Методы обнаружения массово порождаемых неестественных текстов на основе анализа разнообразия тематической структуры текстов

© А.С. Павлов

МГУ им. М.В. Ломоносова,
факультет Вычислительной Математики и Кибернетики
pavvloff@yandex.ru

Аннотация

Данная работа посвящена разработке методов и средств обнаружения массово порождаемых неестественных текстов. В работе предлагается теоретическая обобщенная модель текстов, порожденных по образцам. На основе теоретической модели предлагается алгоритм обнаружения неестественных текстов. Данный алгоритм анализирует тематическую структуру текстов, и определяется по ней неестественные тексты. Предложенный алгоритм проверяется на задаче обнаружения поискового спама.

1. Введение

В связи с ростом объема информации в сети Интернет поисковые машины стали основным средством для эффективного доступа к ней. Задача поисковой машины – на каждый поисковый запрос выдавать ранжированный набор страниц, наиболее соответствующих запросу. Мера соответствия страницы запросу, называемая релевантностью, вычисляется на основе характеристик страниц и запросов.

Появление сайта в первой десятке выдачи популярных поисковых систем по коммерческим запросам обеспечивает большой приток посетителей на сайт. В связи с этим возникает конкуренция между создателями сайтов за попадание на верхние позиции выдачи поисковых систем. Она приводит к тому, что создатели сайтов пытаются повлиять на работу алгоритмов, применяемых в поисковых системах, чтобы незаслуженно повысить оценку релевантности страниц. Это явление получило название поискового спама [1].

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

Поисковый спам ухудшает качество поиска и увеличивает нагрузку на поисковую систему. Он был признан одной из основных угроз для современных поисковых систем [2]. По некоторым оценкам до 20% всего содержимого сети Интернет является поисковым спамом [3], уровень поискового спама в выдаче ведущих поисковых систем составляет 3-6% [4].

Поисковые системы используют различную информацию для ранжирования страниц: содержимое страницы и сайта, на которой она расположена; ссылки между страницами и сайтами и пр. В настоящее время существует несколько разновидностей поискового спама, нацеленных на различные алгоритмы, применяемые внутри поисковых систем. Например, ссылочный спам нацелен на алгоритмы ссылочного ранжирования, такие как PageRank.

Одной из разновидностей поискового спама является массовое порождение неестественных текстов. При использовании такого спама целью спамеров является попадание в выдачу по запросам с малым количеством релевантных страниц. Это позволяет странице с большой вероятностью показаться по этому запросу. Чтобы максимизировать количество переходов пользователей по таким запросам спамерам приходится создавать тысячи страниц, каждая из которых должна показываться по одному или нескольким низкочастотным запросам. Такой спам особенно опасен для поисковых систем, так как спамерские страницы непосредственно показываются в выдаче.

Так как создание большого количества страниц с текстом вручную не представляется возможным, спамеры применяют автоматические алгоритмы массового порождения текстов. При этом им необходимо максимально затруднить обнаружение таких текстов со стороны поисковой системы. 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 [5]. В связи с этим большое распространение получили алгоритмы автоматического порождения текстов.

Данная работа посвящена алгоритмам обнаружения массово порождаемых неестественных текстов.

1.1 Обзор существующих решений

В основе многих методов обнаружения неестественных текстов лежит подход, предложенный в работе [8]. 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 [12]. В данной работе предлагается использовать метод скрытого распределение Дирихле, для определения спаммерских текстов. Данный метод ориентирован на обнаружение текстов определенных тематик, свойственных поисковому спаму. В работе [13] рассматривается алгоритм на основе объединения текстовых характеристик и характеристик ссылочной структуры. При этом используется методика для повышения качества классификации при несбалансированном обучающем наборе.

В работе [11] предлагается подход к определению неестественных текстов, в основе которого 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 жанровыми особенностями, которые впоследствии объединяются в автоматический классификатор.

2. Теоретическая модель неестественных текстов

Одной из целей исследования является изучение теоретических основ обнаружения неестественных текстов, порожденных на основе документов-образцов.

Автоматическая генерация текстов в настоящее время является нерешенной задачей. Естественным текстам свойственно большое количество закономерностей, которые сложно воспроизвести автоматическими методами:

- Локальная связность;
- Единство стиля и жанра;
- Синтаксическая структура предложений;
- Глобальная тематическая связность текста;
- Структура изложения;
- И т.п.

При порождении поискового спама используются различные алгоритмы порождения текстов. Рассмотрим алгоритмы, которые используют выборку естественных текстов для обучения генераторов текстов.

В данной работе был рассмотрен ряд алгоритмов порождения текстов, зачастую применяемых для порождения неестественных текстов:

- Алгоритм на основе модели «мешок слов»;
- Алгоритм на основе цепей Маркова порядка k;
- Алгоритм на основе копирования фрагментов документов-образцов.

Чтобы выработать общий алгоритм обнаружения текстов, порожденных генераторами на основе образцов, важно выделить их общие черты. В данном разделе предлагается обобщенная модель генератора текстов, на основе образцов.

Пусть $D_{образцы}$ – набор текстов-образцов. Рассмотрим множество троек (t): документ (d), номер слова (m), слово, стоящее в данной позиции (v):

$$T = \{t\} = \{(d, m, v)\};$$

$$|T| = \sum_{d \in D_{образцы}} |d|;$$

Опишем процесс порождения текста на основе обобщенной модели. На первом шаге произвольным образом выбирается начальное состояние цепи. Затем на каждом последующем шаге, исходя из матрицы вероятностей, выбирается следующее состояние, при переходе в состояние $t = (d, m, v)$ к порождаемому документу добавляется слово v. Процесс заканчивается, когда порожденный текст достигает определенной длины.

Все вышеперечисленные алгоритмы порождения текстов на основе образцов можно представить в виде однородной цепи Маркова с пространством состояний *T*, у которой переходная матрица *P* определяется разновидностью алгоритма.

2.1 «Мешок слов»

Так как в данной модели вероятность порождения любого слова на любом шаге пропорциональна частоте этого слова в наборе, матрица переходов будет иметь простейший вид:

$$P_{t_i t_j} = P(X_{n+1} = t_j \mid X_n = t_i) = \frac{1}{|T|}; \quad (1)$$

2.2 Алгоритм на основе цепей Маркова

Пусть k – порядок цепи Маркова, тогда элемент матрицы переходов для этого алгоритма не равен нулю, только если предыдущие k слов для двух состояний совпадают. Введем сходства на множестве состояний. Два состояния схожи по k

предыдущим, если предыдущие k слов для этих двух состояний совпадают.

Очевидно, отношение сходства по k предыдущим состояниям является отношением эквивалентности и множество состояний разделяется на классы эквивалентности $T_1^k,..,T_{N_k}^k$. Обозначим $P(T)$ - множество состояний, непосредственно предшествующих состояниям из множества T. Тогда вероятность перехода между состояниями может быть выражена через их классы эквивалентности:

$$P_{t_i t_j} = \begin{cases} \frac{1}{|T_l^k|}, t_j \in T_l^k, t_i \in P(T_l^k); \\ 0, иначе; \end{cases} \quad (2)$$

2.3 Алгоритм на основе фрагментов текстов

Для генераторов на основе фрагментов текстов введем дополнительные обозначения. Пусть B – множество состояний, в которых фрагменты начинаются, а E – множество состояний, в которых фрагменты заканчиваются, тогда элемент матрицы переходов можно выписать в следующем виде:

$$P_{t_i t_j} = \begin{cases} \frac{1}{|B|}, t_i \in E, t_j \in B; \\ 1, d_i = d_j, m_j = m_i + 1, t_i \notin E; \\ 0, иначе; \end{cases} \quad (3)$$

3. Модель тематической структуры текстов

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. При этом в тексте встречаются слова из различных документов-образцов, посвященных разным тематикам. Таким образом, на интуитивном уровне тематика синтетических текстов более разнообразна и расплывчата. Чтобы использовать это наблюдение для обнаружения неестественных текстов, вначале формализуем понятие тематики.

Предлагаемый подход к формализации понятия тематики аналогичен лингвистической теории, сформулированной авторами [6]. 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 $и_i^d$:

$$\theta_i^d = \frac{|\{w : w \in d, w \in \Theta_i\}|}{|d|}; \quad (4)$$

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

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

$$\theta^{порожд} \xrightarrow{P} \frac{\sum_{d \in D_{образцы}} |d| \theta^d}{\sum_{d \in D_{образцы}} |d|}; \quad (5)$$

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.

4. Предлагаемый алгоритм обнаружения

4.1 Моделирование тематик с помощью модели СРД

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) [7].

В модели СРД 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 и, порожденный из распределения Дирихле с вектором параметров б. При этом считается, что каждое слово в документе порождено строго одной тематикой.

Модель СРД также позволяет по имеющемуся набору документов восстановить вероятности слов в тематиках и веса тематик и для каждого документа. Тематики в модели СРД, восстановленные 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.

Описанные свойства позволяют рассматривать веса тематик для документов, полученные в модели СРД, как некоторую модель интуитивного понятия о тематиках документа. В данной работе тематики текстов моделировались с помощью модели СРД, обученной на 10000 документах из коллекции Romip.ByWeb [8]. При этом использовались следующие параметры модели:

- Количество тематик: K=100;
- Вектор параметров распределения Дирихле: б=(0.01,..,0.01);

4.2 Методы оценки разнообразия тематической структуры

Из теоремы 1 следует, что с ростом длины порожденного документа его тематическая структура будет стремиться к усредненной тематической структуре набора документов-образцов.

4.2.1 Оценка разнообразия тематической структуры на основе критерия Пирсона

Для того чтобы оценить естественность тематической структуры можно применить критерий согласия Пирсона. Будем использовать критерий Пирсона, чтобы проверить гипотезу, что наблюдаемые веса тематик подчиняются усредненному распределению тематик.

В реальности усредненные веса тематик документов-образцов могут быть не известны. Чтобы обойти эту проблему воспользуемся особенностью модели СРД. В модели считается, что веса тематик документов подчиняются распределению Дирихле с однородными параметрами $\vec{\alpha}$. Математическое ожидание такой случайной величины равно $\left(\frac{1}{K},..,\frac{1}{K}\right)$. Таким образом, в качестве усредненного веса тематик в документах-образцах можно взять математическое ожидание весов тематик в модели СРД:

$$\chi^2(d) = K^2 \sum_{i=1}^{K} \left(\frac{1}{K} - \theta_i^{\,d} \right)^2; \qquad (6)$$

4.2.2 Оценка разнообразия тематической структуры на основе закона Ципфа

Естественным текстам свойственен ряд статистических закономерностей, таких как закон Ципфа [9]. 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 справедлива аналогичная закономерность – если упорядочить тематики по весу в документе, то вес тематики будет обратно пропорционален ее порядковому номеру. Вес тематики и$_k$ с порядковым номером k подчиняется следующему соотношению:

$$\theta_k(s,c) \approx \frac{c}{k^s}; \qquad (7)$$

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.

Для оценки разнообразия тематик в тексте можно по частотам тематик в тексте оценить параметры s и c. Для вычисления значения s формулу (15) удобно привести к логарифмической шкале:

$$\log(\theta_{\mathrm{k}}(s,c)) \approx \log(c) - s\log(k); \quad (8)$$

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

$$f_k = \log(\theta_{\mathrm{k}}(s,c));$$
$$r_k = \log(k);$$
$$s = -\frac{K\sum_k r_k f_k - \sum_k r_k \sum_k f_k}{K\sum_k (r_k)^2 - \left(\sum_k r_k\right)^2}. \qquad (9)$$

Характеристика разнообразия тематик в тексте, вычисленная по формуле (9) может также использоваться как один из факторов для оценки разнообразия тематик текстов. Чем больше значение параметра Ципфа для текста, тем с менее разнообразна тематическая структура документа.

4.2.3 Алгоритм машинного обучения для обнаружения неестественных текстов

Предлагаемый алгоритм машинного обучения для обнаружения неестественных текстов состоит из двух основных частей:

- Модель СРД, применяющаяся для автоматического построения тематической структуры текстов;
- Алгоритм машинного обучения на основе деревьев решений;

При обучении данного алгоритма вначале строится модель СРД по обучающей выборке документов. По этой модели с помощью формул (6) и (9) можно оценить тематическое разнообразие естественных и неестественных документов, содержащихся в обучающей выборке.

Характеристики тематического разнообразия используются как факторы при построении автоматического классификатора. Наряду с факторами тематического разнообразия применяются многочисленные статистические характеристики, предложенные в работе [10].

В результате обучение происходит дважды: вначале на выборке качественных документов обучается модель СРД, затем по обучающему набору обучается классификатор спама.

Таблица 1. Характеристики классификации неестественных текстов различными вариантами алгоритма

	Точность	Полнота	F-мера
Мешок слов. Базовая версия	98,41%	98,50%	98,45%
ЦМ-2. Базовая версия	96,19%	96,11%	96,15%
ЦМ-3. Базовая версия	94,08%	92,29%	93,18%
Предложения. Базовая версия	92,69%	91,87%	92,28%
Мешок слов. Улучшенная версия	99,70%	99,25%	**99,47%**
ЦМ-2. Улучшенная версия	98,37%	97,93%	**98,15%**
ЦМ-3. Улучшенная версия	97,72%	97,09%	**97,40%**
Предложения. Улучшенная версия	96,23%	97,03%	**96,63%**

В данной работе используется алгоритм классификации, основанный на деревьях решений C4.5 с добавлением процедуры объединения нескольких деревьев путем голосования.

При классификации текстов, с помощью предварительно обученной модели СРД, вычисляется разнообразие тематической структуры текстов, затем вычисляются характеристики, предложенные в работе [10]. В итоге автоматический классификатор оценивает вероятность, что текст неестественный на основе факторов тематического разнообразия и других статистических факторов.

5. Результаты экспериментов

Важной частью работы является экспериментальное подтверждение применимости предлагаемого алгоритма обнаружения неестественных текстов.

В рамках первого эксперимента проверялась способность предложенного алгоритма решать модельную задачу – обнаруживать документы, порожденные каждым из рассмотренных алгоритмов генерации неестественных текстов.

Измерялась точность полнота и F-мера обнаружения неестественных текстов при использовании классификатора, обученного на выборке естественных текстов из коллекции ROMIP.ByWeb и наборе текстов, порожденных различными генераторами текстов:

- Генератор на основе модели «мешок слов» (мешок слов);
- Генератор на основе цепей Маркова порядка 2 (ЦМ-2);
- Генератор на основе цепей Маркова порядка 3 (ЦМ-3);
- Генератор на основе копирования предложений (предложения);

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

В ходе эксперимента было построено два классификатора для каждой тренировочной выборки. В качестве базового был взят классификатор с использованием характеристик, предложенных в работе [10]. Также была построена улучшенная версия классификатора с добавлением характеристик тематического разнообразия предложенных в данной работе.

Разница в точности и полноте классификаторов позволяет оценить выигрыш при использовании характеристик тематического разнообразия. Результаты эксперимента приведены в таблице 1.

Чтобы сравнить предлагаемые алгоритмы обнаружения поискового спама с существующими аналогами был проведен ряд экспериментов на наборе данных WebspamUK-2007 [11]. Этот набор представляет собой набор всех страниц из доменной зоны .uk, собранный за 2007 год. 4000 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 – базовая без характеристик тематической структуры и улучшенная, содержащая характеристики, предложенные в разделе 4. Версии алгоритма сравнивались между собой, а также с лучшими результатами других исследователей на данном наборе. В частности 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 [12].

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 под ROC-кривой, тем в среднем больше полнота алгоритма при фиксированной точности.

Результаты эксперимента и сравнение с существующими аналогами приведено в таблице 2. Как видно, улучшенный алгоритм превосходит, как базовую версию, так и лучшие на текущий момент алгоритмы классификации поискового спама.

Таблица 2. Результаты эксперимента на наборе WebspamUK-2007

Алгоритм	AUC
Базовый алгоритм	0.847
Улучшенный алгоритм	**0.870**
Победитель WSC-2008 [13]	0.850
Linked LDA [12]	0.854

6. Результаты работы

Для решения задачи определения автоматически порожденных неестественных текстов разработан новый алгоритм машинного обучения на основе оценки разнообразия тематик документа.

Теоретически и численно обоснована применимость разработанного алгоритма для обнаружения неестественных текстов, порожденных генераторами текстов на основе цепей Маркова, широко используемых для создания веб-спама.

Разработанный алгоритм апробирован на стандартном наборе данных реальных сайтов WebspamUK-2007. Получены более высокие характеристики классификации веб-спама, по сравнению с известными методами.

Литература

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

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

[3] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna, A reference collection for web spam, ACM SIGIR Forum, v.40 n.2, p.11-24, December 2006.

[4] Анализаторы поисковых машин. http://analyzethis.ru/. 2011.

[5] 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.

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

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

[8] Веб коллекция BY.web. http://romip.ru/ru/collections/by.web-2007.html.

[9] 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.

[10] 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.

[11] 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.

[12] I. Bнry, D. Siklysi, J. Szaby, A. A. Benczъr, 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.

[13] 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.

[14] Web Spam Challenge. http://webspam.lip6.fr/wiki/pmwiki.php, 2008.

Detecting Mass-Generated Unnatural Texts through Topical Diversity Analysis

© A.S. Pavlov

This work is dedicated to development of methods and tools for detecting mass-generated unnatural texts. A generalized model of unnatural texts generated using natural samples is proposed. An algorithm for unnatural texts detection is also proposed. The algorithm is based on theoretical properties of unnatural texts and uses topical diversity analysis to distinguish natural and unnatural texts. The proposed algorithm is evaluated on web spam detection task.