Анализ динамических характеристик поискового спама, создаваемого ссылочными брокерами

© Р.В. Шарапов, Е.В. Шарапова

Муромский институт (филиал) Владимирского государственного университета

info@vanta.ru

Аннотация

Рассматриваются характеристики поискового спама, размещаемого ссылочными брокерами. Исследуется время жизни и ротация ссылок, анализируется тематическая близость ссылок и страниц. Исследуется поведение ссылок в группах ссылок. Рассматривается возможность использования динамических характеристик для обнаружения ссылочного спама.

1 Введение

Последнее десятилетие ознаменовалось бурным развитием глобальной сети интернет. Появилось большое количество разнообразных сайтов, содержащих различную информацию. Для облегчения нахождения пользователями сети нужных сведений появились поисковые системы. По мере своего развития они становились все сложнее, что было вызвано стремлением удовлетворить все более растущие потребности пользователей. Для этого поисковые системы научились использовать не только основное содержание документов, но и дополнительные сведения о них. Появление понятия «авторитетности» ресурсов, связанное с внедрением таких алгоритмов, как PageRank и HITS, привело к активному использованию поисковыми системами ссылочной информации.

Параллельно с поисковыми системами развивался и поисковый спам как средство введения в заблуждение поисковых систем. К настоящему времени поисковый спам имеет множество видов и направлений, таких как клоакинг, дорвеи, спам содержания, спам комментариев и т. д. [5].

Использование поисковыми системами ссылок привело к возникновению нового вида поискового спама, получившего название ссылочный спам.

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 [11].

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

Одним из основных источников размещения ссылочного спама является покупка ссылок через ссылочных (рекламных) брокеров. В настоящее время наибольшую популярность получили такие брокеры, как Sape.ru, MainLink.ru, Xap.ru, Link-Feed.ru, SetLinks.ru, Clx.ru и т. д. Суммарное число страниц, на которых такие системы могут размещать ссылки, превышает полмиллиарда и стремительно растет. Для сравнения, поисковая система Яндекс в настоящее время осуществляет поиск по 8,9 миллиардам страниц. Таким образом, доля страниц, на которых ссылочные брокеры могут размещать ссылки, превышает 6% от размеров поисковой базы Яндекса. В 2008 году эта доля составляла всего 1 – 2%.

Как уже отмечалось ранее [9, 11], ссылки, размещаемые через ссылочных брокеров, имеют целью поисковый спам и не выполняют функции рекламы (как их позиционируют ссылочные брокеры).

Учитывая массовое распространение ссылочных брокеров, размещаемые ими ссылки нуждаются в детальном изучении. Следует подвергнуть исследованию все стороны жизни таких ссылок.

2 Текущее состояние проблемы

Вопросам изучения ссылочного спама посвящено немало работ. Достаточно подробные обзоры состояния проблемы приведены нами в [10, 11].

Ряд работ посвящено изучению ферм ссылок и борьбе с ними. Например, в работе [8] предлагается анализировать веб-граф для определения ссылочного спама. Проводится анализ входящих и исходящих ссылок сайтов, исследуется их пересечение. Рассматривается влияние ссылочного спама на алгоритм HITS.

В работе [3] проводится статистический анализ автоматически сгенерированных страниц со спамом. Авторы рассматривают отклонения от нормального распределения различных свойств страниц, включая имена сайтов, IP-адреса, входящие и исходящие ссылки, содержание страницы и норму изменения.

В [6] рассматриваются различные характеристики страницы (число слов на странице и в заголовке, длина слов, процент видимого текста и т. д.). Даются сведения о процентном содержании поискового спама в различных доменных зонах. Проводится

сравнение выявленных характеристик с их распределением на «обычных» страницах, что способствует выявлению страниц, содержащих спам.

В работе [1] подробно анализируются ссылочные структуры, образующие веб-граф. Исследуются различные характеристики, способствующие обнаружению ссылочного спама.

В работе [2] делается попытка определять ссылочный спам («непотистский» спам). Для решения задачи используется дерево решений C4.5. Авторы рассматривают 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 [7] рассматриваются две группы свойств, характеризующих ссылочный спам (для его обнаружения) – 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 [4] рассматривается понятие массы спама, меры воздействия спам-ссылок на ранг страницы. Рассматриваются вопросы оценки массы спама. Для определения спама активно используется ссылочная структура веб-графа.

Несмотря на все разнообразие работ, подробного исследования ссылок, размещаемых с использованием ссылочных брокеров, не проводилось. Интерес представляет исследование таких ссылок с точки зрения их динамики и содержания, выявление свойств, способных помочь в борьбе с ними.

Цель нашего исследования – изучить характеристики ссылок, размещаемых с помощью ссылочных брокеров. Интерес представляют в первую очередь динамические характеристики ссылок – как долго присутствуют ссылки на станицах, как часто они заменяются на новые ссылки и т. д. Кроме того, нуждается в исследовании тематическая близость размещаемых ссылочными брокерами ссылок и страниц, на которых они размещаются. Интерес представляет так же возможность использования указанных характеристик для обнаружения ссылочного спама.

3 Источники данных

В качестве объекта исследования были выбраны 10 сайтов, размещающих ссылочный спам с использованием ссылочных брокеров. Сайты ежедневно сканировались в течение 7 месяцев (с 1 июня 2009 г. по 31 января 2010 г.). Общее число сканируемых страниц составило около 5000 (число страниц менялось в связи с изменениями сайтов). На сайтах ежедневно ссылочными брокерами размещалось около 5500 ссылок. Информация о факте размещения и месте расположения ссылок была предоставлена нам владельцами сайтов.

Параллельно с этим сайты ежедневно сканировались на наличие (и изменение) обычных ссылок, размещаемых на сайтах их владельцами.

Сайты состояли из различного количества страниц – от 20 до более 2000 – и имели различную тематику (история, спорт, кино, мультфильмы, знаменитости/актеры, здоровье, музыка, мобильные телефоны, интернет-магазин и бизнес-сайт). В период исследования основные показатели сайтов – тематика, индекс цитируемости и PageRank не изменялись, а число страниц изменялось незначительно. По этой причине влияние этих показателей на размещение ссылок в разные периоды времени можно считать минимальным. Таким образом, процесс размещения ссылок через ссылочных брокеров на исследуемых сайтах, можно считать естественным.

Анализ полученных данных позволил выявить основные характеристики и особенности спам-ссылок, а также показатели, характеризующие ссылки, размещаемые посредством брокеров.

4 Характеристики ссылок

4.1 Ротация спам-ссылок

Для анализа ротации спам-ссылок мы использовали два параметра – общее число спам-ссылок, размещаемых на сайте за период исследования L_7 (7 месяцев), и число спам-ссылок, размещенных в настоящее время L_1.

Коэффициент ротации ссылок (K_r) представляет собой отношение разности значений L_7 и L_1 к значению L_1:

$$K_r = \left(L_7 - L_1\right) / L_1.$$

Коэффициент ротации спам-ссылок за месяц (K_{rm}) можно вычислить, разделив коэффициент K_r на количество месяцев, в течение которых проводились исследования:

$$K_{rm} = K_r / 7.$$

Значения коэффициента ротации приведены в табл. 1.

Как можно заметить, коэффициент ротации спам-ссылок K_r изменяется в диапазоне от 1.09 до 7.56. Среднее значение коэффициента ротации

спам-ссылок составило 2.38. Аналогично, коэффициент ротации спам-ссылок в месяц K_{rm} меняется от 0.16 до 1.08, при среднем значении в 0.34.

Для того чтобы сравнить полученные значения коэффициента ротации для спам-ссылок с аналогичными значениями для обычных (не спам) ссылок, мы проведи анализ каждого из исследуемых сайтов. При этом рассматривались все ссылки, за исключением тех, которые были размещены рекламными брокерами. Результаты оказались достаточно интересными. Коэффициент ротации для обычных ссылок K_{r_n} (вычисляется аналогично K_r) для каждого сайта оказался не более 0.01, а для большинства сайтов – вообще 0. Значение коэффициента ротации за месяц K_{rm_n} получилось не более 0.002. Другими словами, на сайтах практически отсутствует ротация обычных ссылок. Однажды попав на сайт, ссылки так и остаются и не заменяются другими.

В связи с тем, что сайты могут существенно отличаться как по структуре, так и по содержанию, мы решили проверить значение коэффициента ротации ссылок на сайтах с наиболее динамичным контентом. Для этого было выбрано несколько наиболее посещаемых форумов, новостных сайтов и популярных блогов. Анализ показал, что, несмотря на всю динамичность содержания, коэффициент ротации ссылок за месяц K_{rm_n} на этих сайтах не превысил 0.05 (новостные сайты, хранящие новости в архиве ограниченное время).

Таким образом, значение коэффициента ротации может способствовать обнаружение ссылочного спама, размещаемого ссылочными брокерами. При коэффициенте ротации ссылок в 0.1 и более можно считать такие ссылки спамом. Аналогичным образом можно ссылки, размещаемые в тех местах страницы, где коэффициент ротации превышает 0.1, также считать спамом.

4.2 Тематическая близость спам-ссылок и сайта

Анализ тематики ссылок, размещаемых с помощью ссылочных брокеров, дал также интересные результаты.

Тематическая ссылка – ссылка, тематика которой совпадает или близка к тематике страницы, на которой она размещается.

Для определения тематической близости была использована методика, применявшаяся нами в [10]. Среди всего числа спам-ссылок L_1 (5476) количество тематических ссылок T оказалось достаточно небольшим – всего 242. В связи с тем, что распределение тематических ссылок по сайтам сильно отличается, интерес представляет относительный показатель – процент тематических ссылок T_{link}, вычисляемый по формуле $T_{link} = (T / L_1)100\%$. Процент тематических ссылок изменяется в диапазоне от 0.7 до 10.6 % (табл. 2). Среднее значение T_{link} составило 4.4 %. Таким образом, в среднем только одна из 22 ссылок, размещаемых с использованием ссылочных брокеров, имеет тематику, совпадающую или близкую с тематикой сайтов.

Анализ аналогичных показателей для обычных ссылок дал несколько двоякие результаты. Для ряда сайтов процент тематических ссылок T_{link_n} (вычисляется аналогично T_{link}) составил более 80%. Для сайтов, содержащих такие разделы как «Каталог ссылок», «Ссылки», «Наши друзья» и т. д. процент тематических ссылок оказался достаточно небольшим (снижается практически прямо пропорционально количеству ссылок в этих разделах). При большом размере указанного раздела процент тематических ссылок может опускаться ниже 1 %.

Таким образом, несмотря на свою показательность, тематическая близость не может являться средством для выявления спам-ссылок, размещаемых ссылочными брокерами (так как вместе с ними будут найдены все не тематические ссылки).

Тем не менее, указанный показатель может применяться именно для комплексного отсева спам-ссылок, причем в данном случае роль тематической близости ссылок и сайта будет основной.

4.3 Тематическая близость в группе спам-ссылок

Спам-ссылки могут размещаться на странице как по одной, как и группами. Расположение ссылок отличается на различных сайтах (табл. 3). Некоторые сайты не содержат ни одной одиночной ссылки, а большинство групп состоит из 4 – 8 ссылок, другие содержат в основном одиночные ссылки, и лишь иногда группы из двух-трех ссылок. Тем не менее, анализ групп показал интересный результат. Из 1023 групп спам-ссылок, только в 178 группах оказалось по одной тематической ссылке (17.4 % от количества групп ссылок), в 16 группах – по две и более тематических ссылок (1.6 %). Из 443 одиночных ссылок только 29 оказались тематическими, что составляет всего 6.5 % от числа одиночных ссылок.

Таким образом, показатель тематической близость является отличительной чертой ссылочного спама, размещаемого ссылочными брокерами. Ссылки различаются по тематике как между собой (при размещении в группах), так с содержанием страницы, где они расположены. При этом различие в тематике – колоссальное. Практически все ссылки имеют совершенно другую тематику. Приведем пример ссылок, размещаемых в период исследования на странице с биографией известного американского актера (рис. 1).

Как можно заметить, только ссылка «Скачать фильмы бесплатно» имеет хоть какое-то отношение к странице с биографией актера (и фильмы, и актер связаны с кино). Все остальные ссылки не имеют ничего общего со страницей, и к тому же вряд ли будут интересны пользователям. Это является прямым доказательством того, что ссылки, размещаемые посредством ссылочных брокеров, являются именно ссылочным спамом и не предназначены для пользователей.

4.4 Время жизни спам-ссылок

Время жизни ссылки (D_{link}) – это период времени, в течение которого ссылка была размещена на странице (до момента ее удаления). Надо заметить, что некоторые спам-ссылки могут кратковременно исчезать со страниц, а затем вновь появляться на них. В этом случае, ссылка считалась удаленной, если она не появлялась вновь в течение 10 суток с момента исчезновения.

На основании собранной статистики было получено распределение спам-ссылок по времени жизни (количество ссылок, существовавших один, два, три и т. д. дней). На рис. 2 показано распределение времени жизни ссылок за 1 год. Число ссылок, имеющих время жизни больше одного года, продолжает уменьшаться, и к концу второго года сокращается до 1 – 2 штук.

Рассмотрим процентный состав времени жизни ссылок, сгруппированных по месяцам (табл. 4). Как можно заметить, подавляющее число ссылок (более 50 %) существует не более 2 месяцев. Практически 90 % ссылок имеют время жизни не более 6 месяцев.

Таким образом, большинство спам-ссылок имеют достаточно небольшое время жизни. Кроме того, ссылки, размещенные на одной странице (группой), также имеют разное время жизни. Поэтому, можно наблюдать ситуацию, когда, скажем, первая и третья ссылки в группе остаются неизменными, а вторая и четвертая ссылка успевают измениться несколько раз. Такие несбалансированные группы являются явным признаком ссылочного спама, размещаемого с использованием ссылочных брокеров.

Анализ времени жизни обычных ссылок показал обратный результат. Не спам ссылки после размещения продолжают находиться на сайте длительное время (многие годы). По этой причине время жизни обычных ссылок D_{link_n} будет приближаться в возрасту страницы, на которой ссылка размещена. Исследование сайтов подтвердило это. Подавляющее большинство ссылок находились на сайтах начиная с момента начала исследований до их окончания.

Такая особенность времени жизни ссылок позволяет использовать ее для обнаружения спам-ссылок, размещаемых рекламными брокерами. Ссылки со временем жизни менее 6 месяцев можно считать ссылочным спамом.

4.5 Перемещение ссылок по сайту

В связи с тем, что на многих сайтах контент может передвигаться со страницы на страницу (форумы, доски объявлений, блоги, новости и т. д.), мы провели анализ перемещения ссылок по сайту. Под фактом передвижения ссылки по сайту понималось ее удаление с одной страницы и появление в течение 10 дней на другой странице сайта. Таких перемещений для ссылок, размещаемых ссылочными брокерами, за период наблюдения было выявлено всего 3. Это позволяет сделать заключение о том, что спам-ссылки не перемещаются по сайту (даже при перемещении основного контента), а привязаны к конкретным страницам.

Анализ обычных ссылок на сайтах с динамическим контентом (новости, форумы, блоги), показал иную ситуацию. Такие ссылки перемещаются по сайту вместе с основным контентом. Таким образом, если не спам ссылка пропадает с одной страницы, есть высокая вероятность, что она появится на другой странице сайта.

5 Использование динамических характеристик для обнаружения ссылочного спама

Для обнаружения ссылочного спама мы использовали разработанный ранее алгоритм, работающий на основе метода опорных векторов [11]. На основе проведенных исследований мы расширили пространство признаков. Ранее применявшиеся признаки имели статический характер. К этим признакам мы добавили ряд признаков, описывающих динамические характеристики ссылок:

- коэффициент ротации ссылок на сайте;
- коэффициент ротации ссылок на странице;
- время жизни ссылки на странице;
- время жизни ссылок на сайте;
- время жизни исследуемой ссылки;
- показатель перемещения ссылки по сайту.

По аналогии с [11] для исследования работы алгоритма нами использовалась собственная коллекция описанная выше. В связи с тем, что алгоритм использовал сведения о динамических характеристиках ссылок, апробация его на коллекциях By.Web и Narod.Ru, содержащих лишь один срез документов, не проводилась.

Для оценки качества работы алгоритма использовалась следующие метрики [9]:

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

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

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

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

Значения метрик приведены в табл. 5. Если сравнить полученные результаты с данными прошлых лет [9, 11], то можно увидеть существенное улучшение значений всех четырех метрик. Показатель Recall вырос на 0.05 по сравнению с лучшим результатом работы алгоритма со статическими признаками. Существенно снизились и показатели ошибочного отнесения ссылок к спаму (FalseSpam) и не спаму (FalseNotSpam).

Таблица 1. Статистика по размещению спам-ссылок

Сайт	**Страниц *P***	**Ссылок за 7 месяцев** L_7	**Ссылок сейчас** L_1	**Коэффициент ротации** K_r	**Коэффициент ротации в месяц** K_{rm}
Сайт об истории	223	3162	1030	2.07	0.30
Сайт о мультфильмах	22	327	144	1.27	0.18
Сайт об актере	58	780	270	1.89	0.27
Сайт о спорте	110	3474	843	3.12	0.45
Сайт о здоровье	163	2252	1077	1.09	0.16
Бизнес сайт	86	1552	393	2.95	0.42
Сайт о музыке	169	1980	775	1.55	0.22
Сайт о телефонах	1322	1289	458	1.81	0.26
Сайт о кино	2316	3201	374	7.56	1.08
Интернет магазин	423	496	112	3.43	0.49
Всего	**4892**	**18513**	**5476**	**2.38**	**0.34**

Таблица 2. Количество тематических ссылок

Сайт	**Ссылок** L_1	**Число тематических ссылок** T	**% тематических ссылок** T_{link}
Сайт об истории	1030	7	0.7
Сайт о мультфильмах	144	7	4.8
Сайт об актере	270	14	5.2
Сайт о спорте	843	33	3.9
Сайт о здоровье	1077	82	7.6
Бизнес сайт	393	42	10.6
Сайт о музыке	775	4	0.5
Сайт о телефонах	458	20	4.3
Сайт о кино	374	24	6.4
Интернет магазин	112	9	8.0
Всего	**5476**	**242**	**4.4**

Таблица 3. Группы ссылок

Сайт	**Страниц *P***	**Одиночных ссылок**	**Одиночных тематических ссылок**	**Групп ссылок**	**Групп с 1 тематической ссылкой**	**Групп с 2 и более тематическими ссылками**
Сайт об истории	223	1	0	222	7	0
Сайт о мультфильмах	22	0	0	22	7	0
Сайт об актере	58	0	0	56	12	1
Сайт о спорте	110	0	0	110	29	2
Сайт о здоровье	163	0	0	163	69	6
Бизнес сайт	86	2	0	84	28	6
Сайт о музыке	169	0	0	169	4	0
Сайт о телефонах	1322	271	11	102	9	0
Сайт о кино	2316	120	12	73	10	1
Интернет магазин	423	49	6	22	3	0
Всего	**4892**	**443**	**29**	**1023**	**178**	**16**

- *аренда погрузчика от фирмы*
- *Оптимизация сайта seo поисковое продвижение сайтов сайт seo-studio.*
- *Триал спорт теннисный стол спорт инвентарь маты.*
- *новый коттедж готовые коттеджи*
- ***Скачать фильмы бесплатно***
- *окна от производителя*
- *Метизы усовершенствованные. Метизы фильтры. Метизы классные. метизы.*
- *купить грунт*
- *Курсы менеджеров, курсы pr менеджеров.*
- *Костюм деда мороза и снегурочки. Заказывать Деда Мороза и Снегурочку.*
- *Wmz sms обменяй с гарантией. Wmr wme обмен дорого.*
- *tehsklad.ru предлагает пилы Makita*
- *Массовая рассылка смс от 1157. Рассылка смс от 1054.*
- *Банки переводов денег. Перевод денег с карты на карту лимит суммы альфа банк.*
- *Автомобили Тула, продажа авто Тула. Продажа б/у авто в городе Тула.*
- *Iso 9000, iso 9001 2008. Международного стандарта iso 9001 2008.*
- *интернет магазин часов копии.*

Рис. 1. Пример ссылок, размещенных ссылочными брокерами на странице с биографией известного американского актера

Таблица 4. Распределение ссылок по времени жизни (месяцев)

Период	Процент ссылок, %
1 месяц	30.619
2 месяца	20.283
3 месяца	17.420
4 месяца	8.532
5 месяцев	7.349
6 месяцев	4.254
7 месяцев	3.252
8 месяцев	2.458
9 месяцев	1.746
10 месяцев	1.291
11 месяцев	0.786
12 месяцев	0.397
13 месяцев	0.546
14 месяцев	0.215
15 месяцев	0.223
16 месяцев	0.207
17 месяцев	0.232
18 месяцев	0.066
19 месяцев	0.074
20 месяцев	0.050

Таблица 5. Результаты работы

Метрика	Значение
Precision	0.96
Recall	0.92
FalseSpam	0.09
FalseNotSpam	0.08

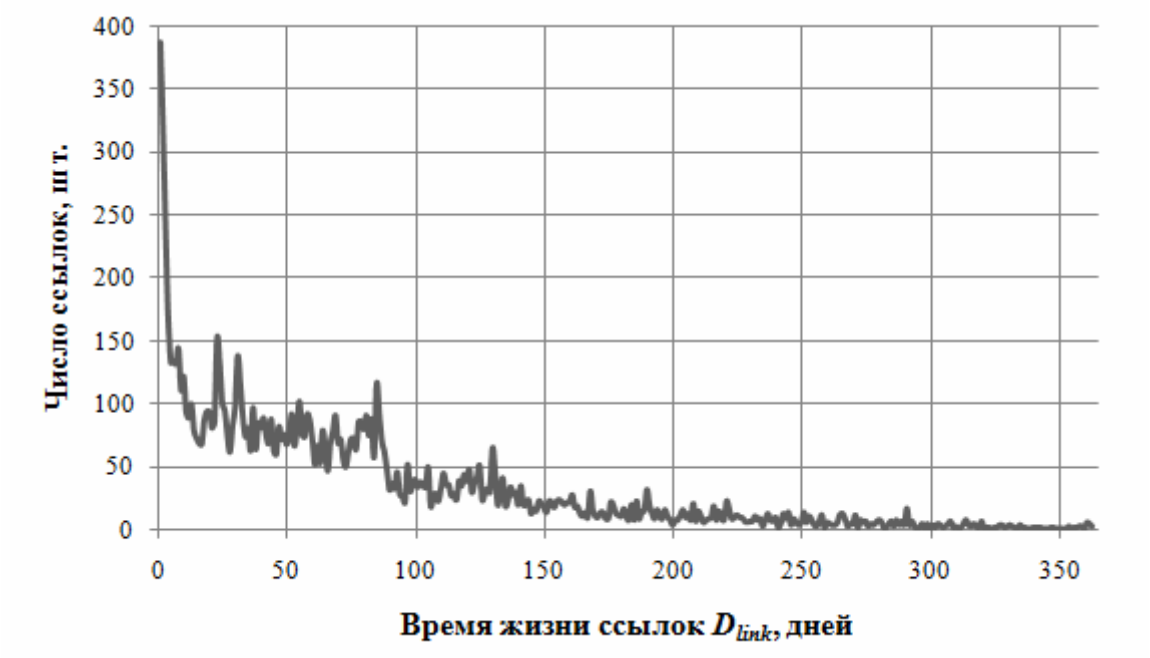

Рис. 2. График распределения времени жизни спам-ссылок

6 Выводы

Таким образом, анализ ссылок, размещаемых с использованием ссылочных брокеров, показал, что они действительно предназначены для спама и не несут полезной информации для пользователей.Спам ссылки часто заменяются одна на другую, т. е. имеют высокий коэффициент ротации. При значении коэффициента ротации более 0.1 ссылки можно считать поисковым спамом.

Спам-ссылки слабо соответствуют тематике страниц, на которых они расположены. В среднем менее 5 % ссылок, размещенных ссылочными брокерами, тематически близки к страницам, на которых они находятся.

Практика показала достаточно невысокое время жизни таких ссылок, размещаемых ссылочными брокерами. Более 90% таких ссылок живут не более 6 месяцев.

Спам-ссылки не перемещаются по сайту вместе с основным контентом. Они жестко привязаны к конкретной странице.

Таким образом, использование динамических характеристик позволяет существенно повысить качество обнаружения ссылочного спама. Можно достичь полноты обнаружения ссылочного спама в 0.92 и точности в 0.96.

Литература

[1] Becchetti L., Castillo C., Donato D., Leonardi S., Baeza-Yates R. Link analysis for web spam detection// ACM Trans. Web 2. – 2008. – V. 1. – P. 1-42.

[2] Davison B.D. Recognizing nepotistic links on the web//AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, 2000. – P. 23-28.

[3] Fetterly D., Manasse M., Najork M. Spam, damn spam, and statistics – using statistical analysis to locate spam web pages//Proc. the 7th Int. Workshop on the Web and Databases (WebDB), Paris, France, 2004.

[4] Gyongyi Z., Berkhin P., Garcia-Molina H., Pedersen J. Link spam detection based on mass estimation//32nd Int. Conf. on Very Large Data Bases (VLDB 2006), September 12 – 15, 2006, Seoul, Korea.

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

[6] Ntoulas A., Najork M., Manasse M., Fetterly D. Detecting spam web pages through content analysis// Proc. of the 15th Int. World Wide Web Conference, Edinburgh, Scotland, May 2006. – P. 83-92.

[7] Gan Q., Suel T. Improving web spam classifiers using link structure// Proc. in Third Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb ’07), May 2007, Banff, Alberta, Canada.

[8] Wu B., Davison B. D. Identifying link farm pages// Proc. of the 14th Int. World Wide Web Conference (WWW), 2005.

[9] Шарапов Р.В., Шарапова Е.В. Обнаружение ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всерос. науч. конф. RCDL'2008, Дубна, Россия, 7 – 11 октября 2008 г. – Дубна: ОИЯИ, 2008. – С. 191-196.

[10] Шарапов Р.В., Шарапова Е.В. Алгоритм обнаружения ссылочного спама // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной межд. конф. «Диалог 2009» (Бекасово, 27 – 31 мая 2009 г). – М: РГГУ, 2009. – Вып. 8 (15). – С. 537-542.

[11] Шарапов Р.В., Шарапова Е.В. Применение метода опорных векторов для обнаружения ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всерос. науч. конф. RCDL'2009, Петрозаводск, Россия, 17 – 21 сентября 2009 г. – Петрозаводск: КарНЦ, 2009. – С. 318-324.

Analysis of dynamic characteristics of web spam placed by link brokers

R.V. Sharapov, E.V. Sharapova

We examine the characteristics of Web spam placed by Link Brokers. We investigate the link lifetime and rotation, analyze the thematic proximity of links and pages. There is given the analysis of links in groups of links. The possibility of dynamic characteristics using for the detection of link spam is considered.