Метод определения массово порождаемых неестественных текстов

A method of detecting mass generated unnatural texts

Павлов А. С. (pavvloff@gmail.com),
Факультет вычислительной математики и кибернетики
МГУ им. М. В. Ломоносова

Добров Б. В. (dobroff@mail.cir.ru),
Научно-исследовательский вычислительный центр
МГУ им. М. В. Ломоносова;

Рассматривается быстрый метод обнаружения автоматически порожденных неестественных текстов на основе сравнения большого количества статистических характеристик для нормального связного текста и текстов массового поискового спама. Исследуется возможность применимости методов для определения поискового спама, массово порождаемого с помощью генераторов на цепях Маркова, для русского и английского языков. Приводятся значимость отдельных характеристик для детектирования указанного вида поискового спама на этих языках.

1. Введение

Одной из основных проблем для современных поисковых систем является деятельность спамеров, которые наполняют Интернет поисковым спамом для того, чтобы увеличить оценку релевантности продвигаемых спамером страниц в поисковой системе. Наиболее вредным видом поискового спама являются «дорвеи». Дорвеи — это сайты и страницы, не содержащие полезной информации, основной целью которых является перенаправление пользователя, пришедшего с поисковой системы.

Для создания успешного дорвея спамеры должны одновременно удовлетворить нескольким требованиям.

Во-первых, дорвей должен находиться по большому количеству запросов, чтобы собирать наибольшее количество переходов с поисковых систем. Для этого часто спамеры стараются размещать на дорвейных страницах тексты, содержащие поисковые запросы, по которым данные страницы будут формально релевантны.

Во-вторых, дорвеи наиболее эффективны, когда создаются массово и автоматически. При этом для массового создания дорвеев спамеры зачастую прибегают к порождению большого количества неестественных текстов. Такие тексты создаются автоматически без участия человека.

В-третьих, спамеры стремятся затруднить обнаружение дорвеев поисковой системой, чтобы та не смогла исключить автоматически дорвеи из результатов поиска. В настоящее время большое распространение получили алгоритмы порождения дорвеев с помощью специальных программ-генераторов на основе цепей Маркова.

Следует учесть, что основным требованием к алгоритмам, применяемым в поисковых системах, является их высокая производительность. Поисковой системе необходимо обрабатывать миллиарды документов, поэтому алгоритм для обнаружения поискового спама также должен отличаться быстротой.

В рамках данной работы мы продолжаем исследовать метод [1] обнаружения автоматически порожденных неестественных текстов на основе

Dialog'2010.indb 368 11.05.2010 16:57:47

сравнения большого количества статистических характеристик для нормального связного текста и текстов, порождаемых марковскими цепями.

Метод проверки является быстрым и может использоваться в работе глобальных поисковых машин.

Рассматривается возможность его применимости для определения поискового спама, массово порождаемого с помощью генераторов на цепях Маркова, для русского и английского языков. Приводятся значимость отдельных характеристик для детектирования указанного вида поискового спама на этих языках.

2. Существующие методы обнаружения поискового спама

Применимость простых статистических характеристик для определения поискового спама изучалась в работе [2]. Эта работа в основном посвящена общим характеристикам поискового спама, и в ней не исследуются свойства искусственных текстов. Некоторые лингвистические характеристики для обнаружения поискового спама исследовались в работе [3].

Подход, основанный на анализе частот пар слов, предлагается в статье [4]. Данный подход ограниченно применим к генераторам на основе цепей Маркова, так как такие генераторы порождают небольшое количество редких пар слов.

Подходы, не зависящие от конкретной лексики и тематики документов, предлагаются в работах [5, 6]. Первая работа посвящена обнаружению спама в блогах, и использует особенности формата блогов, что ограничивает применимость данного метода. Вторая работа основана на анализе стилистических особенностей HTML-кода страниц, в то время как текстовое содержимое не учитывается в принципе.

Помимо методов обнаружения поискового спама на основе содержимого документов широко распространены методы на основе анализа графа ссылок, например [7].

3. Метод обнаружения неестественных текстов

Одним из наиболее распространенных методов порождения тексов являются генераторы на основе цепей Маркова. Данные генераторы основываются на порождающей модели текстов, где каждое следующее слово зависит от конечного числа предыдущих слов.

Вначале генератор обучается на коллекции естественных текстов, а затем, используя собранные статистики употребления слов, синтезирует последовательность слов, внешне напоминающую текст.

Сложность обнаружения неестественных текстов, порожденных с помощью цепей Маркова, заключатся в том, что они по некоторым характеристикам неотличимы от текстов, созданных человеком:

- они могут содержать фрагменты естественных текстов;
- они могут обладать локальной связностью;
- с точки зрения поисковой машины, они могут быть релевантны некоторым запросам;
- часто исходящие ссылки на этих страницах достаточно хорошо поддержаны лексикой спамерских страниц.

В качестве примера, приведем фрагмент поискового спама, порожденного с помощью генератора текстов.

Данный фрагмент текста расположен по адресу http://www.liveinternet.ru/users/ullub/post119168490/:

Бишофит гель оказывает противовоспалительное и анальгезирующее действие. степени ожогов артроз плечевого сустава Артроз успешно лечится! Институт здравоохранения Всё об артрозе степени свободы деформирующий остеоартроз степени сравнения прилагательных Болезни суставов излечимы! Ответы ревматолога на вопросы о здоровье суставов! Читайте на Клео!

Фундаментальная идея рассматриваемого метода учитывает следующие обстоятельства:

- нормальный связный текст является значительно информационно избыточен с формальной точки зрения;
- в настоящий момент не существует успешных реализаций генерации связного текста.

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

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

3.1. Выделяемые характеристики текстов

Для текстов на русском языке выделяется 61 признак. Выделяемые признаки можно разделить на четыре группы:

- глобальные статистические характеристики текста;
- статистика употребления частей речи;
- характеристики разнообразия текста;
- статистика употребления редких оборотов.

Глобальные статистические характеристики зачастую используются при оценке читаемости текстов. Признаки, попавшие в данную группу, были выбраны, потому что их трудно контролировать человеку-автору. Ряд характеристик коррелируют с оценками читаемости текста [8], при этом неестественные тексты практически всегда нечитаемые и лишены смысла в целом.

Признаки, связанные со статистикой употребления частей речи, часто используются при определении авторства текстов [9]. Генераторы текстов на основе цепей Маркова могут нарушать соотношения частей речи, свойственные людям. Также статистика употребления частей речи может быть использована для определения стилистики текстов. Текст, полученный с помощью цепи Маркова, обладает чертами сразу нескольких документов из обучающего набора. Такое смешение стилей в рамках одного документа также может быть потенциально обнаружено по статистике употребления частей речи. Для определения частей речи использовался парсер mystem [10].

Одной из характерных особенностей естественных текстов является ограниченность словаря наиболее используемых слов. При этом частоты употребления слов моделируются законом Ципфа, по которому, если упорядочить слова текста по частотности, то частота каждого слова будет обратно пропорциональна его порядковому номеру.

Частота f слова с порядковым номером k подчиняется следующему соотношению:

$$f(k;s,c) \approx \frac{c}{k^s};$$

где s — параметр, характеризующий разнообразие слов в тексте, c — параметр, характеризующий частоту наиболее популярных слов. Для оценки разнообразия слов в тексте можно по частотам слов в тексте оценить параметры s и c. Также для оценки разнообразия могут применяться такие метрики, как степень сжатия различными алгоритмами сжатия информации.

Некоторым жанрам естественных текстов свойственно ограниченное употребление некоторых частей речи или оборотов [11]. Например, в нормативно-правовых актах трудно представить частицы «ну» и «вот», так как использование этих частиц противоречит стилю, принятому в таких документах. Генераторы текстов напрямую (пока) не моделируют стилистическое единство порождаемого текста, редкие характеристики могут использоваться для обнаружения искусственных текстов. Критерием отнесения того или иного признака к данной группе была выбрана его частота встречаемости.

3.2. Машинное обучение

Выделяемые признаки объединялись с помощью машинного обучения в автоматический классификатор. Для данной задачи был разработан алгоритм машинного обучения на основе деревьев решений. В основе разработанного алгоритма лежит широко распространенный алгоритм C4.5 [12].

Каждое дерево решений представляет собой двоичное дерево. Каждая вершина, не являющаяся листом, помечена номером признака и значением, по которому происходит разбиение набора документов на две части. Листы дерева помечены вероятностями принадлежности документа спаму или неспаму.

Дерево строится с корня. Вначале, в корень дерева помещается часть тренировочного набора. Затем, в каждом листе выбирается такой признак и такое значение разбиения, которые минимизируют информационную энтропию в наборах, полученных после разбиения. В случае если энтропия в наборах, полученных после разбиения, меньше, чем в исходном наборе, для данного листа строится левые и правые поддеревья, и лист помечается номером соответствующего признака и порогом разбиения. Затем набор распределяется по левому и правому поддереву в соответствии с выбранным разбиением.

После построения дерева для каждого листа вычисляется вероятность того, что документы, попавшие в этот лист, являются спамом или неспамом. Для этого документы распределяются по листам построенного дерева, затем для каждого листа вычисляется доли спам и неспам-документов, попавших в данный лист, которые и записывается в лист дерева. Чтобы минимизировать эффект переобучения на тренировочном наборе дерево строится на одной части тренировочного набора, а вероятности вычисляются по другой.

При обучении по одному и тому же набору строится несколько деревьев решений. При построении каждого дерева тренировочный набор произвольным образом делится пополам. Первая половина используется для построения дерева, вторая используется для вычисления вероятностей спама и неспама в каждом листе дерева.

Деревья объединяются в один классификатор с помощью простой процедуры голосования. При классификации документа вычисляется, в какой лист он попадает в каждом дереве. После этого вычисляется сумма вероятностей принадлежности спаму и неспаму по всем деревьям. Документу присваивается та метка, сумма вероятностей которой наибольшая.

3.3. Версия для английского языка

Англоязычный поисковый спам также содержит документы, порожденные с помощью генераторов текстов, например (Данный фрагмент текста

расположен по адресу http://dianajonsson.blogspot.com/2010/01/auto-insurance-in-us.html):

If the auto insurance in california and members the car insurance baltimore that then by all means do. If you already have a lower price for their insurance premium, and the auto insurance in us of insurance that would benefit you most. Go ahead and buy that low priced policy for the other driver's. At any time a driver must find a cheap policy, it's wise to know about the auto insurance in us to see if they would be especially hard if you gather an appropriate number.

Разработанный алгоритм также был адаптирован для обработки английского языка. Вместо парсера mystem, для определения частей речи использовался Stanford Part Of Speech Tagger [13]. Данный парсер использует набор меток частей речи Penn Treebank tag set [14]. Данный инструмент позволяет размечать англоязычные тексты по частям речи с учетом синтаксиса, а также поддерживает богатую классификацию частей речи.

К группе редких характеристик с точки зрения английского языка были отнесены те части речи, которые встречались менее чем в одном проценте предложений тренировочного набора. В эту группу попала статистика употребления модальных глаголов, притяжательных окончаний ('s), частиц и т.п. Как и в случае с русским языком, данные конструкции сильно влияют на стилистику документов.

В итоге набор признаков, применяемый для английского языка, состоит из 91 признака. В то же время, такие группы признаков, как глобальные статистические характеристики и меры разнообразия, не зависят от языка документа и также применялись для англоязычных текстов.

4. Эксперименты

В рамках данной работы мы изучали применимость предлагаемого метода для английского языка, а также оценивали вклад различных групп признаков в зависимости от языка документа.

4.1. Выборки документов

Обучающие и тестовые выборки для русского и английского языков формировались по сходным принципам. Вначале из множества исходных вебстраниц выбиралось 10 000 документов-образцов для генератора текстов. На основе данных образцов порождалось 10 000 неестественных текстов. Тренировочный набор составлялся из 5000 документов из исходной коллекции и 5000 документов порожденных с помощью генератора. Тестовый набор составлялся из других 5000 документов из коллекции и оставшихся 5000 порожденных документов.

Документы из исходной коллекции выбирались так, чтобы множество документов-образцов дл генератора текстов не пересекалось с обучающими и тестовыми выборками. Тренировочный и тестовый наборы строились отдельно для русского и английского языка, а также отдельно для цепей Маркова длины 2 и 3.

Исходной коллекцией для русского языка была коллекция ROMIP By.Web [15]. Для английского языка использовалась коллекция WebspamUK-2007 [16].

4.2. Применимость метода

В рамках первого эксперимента мы использовали предложенный метод для обнаружения документов, порожденных с помощью цепей Маркова. Для оценки качества классификации измерялась точность, полнота и F1-мера обнаружения документов, порожденных с помощью цепей Маркова в тестовом наборе.

В таблице 1 приведены результаты данного эксперимента, и результаты аналогичного эксперимента, проведенного на русскоязычных текстах.

Таблица 1. Результаты эксперимента по обнаружению неестественных текстов

	Точность	**Полнота**	**F1-мера**
Английский, цепь длины 2	96,19%	96,11%	96,15%
Английский, цепь длины 3	94,08%	92,29%	93,18%
Русский, цепь длины 2	94,98%	95,71%	95,34%
Русский, цепь длины 3	91,56%	95,02%	93,25%

Эксперимент подтверждает, что предложенный метод показывает схожие высокие результаты на русскоязычных и англоязычных текстах. Также как и для русского языка, чем больше длина цепи Маркова в генераторе текстов, тем меньше точность обнаружения синтезированных документов.

4.3. 4.3 Оценка качества выделяемых характеристик

В рамках данного исследования мы также сравнивали вклад различных признаков в классификацию при работе с русскоязычными и англоязычными текстами. Для этого была произведена численная оценка качества предложенных признаков. Если построить классификатор, который использует только один признак для обнаружения неестественных текстов, то F-мера классификации может служить индикатором качества данного признака. Чем выше F-мера при использовании одного признака, тем выше его ценность и вклад в обучение.

Таблица 2. Наиболее ценные признаки для классификации англоязычных текстов

№	Название признака	F-мера, %	Тип признака
1	Степень сжатия gzip	89,70	Разнообразие
2	Степень сжатия bz2	85,04	Разнообразие
3	Параметр S в распределении Ципфа для существительных	81,28	Разнообразие
4	Доля слов повторяющихся в соседних предложениях	79,60	Разнообразие
5	Доля глаголов в прошедшем времени	74,49	Части речи
6	Среднее количество знаков экспрессивной пунктуации	73,54	Глобальные
7	Дисперсия доли глаголов в прошедшем времени по предложениям	73,34	Части речи
8	Дисперсия доли модальных глаголов по предложениям	72,88	Редкие
9	Доля предложений с несколькими глаголами	71,27	Глобальные
10	Доля личных местоимений	71,13	Части речи
11	Доля имен собственных	71,06	Части речи
12	Дисперсия доли притяжательных окончаний по предложениям	70,66	Редкие
13	Доля слов из одного слога	70,63	Глобальные
14	Доля модальных глаголов	70,59	Редкие
15	Доля слов не более чем из 2-х слогов	70,56	Глобальные
16	Дисперсия порядковых числительных по предложениям	70,55	Части речи
17	Доля порядковых числительных по предложениям	70,06	Части речи
18	Доля определяющих слов	69,82	Части речи
19	Среднее количество знаков пунктуации на предложение	69,52	Глобальные
20	Доля слов длиннее 7 символов	69,49	Глобальные

Мы оценили вклад ряда признаков в обучение. В таблице 2 приведен список 20 характеристик наиболее ценных для классификации англоязычных текстов. Аналогичный список для русского языка приведен в таблице 3. Также в таблицах указана группа, к которой относится каждый признак.

4.4. Анализ различий для русского и английского языка

Изучение наиболее ценных признаков показывает, что в зависимости от языка важность различных групп признаков изменяется.

Несмотря на то, что максимальный вклад в обеих задачах дают характеристики, описывающие разнообразие документов, очевидно, что при обнаружении неестественных текстов на английском языке их значение гораздо больше. Отчасти это может объясняться тем, что характер распределения Ципфа различается для разных языков [17].

Также для английского языка велико значение повторов слов в соседних предложениях. Согласно нашей гипотезе, в английском языке повторы слов чаще используются для поддержания логической связи между предложениями, чем в русском, поэтому соответствующий признак играет большую роль.

Таблица 3. Наиболее ценные признаки для классификации русскоязычных текстов

№	Название признака	F-мера, %	Тип признака
1	Степень сжатия gz	78,87	Разнообразие
2	Степень сжатия bz2	77,92	Разнообразие
3	Параметр S в распределении Ципфа для существительных	77,67	Разнообразие
4	Дисперсия доли местоименных наречий по предложениям	75,64	Редкие
5	Дисперсия доли междометий по предложениям	75,23	Редкие
6	Дисперсия доли частиц по предложениям	75,22	Части речи
7	Дисперсия доли предлогов по предложениям	75,14	Части речи
8	Доля местоименных наречий	74,94	Редкие
9	Дисперсия доли местоименных прилагательных по предложениям	74,70	Редкие
10	Дисперсия доли местоименных существительных по предложениям	74,43	Редкие
11	Доля местоименных существительных	74,33	Редкие
12	Доля глаголов прошедшего времени	74,32	Части речи

Dialog'2010.indb 372 11.05.2010 16:57:47

№	Название признака	F-мера, %	Тип признака
13	Доля предложений с несколькими глаголами	74,03	Глобальные
14	Максимальное количество слов в предложении	73,94	Глобальные
15	Доля предлогов	73,78	Части речи
16	Дисперсия доли глаголов в предложениях	73,74	Части речи
17	Дисперсия доли существительных по предложениям	73,71	Части речи
18	Среднее количество знаков экспрессивной пунктуации	73,58	Редкие
19	Среднее количество существительных в предложении	73,58	Части речи
20	Среднее количество слов, начинающихся с большой буквы	73,57	Глобальные

Также большие различия между русскоязычными и англоязычными текстами видны на редких оборотах. Редкие обороты для русского языка составляют значительную долю наиболее сильных признаков. В то же время, для английского языка соответствующие признаки оказываются гораздо слабее.

Еще одним важным различием с точки зрения классификации является количество полезных признаков. Для русского языка в двадцатке наиболее полезных характеристик F-мера для каждой из них превосходит 73%. Для английского языка только семь наиболее ценных характеристик достигают такого значения. С точки зрения классификации поискового спама данное обстоятельство ухудшает устойчивость алгоритма к попыткам его обойти, так как для английского языка есть меньше надежных признаков.

4.5. Анализ влияния групп признаков

Для того чтобы лучше показать разницу в работе метода для русского и английского языков, мы также сравнили общую ценность всех четырех групп признаков. Для каждой группы признаков был обучен классификатор, использующий только данную группу для обнаружения неестественного текста. Чем лучше работает классификатор, обученный только по группе признаков, тем больший вклад данной группы признаков в классификацию.

Результаты второго эксперимента приведены на рисунке 1. На их основании можно сделать вывод, что для англоязычных текстов метрики разнообразия имеют большее значение для задачи обнаружения неестественных текстов. В то же время, вклад редких характеристик в классификацию оказывается гораздо ниже.

С точки зрения задачи обнаружения поискового спама, сильное влияние одной группы признаков на обучение является негативным фактором, так как позволяет относительно легко снизить эффективность предложенного метода обнаружения неестественных текстов. Спамерам достаточно порождать тексты, в которых распределение Ципфа для слов будет похожим на соответствующее распределение для естественных документов.

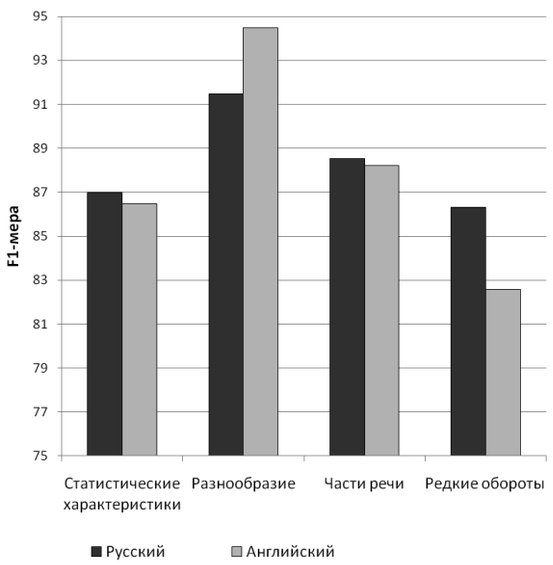

Рис. 1. Вклад групп признаков в классификацию в зависимости от языка

5. Планы

Данное исследование является основой для дальнейшего применения предложенного метода для англоязычных текстов. Мы планируем провести эксперимент по обнаружению поискового спама в коллекции WebspamUK-2007. Данная коллекция специально собрана и размечена для сравнения алгоритмов обнаружения поискового спама.

Для более полного и точного обнаружения спама на данной коллекции потребуется настройка данного алгоритма и на другие виды поискового спама, такие как:

- списки поисковых запросов, вставленные в текст;
- вкрапления ключевых слов в нормальные тексты;
- тексты, составленные из фрагментов из разных источников;
- помимо предложенных в данном методе характеристик также планируется добавление но-

вых признаков, для обнаружения дублированного и украденного содержимого, например, на основе шинглирования [18].

6. Заключение

В данной работе исследовалось применимость алгоритма обнаружения синтезированных текстов, ранее разработанного для русского языка, на англоязычных документах. Было показано, что адаптированный метод позволяет обнаруживать такие тексты с высокой точностью и полнотой.

При этом обнаружено, что вклад различных признаков в работу алгоритма зависит от языка текста. Построенный методом машинного обучения классификатор для английского языка в большей степени полагается на метрики разнообразия текстов, и потенциально менее устойчив к попыткам его обойти.

Разработанный алгоритм также может быть адаптирован для автоматического определения авторства и стиля текстовых документов, а также может быть адаптирован к другим языкам.

Литература

1. *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.
2. *Ntoulas A., Manasse M.* Detecting spam web pages through content analysis // In Proceedings of the World Wide Web conference, ACM Press, 2006. p. 83–92
3. *Piskorski J., Sydow M., Weiss D.* Exploring Linguistic Features for Web Spam Detection: A Preliminary Study // In Proceedings of the 4th international workshop on Adversarial Information Retrieval on the Web, Beijing, China, 2008. p. 25–28.
4. *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.
5. *Mishne G., Carmel D. and Lempel R.* Blocking blog spam with language model disagreement // In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.
6. *Urvoy T., Chauveau E., Filoche P.* Tracking Web Spam with HTML Style Similarities // ACM Transactions on the Web, 2006. Vol. 2, n. 1, Article 3.
7. *Castillo C., Donato D., Murdock V., Silvestri F., Know your neighbors: Web spam detection using the web topology // In Proceedings of SIGIR, ACM, 2007.*
8. *Dubay W. H.* The Principles of Readability // Costa Mesa, CA: Impact Information, 2004.
9. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов // В сб.: Методы количественного анализа текстов нарративных источников. — М.: АН СССР, Ин-т Истории СССР, 1983. с. 86–109.
10. *Парсер* mystem (http://company.yandex.ru/technology/mystem/).
11. *Braslavski P.* Document Style Recognition Using Shallow Statistical Analysis // Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004. p. 1–9.
12. *Quinlan J. R.* C4.5: Programs for Machine Learning // Morgan Kaufmann Publishers, 1993.
13. *Stanford Log-linear Part-Of-Speech Tagger (http://nlp.stanford.edu/software/tagger.shtml)*.
14. *Marcus M. P., Marcinkiewicz M. A., Santorini B.* Building a Large Annotated Corpus of English: the Penn Treebank // Computational Linguistics, 1993. Vol. 19 n. 2
15. *Веб коллекция* BY.Web, http://romip.ru/ru/collections/by.web-2007.html.
16. *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/).
17. *Gelbukh A., Sidorov G.* Zipf and Heaps Laws' Coefficients Depend on Language // Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, 2001. p. 332–335.
18. *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: 2007.

Dialog'2010.indb 374 11.05.2010 16:57:48