику t, которая порождает слово в документе. Согласно модели СРД, эта случайная величина может принимать значения от 1 до N, где N — количество тематик в модели. При этом модель СРД для каждого документа позволяет оценить вероятность встретить слово, из i-ой тематики θ_i. Используем критерий Пирсона, чтобы проверить гипотезу, что наблюдаемые веса тематик в модели СРД на самом деле подчиняются равномерному распределению:

$$\chi^2 = N \sum \frac{(1/N - \theta_i)^2}{1/N}. \tag{4}$$

Чем больше значение критерия, тем с меньшей вероятностью тематики документа распределены равномерно, тем с большей вероятностью документ естественный. Данный критерий также использовался для обнаружения неестественных текстов.

4. Эксперимент

В рамках эксперимента требовалось проверить, насколько предложенные характеристики улучшают обнаружение неестественных текстов. Измерялась точность, полнота и F-мера обнаружения неестественных текстов при использовании классификатора, обученного на выборке естественных текстов из коллекции ROMIP.ByWeb (http://romip.ru) и наборе текстов, порожденных цепями Маркова длины 2 и 3.

Обучающие выборки составлялись из 10000 документов из коллекции ROMIP.ByWeb в качестве примеров естественных текстов, и 10000 документов, порожденных цепями Маркова, обученными на текстах из той же коллекции. Тестовые выборки составлялись аналогичным образом и не содержали пересечения с обучающими.

Также исследовалась возможность обнаружения текстов, порожденных генератором дорвеев Doorway.Su (http://doorway.su), который в действительности применяется для порождения поискового спама. В качестве алгоритма классификации использовался алгоритм на основе деревьев решений, предложенный в работе [7].

В ходе эксперимента было построено два классификатора для каждой тренировочной выборки. В качестве базового был взят классификатор с использованием характеристик, предложенных в работе [7]. Также была построена улучшенная версия классификатора с добавлением характеристик тематического разнообразия, вычисленных по формулам (3) и (4). Разница в точности и полноте классификаторов позволяет оценить выигрыш при использовании характеристик разнообразия. Результаты эксперимента приведены в таблице 1.

Результаты проведенного эксперимента показывают, что применение характеристик тематического разнообразия позволяют повысить полноту и точность классификатора неестественных текстов. При этом уровень ошибок первого и второго рода снижается вдвое по сравнению с вариантом алгоритма, не использующим характеристики разнообразия. Таким образом, предложенные характеристики вносят существенный вклад в задачу обнаружения неестественных текстов. Экспери-

Таблица 1: Характеристики классификации неестественных текстов различными вариантами алгоритма

	Точность	Полнота	F-мера	Ошибки 1-го рода	Ошибки 2-го рода
Цепи Маркова длины 2. Базовая версия [7]	96,19%	96,11%	96,15%	3,81%	3,89%
Цепи Маркова длины 3. Базовая версия [7]	94,08%	92,29%	93,18%	5,92%	7,57%
Генератор Doorway.Su. Базовая версия [7]	95,89%	95,11%	95,50%	4,11%	4,85%
Цепи Маркова длины 2. Улучшенная версия	98,37%	97,93%	98,15%	1,63%	2,06%
Цепи Маркова длины 3. Улучшенная версия	97,72%	97,09%	97,40%	2,28%	2,89%
Генератор Doorway.Su. Улучшенная версия	98,12%	97,56%	97,84%	1,88%	2,43%

мент также подтверждает, что данный подход может применяться и для борьбы с распространенными генераторами неестественных текстов, таких как Doorway.Su.

Заключение

В данной работе предлагается улучшение метода обнаружения неестественных текстов за счет использования метрик разнообразия тематики текстов. Тематики текста моделировались с помощью статистического метода СРД.

Предложено два подхода к оценке разнообразия тематической структуры. Эксперименты показали, что применение предложенных характеристик позволяет существенно уменьшить ошибки 1-го и 2-го рода при обнаружении неестественных текстов.

Список литературы

[1] 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.

[2] Gyongyi, Z., Garcia-Molina, H. Web Spam Taxonomy // Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.

[3] Henzinger M., Motwani R., Silverstein C. Challenges in Web Search Engines // SIGIR Forum 36(2), 2002.

[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, May 23-26, 2006, Edinburgh, Scotland.

[5] Biro I., Szabo J., Benczur A. A. Latent Dirichlet allocation in web spam filtering // Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, April 22, 2008, Beijing, China.

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

[7] 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.

[8] Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М. Поиск неестественных текстов // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск: 2009.

[9] Abernethy J., Chapelle O., Castillo C. WITCH: A New Approach to Web Spam Detection // Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2008.

[10] Biro I., Siklosi D., Szabo J., Benczur A. A. Linked latent Dirichlet allocation in web spam filtering // Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, April 21-21, 2009, Madrid, Spain.

[11] ван Дейк Т.А., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. Вып.23. М.: Прогресс. 1988. С.153-211.

[12] Gelbukh A., Sidorov, G. Zipf and Heaps Laws' Coefficients Depend on Language // In Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001), February 18–24, 2001.

[13] Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., Vigna S. A reference collection for web spam // SIGIR Forum 40, № 2. ACM, 2006. 11–24.