Применение метода опорных векторов для обнаружения ссылочного спама

© Шарапов Руслан Владимирович © Шарапова Екатерина Викторовна

Муромский институт (филиал) Владимирского государственного университета
info@vanta.ru

Аннотация

В статье рассматриваются подходы к выявлению ссылочного спама методами машинного обучения. Приводится обзор существующих методов борьбы с поисковым спамом. Анализируются значимые признаки, способствующие выявлению ссылочного спама. Дается алгоритм выявления спама на основе метода опорных векторов и приводятся результаты его работы.

1 Введение

С ростом популярности сети интернет повышается интерес к поисковым системам, как средствам быстрого поиска нужной информации. Вместе с тем, увеличивается число попыток манипулирования поисковыми системами посредством поискового спама.

Поисковый спам можно разделить на две большие группы: спам содержания (контента) и ссылочный спам [13]. К спаму содержания относятся методы искусственного добавления ключевых слов на страницу (в заголовки, метатеги, тексты ссылок, названия URL и текст страниц). Ссылочный спам заключается в формировании ссылочных структур, способных повлиять на алгоритмы работы поисковых систем с целью достижения более высоких позиций в результатах поиска по пользовательским запросам.

Поисковые системы активно используют ссылки. Большинство систем, так или иначе, учитывают ссылки на страницы для более эффективного ранжирования результатов поиска. В основе этого лежит постулат о том, что ссылка является воплощением желания поделиться полезной информацией с другими людьми, своего рода голосом за ресурс, на который ведет ссылка. Поэтому сайт, на который ведет много ссылок, вероятно, будет более полезен и интересен пользователям, чем сайт, на который никто не ссылается. Кроме того, ссылки с известных и популярных ресурсов считаются более весомыми, чем с никому не известных сайтов. Все это используется современными алгоритмами поисковых систем (PageRank, HITS, индекс цитирования), чтобы предоставлять пользователям более нужную и полезную информацию по поисковым запросам.

Этими же принципами пользуются при размещении ссылочного спама – намеренно размещая большое число ссылок на сайтах с возможностью простого добавления информации (форумах, гостевых книгах, комментариях в блогах и т.д.). Такие ссылки предназначены в первую очередь для поисковых систем, а не для человека. В результате набираются искусственные “голоса” в пользу сайтов, на которые ведут эти спам-ссылки и сайты начинают лучше “искаться” поисковыми системами, оттесняя качественным и интересные ресурсы на второй план.

Существует несколько способов размещения большого количества ссылочного спама. Несколько лет назад основными способами являлись обмен ссылками и создание ферм ссылок. Методам борьбы с ними посвящено множество алгоритмов, которые успешно используются поисковыми системами. В настоящее время на смену им пришли автоматизированные средства массового размещения ссылок. К таким средствам относятся специализированные программные продукты, позволяющие автоматически добавлять ссылки в каталоги, гостевые книги, форумы, блоги и т.д. Например, с помощью программы Allsubmitter можно за несколько часов поместить ссылки на десятках (а то и сотнях) тысяч сайтов. С такими ссылками можно бороться путем выявления сайтов с возможностью свободного, немодерируемого добавления информации (ссылок). Еще большей проблемой являются системы пакетной покупки ссылок через рекламных брокеров. Такие системы могут размещать ссылки на миллионах страниц. Например, самая популярная система купли-продажи ссылок – Sape.ru имеет возможность размещать ссылки на более чем 73 миллионах страниц. В прошлом году это число составляло 35 миллионов страниц. Рост аудитории более чем в 2 раза свидетельствует о повышающейся популярности этой системы. Система MainLink.ru также размещает ссылки на 40 миллионах страниц, LinkFeed.ru – на 14 миллионах.

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

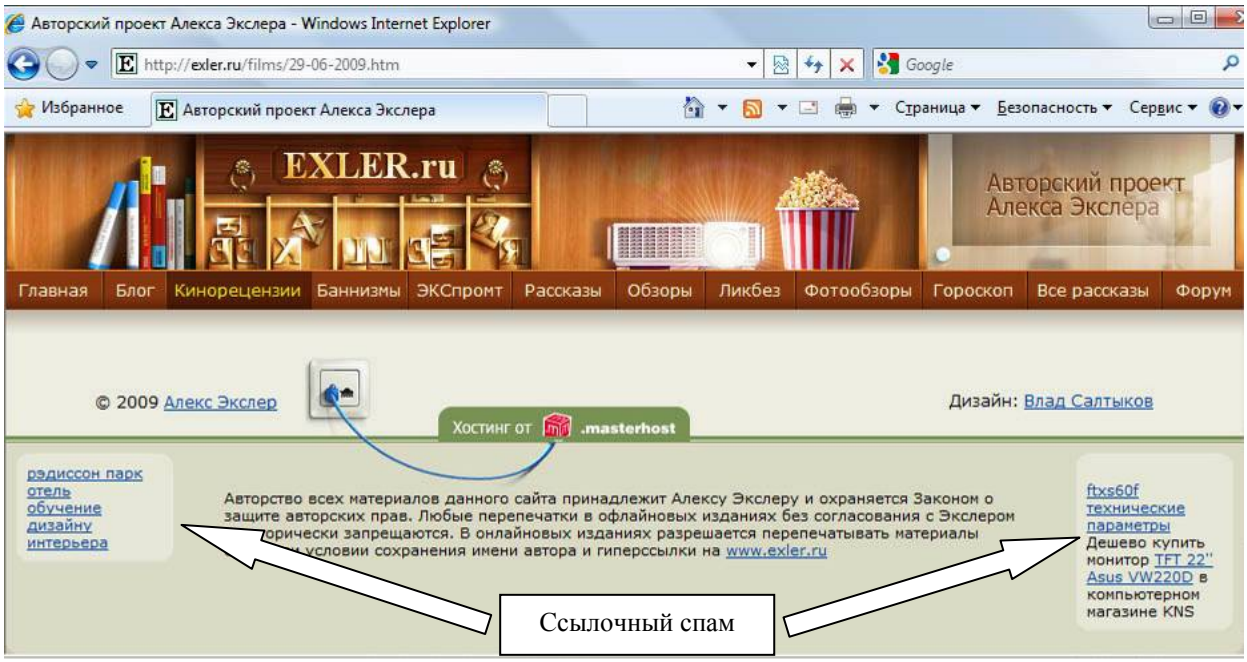

Рис. 1. Пример ссылочного спама на сайте Алекса Экслера

Таблица 1. Характеристики рекламных брокеров

Система	Страниц	Сайтов
Sape.ru	75 702 512	206 325
MainLink.ru	39 253 465	92 929
Xap.ru	32 608 976	80 000 (?)
LinkFeed.ru	14 068 832	35 302
SetLinks.ru	7 600 136	28 991

Среди сайтов, размещающих у себя ссылки через рекламных брокеров имеется множество популярных и авторитетных сайтов. Например, один из рекламных брокеров (Prospero.ru) размещает ссылки на сайтах с PR=7 и индексом цитируемости до 27000. В списке доступных для размещения площадок у него мы нашли такие сайты, как http://www.auto.ru, http://www.foto.ru, http://www.vkontakte.ru, http://www.interfax.ru/, http://news.rin.ru и т.д.

Влияние спам-ссылок на алгоритмы поисковых систем может быть существенным, учитывая их массовый характер. Особую опасность представляют ссылки с авторитетных ресурсов. Поэтому выявление ссылочного спама является актуальной задачей.

Ранее [27] мы предлагали алгоритм, основанный на эмпирических методах определения ссылочного спама. Сущность метода заключалась в обнаружении ряда признаков ссылочного спама, за каждый из которых ссылкам назначались штрафы. При превышении суммы штрафов некоторого порога, ссылки признавались спамом. Недостатком алгоритма является ручной подбор значений штрафов и некоторые неточности в работе.

В текущей работе рассматривается возможность применения методов машинного обучения для задач обнаружения ссылочного спама.

2 Текущее состояние проблемы

Вопросу выявления ссылочного спама посвящено несколько направлений исследований.

В работе [10] предлагается статистический анализ для выявления автоматически сгенерированных страниц со спамом. О спаме может свидетельствовать: отклонение от нормального распределения различных свойств страниц, включая имена и IP-адреса, входящие и исходящие ссылки, содержание страницы и норму изменения.

Множество работ посвящено анализу ссылочной информации – в первую очередь взаимосвязях страниц, объединяемых ссылками и текстам самих ссылок. Ряд разработчиков предлагают алгоритмы, построенных на основе PageRank.

В [16] рассматривается алгоритм Anti-Trust Rank. Алгоритм основан на ручном отборе страниц с и без спама. Дальнейший анализ структуры вэб-графа, построенного на основе ссылочных структур, позволяет выявить страницы, использующие спам. Алгоритм показывает высокую точность обнаружения спама, в том числе для страниц с высоким PageRank.

В [2] предлагается алгоритм SpamRank. Алгоритм основан на понятии персонализированного PageRank и обнаруживает страницы с незаслуженным высоким значением PageRank без использования любого вида белых или черных списков или других средств вмешательства человека.

В работе [11] описывается алгоритм TrustRank для борьбы со спамом. Принцип TrustRank строится на том, что “хорошие” страницы обычно ссылаются на “хорошие” страницы и редко используют ссылки

для спама. Сначала выбирается набор “хороших” страниц и им назначается высокий вес. Далее используется подход, аналогичный PageRank: вес разделяется на исходящие ссылки к другим страницам. Наконец, после конвергенции, страницы с высоким весом принимаются за хорошие страницы. Авторы считают, что использование алгоритма TrustRank дает более качественные результаты, чем PageRank.

В работе [23] предлагается анализировать вэб-граф для определения ссылочного спама (в частности, ферм ссылок). Алгоритм основан на анализе входящих и исходящих ссылок сайтов. В случая обнаружения пересечения входящих и исходящих ссылок больше определенного порога, страницам назначается штраф. Эта операция выполняется для всех страниц.

В [12] рассматривается алгоритм определения страниц, повышающих свой PageRank с помощью ссылочного спама. Используется понятие массы спама, меры воздействия спам-ссылок на ранг страницы. Рассматриваются вопросы оценки массы спама. Для определения спама активно используется ссылочная структура вэб-графа.

В работе [8] предлагается алгоритм HostRank (PageRank, вычисленный по графу хостов), который более гибок по отношению к ссылочному спаму. Алгоритм позволяет сократить число сомнительных сайтов в результатах поиска, что достигается уменьшением веса, получаемого сайтами от ссылочного спама.

Еще одно направление работ – применение методов машинного обучения.

В работе [5] делается попытка определять ссылочный спам (“непотистский” спам). Для решения задачи используется дерево решений C4.5. Всего авторы выделяют 75 свойств, используемых для классификации. Эти свойства позволяют определять: совпадение заголовка и описания страницы, описание пересекается с текстом страницы, совпадение имен хостов, совпадение доменов, совпадение адресов страниц без доменов, совпадение некоторых частей IP адресов, одинаковые контактные E-mail домены и т.д.

В работе [18] для определения страниц со спамом так же применяется дерево решений C4.5. Авторы считают, что деревья решений имеют преимущество перед нейронными сетями, системами, основанными на правилах и методам опорных векторов. В качестве свойств для классификации используются: число слов на странице, число слов в заголовке, средняя длина слов, количество текста в ссылках, процент видимого содержания, величина сжатия страницы, процент страницы, описанный в списке популярных слов, независимая вероятность n-грамм и т.д.

В работе [19] для задач классификации используются дерево решений C4.5, входящее в пакет Weka 3.4.4. В качестве основы для классификации используются две группы свойств – связанные с содержанием и со ссылочной структурой. К первой группе относятся: число слов на странице, средняя длина слов на странице, процент слов из списка популярных слов, процент видимого содержания страницы, число слов в заголовке страницы и т.д. Во второй группе относятся: процент страниц на наиболее популярном уровне, число входящих ссылок на страницу, число исходящих ссылок на страницу, отношение числа входящих и исходящих ссылок, число ссылок с главных страниц, процент входящих ссылок на наиболее популярные страницы, процент исходящих ссылок на наиболее популярные страницы, перекрестные ссылки на страницу, средний уровень страниц на сайте и т.д.

В [3] также применяется дерево решений C4.5. Как и в [19] выделяются две группы свойств. Ссылочные свойства включают: 16 свойств степени близости (входящие и исходящие ссылки, число взаимных ссылок и т.д.), 11 свойств, основанных на PageRank (различные меры, связанные с PageRank страницы и PageRank со ссылающихся на нее страниц), Truncated PageRank и т.д. 24 свойства, зависящие от содержания, аналогичны [18]. Предложенный авторами алгоритм позволяет определять 88,4% спама. Еще одно применение дерева решений C4.5 описано в [1]. В качестве основы для классификации используются ссылочные структуры вэб-графа. Работа является продолжением исследований [3].

Работа [6] посвящена обнаружению ссылочного спама. Задача сводится к разбиению страниц на два класса – “spam” (спам) и “ham” (не спам). Для этого используется метод опорных векторов (пакет SVM-light со стандартными параметрами). Для каждой страницы выделяются 89 свойств и TF-IDF вектор. Авторы упоминают о следующих свойствах, используемых для классификации: число слов в метатегах keyword, description и заголовке, редирект на страницу, число входящих и исходящих ссылок, число символов в URL и домене, число поддоменов в URL, длина страницы, домены в зонах .edu, .org, .biz, .com, одинаковые IP-адреса, одинаковый размер страниц и т.д. Для обнаружения одинаковых страниц авторы предлагают использовать MD5 хэш-коды. Авторы анализируют различные варианты ядер SVM и выявляют наиболее важные для классификации свойства.

В работе [21] рассматривается один из подходов борьбы с поисковым спамом. Задача сводится к классификации страниц с использованием метода опорных векторов (SVM). За основу берется функция линейного ядра. В качестве базиса для классификации используется 360 свойств (зависящих и независящих от запросов). Среди независящих от запроса свойств выделяются: свойства страницы (статический ранг, самый частый терм, число уникальных термов, общее количество термов, число слов в адресе страницы, число слов в заголовке), свойств домена (ранг домена, среднее число слов, уровень домена), популярность (число

посещений домена, пользователи домена, число посещений страницы, пользователи страницы), время (дата посещения пауком, время затраченное на получение страницы, последняя дата изменения) и т.д. К зависящим от запроса свойствам относятся: количество слов запроса в заголовке, частота слов запроса в документе, частота слов запроса во всех документах, количество документов, содержащих слова запроса, n-граммы по словам запроса / документу и т.д. Эти свойства вычисляются отдельно для каждой пары (запрос, страница).

В [20] рассматривается подход к обнаружению E-mail спама и спама блогов (как ссылочного, так и спама контента) с помощью ослабленного онлайн SVM (Relaxed Online SVM). Демонстрируются возможности работы с большими наборами данных и обсуждаются возможности существенного снижения вычислительной нагрузки.

В [15] также описывается применение SVM для определения ссылочного спама в блогах. Большое внимание уделяется локальным свойствам, зависящим только от содержания каждой конкретной страницы. Проводится сравнение классификации на основе локальных свойств для линейного ядра, ядра радиальной базисной функции и сигмоида. Далее рассматриваются глобальные свойства и проводится сравнение классификации для линейного, полиномиального ядра, ядра радиальной базисной функции.

В [24] кроме содержания страницы, учитываются временные характеристики и ссылочная структура вэб-графа (на основе HITS-алгоритма) для определения спама в блогах. Для классификации используется SVM с полиномиальным ядром.

В [7] проводится сравнение метода опорных векторов, деревьев решений C4.5 и алгоритма Rocchio. Авторы указывают, что SVM работает быстрее, чем C4.5, как при обучении, так и классификации.

Анализ показывает, что существующие алгоритмы базируются на анализе структуры сети ссылок, выявлении спам-страниц и сайтов и т.д. Но они практически не предназначены для обнаружения “хороших” и “спам” ссылок на каждой отдельной странице.

Цель нашего исследования – определение спам-ссылок на любых веб-сайтах, в том числе авторитетных. На каждой отдельной странице могут присутствовать и обычные и спам-ссылки.

3 Метод исследования

Анализ показал, что среди методов машинного обучения высокие перспективы имеем метод опорных векторов [4, 22]. Этот метод позволяет добиться высокого качества в области классификации.

Работа метода опорных векторов включает в себя два основных этапа – обучение на тренировочных данных и непосредственно классификация.

Для работы метода необходимо определение пространства признаков, по которым будет проходить выявление ссылочного спама.

4 Признаки ссылочного спама

Рассмотрим признаки ссылочного спама, которые можно выделить на основе анализа содержания страницы или всего сайта. Признаки, основанные на свойствах вэб-графа (ссылочной структуре, образуемой сайтами интернет), в рамках данной работы рассматриваться не будут.

Из применяемых в настоящее время признаков (см. пункт 2), львиная доля предназначены для выявления страниц и сайтов со спамом. Для идентификации на странице конкретных спам-ссылок такие признаки использоваться практически не могут. По этой причине нами были выделены новые признаки, способные справиться с поставленной задачей. Надо заметить, что признаки не претендуют на полноту и их перечень будет расширяться по мере развития и совершенствования работы.

Все признаки ссылочного спама мы разбили на две группы [9, 17, 26, 27].

Группа 1. Свойства ссылки.

1.1. Тематическая близость ссылки и страницы.

Указывает, насколько отличается текст ссылки от тематики страницы, на которой ссылка расположена.

1.2. Тематическая близость сайта, на который ведет ссылка ссылки и страницы, на которой ссылка расположена.

1.3. Тематическая близость соседних ссылок.

Сигнализирует, совпадает ли тематика ссылки с тематикой соседних ссылок (в случае наличия блока ссылок).

1.4. Расположение ссылки в блоке ссылок.

Указывает, является ли ссылка одиночной, либо расположенной в области с повышенной плотностью ссылок на небольшом участке страницы (блоке ссылок).

1.5. Место расположения ссылок.

Указывает положение ссылки на странице – в начале или конце страницы, по центру, в левом или правом столбце и т.д. Также указывает на расстояние ссылки от основного содержания страницы.

1.6. Пометка ссылки как рекламного объявления.

Сигнализирует, есть ли в окрестностях ссылки пометки "Реклама", "Спонсоры", "Наши Партнеры", и т.д.

1.7. Наличие похожих ссылок на сайте.

Указывает, встречаются ли еще на сайте ссылки, похожие на анализируемую ссылку.

1.8. Наличие ссылки в спам-списке.

Спам-список [27] содержит ссылки, отобранные вручную и определенные раннее как спам.

1.9. Признак размещения ссылки рекламным брокером

Используется способ, описанный в [25]. Суть его заключается в том, что ссылки от рекламных брокеров устанавливаются для определенных страниц, и чаще всего с помощью одного и того же кода. Соответственно, рекламный брокер узнает о том, какой код разместить на странице, анализируя строку адреса страницы, например, http://www.site.ru/index.php?cat=1&page=11. Тогда, передав дополнительный параметр (например, http://www.site.ru/index.php?cat=1&page=11&**aa=bb**) можно ввести в заблуждение рекламного брокера, и он не установит рекламные ссылки на страницу. Сравнив содержание страницы в первом и втором случае, появляется возможность выявить такие ссылки.

Группа 2. Свойства страницы/сайта.

2.1. Наличие спам-ссылок на сайте.

2.2. Наличие спам-ссылок на странице.

2.3. Сайт продает ссылки.

Указывает, если ли на сайте информация о том, как можно купить ссылки.

2.4. Наличие на сайте признаков кода рекламных брокеров.

Многие автоматизированные системы установки ссылок (биржи, обменники, брокеры) устанавливают код автоматически по шаблону. Наличие блока идентичных по коду ссылок может указывать на их спамерское происхождение.

2.5. Наличие на странице признаков кода рекламных брокеров.

2.6. Наличие на сайте ссылки на рекламного брокера.

2.7. Наличие на странице ссылки на рекламного брокера.

2.8. Отношение числа внешних ссылок на странице к среднему числу внешних ссылок на сайте.

2.9. Процент контента страницы, занятого внешними ссылками.

2.10. Совпадение IP-адресов сайтов.

Указывает, совпадают ли IP-адреса сайта, на котором размещена ссылка, с сайтом, на который она указывает.

2.11. Совпадение контактных E-mail сайтов.

Указывает, совпадает ли контактный E-mail (указанный при регистрации домена) сайта, на котором размещена ссылка, с контактным E-mail сайта, на который ссылка указывает.

5 Набор данных

В качестве тестовых наборов мы использовали собственную коллекцию RV, коллекции narod.ru и By.Web семинара РОМИП. В каждой коллекции были выделены ссылки, для которых установлены метки “спам” и не “спам”.

В коллекцию RV вошли ссылки с 20 сайтов, размещающих спам-ссылки (информация о местах размещения платных ссылок были предоставлены нам владельцами сайтов). Число страниц на каждом сайте – от 100 до 5000 [27]. Всего было размечено (в автоматическом режиме) 23000 спам-ссылок и 8000 обычных ссылок.

В связи с тем, что коллекция narod.ru содержит сайты 2003 года, когда ссылочный спам только начинал свое массовое распространение (первая биржа ссылок clx.ru появилась в середине 2002 года), в ней отсутствуют некоторые признаки, присущие современному ссылочному спаму. Мы произвольно выбрали из коллекции набор страниц, на которых вручную провели разметку ссылок. Всего было размечено 2000 ссылок, из которых спам-ссылок 500, обычных ссылок 1500.

Коллекция By.Web оказалась более современной и интересной. В ней ссылочный спам представлен достаточно ярко и разносторонне. Из-за ограниченности в ресурсах, мы выбрали по 3500 спам и обычных ссылок.

6 Результаты исследований

В качестве основы для исследования был взят пакет SVM-Light [14]. Мы использовали линейное ядро с параметрами по умолчанию.

Для коллекции RV мы выбрали 4000 ссылок для обучения (по 2000 спам и не спам). Для классификации было использовано 21000 спам и 6000 не спам ссылок.

Для коллекции Narod.ru мы выбрали 200 ссылок для обучения (по 100 спам и не спам). Для классификации было использовано 400 спам-ссылок и 1400 не спам.

Для коллекции By.Web мы выбрали по 1750 спам и не спам ссылок для обучения. Для классификации было использовано также по 1750 ссылок (всего 3500).

Таким образом, для обучения во всех трех коллекциях использовалось одинаковое число положительных и отрицательных примеров (в нашем случае, спам и не спам ссылки).

Для оценки качества работы алгоритма использовалась следующие метрики [27]:

$$\text{Precision} = \frac{\text{Число спам - ссылок, отмеченных как спам}}{\text{Число ссылок,отмененных как спам}}$$

$$\text{Recall} = \frac{\text{Число спам - ссылок, отмеченных как спам}}{\text{Общее число спам - ссылок}}$$

$$\text{FalseSpam} = \frac{\text{Число обычных ссылок, отмеченных как спам}}{\text{Общее число обычных ссылок}}$$

$$\text{FalseNotSpam} = \frac{\text{Число спам - ссылок, отмеченных как не спам}}{\text{Общее число спам - ссылок}}$$

Значения метрик для обеих коллекций приведены в таблице 2.

Качество определения спам-ссылок для коллекции RV значительно лучше, чем для коллекций Narod.ru и By.Web. Разницу можно объяснить существенным различием в виде

тестовых данных. В коллекции RV преобладают ссылки с явно выраженными признаками спама (сайты массово размещают ссылки через рекламных брокеров). В коллекции Narod.ru преобладают обычные ссылки. Невысокое значение Precision (0.53) объясняется ошибочным отнесением хороших ссылок в разряд спама. В коллекции By.Web значение Precision несколько выше (0.72). В тоже время количество ссылок, ошибочно отнесенных к разряду спама в этой коллекции несколько выше (FalseSpam = 0.3).

Таблица 2 – Результаты работы

	RV	Narod.ru	By.Web
Precision	0.95	0.53	0.72
Recall	0.87	0.77	0.8
FalseSpam	0.13	0.20	0.3
FalseNotSmap	0.13	0.23	0.2

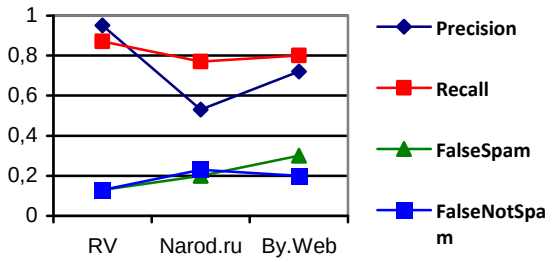

Рис. 2. Сравнение результатов для разных коллекций.

Таким образом, предложенный метод обнаружения ссылочного спама демонстрирует приемлемые результаты. Продолжение работы мы видим в расширении пространства признаков, исследовании их влияния на качество классификации, оптимизации параметров SVM-Light.

Литература

[1] Becchetti L., Castillo C., Donato D., Leonardi S., Baeza-Yates R. Link Analysis for Web Spam Detection. ACM Trans. Web 2, 1, 1-42, 2008

[2] Benczur A. A., Csalogany K., Sarlos T., Uher, M. Spamrank - fully automatic link spam detection. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[3] Castillo C., Donato D., Gionis A., Murdock V., Silvestri F. Know Your Neighbors: Web Spam Detection Using the Web Topology. SIGIR'07, May, 2007.

[4] Cristianini N., Shawe-Taylor J. "An introduction to Support Vector Machines", Cambridge, 2000.

[5] Davison B. D. Recognizing nepotistic links on the web. In AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, pages 23–28, July 30 2000.

[6] Drost, I., Scheffer, T. Thwarting the Nigritude Ultramarine: Learning to Identify Link Spam. in 16th European Conference on Machine Learning, (Porto, 2005).

[7] Drucker H., Wu D., Vapnik, V. Support vector machines for Spam categorization. IEEE-NN 10(5):1048–1054, 1999.

[8] Eiron N., McCurley K. S., Tomlin J. A. Ranking the web frontier. In Proceedings of the 13th International World Wide Web Conference (WWW), pages 309–318, New York, NY, USA, 2004. ACM Press.

[9] Enge E. 15 Methods for Paid Link Detection http://www.stonetemple.com/blog/?p=167

[10] Fetterly D., Manasse M., Najork M. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In Proceedings of the 7th International Workshop on the Web and Databases (WebDB), Paris, France, 2004.

[11] Gyongyi Z., Garcia-Molina H., Pedersen J. Combating web spam with TrustRank. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada, 2004.

[12] Gyongyi Z., Berkhin P., Garcia-Molina H., Pedersen J. Link Spam Detection Based on Mass Estimation. In: 32nd International Conference on Very Large Data Bases (VLDB 2006), September 12-15, 2006, Seoul, Korea

[13] Gyongyi Z., Garcia-Molina H. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), May 10-14, 2005, Chiba, Japan.

[14] Joachims T. Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press, 1998.

[15] Kolari P., Java A., Finin T., Oates T., Joshi A. Detecting Spam Blogs: A Machine Learning Approach. AAAI '06, 2006.

[16] Krishnan V., Raj R. Web Spam Detection with Anti-Trust-Rank. In the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '06), August 2006.

[17] Nash T. How to find a paid link? http://paymentblogger.com/2007/10/07/how-to-find-a-paid-link/

[18] Ntoulas A., Najork M., Manasse M., Fetterly D.. Detecting spam web pages through content analysis. In WWW, pages 83–92, Edinburgh, Scotland, May 2006.

[19] Qingqing Gan, Torsten Suel. Improving web spam classifiers using link structure. Proceedings in Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07), May 2007, Banff, Alberta, Canada.

[20] Sculley D., Gabriel M. Wachman. Relaxed Online Support Vector Machines for Spam Filtering, SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference, 2007

[21] Svore, K., Wu, Q., Burges, C. and Raman, A. Improving Web Spam Classification using Rank-time Features. Proceedings in Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07), May 2007.
[22] Vapnik V. "An Overview of Statistical Learning Theory", IEEE Transactions on Neural Networks, 1999.
[23] Wu B., Davison B. D. Identifying link farm pages. In Proceedings of the 14th International World Wide Web Conference (WWW), 2005.
[24] Yu-Ru Lin, Hari Sundaram, Yun Chi, Jun Tatemura, Belle Tseng. Splog Detection Using Content, Time and Link Structures,Proc. International Conference and Multimedia Expo (ICME) 2007, July 2007, Beijing, China.
[25] Детектор продажных ссылок, 2008. http://venality.name/
[26] Кравцов Алексей. Ссылочный спам: найти и обезвредить http://www.kravcov.ru/2007/03/11/nnueiiue-niai-e-eae-n-iei-aidhiouny/
[27] Шарапов Р.В., Шарапова Е.В. Обнаружение ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции «RCDL'2008» (Дубна, Россия, 7-11 октября 2008 г.). - Дубна: ОИЯИ, 2008. С. 191-196.

The Using of Support Vector Machines for Link Spam Detection

Sharapov R.V., Sharapova E.V.

In article approaches to detecting of Link Spam by methods of machine learning are considered. The significant signs helping Link Spam detection are analyzed. The algorithm of detecting a spam-links on basis of Support Vector Machine is given and results of its work are considered.