Алгоритм обнаружения ссылочного спама

An algorithm of link spam detection

Шарапов Р. В. (info@vanta.ru), **Шарапова Е. В.** (mivlgu@mail.ru)

Муромский институт (филиал)
Владимирского государственного университета

В статье рассматриваются подходы к выявлению ссылочного спама на основе анализа содержания страницы. Основное место в работе посвящено выявлению рекламных (платных) ссылок. Анализируются признаки, характерные для рекламных ссылок. Дается алгоритм выявления спам-ссылок и приводятся результаты его работы.

1. Введение

Количество информации, доступной пользователям глобальной сети Интернет, с каждым годом становится все больше и больше. Постоянно растет число сайтов, увеличивается число их страниц. Например, поисковая система Яндекс, на момент написания статьи, осуществляла поиск по 17 миллионам сайтов и 4,5 миллиардам веб-страниц [15]. Интернет становится не только средством получения информации и общения, но и средством ведения бизнеса. Естественно, нахождение сайта по наиболее популярным поисковым запросам и положение сайта в списке результатов поиска является актуальной проблемой для большинства владельцев сайтов. В связи с тем, что сайтов одной тематики иногда бывает слишком много, и каждый хочет быть на вершине списка результатов поиска по ключевым запросам, владельцы начинают прибегать к различным ухищрениям, чтобы поднять свои сайты как можно выше к началу списка. Простым выходом кажется применение различных технологий манипулирования поисковыми системами, например с помощью поискового спама. Существует большое количество методик, используемых для того, чтобы ввести в заблуждение поисковые системы [8]. Рассмотрим одну из них — ссылочный спам.

Увеличение числа ссылок на сайты стало одним из основных методов манипулирования поисковыми системами в последнее время. Масштабы манипуляции постоянно растут. Если несколько лет назад основным способом являлся так называемый обмен ссылками, который проводился вручную, то теперь ему на смену пришли различные способы автоматического размещения ссылок. Можно выделить несколько вариантов такого размещения [17]:

1. Использование специализированных программ для автоматического добавления ссылок в каталоги, гостевые книги, форумы и т.д.
2. Покупки ссылок у рекламных брокеров.

С первым вариантом поисковые системы научились бороться, выявляя ресурсы, где есть возможность простого, немодерируемого добавления ссылок. Вес ссылок с таких ресурсов сильно снижается. Размещение же ссылок с использованием рекламных брокеров представляет для поисковых систем много большую проблему.

В настоящее время в русском сегменте интернет действует около десятка крупных рекламных брокеров, занимающихся продажей текстовых ссылок. Только один из них, Sape.ru, имеет возможность размещать ссылки на более чем 55 миллионах страниц (за прошедшие 6 месяцев число страниц увеличилось на 20 миллионов) [11]. Несмотря на то, что ссылки в таких системах называют «рекламными», их основная цель — не реклама с целью привлечения посетителей (ссылки часто размещаются в самых неприметных местах страницы и пользователь их просто не замечает), а улучшение своего положение в поисковых системах. Стоимость такой «рекламы» также часто бывает номинальной, иногда всего 0.01$ за месяц размещения. Ссылки, размещенные с помощью рекламных брокеров, будем называть «платными», подчеркивая что ссылки имеют искусственное происхождение (т. е. не имеют никакого отношения к содержанию страницы). Они размещаются владельцем страницы за деньги, а не из «уважения» к сайту, на который ссылаются.

В чем же основная опасность крупномасштабного ссылочного спама, наблюдаемого последние несколько лет? Опасность заключается в том, что ссылки активно используются современными поисковыми системами для ранжирования результатов поиска. Со ссылками связано и понятия Индекса цитируемости в Яндекс и определение PageRank в Google. Массовое увеличение ссылок неестественного происхождения (ссылочного спама) может сильно «испортить» эффективность их работы. Ситуация осложняется тем, что «платные» ссылки могут размещаться на любых сайтах, в том числе и на очень уважаемых и популярных ресурсах. Таким образом, становится невозможным простое деление страниц на «хорошие» и страницы для ссылочного спама [17].

2. Текущее состояние проблемы

В настоящее время существует несколько подходов к определению поискового спама. 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. Например, в работе [7] 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 [4] 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.

В работе [1] извлекаются особенности, основанные на связанных образцах сайтов. Кластеризация пространства особенностей позволяет выделить кластеры, сайты которых принадлежат одной и той же группе спам-сайтов.

Для выявления ссылочного спама с помощью Truncated PageRank [2] предлагается анализировать топологию сети ссылок. На основе вычисляемых алгоритмом атрибутов производится классификация ссылок на предмет спама.

В работе [12] предлагается идентифицировать страницы с «ферм ссылок», основываясь на наблюдении, что входящее и исходящее их окружение имеет тенденцию пересекаться. Набор «плохих» страниц многократно расширяется и ссылки между ними отбрасываются.

Другая группа работ основана на анализе содержания страниц.

В [10] рассматриваются различные характеристики страницы (число слов на странице и в заголовке, длина слов, процент видимого текста и т.д.). Проводя сравнение выявленных характеристик с их распределением на «обычных» страницах можно выявить страницы, содержащие спам.

В работе [6] предлагается статистический анализ для выявления автоматически сгенерированных страниц со спамом. Отклонения от нормального распределения различных свойств страниц, включая имена и IP-адреса, входящие и исходящие ссылки, содержание страницы и норму изменения, — все это может свидетельствовать о спаме.

В работе [3] предлагается применять дерево решений для отделения спам-ссылок от обычных.

Таким образом, существующие алгоритмы базируются на анализе структуры сети ссылок, выявлении спамерских страниц и сайтов и т.д. Но существующие алгоритмы практически не предназначены для обнаружения «хороших» и «спамерских» ссылок на каждой отдельной странице [17].

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. Выявление ссылочного спама

Рассмотрим признаки определения рекламных/платных ссылок [5, 9, 17]:

3.1. Ссылки, отмеченные как рекламные объявления

Для этого необходимо просмотреть окрестность ссылки (текст, соседствующий ссылке). Признаки платной ссылки — слова: «Реклама», «Спонсоры», «Наши Партнеры», и т.д.

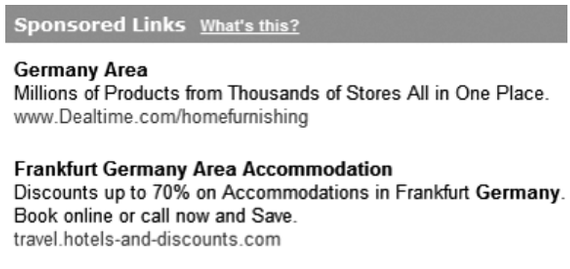


Рис. 1. Пример рекламных объявлений

3.2. Большой блок ссылок

Повышенная плотность ссылок на небольшом участке страницы (блок ссылок) может свидетельствовать об их неестественном происхождении.

3.3. Ссылки на агентства по продаже ссылок/рекламы

Часто вблизи рекламных блоков можно увидеть ссылки на рекламных брокеров.

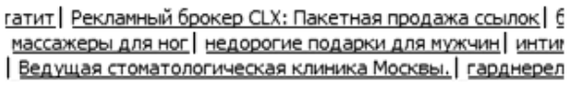
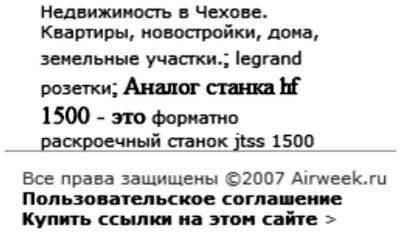

гатит | Рекламный брокер CLX: Пакетная продажа ссылок | б
массажеры для ног | недорогие подарки для мужчин | интим
| Ведущая стоматологическая клиника Москвы. | гарднерел

Рис. 2. Пример ссылки на рекламного брокера в блоке ссылок

3.4. На сайте есть информация о том, как можно купить ссылки

Если на сайте или около блока ссылок содержится такая информация, это является фактом, заставляющим усомниться в том, что ссылки являются естественными, а не рекламными.


Недвижимость в Чехове.
Квартиры, новостройки, дома,
земельные участки.; legrand
розетки; **Аналог станка hf**
1500 - это форматно
раскроечный станок jtss 1500

Все права защищены ©2007 Airweek.ru
Пользовательское соглашение
Купить ссылки на этом сайте >

Рис. 3. Пример предложения о покупке ссылок

3.5. Тематическая близость ссылки

Если текст ссылки или тематика сайта, на который ведет ссылка сильно отличается от тематики страницы, на которой ссылка расположена, то ссылку можно считать спамом.

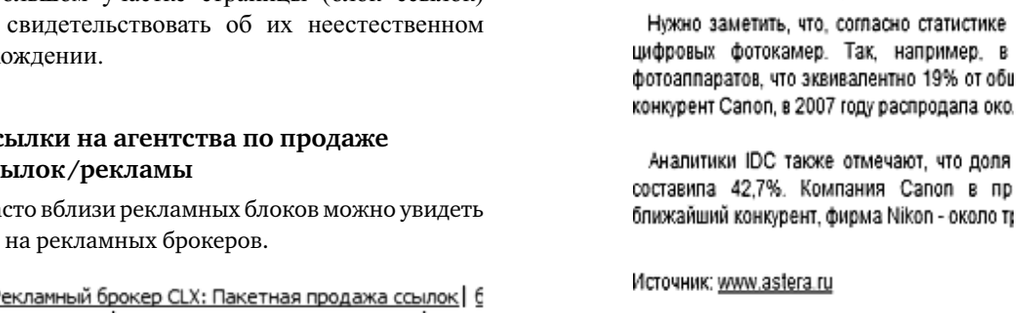
Однако определение тематики ссылки не всегда является тривиальной задачей. Ссылка может располагаться внутри предложения (хотя и не быть частью основного текста страницы). Поэтому ориентироваться на текст в непосредственной близи к ссылке не всегда оправдано.

Агентство дает ипотечный кредит под 9 процентов.

Агентство дает ипотечный кредит **под 9 процентов**.

Рис. 4. Пример ссылки с текстом по краям

Часто ссылки указывают на ресурс с достаточно общей тематикой (например, при ссылках на источник новостей или сайт автора какой-либо статьи).


Нужно заметить, что, согласно статистике
цифровых фотокамер. Так, например, в
фотоаппаратов, что эквивалентно 19% от общ
конкурент Canon, в 2007 году распродала око

Аналитики IDC также отмечают, что доля
составила 42,7%. Компания Canon в пр
ближайший конкурент, фирма Nikon - около т

Источник: www.astera.ru

Рис. 5. Пример ссылки на источник

Для правильного определения тематики ссылки может помочь глубокий анализ тематика сайта, на который ведет ссылка. Понятно, что задача эта трудоемка и требует большого количества времени.

3.6. Тематическая близость соседних ссылок

Для этого необходимо проанализировать тематику группы ссылок, размещенных на странице. Если ссылки не являются тематическими и имеют явно выраженный разброс тематики, то они — рекламные.

3.7. Место расположения ссылок

Для этого необходимо проанализировать расположение ссылок на странице. Чем дальше ссылки от основного содержания страницы, тем более вероятно, что они являются рекламными. Например, часто такие ссылки размещаются внизу страницы или в правом столбце, когда основной текст располагается посередине.

3.8. Код ссылок

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.

3.9. Динамичность/Время жизни ссылок

Частое изменение ссылок на страницах без изменения остального содержания может свидетельствовать о неестественном их происхождении. Ссылки могут либо просто на время исчезать со страниц (в случае неполадок систем по автоматическому размещению ссылок), либо часть их может заменяться на новые.

3.10. Сообщение о платных ссылках

О платных ссылках могут сообщить конкуренты, бывшие покупатели ссылок, бывшие сотрудники и т.д.

3.11. Просмотр страницы человеком

Просмотр страниц модератором и выявление ссылочного спама вручную.

4. Алгоритм обнаружения ссылочного спама

Теперь рассмотрим алгоритм, способный выявить спамерские ссылки. Он состоит из нескольких этапов.

Этап 1: Формирование предварительного набора спам-ссылок *S*. Набор формируется из следующих ссылок:

- выбранных вручную;
- определенных алгоритмом раннее, как спам;
- определенных на основе анализа кода рекламных брокеров для автоматического размещения ссылок.

Наибольший интерес представляет именно последний способ. У некоторых рекламных брокеров можно обнаружить отличительные особенности в размещении кодов, которые могли бы помочь их идентифицировать. Например, при автоматическом размещении кода Prospero.ru можно заметить следующее:

```
<a class=prospero href="http://www.logipark.
ru">таможенное оформление Япония</a>
<a class=prospero href="http://www.svadbaexclusive.
com/">ЗАГСы Москвы, организация свадьбы
в Москве</a>
```

При кэшировании рекламных ссылок иногда можно увидеть такой код:

```
<!--from cache 14:18:25 13.04.2008-->
<a href="http://www.clinicsex.ru/" target=_
blank>цитомегаловирус затем гарднереллез
анализы мочи</a>
<a href="http://zemnovosti.ru" target=_
blank>Статьи земельная тематика</a>
<!--/from cache-->
```

Оригинальный способ предлагается на [14]. 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 платные ссылки.

Еще один метод заключается в отслеживании динамики изменения содержания страницы. Если в течение времени на странице изменяется только группа ссылок, то эта группа может являться платными ссылками. Аналогичные выводы можно сделать, увидев на странице в какой-то момент времени следующее сообщение:

```
<b>Warning</b>: mysql_connect(): Too many
connections in <b>/home/clx/inc/conf.inc</b>
on line <b>56</b><br />
```

Надо заметить, что не все ссылки, определенные алгоритмом как спам стоит заносить в набор *S*, а только те, чьи признаки спама носят явно выраженный характер (чтобы исключить случайного попадания ссылок в разряд спама).

На этапе 1 можно использовать различные алгоритмы классификация и машинного обучения.

Этап 2: Выявление спам-ссылок на основе содержания страницы. Основная идея состоит в анализе содержания страницы и выявлении признаков спама. За каждый признак спама на ссылку налагается штраф q_i. Если суммарный штраф превышает определенный порог, ссылка признается спамом.

Шаг 1. Страница сканируется на наличие ссылок S_b, занесенных в список *S*, сформированный на Этапе 1. При обнаружении таких ссылок сканируется область вокруг них. Если ссылки обнаружены, то им назначается штраф q_1, величина которого снижается по мере удаления от ссылки S_b.

Шаг 2. Страница сканируется на наличие признаков рекламного блока. Признаком могут служить слова «Реклама», «Спонсоры», «Наши Партнеры» и т.д. При обнаружении признаков рекламного блока, ссылкам в его окрестностях назначается штраф q_2.

Шаг 3. Страница сканируется на наличие ссылок на рекламного брокера. При обнаружении таких признаков рекламного блока, ссылкам в его окрестностях назначается штраф q_3.

Шаг 4. Страница сканируется на наличие информации о продаже ссылок (и о том, каких можно купить). При обнаружении таких признаков, ссылкам в их окрестностях назначается штраф q_4.

Шаг 5. Страница сканируются на наличие большого блока ссылок. Если количество ссылок в блоке больше определенного порога, им назначается штраф q_5.

Шаг 6. Ссылки сканируются на признаки кода рекламного брокера, в случае обнаружения которого ссылкам назначается штраф q_6.

Шаг 7. Проверяется соответствие тематики ссылки и общей тематики страницы. В случае несоответствия, ссылке назначается штраф q_7. Для проверки тематики часто бывает достаточно просто просканировать текст страницы на совпадение слов с текстом ссылки.

Шаг 8. Проверяется соответствие тематики ссылки и тематики ссылок в ее окрестностях. В случае несоответствия, ссылке назначается штраф q_8.

Шаг 9. Проверяется место размещения ссылки. Если ссылка находится в самом конце страницы, ей назначается штраф q_9.

Этап 3. Анализ структуры сайта с целью выявления спама. Этот этап является самым сложным. Его цель — выявить особенности структуры сайта и места на страницах, где встречаются «платные» ссылки.

Для этого из страниц сайта удаляется весь изменяющийся контент (кроме ссылок). Далее производится объединение страниц с одинаковым шаблоном в кластеры. Следующий этап: для каждого кластера удаляются повторяющиеся ссылки и идентифицируются области, где ссылки меняются на каждой страницы кластера. Для ссылок, входящих в такие области назначается штраф q_r.

Этап 4. Для каждой ссылки все начисленные штрафы суммируются. Если сумма превышает определенный порог, делается вывод, что ссылка — спам. В этом случае ссылка заносится в список *S*.

5. Результаты исследований

Для формирования начального набора спам-ссылок *S* были просканированы 20 сайтов, размещающих платные ссылки (информация о местах размещения платных ссылок были предоставлены нам владельцами сайтов). Число страниц на каждом сайте — от 100 до 5000. После удаления дубликатов был получен набор из 15000 платных ссылок.

В предыдущих исследованиях [17] были реализованы этапы 1,2 и 4. В текущей работе мы продолжили реализацию этапа 3.

При слабо выраженных иных факторах, тематическая близость становится наиболее значимым мерилом спама. Особо это актуально для одиночных ссылок. В наших предыдущих исследованиях [17] мы использовали «наивный» подход для определения тематической близости, основанный на совпадении набора слов. Конечно, такой подход являлся достаточно приближенным и не точным. Кроме того, в рамках наших исследований мы остановились лишь на анализе страницы, на которую ведет ссылка, а не всего сайта. Это позволило ускорить задачу анализа страниц. Недостатком такого подхода явилось возникновение неточностей в определении тематической близости (так как часто ссылки ведут на главную страницу, а не на подраздел сайта). В частности такой подход вызвал ошибочное отнесение ряда ссылок в разряд спама.

Теперь мы усовершенствовали методику определения тематической близости. Некоторые идеи были почерпнуты их [16].

Для оценки качества работы алгоритма использовалась методика, описанная в [2].

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

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

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

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

Для тестирования были вручную отобраны 100 страниц с числом внешних ссылок от 1 до 30 на каждой. Общее количество ссылок составило 783. Для каждой страницы были вручную отмечены спам-ссылки, которых оказалось 519. В результате работы алгоритма 488 ссылок были отмечены как спам, их которых 461 действительно были спам-ссылками (совпали с отобранными вручную). Результаты оценки качества работы алгоритма приведены в таблице 1.

Таблица 1. Результаты тестирования алгоритма

Precision	0,94
Recall	0,89
FalseSpam	0,102
FalseNotSmap	0,112

Ряд ошибок [17] в выявлении спам-ссылок возникало из-за неглубокого анализа тематической близости. Усовершенствование алгоритма определения тематического подобия позволило повысить Precision на 2%, Recall на 3%, снизить FalseSpam на 5% и FalseNotSmap 3%.

Одиночные рекламные ссылки (в основном, размещенные вручную) близкой к странице тематики не были выявлены как спам. Это связано с тем, что у таких ссылок признаки спама часто отсутствуют. Решением может служить только анализ структуры страниц сайта и выявление мест размещения рекламы (планируется осуществить в будущем), а также анализ времени жизни ссылок, для чего необходим длительный мониторинг страниц.

Таким образом, предложенный алгоритм демонстрирует достаточно неплохие результаты в определении спам-ссылок.

Литература

1. *Amitay E., Carmel D., Darlow A., Lempel R., Soffer A.* The connectivity sonar: Detecting site functionally by structural patterns. In Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia, Aug. 2003, pages 38–47.
2. *Becchetti L., Castillio C., Donato D., Leonardi S., Baeza-Yates R.* Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection. Technical report, DELIS — Dynamically Evolving, Large-Scale Information Systems, 2006.
3. *Davison B. D.* Recognizing nepotistic links on the web. In AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, July 30 2000, pages 23–28.
4. *Eiron N., McCurley K. S., Tomlin J. A.* Ranking the web frontier. In Proceedings of the 13th International World Wide Web Conference (WWW), New York, NY, USA, 2004. ACM Press, pages 309–318.
5. *Enge Eric.* 15 Methods for Paid Link Detection http://www.stonetemple.com/blog/?p=167
6. *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.
7. *Gyöngyi 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.
8. *Gyöngyi Z., Garcia-Molina H.* Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.
9. *Nash Tim.* How to find a paid link? http://paymentblogger.com/2007/10/07/how-to-find-a-paid-link/
10. *Ntoulas A., Najork M., Manasse M., Fetterly D.* Detecting spam web pages through content analysis. In Proceedings of the World Wide Web conference, Edinburgh, Scotland, May 2006, pages 83–92.
11. *Sape.ru* — главная страница, 2009. http://www.sape.ru/
12. *Wu B., Davison B. D.* Identifying link farm pages. In Proceedings of the 14th International World Wide Web Conference (WWW), 2005.
13. *Кравцов Алексей.* Ссылочный спам: найти и обезвредить http://www.kravcov.ru/2007/03/11/nnueiiue-niai-e-eae-n-iei-aidhiouny/
14. *Детектор* продажных ссылок, 2008. http://venality.name/
15. *Компания* Яндекс, 2009. http://company.yandex.ru/
16. *Некрестьянов И. С.* Тематико-ориентированные методы информационного поиска: Диссертационная работа к.т.н.: 05.13.11 // Санкт-Петербургский государственный университет. СПб., 2000. 80 с.
17. *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.