# Метод обнаружения поискового спама, порожденного с помощью цепей Маркова

© Павлов А.С.

Факультет Вычислительной математики и кибернетики
МГУ имени М.В.Ломоносова
pavvloff@gmail.com

© Добров Б.В.

Научно-исследовательский вычислительный центр
МГУ имени М.В.Ломоносова;
АНО Центр информационных исследований
dobroff@mail.cir.ru

**Аннотация**

В работе предложен метод обнаружения поискового спама, порожденного генераторами текста. Данный метод основывается на анализе частотных распределений стилистических и жанровых особенностей текста. Для автоматического обнаружения спама на основе выделенных характеристик используются методы машинного обучения. Проведены эксперименты, в которых показана возможность обнаружения текстов, порожденных генераторами на основе цепей Маркова, с помощью предложенного подхода.

## 1 Введение

В связи с большим количеством информации в сети Интернет пользователи чаще всего используют веб поиск для нахождения интересующих их данных. В настоящее время одной из основных проблем информационного поиска является распространение поискового спама.

Поисковый спам создается в результате намеренных действий, направленных на завышение оценки страницы в поисковой системе, по сравнению с ее истинной ценностью. В соответствии с современными оценками поисковый спам составляет около 22% всего содержимого сети Интернет [11]. На настоящий момент поисковый спам остается эффективным методом выведения сайта на верхние позиции в выдаче поисковых систем. Поисковый спам ухудшает качество поиска и мешает нормальной работе поисковых систем.

Поисковый спам наиболее эффективен при массовом автоматическом создании спам-страниц. Одним из распространенных способов автоматического создания большого количества текстов является генерация текстов на основе цепей Маркова. При использовании генерации текстов на основе цепей Маркова сначала на отобранных текстах производится обучение, затем можно породить большое количество в целом бессмысленных, но локально связных текстов. Учитывая то, что в качестве исходных текстов часто берутся релевантные определенной тематике документы, то и результаты генерации текстов также отражают статистические тематические характеристики.

Отметим, что в настоящее время не существует полной теории, описывающей законы порождения связных осмысленных текстов. Как следствие отсутствуют в общем случае методы порождения текстов, не отличимых от созданных человеком.

Тем не менее, известны многие закономерности, характерные естественным текстам - единство стиля, следование законом жанра, локальная связность, глобальная тематическая связность и т.п.

Цепи Маркова позволяют моделировать лишь локальную связность текста и общие тематические характеристики.

Основная идея настоящей работы состоит в том, чтобы с помощью учета статистических характеристик стилистических и жанровых особенностей естественных текстов обнаруживать тексты, обладающие локальной связностью, но нарушающие другие свойства естественных текстов.

## 2 Обзор существующих методов

### 2.1 Методы детектирования поискового спама

Применимость простых статистических характеристик для определения поискового спама изучалась в работе [14]. При этом наибольшее внимание уделялось ссылочным характеристикам, в то время как статистические характеристики текста практически не рассматривались.

**Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.**

Исследование лингвистических характеристик для обнаружения поискового спама исследовалось в работе [19]. Данный подход основывается на применении специального словаря, на основе которого вычисляется более 200 статистических признаков. Применение словаря означает, что генераторы текстов могут обойти предлагаемые методы.

Еще один интересный подход к анализу содержимого документов для обнаружения поискового спама предлагается в [9]. Эта работа основывается на определении коммерческой направленности текстов по нескольким статистическим атрибутам. Данные атрибуты в основном основаны на анализе логов поисковых запросов или логов систем контекстной рекламы. Использование характеристик, ограниченных некоторым словарем, позволяет генераторам на основе цепей Маркова обходить предложенные алгоритмы.

Подходы, не зависящие от конкретной лексики и тематики документов, предлагаются в работах [18, 21]. Первая работа посвящена обнаружению спама в блогах, и использует особенности формата блогов, например, наличие комментариев, 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.

### 2.2 Использование несловарных характеристик при анализе текстов

Существует постоянный интерес исследователей к анализу статистических, трудно контролируемым автором, характеристик естественного текста.

Метрики читаемости (readability, также употребляется термин «читабельность») текста подробно описаны в [12, 13]. Применение простых статистических характеристик, таких как длина предложений и длина слов, широко используется в США для оценки простоты восприятия текста.

Задача определения жанра текста по простым статистическим характеристикам решается в [10]. В этой работе показано как небольшое число характеристик текста позволяют с неплохой точностью определять наиболее распространенные стили и жанры.

Определение авторства текста, точнее специфического авторского стиля, также может основываться на глобальных статистических закономерностях текстов. Большое количество работ в данном направлении основывается на статье [6]. Анализ статистики употребления частиц, предлогов, а также длин предложений и слов позволяет формулировать критерии принадлежности текста конкретному автору.

### 2.3 Определение дубликатов

Задача обнаружения дубликатов текста является смежной с задачей обнаружения текстов, созданных цепями Маркова, так как в зависимости от длины цепи порожденный текст может копировать большие куски документов-образцов.

Обзор методов обнаружения дубликатов приведен в [3]. В статье [15] описано применение шинглирования для обнаружения спам-текстов, порожденных из отрывков естественных текстов.

## 3 Генераторы текстов на основе цепей Маркова

### 3.1 Методы порождения текстов

Создание поискового спама сопряжено с созданием большого количества текстов для автоматического наполнения сайтов. В настоящий момент существует несколько подходов к созданию текстов для спам-сайтов [16]:

- Создание текстов вручную;
- Копирование текстов из других источников;
- Автоматическая генерация текстов;
- Автоматическая модификация существующих текстов.

Создание текстов вручную является трудоемким и дорогостоящим процессом, поэтому редко применяется для массового поискового спама. Копирование содержимого других сайтов является довольно распространенным явлением, но в настоящее время существуют достаточно эффективные способы определения скопированного текста, например, на основе шинглирования [3].

В итоге на данный момент наиболее эффективными являются методы, которые позволяют автоматически получать уникальные тексты.

Генератор текста — компьютерная программа, способная генерировать последовательности символов, внешне похожие на текст, но при этом, как правило, лишённые смысла. Такие тексты не представляют никакой ценности для пользователей поиска. При генерации текста спамеры также стараются оптимизировать его под некоторый набор запросов, чтобы повысить вероятность попадания сайта с этим содержимым в выдачу поисковой системы.

### 3.2 Цепи Маркова

Распространенным видом генераторов текста являются генераторы текста на основе цепей Маркова. Цепью Маркова с дискретным временем называется последовательность случайных величин, для которой условное распределение каждой величины зависит только от значения предыдущих величин.

Цепь Маркова описывается множеством значений случайных величин, которое называется пространством состояний; а также матрицей переходных вероятностей между состояниями. Матрица переходных состояний определяет вероятность перехода в следующее состояние, с учетом текущего. В случае если матрица

переходных вероятностей не зависит от шага, она называется однородной, именно однородные матрицы чаще всего применяются для порождения текстов.

### 3.2 Порождение поискового спама с помощью цепей Маркова

Когда цепи Маркова применяются для порождения искусственных текстов, пространством состояний становится множество всех слов и знаков препинания. Переходная матрица обычно формируется по некоторому множеству текстов-образцов. По образцу оценивается вероятность порождения нового слова после последовательности уже порожденных слов. Последовательность событий, произведенная такой цепью Маркова, представляет собой набор слов и знаков препинания, внешне напоминающий связный текст.

Важной характеристикой таких генераторов является порядок цепи Маркова – то есть количество слов, учитывающихся при порождении следующего слова. С ростом порядка цепи растет длина локально связных фрагментов текста, в то же время с ростом длины цепи генератор начинает повторять все большие куски исходного текста.

Тексты, созданные с помощью цепей Маркова, обладают рядом свойств, благодаря которым этот метод порождения текстов стал популярен при создании поискового спама. Во-первых, порожденный текст содержит ту же лексику, что и исходный образец. Это позволяет использовать в качестве образца существующие тексты, которые высоко ранжируются поисковыми системами, например, брать образцы текстов из сниппетов поисковых систем, и получать на выходе тексты, оптимизированные под конкретные запросы. Во-вторых, порожденный текст является с высокой вероятностью уникальным. Это затрудняет обнаружение таких текстов методами обнаружения дубликатов.

В качестве примера приведем фрагмент текста, порожденного по данной статье:

«*Генераторы текстов из отдельных документов. Для проверки возможности обнаружения дубликатов Задача определения текстов, порожденных генераторами текстов из рассматриваемых генераторов текстов. Большое количество слов, начинающихся не являющаяся листом, помечена номером признака и авторства. Проверялась эффективность данного метода опорных векторов позволяет формулировать критерии принадлежности спаму или неспаму.*»

Применение автоматических генераторов текстов на основе цепей Маркова часто используется в таком виде спама как дорвеи. Функция дорвея перенаправить пользователя на некоторый целевой сайт, при этом само содержимое такого сайта никакой ценности для пользователя не несет. Дорвеи должны попадать в выдачу по популярным запросам, поэтому эффективное обнаружение такого вида спама может сократить количество спама в выдаче поисковых систем.

## 4 Предлагаемый метод

В основе предлагаемого подхода лежат методы, ранее использовавшиеся для определения жанра текста и авторства.

В рамках данной работы выделяется набор трудно контролируемых автором статистических признаков текстового документа.

Затем, на основе полученных характеристик и машинного обучения строится автоматический классификатор, который позволяет обнаруживать неестественные тексты.

Данный метод основывается на предположении, что по данным характеристикам тексты, полученные с помощью генератора на основе цепей Маркова, будут отличаться от естественных текстов.

### 4.1 Рассматриваемые характеристики

Рассматриваемые характеристики можно условно разделить на следующие пересекающиеся группы:

- Признаки, связанные с читабельностью текста;
- Стилистические особенности текста;
- Жанровые особенности текста [10];
- Глобальные статистические характеристики;
- Морфологические особенности слов текста (для анализа морфологической информации применялся парсер mystem [4]);
- Статистика употребления знаков препинания.

Признаки выбирались из условия, что автору-человеку сложно проконтролировать их значение. Каждому автору и каждому жанру свойственен особенный стиль, который отражается на значении некоторых характеристик.

Предположительно, генераторы текстов на основе цепей Маркова не способны эмулировать некоторые из этих характеристик. Таким образом, поисковый спам можно выделить в качестве отдельного жанра документов, и рассматривать задачу идентификации этого жанра.

Важным свойством выбранных характеристик является то, что они не основываются на тематике или лексике текстов, таким образом, они позволят определять искусственно порожденные тексты вне зависимости от их тематики.

Ниже приведен полный список выделявшихся характеристик, некоторым пунктам соответствует несколько характеристик:

1. Среднее количество слов в предложениях;
2. Среднее количество символов в словах;
3. Среднее количество слогов в слове;
4. Доля слов длиннее 7 символов;

5. Доля слов более чем из 7 слогов;
6. Доля слов из слога;
7. Доля слов из двух слогов;
8. Минимальное количество слогов в одном предложении;
9. Максимальное количество слогов в одном предложении;
10. Количество частиц «бы»;
11. Количество частиц «ну», «вот», «ведь»;
12. Среднее количество знаков пунктуации на предложение;
13. Среднее количество знаков экспрессивной пунктуации («!», «?», «…»);
14. Среднее количество слов, начинающихся с заглавной буквы;
15. Доля различных частей речи:
    a. Доля глаголов среди слов;
    b. Доля прилагательных среди слов;
    c. Доля существительных среди слов;
    d. Доля числительных среди слов;
    e. Доля порядковых числительных среди слов;
    f. Доля наречий среди слов;
    g. Доля частиц среди слов;
    h. Доля предлогов среди слов;
    i. Доля частиц среди слов;
    j. Доля междометий среди слов;
16. Дисперсии количества различных частей (из п.15) речи по предложениям;
17. Доля местоимений первого лица;
18. Доля местоимений второго лица;
19. Доля глаголов по временам:
    a. Доля глаголов настоящего времени;
    b. Доля глаголов прошедшего времени;
    c. Доля глаголов не прошедшего времени;
20. Доля существительных по родам:
    a. Доля существительных мужского рода среди слов и среди существительных;
    b. Доля существительных женского рода среди слов и среди существительных;
    c. Доля существительных среднего рода среди слов и среди существительных;
21. Сжатие текста различными алгоритмами:
    a. bz2;
    b. zlib2;
22. Частотность употребления слов (оценка распределения по гистограмме частотностей).

Всего исследуется 61 характеристика текста. Каждая характеристика представляет собой положительное вещественное число. Таким образом, каждому документу ставится в соответствие вектор признаков из 61 элемента.

### 4.2 Методы машинного обучения

Для определения текстов, созданных с помощью генераторов, предлагается объединить выделенные характеристики в классификатор с помощью алгоритмов машинного обучения.

Предлагается обучать классификатор с использованием тренировочного набора, состоящего из естественных документов и документов, которые предположительно созданы с помощью генераторов текстов. В случае если классификатору удастся построить зависимость между выделяемыми характеристиками и методом порождения документа, данный классификатор можно будет использовать для обнаружения поискового спама.

В данной работе рассматривались два распространенных алгоритма машинного обучения:

- Классификатор на основе машины опорных векторов с использованием линейного ядра SVMLight [17];
- Классификатор на основе деревьев решений Dtree.

#### 4.2.1 Алгоритм машинного обучения SVMLight

Алгоритмы классификации на основе метода опорных векторов строят гиперплоскость, разделяющую разные классы объектов в пространстве признаков. При этом метод опорных векторов позволяет максимизировать зазор между классами, что способствует более качественной классификации.

В данной работе использовалась одна из распространенных реализаций метода опорных векторов SVMLight [17].

#### 4.2.2 Алгоритм машинного обучения DTree

В основе используемого метода лежит алгоритм построения деревьев решений C4.5 [20]. 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

| | DTree | | | SVMLight | | |
|---|---|---|---|---|---|---|
| | Точность | Полнота | F-мера | Точность | Полнота | F-мера |
| Markov2 | 91,30% | 93,27% | 92,27% | 72,9% | 74,83% | 73,85% |
| Doorway_Su | 92,11% | 89,98% | 91,03% | 69,41% | 68,24% | 68,82% |
| Rusadult | 99,19% | 99,26% | 99,23% | 89,65% | 89,83% | 89,74% |

Таблица 1. Точность и полнота обнаружения текстов, порожденных различными генераторами текстов

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.

# 5 Эксперименты

Эксперимент проводился на коллекции веб страниц ROMIP By.Web [1]. Из документов коллекции была удалена вся HTML разметка, включая содержимое тегов <script>.

В первом эксперименте оценивалась возможность обнаружения текстов, порожденных цепями Маркова, а также популярными генераторами поискового спама.

Во втором эксперименте исследовалась возможность обнаружения реального поискового спама.

## 5.1 Рассматриваемые генераторы текстов

В первом эксперименте было взято три генератора текстов:

- Собственная реализация генератора текста на основе цепей Маркова с длиной цепи 2 (далее - markov2);
- Генератор дорвеев Doorway.su [2] (далее - doorway_su);
- Генератор дорвеев Rusadult [5] (далее - rusadult).

Каждый документ, представленный в виде вектора признаков, может принадлежать одному из классов: {spam, good}. К классу spam относятся документы, порожденные генератором текста, к классу good относятся документы, из набора ROMIP By.Web. Изучалась возможность обнаружения текстов, созданных каждым из рассматриваемых генераторов.

## 5.2 Построение тренировочных наборов

Тренировочные наборы для классификаторов строились следующим образом:

1. Из набора By.Web были взяты 20000 произвольных документов, помеченных как good (GoodSet);
2. Из набора By.Web были взяты другие 20000 произвольных документов (ExampleSet).
3. Используя набор документов ExampleSet в качестве образца для алгоритма генерации текстов, были порождены следующие наборы:
   a. 20000 текстов были порождены алгоритмом markov2 и помечены как спам (SpamMarkov2);
   b. 20000 текстов были порождены генератором дорвеев Doorway.Su и помечены как спам (SpamDoorwaySu);
   c. 20000 текстов были порождены генератором дорвеев Rusadult и помечены как спам (SpamRusadult).

Затем каждый из наборов GoodSet, SpamMarkov2, SpamDoorwaySu и SpamRusadult был разбит пополам – первые 10000 документов из каждого набора использовались в качестве тренировочных наборов, вторые 10000 документов использовались при тестировании построенных классификаторов. Для проверки возможности обнаружения каждого из рассматриваемых генераторов по отдельности было составлено три обучающих множества, каждый состоит из 10000 документов с пометкой good и 10000 документов, порожденных одним из генераторов.

Каждый классификатор тестировался на 10000 документах, помеченных good, и 10000 документах, порожденных соответствующим генератором.

## 5.3 Применение методов машинного обучения

При обучении классификатора SVMLight все параметры брались по умолчанию. После процедуры обучения производился подбор порога классификации таким образом, чтобы сбалансировать точность и полноту классификатора на тренировочном наборе.

| | DTree | | SVMLight | |
|---|---|---|---|---|
| | Полнота по реальному спаму | F-мера | Полнота по реальному спаму | F-мера |
| Markov2 | 65,66% | 91,57% | 44,33% | 71,69% |
| Markov2+RealSpam | 87% | 92,61% | 49,33% | 72,46% |

Таблица 2. Полнота обнаружения реального спама

При классификации с использование деревьев решений для каждого дерева строилось не более 50 вершин, 100 деревьев объединялись в один классификатор голосованием.

В ходе эксперимента измерялась точность, полнота и F1-мера при обнаружении спама. Результаты эксперимента для трех генераторов текстов и двух алгоритмов классификации приведены в таблице 1.

### 5.4 Эксперимент по обнаружению реального спама

Целью второго эксперимента было проверить возможность обнаружения реального спама, при условии отсутствия образцов таких документов. Для этого вручную было отобрано 600 документов, являющихся поисковым спамом, предположительно порожденным с помощью цепей Маркова (RealSpam). Данные документы были обнаружены в Поиске по блогам компании Яндекс [8], по коммерческим запросам.

Обнаруженный вид спама ориентирован на попадание в выдачу по низкочастотным запросам, например:

«*Deo автомобиль Deo*
*Продажа автомашин. Каталог цен! автомобильная акустика Большой выбор, хорошие цены Модель Daewoo Nexia представляет собой последнюю модификацию модели Opel Kadett E. В 1986 году лицензированное производство этого автомобиля началось в автомобильный видеорегистратор Купля-продажа авто в Тюмени Много частных объявлений о продаже автомобилей в Тюмени на Новые и. автомобили.*»[1]

По всей видимости, для порождения данного текста спамеры использовали несколько текстов на автомобильную тематику. Целью спамеров было привлечь посетителей на данную страницу, с расчетом, что те перейдут по ссылкам, размещенным на ней.

В ходе эксперимента было рассмотрено два тренировочных набора:

- Набор, состоящий из 10000 документов из коллекции ROMIP By.Web, и 10000 документов, порожденных алгоритмом markov2;
- Набор, идентичный предыдущему, к которому было добавлено 300 образцов реального спама.

Тестовый набор состоял из 10000 документов из коллекции ROMIP By.Web, 10000 документов, порожденных алгоритмом markov2 и 300 других документов из набора RealSpam.

Также как и в первом эксперименте использовалось два алгоритма машинного обучения: DTree и SVMLight. В ходе эксперимента измерялась полнота обнаружения реального спама в тестовом наборе, а также общая F1-мера классификации. Результаты эксперимента приведены в таблице 2.

## 6 Выводы

На основании проведенных экспериментов можно сделать вывод, что анализ трудно контролируемых автором признаков может быть использован для обнаружения поискового спама, порожденного с помощью различных генераторов текста. Классификатор на основе деревьев решений показывает лучшие результаты при использовании большого количества разнородных характеристик.

На основании полученных результатов можно сделать некоторые выводы относительно алгоритмов порождения текстов, применяемых в рассмотренных генераторах поискового спама.

Скорее всего, генератор Doorway.Su использует цепи Маркова для порождения текстов, так как точность обнаружения такого рода спама близка к точности обнаружения текстов, созданных модельным алгоритмом цепей Маркова.

Результаты второго эксперимента показывают, что, обучаясь на искусственно созданных образцах поискового спама, можно обнаруживать реальный спам. Как следствие, применение данного подхода для обнаружения спама потребует меньше трудозатрат на составление адекватного тренировочного набора.

При этом добавление даже небольшого количества образцов реального спама (в данном случае 300 документов) ведет к значительному улучшению полноты обнаружения данного типа спама, без значительных изменений в общем качестве обнаружения спама.

Подчеркнем, что обнаружение текстов, порожденных цепями Маркова, возможно даже в том случае, когда образцы, по которым они созданы, неизвестны.

[1] Данный пример можно найти по адресу: *http://eltec15.livejournal.com/6698.html*

## 7 Заключение

В данной работе представлен подход к обнаружению поискового спама, созданного с помощью генераторов текста. Подход основан на выделении характеристик текста, применявшихся ранее для задач жанровой классификации и определения авторства.

Проверялась эффективность данного метода при обнаружении текстов, созданных с помощью генератора текстов на основе цепей Маркова, и двух распространенных в российском сегменте Интернета генераторов поискового спама. Эксперименты показали применимость предлагаемого подхода для задачи обнаружения поискового спама.

В дальнейшем планируется исследовать другие подходы к анализу текстов, а также исследовать возможность обнаружения других распространенных видов поискового спама.

## Литература


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

[2] Генератор дорвеев Doorway.Su, http://doorway.su/.

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

[4] Парсер mystem http://company.yandex.ru/technology/mystem/.

[5] Серверный генератор дорвеев от RUSADULT.com, http://doorways.rusadult.com/ru/.

[6] Фоменко В.П., Фоменко Т.Г., Авторский инвариант русских литературных текстов, 1981.

[7] Чжун Кай-лай, Однородные цепи Маркова. Перев. с англ. — М.: Мир, 1964. — 425 с.

[8] Яндекс.Поиск по блогам, http://blogs.yandex.ru/.

[9] Benczúr, A. A., Bíró, I., Csalogány, K. and Sárlós, T. Web Spam Detection via Commercial Intent Analysis. In Proceedings of the 3rd international workshop on Adversarial Information Retrieval on the Web, Banff, Alberta, Canada, May 8th, 2007. Pages: 89–92.

[10] Braslavski P. Document Style Recognition Using Shallow Statistical Analysis. In Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004, p. 1–9.

[11] Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., Vigna, S. A Reference Collection for Web Spam. ACM SIGIR Forum Volume 40, Issue 2 (December 2006) Pages: 11–24.

[12] Dale, E. and J. S. Chall. 1949. "The concept of readability." Elementary English 26: 23.

[13] Dubay, W.H.. 2004. The Principles of Readability. Costa Mesa, CA: Impact Information

[14] Fetterly, D., Manasse, M., Najork, M. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In Proceedings of WebDB'04, New York, USA, 2004.

[15] Fetterly, D., Manasse, M., Najork, M. Detecting phrase-level duplication on the World Wide Web. In Proceedings of SIGIR'05, pages 170–177, New York, NY, USA, 2005. ACM.

[16] Gyöngyi, Z. and Garcia-Molina H., Web Spam Taxonomy. In Proceedings of AIRWeb 2005, May 2005.

[17] Joachims, T. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schцlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[18] Mishne, G., Carmel, D., and Lempel, R. Blocking blog spam with language model disagreement. In Proceedings of AIRWeb 2005, May 2005.

[19] 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, Pages 25-28.

[20] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[21] Urvoy T., Chauveau E., Filoche, P. Tracking Web Spam with HTML Style Similarities. ACM Transactions on the Web, Vol. 2, No. 1, Article 3.


## Detecting Web Spam Created With Markov Chains Text Generators


Anton S. Pavlov, Boris V. Dobrov



In this paper we introduce an approach to detection of web spam generated by text generators. This method is based on text style and genre analysis. Machine learning algorithms are applied to automate spam detection, using the extracted features. Experiments prove possibility of proposed method to be applied to detect texts, created by Markov chains text generators.