Исследование поискового спама, размещаемого посредством ссылочных брокеров

Research of web spam placed by link brokers

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

Муромский институт (филиал)
Владимирского государственного университета

В статье рассматриваются характеристики ссылочного спама, размещаемого ссылочными брокерами. Исследуется время жизни и ротация ссылок, анализируется тематическая близость ссылок и страниц.

Введение

Последнее десятилетие ознаменовалось бурным развитием глобальной сети Интернет и поисковых систем, позволяющих искать информацию в ней. Стремясь повысить качество поиска, поисковые системы стали использовать дополнительные сведения о документах, в том числе ссылки на них.

Ссылки используются 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 наиболее нужную и полезную информацию по поисковым запросам.

В то же время, использование поисковыми системами ссылок привело к возникновению нового вида поискового спама, получившего название ссылочный спам [5]. 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.

Ссылочный спам проявляется в размещении большого числа 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.

Еще большей проблемой являются системы пакетной покупки ссылок через ссылочных (рекламных) брокеров (таких как MainLink.ru, Xap.ru, Sape.ru, LinkFeed.ru, SetLinks.ru, Clx.ru и т. д.). Такие системы могут размещать ссылки на сотнях миллионов страниц. Массовое появление ссылок, размещаемых ссылочными брокерами, может оказать существенное влияние на алгоритмы поисковых систем [9]. Несмотря на то, что такие ссылки позиционируются как рекламные, считать их таковыми нельзя. Ссылки размещаются в неприметных местах страницы, чаще всего в самом ее низу, отображаются мелким шрифтом. Таким образом, функцию рекламы они выполнять не могут, так как пользователи такие ссылки просто не замечают. Следовательно, функцией ссылок, размещаемых ссылочными брокерами, является именно ссылочный спам.

1. 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] предла-

гается анализировать вэб-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 вэб-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 вэб-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. В первую очередь нас интересовали динамические характеристики ссылок — как долго присутствуют ссылки на станицах, как часто они заменяются на новые ссылки и т. д. Кроме того, исследовалась тематическая близость ссылок, размещаемых ссылочными брокерами, и страниц, на которых они размещаются.

2. Источники данных

В качестве объекта исследования были выбраны 10 сайтов, размещающих ссылочный спам с использованием ссылочных брокеров. Сайты ежедневно сканировались в течение 7 месяцев (начиная с июня 2009 г.). Общее число сканируемых страниц составило около 5000 (число страниц менялось в связи с изменениями сайтов). На сайтах было размещено ежедневно около 5500 ссылок. Под «ссылкам» в данной работе мы будем понимать исключительно спам-ссылки, размещаемые с использованием ссылочных брокеров, исключая из рассмотрения естественные ссылки, также присутствующие на сайтах. Информация о факте размещения и месте расположения ссылок были предоставлены владельцами сайтов.

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

3. Ротация спам-ссылок

Коэффициент ротации ссылок (K_r) представляет собой отношение общего числа спам-ссылок за период исследования L_7 (7 месяцев) к числу ссылок, размещенных в настоящее время L_1:

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	3,07	0,44
Сайт о мультфильмах	22	327	144	2,27	0,32
Сайт об актере	58	780	270	2,89	0,41
Сайт о спорте	110	3474	843	4,12	0,59
Сайт о здоровье	163	2252	1077	2,09	0,30
Бизнес сайт	86	1552	393	3,95	0,56
Сайт о музыке	169	1980	775	2,55	0,36
Сайт о телефонах	1322	1289	458	2,81	0,40
Сайт о кино	2316	3201	374	8,56	1,22
Интернет магазин	423	496	112	4,43	0,63
Всего	**4892**	**18513**	**5476**	**3,38**	**0,48**

$$K_r = L_7 / L_1$$

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

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

Как можно заметить (таблица 1), коэффициент ротации спам-ссылок K_r изменяется в диапазоне от 2,09 до 8,56. Это означает, что за семь месяцев ссылки меняются от 2 до 8 раз. Среднее значение коэффициента ротации спам-ссылок составило 3,38. Аналогично, коэффициент ротации спам-ссылок в месяц K_{rm} меняется от 0,30 до 1,22, при среднем значении в 0,48.

4. 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 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:

$$T_{link} = T / L1 \times 100\ \%$$

Процент тематических ссылок изменяется в диапазоне от 0,7 до 10,6 %. 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.

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**

5. Тематическая близость в группе спам-ссылок

Спам-ссылки могут размещаться на странице как по одной, как и группами. Расположение ссылок отличается на различных сайтах. 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 ссылок, только в 178 группах оказалось по одной тематической ссылке (17,4 % от количества групп ссы-

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**

лок), 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 (таблица 3).

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:

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

Как можно заметить, только ссылка «Скачать фильмы бесплатно» (5) 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.

6. 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 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.

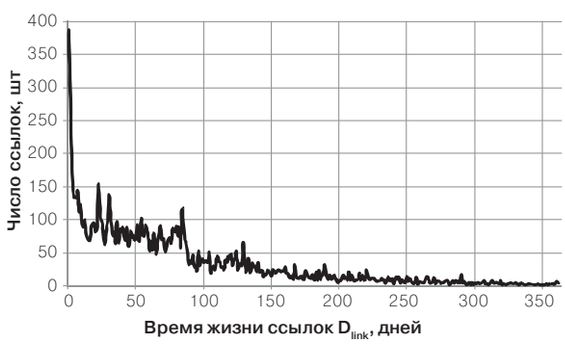

Рис. 1. График распределения времени жизни ссылок

На основании собранной статистики было получено распределение ссылок по времени жизни (количество ссылок, существовавших один, два, три и т. д. дней). На рисунке 1 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.

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

Рассмотрим процентный состав времени жизни ссылок, сгруппированных по месяцам. 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 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.

7. Выводы

Таким образом, анализ ссылок, размещаемых с использованием ссылочных брокеров, показал, что они действительно предназначены для спама и не несут полезной информации для пользователей. Кроме того, практика показала достаточно невысокое время жизни таких ссылок. По этой причине ссылки со временем жизни более 6 месяцев, скорее всего, будут предназначены для пользователей, а не для поисковых систем.

Выявленные характеристики ссылок могут использоваться как исходный материал в алгоритмах обнаружения ссылочного спама и нейтрализации его действия на поисковые системы.

Мы планируем использовать полученные сведения в разрабатываемых нами алгоритмах обнаружения ссылочного спама [10, 11]. Выявленные характеристики ссылочного спама, размещаемого с использованием ссылочных брокеров, позволят расширить число параметров алгоритмов, что будет способствовать более точному определению спам-ссылок.

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. *Davison B. D.* Recognizing nepotistic links on the web. In AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, July 30 2000, p. 23–28.
3. *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.
4. *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.
5. *Gyöngyi Z., Garcia-Molina H.* Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005, Chiba, Japan.
6. *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, p. 83–92.
7. *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
8. *Wu B., Davison B. D.* Identifying link farm pages. In Proceedings of the 14th International 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 Всероссийской научной конференции «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 международной конференции «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 Всероссийской научной конференции «RCDL'2009» (Петрозаводск, Россия, 17–21 сентября 2009 г.) Петрозаводск: КарНЦ, 2009. С. 318–324.