УДК 681.513.7

# МЕТОД ОБНАРУЖЕНИЯ МАССОВО ПОРОЖДЕННЫХ НЕЕСТЕСТВЕННЫХ ТЕКСТОВ НА ОСНОВЕ АНАЛИЗА ТЕМАТИЧЕСКОЙ СТРУКТУРЫ

**А. С. Павлов**[1], **Б. В. Добров**[2]

Поисковый спам — одна из основных угроз для современных поисковых систем. Спамеры используют разнообразные алгоритмы для массового порождения неестественных текстов. Рассматривается обобщенная теоретическая модель текстов, порождаемых на основе документов-образцов, а также предложен новый алгоритм обнаружения неестественных текстов на основе анализа тематической структуры текстов. Предложенный алгоритм апробирован на синтетических и реальных данных.

**Ключевые слова:** поисковый спам, тематическая структура, моделирование.

**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 ему. Мера соответствия страницы запросу, называемая релевантностью, вычисляется на основе сопоставления характеристик страниц и запросов.

Поиск товаров и услуг — одна из значительных потребностей пользователей поисковых систем. Появление сайта в первой десятке выдачи популярных поисковых систем по таким запросам обеспечивает значительный приток посетителей на сайт и как следствие большое количество покупателей. В связи с этим возникает конкуренция между создателями сайтов за попадание на верхние позиции выдачи поисковых систем. Она приводит к тому, что некоторые создатели сайтов пытаются повлиять на результаты работы алгоритмов, применяемых в поисковых системах, чтобы незаслуженно повысить оценку релевантности страниц. Это явление получило название поискового спама [1].

Поисковый спам зачастую ухудшает качество результатов 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 котором она расположена; ссылки между страницами и сайтами и пр. В настоящее время существует несколько разновидностей поискового спама, нацеленных на дискредитацию различных алгоритмов, применяемых внутри поисковых систем. Например, ссылочный спам нацелен на обман алгоритмов ссылочного ранжирования, таких как PageRank [5]. Данная работа посвящена исследованию методов противодействия другой разновидности поискового спама — массово порождаемым неестественным текстам.

При порождении текстов целью спамеров является попадание в выдачу по запросам с малым количеством релевантных страниц. 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 такие страницы с большой вероятностью попадают в выдачу.

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 [1]: копирование существующих естественных текстов и 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 [6]. В связи с этим широкое распространение получили алгоритмы автоматического порождения текстов.

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 119991, Москва; аспирант, e-mail: pavvloff@gmail.com

[2] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, Ленинские горы, д. 1, стр. 4, 119991, Москва; зав. лаб., e-mail: dobroff@mail.cir.ru

Данная работа посвящена алгоритмам обнаружения массово порождаемых неестественных текстов. Раздел 2 посвящен обзору существующих методов обнаружения поискового спама и неестественных текстов. Раздел 3 посвящен описанию теоретической модели текстов, порожденных на основе образцов естественных документов. Раздел 4 описывает новый алгоритм обнаружения неестественных текстов. В разделе 5 приведены результаты исследования применимости предложенных алгоритмов на модельных данных. Раздел 6 посвящен апробации предложенного алгоритма на реальной коллекции.

**2. Обзор существующих решений.** Поисковый спам нацелен на дискредитацию различных алгоритмов поисковой системы и разделяется на несколько направлений. Ссылочный спам создается, чтобы повлиять на вес ссылочного ранжирования для определенной страницы. Примерами ссылочного спама являются специальные сети страниц, связанных ссылками. Такие структуры нацелены на алгоритмы ссылочного ранжирования, аналогичные PageRank [5]. Существует обширный класс алгоритмов, нацеленных на борьбу со ссылочным спамом, например [7].

Другим важным направлением в борьбе с поисковым спамом является обнаружение дубликатов текстов. Обзор методов обнаружения дубликатов приведен в работе [6]. В основе большинства методов обнаружения дубликатов 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 этой работе предлагается использовать метод скрытого распределения Дирихле для определения спамерских 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. В работе предлагается алгоритм для подсчета доли редких пар слов и показывается, что эта характеристика улучшает качество определения поискового спама.

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.

Подход, предлагаемый в настоящей статье, опирается на работу [11], однако существенно расширяет его на основе учета свойств рассматриваемой модели тематической структуры текстов для определения неестественных текстов.

**3. Модель массово порождаемых неестественных текстов.** Автоматическая генерация естественных текстов в настоящее время является нерешенной задачей. Естественным текстам свойственно большое количество закономерностей, которые сложно одновременно воспроизвести автоматическими методами [12]: локальная связность; единство стиля и жанра; синтаксическая структура предложений; глобальная тематическая связность текста; структура изложения и т.п.

Исследования автоматических методов построения аннотаций показывают, что даже специализированные алгоритмы плохо воспроизводят все характеристики естественных текстов [12].

Идея порождения текстов по образцам основывается на том, что воспроизведение даже некоторых свойств естественных текстов резко усложняет задачу обнаружения синтетических текстов. Алгоритмы генерации текстов по образцам используют набор естественных текстов в качестве обучающего набора и при генерации воспроизводят некоторые их свойства.

Далее используются следующие условные обозначения: $V$ — множество всех слов во всех текстах; $v_i$ — $i$-е слово из $V$; $d$ — документ, конечная последовательность слов; $|d|$ — длина документа, количество слов в последовательности; $D_{\text{templ}}$ — множество документов-образцов для алгоритмов генерации текстов; $d_{\text{gen}}$ — порожденный документ; $w_{dj}$ — $j$-я словопозиция в документе $d$; $C\big(\{v_1,\ldots,v_k\};d\big)$ — количество вхождений последовательности слов $\{v_1,\ldots,v_k\}$ в документ $d$.

**3.1. Обзор методов порождения неестественных текстов на основе образцов.** При генерации поискового спама используются различные алгоритмы порождения текстов. В данном разделе мы рассмотрим алгоритмы, которые используют обучающую выборку естественных текстов для построения автоматического генератора текстов.

**3.1.1. Модель “мешок слов”.** Наиболее простой моделью для порождения искусственных текстов является модель на основе “мешка слов”. В рамках данной модели на основе обучающей выборки естественных текстов собирается частота употребления различных слов. Вероятность того, что при порождении

$n$-й словопозиции в документе $d_{gen}$ будет порождено слово $v$, равна доле слов $v$ в наборе документов-образцов: $P(w_{d_{\text{gen}},n} = v) = \dfrac{\sum\limits_{d \in D_{\text{templ}}} C\big(\{v\}; d\big)}{\sum\limits_{d \in D_{\text{templ}}} |d|}$. Данный метод крайне прост в реализации и уже воспроизводит некоторые особенности естественных текстов. Например, популярные алгоритмы классификации текстов, которые опираются только на представление текста в виде "мешка слов" , будут малоэффективны при обнаружении таких текстов. В остальном данный метод порождения текстов нарушает практически все свойства, присущие естественным текстам.

**3.1.2. Генераторы на основе цепей Маркова.** Одним из наиболее распространенных подходов к массовому порождению текстов являются генераторы текстов на основе цепей Маркова. Их основное отличие от простой модели мешка слов заключается в том, что при генерации слов в тексте учитывается локальный контекст. При использовании цепей Маркова порядка $k$ при порождении слова учитываются $k$ предыдущих порожденных слов. Вероятность порождения слова $v$ выражается формулой:

$$P(w_{d_{\text{gen}},n} = v) = \frac{\sum\limits_{d \in D_{\text{templ}}} C\big(\{w_{d_{\text{gen}},n-k}, \ldots, w_{d_{\text{gen}},n-1}, v\}; d\big)}{\sum\limits_{d \in D_{\text{templ}}} C\big(\{w_{d_{\text{gen}},n-k}, \ldots, w_{d_{\text{gen}},n-1}\}; d\big)}. \tag{1}$$

Порождение текстов на основе такой цепи Маркова происходит по следующей процедуре. Вначале выбирается произвольное состояние в цепи, с которого начинается порождение. Затем на каждом шаге выбирается один из возможных переходов из текущего состояния, при этом порождается слово, соответствующее этому состоянию. Порождение заканчивается, когда порожденный текст достигает необходимой длины.

Ниже приведен пример текста, порожденного генератором на основе цепей Маркова длины 2, обученным на данной статье:

> *Генераторы текстов из отдельных документов. Для проверки возможности обнаруженя дубликатов Задача определения текстов, порожденных генераторами текстов из рассматриваемых генераторов текстов. Проверялась эффективность данного метода позволяет формулировать критерии принадлежности спаму или неспаму.*

Данный метод часто используется спамерами, так как тексты, порожденные таким способом, обладают локальной связностью, присущей документам-образцам. Например, пара слов никогда не встретится в порожденном тексте, если она не встречалась в одном из документов-образцов.

Порядок цепи Маркова $k$ позволяет регулировать поведение алгоритма генерации текстов. Чем больше $k$, тем длиннее контекст, учитываемый алгоритмом, тем меньше различных слов можно породить на каждом шаге. При больших $k$ алгоритм копирует куски исходных текстов, что увеличивает вероятность обнаружения порожденного документа методами анализа дубликатов. В то же время, при уменьшении $k$ ухудшается локальная связность. Метод на основе "мешка слов" можно рассматривать как вырожденный случай генератора на основе цепей Маркова нулевого порядка. На практике чаще всего применяются цепи Маркова длины 2 и 3.

При порождении слов согласно формуле (1) возникают ситуации, когда не существует слова с ненулевой вероятностью порождения. Это происходит, если на конце документа встретилась уникальная последовательность из $k$ слов. Такие состояния в цепи Маркова назовем тупиковыми. В реальных генераторах применяются различные техники, для того чтобы продолжить генерацию текстов при достижении тупикового состояния.

В рамках теоретического исследования для простоты мы будем придерживаться алгоритма, при котором документы "закольцовываются" — считается, что после последнего слова в тексте идет первое слово. В разделе 5.4 будет показано, что предлагаемые алгоритмы обнаружения порожденных текстов применимы для различных методов обработки тупиковых состояний.

В качестве примера рассмотрим два документа-образца: $d_1 = \{v_1, v_2, v_3\}$, $d_2 = \{v_1, v_1, v_2\}$. Обучим по ним генератор текстов на основе цепи Маркова порядка 1. На рис. 1 приведена получившаяся цепь Маркова. Веса на ребрах графа обозначают вероятности перехода между состояниями. Переходы из состояния $v_3$ в состояние $v_1$ и из $v_2$ в $v_1$ возникают из-за того, что документы "закольцованы".

**3.1.3. Метод на основе фрагментов текстов.** Еще один распространенный метод генерации уникальных текстов заключается в объединении различных фрагментов документов-образцов в один документ. Один из вариантов такого алгоритма заключается в разбиении исходных текстов на предложения и составлении нового текста из произвольной последовательности предложений.

Данный метод, так же как и метод на основе цепей Маркова, порождает тексты, обладающие локальной связностью. Еще одним важным свойством, которое воспроизводит данный метод, является синтаксическая структура естественных предложений. Даже человек не сможет отличить такой текст от естественного, не вчитываясь в него. Основной недостаток такого подхода — более высокая вероятность обнаружения таких текстов методами анализа дубликатов.

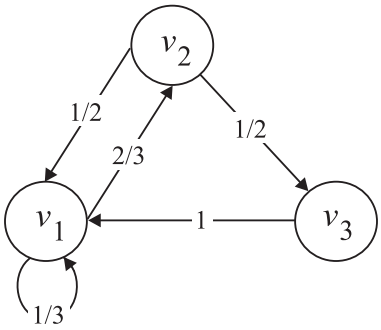


Рис. 1

**3.1.4. Обобщенная модель генератора текстов на основе 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 $D_{\text{templ}}$ — набор текстов-образцов. Рассмотрим множество троек $(t)$, документ $(d)$, номер словопозиции $(m)$, слово, стоящее в данной позиции $(v)$: $T = \{t\} = \big\{(d, m, v)\big\}$, $|T| = \sum\limits_{d \in D_{\text{templ}}} |d|$.

Любой из перечисленных алгоритмов порождения текстов на основе образцов можно представить в виде однородной цепи Маркова с пространством состояний $T$, у которой переходная матрица определяется разновидностью алгоритма.

Пусть $t_i = (d_i, m_i, v_i)$, $t_j = (d_j, m_j, v_j)$ — два элемента из пространства состояний. Обозначим через $P_{t_i t_j}$ вероятность перехода между двумя этими состояниями, где $P$ — матрица переходов для данной цепи. Выпишем $P_{t_i t_j}$ для алгоритмов порождения текстов, описанных выше.

**Модель “мешок слов”.** Так как в данной модели вероятность порождения любого слова на любом шаге пропорциональна количеству словопозиций, содержащих это слово, то матрица переходов будет иметь простейший вид:

$$P_{t_i t_j} = P\big(X_{n+1} = t_j \,|\, X_n = t_i\big) = \frac{1}{|t|}\,. \tag{2}$$

**Генераторы на основе цепей Маркова.** Элемент матрицы переходов для этого алгоритма не равен нулю, только если предыдущие $k$ слов для двух состояний совпадают:

$$P_{t_i t_j} = \begin{cases} \dfrac{1}{\sum\limits_{d \in D_{\text{templ}}} C\big(\{w_{d_i, m_i - k + 1}, \ldots, w_{d_i m_i}\}; d\big)}, & \text{если} \quad \forall l \in [1; k] : w_{d_j, m_j - l} = w_{d_i, m_i - (l-1)}; \\ 0, & \text{иначе.} \end{cases} \tag{3}$$

Для того чтобы отразить “закольцованность” документа с точки зрения генератора текстов, в формуле (3) допускаются нулевые или отрицательные номера словопозиций в документе, и такие номера словопозиций отсчитываются в обратном порядке с конца документа.

**Генераторы 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:

$$P_{t_i t_j} = \begin{cases} \dfrac{1}{|B|}\,, & \text{если} \quad t_i \in E,\ t_j \in B; \\ 1, & \text{если} \quad d_i = d_j,\ m_j = m_i + 1,\ t_i \notin E; \\ 0, & \text{иначе.} \end{cases} \tag{4}$$

Рассмотрим построение обобщенных цепей на примере, приведенном в разделе 3.1.2. Воспользуемся формулами (3) и (4) и построим графы переходов между состояниями обобщенных цепей для различных вариантов генераторов. На рис. 2 приведен граф переходов для генератора на основе цепей Маркова длины 1, обученного на документах $d_1 = \{v_1, v_2, v_3\}$ и $d_2 = \{v_1, v_1, v_2\}$. На рис. 3 — для генератора на основе фрагментов, обученного на тех же документах. Все ребра, выходящие из одной вершины, имеют одинаковые веса (соответствующие вероятностям перехода) и в сумме дают единицу.

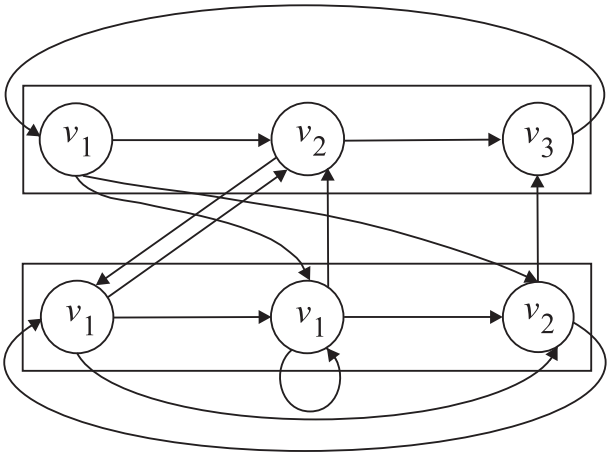


Рис. 2

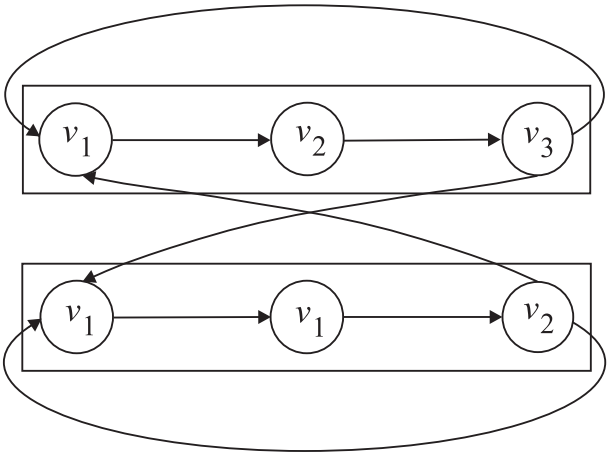


Рис. 3

Опишем процесс порождения текста на основе обобщенной модели. На первом шаге произвольным образом выбирается начальное состояние цепи $t_0 = (d_0, m_0, v_0)$. На каждом последующем шаге, исходя из матрицы вероятностей $P$, выбирается следующее состояние: при переходе в состояние $t = (d, v, m)$ к порождаемому документу добавляется слово $v$. Процесс заканчивается, когда порожденный текст достигает определенной длины.

Рассмотрим некоторые свойства цепей Маркова для обобщенных генераторов текстов.

1. *Невозвратные состояния.* Цепь Маркова для обобщенного генератора не может содержать невозвратных состояний по построению.

2. *Однородность.* Очевидно, что по построению матрица переходов цепи Маркова для обобщенного генератора не зависит от времени, т.е. цепь однородная.

3. *Неразложимые классы.* Если цепь Маркова содержит несколько неразложимых классов, то будем рассматривать каждый из них по отдельности, так как каждый искусственный документ будет порожден одним из классов. При этом каждый неразложимый класс будет состоять из отдельного набора документов-образцов. Во всех дальнейших рассуждениях будем рассматривать только неразложимые цепи Маркова и множества документов-образцов, соответствующих им.

Таким образом, можно считать, что при порождении текстов на основе документов-образцов используются однородные неразложимые цепи Маркова.

**Лемма 1.** *Пусть задана цепь Маркова с множеством состояний* $T = \{t\} = \{(d, m, v)\}$ *и матрицей перехода, заданной одной из формул* (2)–(4), *тогда существует единственное равновесное распределение вероятностей состояний* $\pi = \left(\dfrac{1}{|T|}, \ldots, \dfrac{1}{|T|}\right)$.

**Доказательство.** Во всех трех рассматриваемых случаях граф переходов для цепи Маркова ориентированно связан, следовательно, рассматриваемые цепи Маркова эргодические и у них существует единственное равновесное состояние. Осталось показать, что если на шаге $n$ распределение вероятностей состояний равно $\pi = \left(\dfrac{1}{|T|}, \ldots, \dfrac{1}{|T|}\right)$, то на $n+1$ шаге оно будет таким же.

Очевидно, что для формулы (2) это условие выполняется. Покажем, что оно выполняется для формулы (3). Для этого рассмотрим произвольное состояние $t = (d, v, m)$. Рассмотрим множество состояний $T' = \{t'\}$, из которых можно попасть в состояние $t$. Вероятность перехода из каждого состояния этого множества выражается по формуле (3):

$$P_{t't} = \frac{1}{\sum\limits_{d \in D_{\text{templ}}} C\big(\{w_{d',m'_i-k+1}, \ldots, w_{d'm'}\}; d\big)}\,. \tag{5}$$

Заметим, что по построению цепи Маркова знаменатель в формуле (5) равен количеству состояний, из которых можно попасть в состояние $t$: $P_{t't} = \dfrac{1}{|T'|}$.

Пусть вероятность каждого состояния из $T'$ на шаге $n$ равна $\dfrac{1}{|T|}$, выпишем вероятность перехода в

состояние $t$ на шаге $n+1$: $P(X_{n+1}=t)=\sum\limits_{t'\in T'}P_{t't}P(X_n=t')=\sum\limits_{t'\in T'}\frac{1}{|T'|}\frac{1}{|T|}=\frac{1}{|T|}$.

В силу произвольности выбора $t$, $\pi=\left(\frac{1}{|T|},\ldots,\frac{1}{|T|}\right)$ является равновесным состоянием для цепи Маркова.

Проведем аналогичные рассуждения для матрицы переходов, построенной по формуле (4). Пусть на шаге $n$ распределение вероятностей состояний равно $\pi=\left(\frac{1}{|T|},\ldots,\frac{1}{|T|}\right)$. Рассмотрим множество состояний $T'$, из которых можно попасть в состояние $t$.

Рассмотрим два случая. Пусть $t$ не принадлежит множеству состояний $B$, которое соответствует множеству начал фрагментов. Тогда $T'$ состоит ровно из одного состояния $t'$, вероятность которого на шаге $n$ равна $\frac{1}{|T|}$; следовательно, вероятность состояния $t$ на шаге $n+1$ также будет равна $\frac{1}{|T|}$.

Рассмотрим второй случай, когда $t$ принадлежит множеству состояний начал фрагментов $B$. В этом случае множество $T'$ совпадает с множеством состояний концов фрагментов $E$. Очевидно, что $|B|=|E|$, следовательно: $P(X_{n+1}=t)=\sum\limits_{t'\in E}P_{t't}P(X_n=t')=\sum\limits_{t'\in E}\frac{1}{|B|}\frac{1}{|T|}=\frac{|E|}{|T|\,|B|}=\frac{1}{|T|}$.

Таким образом, для цепи Маркова, построенной в соответствии с формулой (4), также выполняется утверждение леммы.

**Лемма 2.** *При порождении текстов с помощью обобщенной цепи Маркова доля слов, порожденных из любого состояния $t$, сходится по вероятности к $\frac{1}{|T|}$.*

**Доказательство.** Рассмотрим два случая: апериодическую цепь Маркова и цепь Маркова с периодом $q$.

Пусть цепь Маркова апериодична. Тогда согласно лемме 1 для любого изначального состояния $\tau$ верно: $\pi=\left(\frac{1}{|T|},\ldots,\frac{1}{|T|}\right)=\lim\limits_{n\to\infty}\tau P^n$.

Выпишем вероятность того, что в итоговом порожденном документе длины $l$ встретится слово, порожденное из состояния $t$: $P(v_t\in d_{\text{gen}})=\frac{1}{l}\sum\limits_{n=1}^{l}P(X_n=t)=\frac{1}{l}\sum\limits_{n=1}^{l}\left[\tau P^n\right]_t$.

Очевидно, что при $l\to\infty$ вероятность встретить слово, порожденное из состояния $t$, будет сходиться к $\frac{1}{|T|}$; следовательно, и доля слов будет сходиться к $\frac{1}{|T|}$ по вероятности.

Пусть цепь Маркова периодична с периодом $q$. Рассмотрим цепь Маркова с тем же пространством состояний, но с матрицей переходов $P^q$. Эта цепь Маркова будет апериодична и разложима на $q$ неразложимых классов, при этом по построению каждый из классов будет состоять ровно из $\frac{|T|}{q}$ состояний. Для каждого из неразложимых классов утверждение леммы будет справедливо.

Порождение текста исходной цепью Маркова можно рассматривать как поочередное порождение слов из каждого неразложимого класса. Таким образом, доля слов, порожденных из состояния $t$, будет выражаться через долю слов, порожденных неразложимым классом, содержащим это состояние $(1/q)$, и долю таких слов в этом классе (сходится по вероятности к $q/|T|$), т.е. будет сходиться по вероятности к $\frac{1}{|T|}$, что и требовалось доказать.

**Следствие.** *Доля слов в порожденном документе из документа-образца $d$ сходится по вероятности к $\frac{|d|}{\sum\limits_{d'\in D_{\text{templ}}}|d'|}$ с ростом длины порожденного документа.*

**Доказательство.** Выпишем вероятность порождения слова из документа $d$ на $n$-м шаге. Ее можно представить выражением: $P\big(X_n=t\,:\,t=(d,m,v)\big)=\sum\limits_{t'\in d}\frac{1}{|T|}=\frac{|d|}{|T|}=\frac{|d|}{\sum\limits_{d'\in D_{\text{templ}}}|d'|}$.

Что и требовалось доказать.

**Определение 1.** Если для генератора текстов на основе образцов можно построить обобщенную цепь Маркова, которая будет удовлетворять лемме 2, то такой генератор будем называть *смешивающим.*

Особенность таких генераторов текстов в том, что они смешивают слова из документов-образцов, в результате чего нарушаются некоторые закономерности естественных текстов. В дальнейшем мы будем рассматривать только смешивающие генераторы текстов.

**3.2. Тематическая структура текстов.** Изучение неестественных текстов показывает, что они зачастую бессмысленны и лишены единой темы. 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.

Одной из важных характеристик естественных текстов является глобальная тематическая связность текстов. У текстов чаще всего есть одна основная тема и несколько второстепенных. Предлагаемый подход к формализации понятия тематики аналогичен лингвистической теории, сформулированной в статье [13]. 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. Кроме того, предлагаемая модель тематик текстов основывается на порождающих тематических моделях текстов [14], в которых утверждается, что каждому естественному документу соответствует некоторый набор весов тематик, из которых порождаются все слова документа.

В данной статье рассматривается модель тематик текста, в которой каждая словопозиция в тексте относится к одной тематике. Одно и то же слово на разных позициях может соответствовать разным тематикам. В такой модели вклад тематики в документ измеряется долей словопозиций в этом документе, которым соответствует эта тематика.

Кроме того, в модели предполагается, что тематик конечное число. Пусть $\Theta = \{\Theta_1, \ldots, \Theta_K\}$ — множество всех тематик, тогда вектор $\theta^d = \left(\theta_1^d, \ldots, \theta_K^d\right)$ называется *тематической структурой* документа $d$, если доля словопозиций, принадлежащих тематике $i$, равна $\theta_i^d$: $\theta_i^d = \dfrac{\left|\{w : w \in d, w \in \Theta_i\}\right|}{|d|}$.

**3.3. Свойства тематической структуры порожденных текстов.** Тематическая структура текстов может применяться для обнаружения неестественных текстов. Покажем, что генераторы текстов на основе образцов нарушают тематическую структуру естественных текстов.

**Теорема 1.** *Пусть $D_{\text{templ}}$ — набор документов-образцов для генератора текстов и задана тематическая структура каждого документа. Пусть на их основе обучен смешивающий генератор текстов и с помощью него порождается документ $d_{\text{gen}}$ длины $l$. Тогда с ростом $l$ тематическая структура порожденного документа $\theta^{\text{gen}}$ сходится по вероятности к усредненной тематической структуре документов-образцов:*

$$\theta^{\text{gen}} \xrightarrow{\mathbb{P}} \frac{\sum\limits_{d \in D_{\text{templ}}} |d| \theta^d}{\sum\limits_{d \in D_{\text{templ}}} |d|}. \tag{6}$$

**Доказательство.** Выпишем случайную величину — долю словопозиций в порожденном тексте, которые принадлежат тематике $\Theta_i$:

$$\theta_i^d = \frac{\left|\{v : v \in \Theta_i\}\right|}{l} = \sum_{d \in D_{\text{templ}}} \frac{\left|\{v : v \in d\}\right|}{l} \theta_i^d. \tag{7}$$

Нетрудно заметить, что $\dfrac{\left|\{v : v \in d\}\right|}{l}$ — это доля словопозиций в порожденном документе из документа-образца $d$. Поэтому можно применить следствие к лемме 2, откуда и получается утверждение теоремы.

**4. Метод обнаружения неестественных текстов.** В предыдущем разделе была показана связь между тематической структурой документов-образцов и тематической структурой документов, порождаемых из них. В данном разделе предлагается алгоритм для обнаружения неестественных текстов, который основывается на утверждениях теоремы 1.

**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) [14].

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 нену-

левую вероятность порождения в разных тематиках. В данной модели каждому документу ставится в соответствие вектор вероятностей тематик $\theta$. При этом считается, что каждой словопозиции в документе соответствует строго одна тематика.

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

1. Задается матрица вероятностей порождения каждого слова в каждой тематике.

2. Для каждого документа выбирается набор весов тематик $\theta$ на основе распределения Дирихле с вектором параметров $\alpha$. При порождении каждой словопозиции в документе выбираются тематика $\Theta_i$ на основе полиномиального распределения с вектором параметров $\theta$ и слово, которое надо породить, на основе распределения вероятностей порождения слов в выбранной тематике.

| Ученые выяснили, какие отделы мозга активируются при чтении и письме - навыках, которые человечество приобрело совсем недавно по эволюционным меркам и которые не могли привести к физическим изменениям в организации коры. Работы специалистов опубликованы в журнале Science, а их краткое содержание доступно на портале ScienceNow. | | |
|---|---|---|
| Тематика 1 (вес 0.5): | Тематика 2 (вес 0.3): | Тематика 3 (вес 0.2): |
| Наука | Биология | Наука |
| Ученый | Мозг | Ученый |
| Журнал | Эволюция | Журнал |
| Публикация | Отбор | Публикация |
| Статья | Вид | Статья |
| ... | ... | ... |

Рис. 4

Вектор параметров $\alpha$ — это параметр модели. Если все $\alpha$ меньше 1, то из-за свойств распределения Дирихле в документах выделяется небольшое количество главных тематик, к которым относится большая часть словопозиций. Если же принять $\alpha$ больше 1, то документам в большей степени свойственно смешивать веса различных тематик. Исходя из естественного предположения, что в большинстве документов превалирует одна основная тема, авторы СРД предлагают использовать значения $\alpha$ меньше 1 [14].

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

1) слова, которые часто встречаются вместе в одних и тех же текстах, получают высокий вес в одних и тех же тематиках;

2) любое слово может порождаться разными тематиками с разной вероятностью;

3) часто употребляемые слова, такие как предлоги и союзы, будут иметь высокую вероятность порождения в любой тематике.

На рис. 4 приведен пример тематической структуры текста, полученной с помощью модели СРД. Вначале модель была обучена на наборе из 10000 документов из коллекции ROMIP.ByWeb (http://romip.ru/ru/collections/by.web-2007.html). Затем для одного из документов с помощью модели все словопозиции 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.

**4.2. Критерии обнаружения неестественных текстов.** В разделе 3.3 было показано, что порождение текстов на основе образцов усредняет тематическую структуру в неестественных текстах. В данном разделе предлагаются две метрики, по которым можно отличить естественные и неестественные тексты.

**4.2.1. Нарушение тематической структуры текстов.** В основе предлагаемого метода обнаружения неестественных текстов лежит предположение, что распределение тематик в естественных текстах близко модели СРД [14]. Если это условие выполняется, то веса тематик в документах подчиняются распределению Дирихле с вектором параметров $\alpha$, все элементы которого меньше 1 и равны между собой.

Проиллюстрируем нарушение тематической структуры на примере, когда количество тематик равно 3. Пусть веса тематик естественных документов подчиняются распределению Дирихле с параметрами $\alpha = (0.5, 0.5, 0.5)$. В ходе эксперимента было получено 10000 векторов тематик, исходя из распределения Дирихле. На рис. 5 приведена плотность распределения вероятностей векторов тематик $\theta = (\theta_1, \theta_2, \theta_3)$ для естественных документов, удовлетворяющих модели СРД.

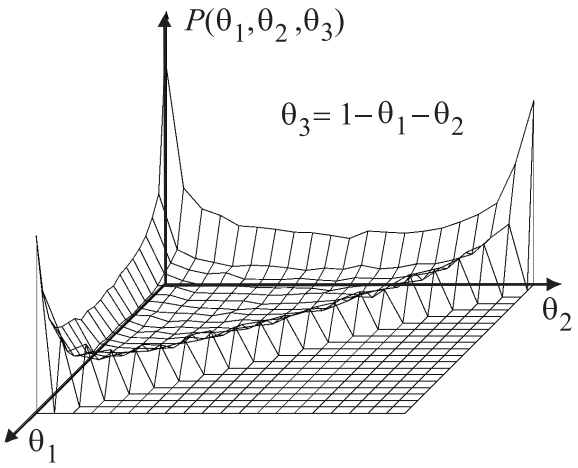


Рис. 5

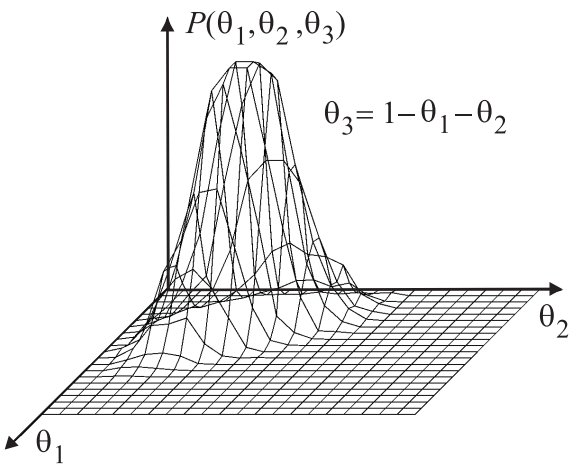


Рис. 6

Затем были порождены вектора тематик, соответствующие неестественным документам. Каждый вектор тематик неестественного документа был получен как среднее арифметическое 10 векторов естественных документов. Таким образом, имитируется ситуация, когда с помощью смешивающего генератора текстов на основе 10 естественных документов-образцов порождается неестественный документ. На рис. 6 приведена плотность распределения вероятностей векторов тематик $\theta = (\theta_1, \theta_2, \theta_3)$ для неестественных документов, порожденных из 10 документов-образцов.

По рис. 5 и 6 хорошо видно, что если естественные тексты подчиняются модели СРД, то распределения весов тематик для естественных и неестественных документов значительно отличаются. В данной работе тематики текстов моделировались с помощью модели СРД, обученной на 10000 документах из коллекции Romip.ByWeb. При этом использовались следующие параметры модели: количество тематик $K = 100$; вектор параметров распределения Дирихле: $\theta = (0.01, \ldots, 0.01)$.

В разделах 4.2.2 и 4.2.3 предлагаются численные характеристики, которые позволяют оценить степень нарушения тематической структуры документов.

**4.2.2. Критерий Пирсона.** Из теоремы 1 следует, что с ростом длины документа, порожденного смешивающим генератором текстов, его тематическая структура будет стремиться к усредненной тематической структуре набора документов-образцов. 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. Пусть $\overline{\theta}$ — усредненный вес тематик в документах-образцах, вычисленный по формуле (6), тогда

$$\chi^2(d) = K \sum_i \frac{\left(\overline{\theta}_i - \theta_i^d\right)^2}{\overline{\theta}_i}.$$

В реальности усредненные веса тематик документов-образцов могут быть не известны. Чтобы обойти эту трудность, воспользуемся особенностью модели СРД. В модели считается, что веса тематик документов подчиняются распределению Дирихле с однородными параметрами $\alpha$. Математическое ожидание такой случайной величины равно $\left(\frac{1}{|K|}, \ldots, \frac{1}{|K|}\right)$. Таким образом, в качестве усредненного веса тематик в документах-образцах можно взять математическое ожидание весов тематик в модели СРД: $\chi^2(d) = K^2 \sum\limits_{i=1}^{K} \left(\frac{1}{K} - \theta_i^d\right)^2$. Чем больше значение величины $\chi^2$, тем с меньшей вероятностью тематики документа распределены равномерно и тем с большей вероятностью документ естественный.

**Утверждение 1.** *При увеличении длины документа, порождаемого смешивающим генератором, и количества документов-образцов генератора величина $\chi^2$ по вероятности сходится к 0.*

**Доказательство.** При произвольном выборе документов-образцов усредненный вес тематик СРД в документах-образцах будет по вероятности сходиться к $\left(\frac{1}{|K|}, \ldots, \frac{1}{|K|}\right)$ при $|D_{\text{templ}}| \to \infty$.

Согласно теореме 1, при увеличении длины порожденного документа распределение тематик в порожденном документе сходится по вероятности к усредненному весу тематик в документах-образцах.

Таким образом, распределение тематик в порожденном документе будет сходиться по вероятности к $\left(\frac{1}{|K|},\ldots,\frac{1}{|K|}\right)$, при этом, согласно формуле (7), величина $\chi^2$ будет по вероятности сходиться к 0.

**4.2.3. Закон Ципфа для тематической структуры.** Естественным текстам свойственен ряд статистических закономерностей, таких как закон Ципфа [15]. Закон Ципфа утверждает, что если упорядочить слова текста по частотности, то частота каждого слова будет обратно пропорциональна его порядковому номеру.

Предлагаемый подход опирается на эмпирические наблюдения, 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 $\theta_k$ с порядковым номером $k$ подчиняется следующему соотношению:

$$\theta_k(s,c) \approx \frac{c}{k^s}\,, \tag{8}$$

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 (8) удобно привести к логарифмической шкале: $\log\bigl(\theta_k(s,c)\bigr) \approx \log(c) - s\log(k)$. Чтобы из этого уравнения получить приближенное значение $s$ для текста, воспользуемся методом наименьших квадратов:

$$f_k = \log\bigl(\theta_k(s,c)\bigr),\quad r_k = \log(k),\quad s = -\frac{K\sum\limits_k r_k f_k - \sum\limits_k r_k \sum\limits_k f_k}{K\sum\limits_k (r_k)^2 - \Bigl(\sum\limits_k r_k\Bigr)^2}\,. \tag{9}$$

Характеристика разнообразия тематик в тексте, вычисленная по формуле (9), может также использоваться как один из факторов для определения неестественных текстов. Чем больше значение параметра Ципфа для текста, тем с большей вероятностью документ естественный.

**Утверждение 2.** *При увеличении длины порождаемого документа и количества документов-образцов величина $s$ по вероятности сходится к 0.*

**Доказательство.** Аналогично доказательству утверждения 1.

**5. Вычислительные эксперименты.** В предыдущем разделе были предложены характеристики, с помощью которых можно классифицировать неестественные тексты. Из утверждений 1 и 2 следует, что при порождении достаточно больших документов по достаточно большому количеству образцов неестественные тексты можно обнаружить, так как характеристики $\chi^2$ и распределение Ципфа будут существенно отличаться у естественных текстов и текстов, порожденных смешивающими генераторами.

На практике спамеры порождают документы различной длины по различному количеству образцов. Таким образом, возникает необходимость исследовать вероятность обнаружения неестественных текстов в зависимости от длины документа и количества образцов. В данном разделе изучается влияние длины порожденного документа, величины набора документов-образцов и метода порождения текстов на вероятность обнаружения неестественных текстов с помощью характеристик, описанных в разделе 4.2.

**5.1. Методология исследования.** Качество предложенных характеристик оценивается на задаче классификации естественных текстов и текстов, порожденных смешивающими генераторами. При этом оценивается точность, полнота и F-мера классификации при использовании только одной характеристики. Точность $P$ — это доля неестественных документов среди всех документов, на которых сработал классификатор. Полнота $R$ — это доля неестественных документов, обнаруженных классификатором. F-мера — это среднее гармоническое между точностью и полнотой.

Пусть $D$ — множество всех документов, $D_{\text{spam}}$ — множество документов, порожденных генератором текстов, $D'_{\text{spam}}$ — множество документов, которые алгоритм классификации отнес к порожденным, тогда точность, полнота и F-мера выражаются по следующим формулам:

$$P = \frac{\bigl|D_{\text{spam}} \cap D'_{\text{spam}}\bigr|}{\bigl|D'_{\text{spam}}\bigr|}\,,\quad R = \frac{\bigl|D_{\text{spam}} \cap D'_{\text{spam}}\bigr|}{|D_{\text{spam}}|}\,,\quad \mathrm{F} = \frac{2PR}{P+R}\,.$$

При классификации использовались простые решающие правила — если значение характеристики меньше порога, то документ считался неестественным, в противном случае он классифицировался как

естественный. Чтобы метод классификации не зависел от выбранного порога в каждом эксперименте выбирался порог, при котором F-мера максимальна.

Для того чтобы оценить вероятность обнаружения текстов, порожденных определенным образом, был проведен ряд численных экспериментов. Каждый из экспериментов проводился по единой схеме.

1. Выбирается 10 тысяч произвольных естественных документов из коллекции Romip.ByWeb.

2. Порождается 10 тысяч документов, по следующему алгоритму: выбирается $N$ произвольных документов-образцов из коллекции Romip.ByWeb; по документам-образцам обучается алгоритм $A$ генерации текстов; генератор текстов порождает документ длины $M$.

3. Вычисляется максимальная F-мера при классификации по характеристике.

**5.2. Зависимость скорости сходимости от количества документов-образцов.** В рамках первого эксперимента проводилось исследование зависимости качества предложенных характеристик от количества документов-образцов. Все порожденные документы в данном эксперименте были фиксированной длины в 6400 слов. При этом было порождено по 10000 документов по 10, 100, 1000 и 10000 документам-образцам. Использовалось 4 различных алгоритма для порождения текстов: на основе “мешка слов” (мешок слов), на основе цепей Маркова порядка 2 (ЦМ-2), на основе цепей Маркова порядка 3 (ЦМ-3), на основе копирования предложений текстов-образцов (предложения).

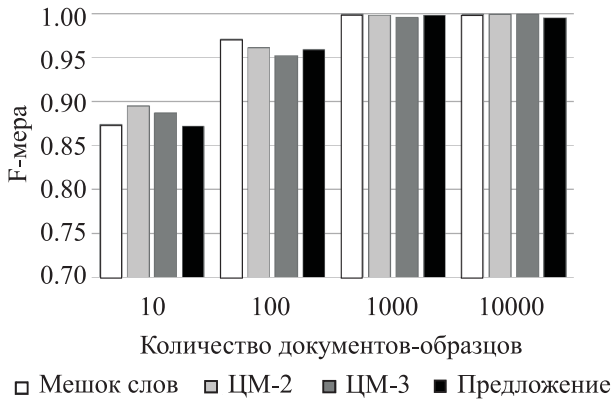


Рис. 7

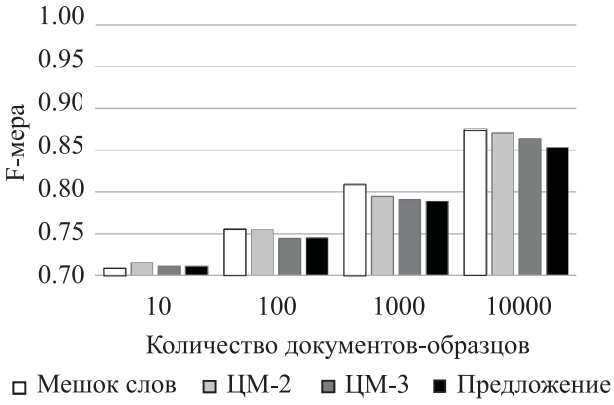


Рис. 8

На рис. 7 приведена зависимость F-меры классификации от количества документов-образцов и метода порождения текстов при использовании статистики $\chi^2$. На рис. 8 приведена зависимость F-меры классификации от количества документов-образцов и метода порождения текстов при использовании закона Ципфа для тематик. Графики позволяют оценить скорость, с которой возрастает качество классификации по одному критерию. Очевидно, что статистика $\chi^2$ существенно превосходит закон Ципфа, так как она достигает почти идеального качества классификации уже при 1000 документов-образцов.

**5.3. Зависимость скорости сходимости от количества слов в документе.** Следующий эксперимент был направлен на исследование качества предложенных характеристик в зависимости от длины порожденных текстов. Все документы в данном эксперименте были порождены по 10 000 документам-образцам, при этом изучались порожденные документы различной длины: 100, 400, 1600 и 6400 слов. Как и в предыдущем эксперименте, использовались различные алгоритмы порождения текстов.

На рис. 9 приведена зависимость F-меры классификации от длины порожденного документа и метода порождения текстов при использовании статистики $\chi^2$. На рис. 10 приведена зависимость F-меры классификации от длины порожденного документа и метода порождения текстов при использовании закона Ципфа для тематик. Качество предложенного алгоритма на малых текстах ниже объясняется, в частности, тем, что порожденный документ длины $N$ не может содержать слова более чем из $N$ документов-образцов.

Кроме того, на малых документах гораздо сильнее заметно влияние алгоритмов порождения. Чем бóльшие куски текстов копируют алгоритмы порождения текстов, тем меньше различных документов-образцов участвуют в порождении одного текста и тем сложнее отличить такой текст от естественного.

**5.4. Применимость критериев для различных генераторов на основе цепей Маркова.** В ходе теоретического исследования методов порождения неестественных документов рассматривались генераторы текстов на основе цепей Маркова, в которых тупиковые состояния удалялись, “закольцовывая” документы. В данном разделе мы исследуем применимость предложенных критериев обнаружения по-

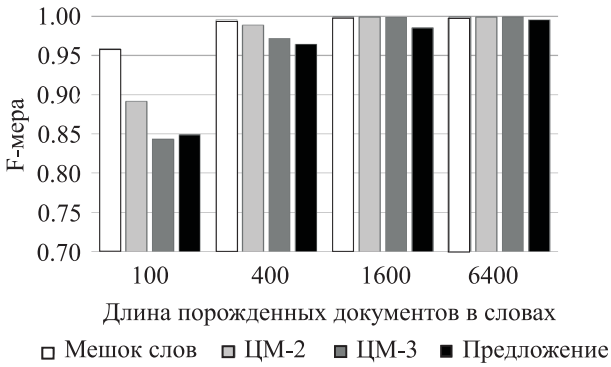


Рис. 9

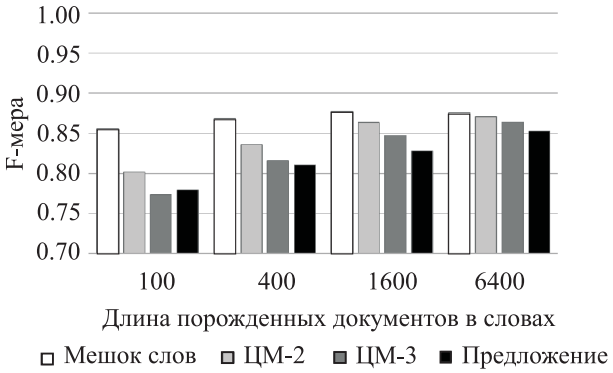


Рис. 10

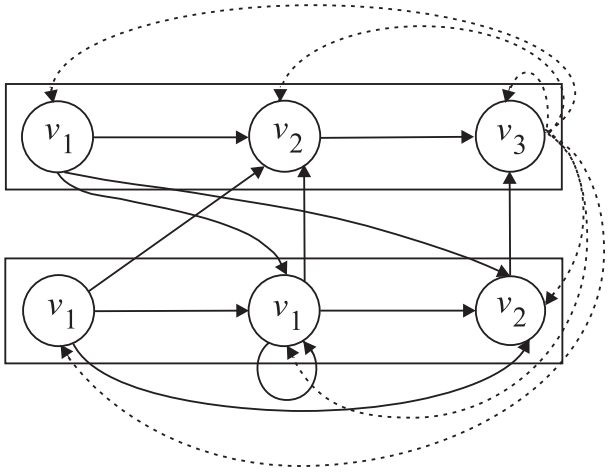


Рис. 11

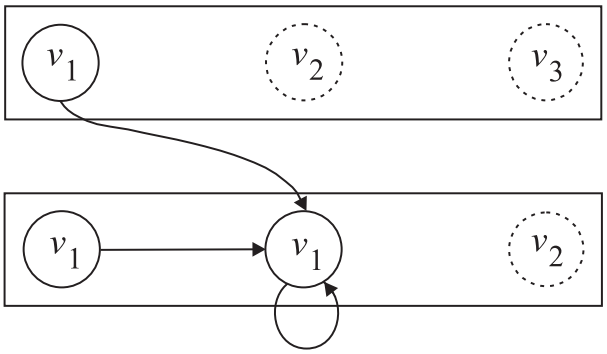


Рис. 12

рожденных документов для других методов учета тупиковых состояний:

1) метод на основе “закольцовывания” документа, рассмотренный в разделе 3.1.2;

2) при достижении тупикового состояния следующее состояние выбирается произвольным образом из множества всех состояний;

3) все тупиковые состояния удаляются.

Покажем, что при использовании таких методов учета тупиковых состояний лемма 1 не выполняется. Для этого построим граф переходов для примера, рассмотренного в разделе 3.1.2. В данном примере состояние со словом $v_3$ является тупиковым.

На рис. 11 приведен граф переходов в случае, когда из тупикового состояния возможен переход в любое другое. Пунктиром обозначены переходы из тупикового состояния. Очевидно, что при генерации слов частота появления слова $v_3$ не может быть меньше частоты слова $v_2$, следовательно, равновесное состояние не может удовлетворять лемме 1.

На рис. 12 приведен граф переходов в случае, когда все тупиковые состояния удалены. Пунктиром обозначены удаленные состояния. В ходе удаления тупиковых состояний помимо состояния $v_3$ были удалены и состояния $v_2$. Получившаяся цепь Маркова также не удовлетворяет условиям леммы 1.

Таким образом, генераторы текстов на основе цепей Маркова, которые используют альтернативные способы учета тупиковых состояний, не являются смешивающими генераторами.

Теоретическое исследование различных моделей учета тупиковых состояний выходит за рамки данного исследования. В то же время относительно несложно эмпирически проверить качество обнаружения текстов, порожденных с использованием различных вариантов генератора на основе цепей Маркова. В ходе эксперимента измерялась максимальная F-мера при обнаружении текстов, порожденных цепями Маркова длины 2 и 3, и различными методами учета тупиковых состояний. Каждым методом было порождено 10 000 документов длиной 6400 слов, каждый документ был порожден с использованием 10 документов-образцов. В качестве критерия классификации использовался критерий $\chi^2$.

Зависимость качества критерия $\chi^2$ от метода учета тупиковых состояний в генераторе на основе цепей Маркова приведена в табл. 1. Эксперимент подтверждает, что различные методы учета тупиковых состояний мало влияют на качество обнаружения текстов, порожденных с помощью генераторов на основе цепей Маркова. Таким образом, предложенный алгоритм обнаружения неестественных текстов может успешно применяться не только для смешивающих генераторов текстов.

Таблица 1

| | Закольцованные документы | Удаление тупиковых состояний | Переход в произвольное состояние |
|---|---|---|---|
| ЦМ-2 | 0.89 | 0.90 | 0.89 |
| ЦМ-3 | 0.89 | 0.88 | 0.87 |

**6. Апробация на реальных данных.** В предыдущем разделе были предложены статистические характеристики, которые могут быть использованы для обнаружения неестественных текстов. Данный раздел посвящен апробации предложенных алгоритмов на данных, максимально приближенных к реальным.

**6.1. Метод на основе большого количества характеристик.** В качестве базового алгоритма классификации неестественных текстов применялся алгоритм на основе большого числа статистических характеристик, предложенный в работе [11]. В основе данного алгоритма лежит гипотеза, что на настоящий момент не существует генератора текстов, который мог бы воспроизвести все закономерности, свойственные естественным текстам.

Метод основывается на выделении большого числа статистических характеристик, которые охватывают различные закономерности естественных текстов. В частности, предлагаются признаки текстов, позволяющие оценить читаемость, единство жанра и стиля и другие характеристики текстов. Многочисленные признаки объединяются в один автоматический классификатор с помощью алгоритма машинного обучения.

Предлагаемый в работе [11] алгоритм машинного обучения опирается на широко распространенный алгоритм построения деревьев решений C4.5. В данном алгоритме строится множество деревьев решений C4.5. Каждое дерево строится по произвольной половине тренировочного набора, а другая половина набора используется, чтобы проставить веса классов в листьях дерева. Такое разделение наборов позволяет уменьшить эффект от переобучения, свойственного алгоритмам на основе деревьев решений.

Деревья строятся до тех пор, пока они улучшают F-меру классификации на тренировочном наборе. На этапе классификации вычисляется вероятность, что документ неестественный, по каждому из деревьев, итоговым результатом классификации является среднее арифметическое вероятностей для различных деревьев. Документ считается неестественным, если алгоритм выдает для него вероятность больше 0.5 .

**6.2. Обнаружение существующих генераторов.** 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 и наборе текстов, порожденных различными генераторами текстов:

а) генератор на основе модели “мешок слов” (мешок слов);

б) генератор на основе цепей Маркова порядка 2 (ЦМ-2);

в) генератор на основе цепей Маркова порядка 3 (ЦМ-3);

г) генератор на основе копирования предложений (предложения).

Кроме того, исследовалась возможность обнаружения текстов, порожденных генераторами дорвеев Doorway.Su (http://doorway.su/) и Rusadult (http://doorways.rusadult.com/ru/), которые в действительности применяются для порождения поискового спама. При порождении текстов использовалось произвольное количество документов-образцов в интервале от 10 до 1000. Длина порождаемых текстов выбиралась произвольным образом из множества длин естественных документов; таким образом, распределение длин неестественных текстов совпадало с распределением длин естественных текстов.

Обучающие выборки составлялись из 10 000 документов из коллекции ROMIP.ByWeb в качестве примеров естественных текстов и 10 000 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 [11]. Кроме того, была построена улучшенная версия классификатора с добавлением характеристик, предложенных в разделе 4. Разница в точности и полноте классификаторов позволяет оценить выигрыш при

использовании характеристик разнообразия. В табл. 2 приведены результаты экспериментов по обнаружению поискового спама, порожденного различными генераторами текстов с помощью различных вариантов алгоритма.

Таблица 2

| | Версии | Точность | Полнота | F-мера | Ошибки 1-го рода | Ошибки 2-го рода |
|---|---|---|---|---|---|---|
| Мешок слов | Базовая | 98.41% | 98.50% | 98.45% | 1.59% | 1.50% |
| | Улучшенная | 99.70% | 99.25% | **99.47%** | **0.30%** | **0.75%** |
| ЦМ-2 | Базовая | 96.19% | 96.11% | 96.15% | 3.81% | 3.89% |
| | Улучшенная | 98.37% | 97.93% | **98.15%** | **1.63%** | **2.06%** |
| ЦМ-3 | Базовая | 94.08% | 92.29% | 93.18% | 5.92% | 7.57% |
| | Улучшенная | 97.72% | 97.09% | **97.40%** | **2.28%** | **2.89%** |
| Предложения | Базовая | 92.69% | 91.87% | 92.28% | 7.31% | 8.06% |
| | Улучшенная | 96.23% | 97.03% | **96.63%** | **3.77%** | **2.99%** |
| Doorway.Su | Базовая | 95.89% | 95.11% | 95.50% | 4.11% | 4.85% |
| | Улучшенная | 98.12% | 97.56% | **97.84%** | **1.88%** | **2.43%** |
| Rusadult | Базовая | 98.16% | 98.40% | 98.28% | 1.84% | 1.60% |
| | Улучшенная | 99.79% | 99.80% | **99.79%** | **0.21%** | **0.20%** |

Важно заметить, что предложенные характеристики работают как на модельных реализациях алгоритмов, так и на реальных генераторах текстов, применяющихся для порождения поискового спама. Кроме того, добавление характеристик тематического разнообразия снижает уровень ошибок первого и второго рода в два раза по сравнению с базовой версией алгоритма.

**6.3. Апробация на наборе WebspamUK-2007.** Чтобы сравнить предлагаемые алгоритмы обнаружения поискового спама с существующими аналогами был проведен ряд экспериментов на наборе данных WebspamUK-2007 [16]. 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.

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.

Если классификатор выдает вероятность принадлежности документа спаму, то площадь под ROC-кривой позволяет оценить качество классификатора без учета порога вероятности, при котором документ признается спамом. Площадь под ROC-кривой равна вероятности того, что для произвольного спамерского документа вероятность, выдаваемая классификатором, будет выше, чем для произвольного неспамерского документа. Таким образом, AUC=1 достигается на классификаторе, у которого существует порог, однозначно разделяющий спам и неспам.

В связи с небольшим размером обучающей выборки алгоритм машинного обучения, основанный на деревьях решений, быстро переобучается, поэтому в данном эксперименте применялся алгоритм логистической регрессии с линейной функцией классификации.

Таблица 3

| Алгоритм | AUC |
|---|---|
| Победитель WSC-2008 [17] | 0.850 |
| Linked LDA [9] | 0.854 |
| Базовый | 0.847 |
| Улучшенный | **0.870** |

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 [17], а также с лучшим результатом на данном

наборе, показанным алгоритмом Linked LDA [9].

Результаты эксперимента на коллекции WebspamUK-2007 и сравнение с существующими аналогами приведены в табл. 3, из которой следует, что улучшенный алгоритм превосходит как базовую версию, так и лучшие на текущий момент алгоритмы классификации поискового спама. Результаты показывают, что алгоритм, предложенный в данной статье, позволяет снизить вероятность того, что классификатор выдаст больший вес неспамерскому документу, чем спамерскому, с 14.6% до 13%. Иными словами, предлагаемый алгоритм ошибается на 10% реже по сравнению с лучшим из существующих аналогов.

**7. Заключение.** В рамках данной работы была разработана и исследована обобщенная модель неестественных текстов, порожденных генераторами на основе образцов. В ходе исследования было теоретически показано, что тексты, порожденные с помощью таких генераторов, нарушают тематическую структуру естественных текстов.

Исходя из теоретической модели, был предложен алгоритм обнаружения неестественных текстов на основе анализа тематической структуры текстов. Построенная теоретическая модель была подтверждена численно в ходе экспериментов. Алгоритм обнаружения порожденных текстов был также протестирован на текстах, синтезированных реальными генераторами поискового спама, а также показал высокое качество классификации реального поискового спама на наборе данных WebspamUK-2007.

Авторы выражают благодарность М. С. Агееву и Н. В. Лукашевич за продуктивную дискуссию и ценные советы.

СПИСОК ЛИТЕРАТУРЫ

1. *Gyongyi Z., Garcia-Molina H.* Web spam taxonomy // Proc. of the 1st Int. Workshop on Adversarial Information Retrieval on the Web. Chiba: ACM, 2005. 39–47.
2. *Henzinger M., Motwani R., Silverstein C.* Challenges in web search engines // SIGIR Forum. 2002. **36**, N 2. 11–22.
3. *Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., Vigna S.* A reference collection for web spam // SIGIR Forum. 2006. **40**, N 2. 11–24.
4. *Ашманов И.С.* Анализатор спама в поисковой выдаче. 2011 (http://analyzethis.ru/?analyzer=spam&location=ru&lang=ru).
5. *Page L., Brin S., Motwani R., Winograd T.* The Pagerank citation ranking: bringing order to the web // World Wide Web Internet And Web Information Systems. Stanford InfoLab. Stanford, 1998. 1–17.
6. *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. Т. 1. Переславль, 2007. 166–174.
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. 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. 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. 37–40.
10. *Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М.* Поиск неестественных текстов // Тр. XI Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Петрозаводск, 2009. 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. 311–317.
12. *Dang H.* Overview of DUC 2006 // Proc. of the Document Understanding. New York: ACM, 2006. 1–10.
13. *ван Дейк Т.А., Кинч В.* Стратегии понимания связного текста // Новое в зарубежной лингвистике. Вып. 23. М.: Прогресс, 1988. 153–211.
14. *Blei D., Ng A., Jordan M.* Latent Dirichlet allocation // J. of Machine Learning Research. 2003. **3**, N 5. 993–1022.
15. *Gelbukh A., Sidorov G.* Zipf and Heaps laws' coefficients depend on language // Proc. of the Second Int. Conference on Computational Linguistics and Intelligent Text Processing. London: Springer, 2001. 332–335.
16. Yahoo! Research: "Web Spam Collections". Milan, 2007 (http://barcelona.research.yahoo.net/webspam/datasets/uk-2007/).
17. *Geng G., Jin X., Wang C.-H.* CASIA at Web spam challenge 2008 Track III // Proc. of the 4th Int. Workshop on Adversarial Information Retrieval on the Web. Beijing: ACM, 2008. 32–33.

Поступила в редакцию
22.06.2011