WELCOME, CREATIVE!

Конференция WWW 2009 в Мадриде

Использование методов категоризации текстовых привязок и анализа графов для идентификации платных ссылок.

Стендовый доклад 22 апреля 2009 года

АНОНС

Один из способов искусственного повышения рейтингов коммерческих страниц в индексах информационно-поисковых систем — размещение платных ссылок на эти страницы на других веб-сайтах. Умение выявлять платные ссылки повышает эффективность поисковой машины. В этой статье описывается новый метод идентификации платных ссылок. Он предусматривает, во-первых, обучение классификатора текстовых привязок и анализ исходящих коммерческих ссылок с различных веб-страниц; и во-вторых, анализ графа ссылок Рунета на основе полученных данных для выявления платных ссылок и сайтов, их продающих и покупающих. Проверка алгоритма на сформированных вручную тестовых выборках доказала его высокую эффективность.

Категории и тематики

H.3.3 [Поиск и обнаружение информации]: Фильтрация данных.

Основные термины

Алгоритмы, Разработка, Эксперимент.

Ключевые слова

Поисковые машины, модель языка, категоризация, анализ ссылок, машинное обучение, поиск данных в Интернете.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

WELCOME, CREATIVE!

Конференция WWW 2009 в Мадриде

1. ВВЕДЕНИЕ

По нашим наблюдениям, основной метод оптимизации сайтов для поисковых машин (SEO), принятый в Рунете, — использование платных ссылок. Платные ссылки действительно влияют на рейтинги в индексах поисковиков, но не считаются спамом, поскольку встречаются в основном на вполне приличных страницах наряду с другими полезными ссылками и часто указывают на ценные коммерческие сайты. Платные ссылки, как правило, стоят больших денег, поэтому к их подготовке подходят особенно тщательно и непременно включают в текст привязки термины, соответствующие популярным коммерческим запросам с использованием актуальных для данного сайта ключевых слов. Создание платных ссылок вручную — сложная и кропотливая работа, поэтому неудивительно, что в них действительно содержатся актуальные сведения о целевом сайте. Тем не менее, умение выявлять платные ссылки значительно повышает эффективность рейтингов поисковых машин.

Этот процесс осуществляется в два этапа. Сначала проводится анализ текста и классификация тематик, а затем формируется стартовое множество страниц различной тематики и строится граф ссылок с использованием модифицированного алгоритма HITS [1], где «посредники» — это сайты, продающие ссылки, а «лидеры» — сайты, покупающие ссылки. Главная задача алгоритма — выявление непосредственно платных ссылок, а не сайтов, их продающих и покупающих.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

2. АЛГОРИТМ

2.1 Классификатор текстов с точки зрения поисковой оптимизации

Параметр, определяющий коммерческую привлекательность текстового фрагмента, мы назвали показателем оптимизации текста. Мы взяли стартовое множество оптимизированных запросов с одного популярного SEO-сайта и создали на его основе простой классификатор текстов с точки зрения поисковой оптимизации, подобный описанному в статье [2], где использовались только две разновидности тематических запросов — оптимизированные и не оптимизированные. Воспользовавшись методом итерации, аналогичным тому, что описан в пункте 2.2, мы получили длинный список текстовых юниграмм (300 000) и биграмм (1 500 000), типичных для текстовых привязок на оптимизированных сайтах. Затем мы воспользовались пулом новостных текстов для генерации естественных текстовых юниграмм и биграмм и использовали полученные данные для создания улучшенного байесовского классификатора текстов с точки зрения поисковой оптимизации.

2.2 Классификатор тематик с точки зрения поисковой оптимизации

Чтобы создать алгоритм идентификации оптимизированных тематических запросов, мы отобрали 22 тематики, наиболее характерные для коммерческих сайтов (недвижимость, финансы, грузоперевозки и т. д.). Алгоритм идентификации тематических запросов состоит из двух частей. Для начала мы составили стартовое множество из 3350 монотематических ключевых слов, отобранных вручную.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

WELCOME, CREATIVE!

Конференция WWW 2009 в Мадриде

У каждого ключевого слова имеется собственный тематический спектр (ТС). Затем, используя тексты привязок с показателем оптимизации выше нуля, мы вычислили ТС для всех остальных ключевых слов, исходя из степени вероятности совпадения с другими ключевыми словами из стартового множества этого же текста привязки. Таким путем мы получили 64 000 ТС, которые затем использовали для категоризации текстовых привязок по принципу, который аналогичен описанному в статье [2].

На следующем этапе мы воспользовались упрощенным хост-графом, в котором насчитывается 20 миллионов ребер, содержащих текстовые привязки с показателем оптимизации выше нуля. С помощью вышеупомянутого алгоритма мы определили по две самых вероятных тематики для каждого ребра. Затем мы рассчитали ТС для целевых вершин на основе входящих ребер, чтобы сузить спектр для большинства целевых сайтов. Полученные тематики мы распространили на текстовые привязки всех входящих ссылок и исходя из этого составили новый словарь, насчитывающий около 200 000 слов и 800 000 словосочетаний. Такое обилие терминов позволило нам создать новый, более эффективный классификатор тематик на основе цепи Маркова первого порядка [3].

После этого словарь был вручную скорректирован с учетом грубых ошибок. Таким образом, составление словаря подобных объемов почти не требует человеческого вмешательства. Фактически, мы создали его автоматически, опираясь на работу, проделанную до нас SEO-оптимизаторами.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

2.3 Классификаторы входящих и выходящих ссылок с точки зрения поисковой оптимизации

Для дальнейшего анализа мы воспользовались алгоритмом типа BHITS [4]. Ранее алгоритм HITS и его различные модификации уже использовались для выявления спама [5] [6], а теперь мы решили использовать его для обнаружения платных ссылок. Мы воспользовались двусторонним графом ссылок (страницы-источники слева, целевые хосты справа), убрав из него все известные спам-страницы, ссылки с линкферм и т. п. Мы усовершенствовали стандартный механизм подготовки ссылок HITS и удалили все ссылки, принадлежащие одному владельцу (владелец — это домен второго уровня, не являющийся хостом, или домен третьего уровня, расположенный на сервере хоста). Таким образом мы получили граф ссылок, насчитывающий 300 миллионов ребер, 50 миллионов страниц-источников и 19 миллионов целевых сайтов. Проанализировав ребра графа с помощью классификатора тематик (пункт 2.2) мы получили 1 миллион монотематических целевых сайтов.

В нашем алгоритме используются понятия оптимизированных входящих и исходящих ссылок, аналогичные соответственно «посредникам» и «лидерам» в классическом алгоритме HITS. Показатель оптимизации исходящих ссылок определяет вероятность того, является ли сайт продавцом ссылок. Показатель оптимизации входящих ссылок определяет вероятность того, продвигается ли сайт с помощью платных ссылок. Сайты с высокими показателями оптимизации входящих ссылок — это коммерческие ресурсы, использующие дорогостоящие средства SEO для повышения своих рейтингов в результатах запросов поисковых систем.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

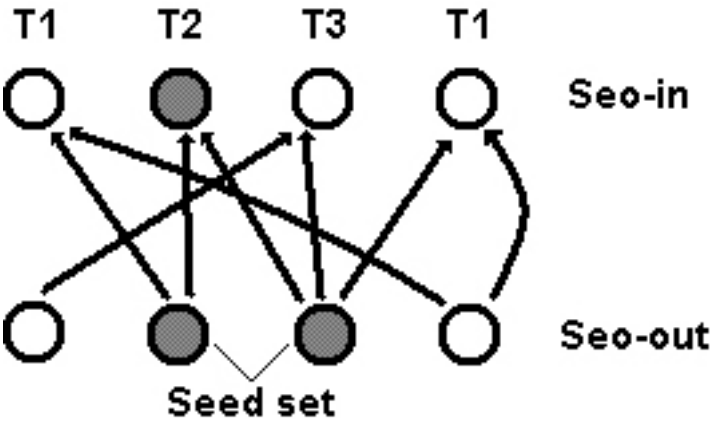
Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

WELCOME, CREATIVE!

Конференция WWW 2009 в Мадриде

Страница, указывающая на целевые сайты различной тематики, вероятнее всего является продавцом ссылок. В качестве стартового множества мы отобрали 3 миллиона мультитематических страниц, которые имеют высокие показатели оптимизации исходящих ссылок и соответствуют ряду других параметров. Показатели оптимизации входящих и исходящих ссылок рассчитываются по стандартному алгоритму HITS (рис. 1) на основе двух итераций. На данном этапе нашей задачей было получить список целевых сайтов с высокими показателями оптимизации входящих ссылок. В результате в полученном списке насчитывается около 500 000 таких сайтов.



Seo-in — показатель оптимизации входящих ссылок; Seo-out — показатель оптимизации исходящих ссылок; Seed set — стартовое множество.

Рисунок 1. Вычисление показателей оптимизации входящих ссылок на основе показателей оптимизации исходящих ссылок стартового множества мультитематических страниц по двухчастному графу ссылок с использованием алгоритма HITS (показана одна итерация, T1, T2, T3 — тематики целевых сайтов).

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

WELCOME, CREATIVE!

Конференция WWW 2009 в Мадриде

2.4 Классификатор ссылок с точки зрения поисковой оптимизации

Показатель оптимизации ссылки определяет вероятность того, является ли она оплачиваемой. Ниже описан простой алгоритм, позволяющий определить этот показатель в один проход. Для начала нужно определить вероятность того, что на странице содержатся платные ссылки (SEO-out), просуммировав следующие показатели: средний показатель оптимизации входящих ссылок целевого сайта (AvgSEOin), средний показатель оптимизации текстовых привязок (AvgSEOtext), количество целевых тематик (Nth) и некоторые другие параметры страницы по следующей формуле:

$$SEOout = k1 \times AvgSEOin + k2 \times AvgSEOtext + k3 \times NTh + \ldots \quad (1)$$

Затем на основе этих данных (показатель оптимизации текстовых привязок, показатель оптимизации исходящих ссылок страницы-источника, показатель оптимизации входящих ссылок целевого сайта и некоторые другие параметры ссылки) вычисляется показатель оптимизации интересующей нас ссылки:

$$SEOlink = l1 \times SEOtext + l2 \times SEOin + l3 \times SEOout + \ldots \quad (2)$$

Показатели ki и li были получены на основе обучающей выборки из 2500 ссылок, отобранных вручную, и около 10 000 ссылок, взятых из Википедии и с сайтов, являющихся продавцами ссылок.

Эти вычисления отнимают совсем не много времени и ресурсов и могут быть выполнены с помощью любой программы обработки баз ссылок.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

WELCOME, CREATIVE!

Конференция WWW 2009 в Мадриде

3. РЕЗУЛЬТАТЫ

Чтобы определить точность и эффективность описанных алгоритмов, мы воспользовались тестовыми выборками, составленными при участии восьми экспертов. Для оценки алгоритма категоризации мы отобрали 2200 сайтов по выбранным тематикам (по 100 самых популярных сайтов на каждую), опираясь на данные популярного среди SEO-специалистов сайта рейтингов, и сформировали список входящих текстовых привязок с показателями оптимизации выше нуля методом случайной выборки. Если принадлежность текстовой привязки к той или иной тематике была очевидна, эксперты приписывали ее к одной из 22 тематик. Часть тестовой выборки (12 100 привязок) была использована для проверки и корректировки алгоритмов. Другая часть (3 800 привязок) — для оценки эффективности. В результате было установлено, что точность описанных алгоритмов достигает 94 %, а эффективность — 97 %.

Для проверки алгоритма идентификации платных ссылок мы использовали две тестовых выборки (табл. 1). Первая включает в себя около 1700 полезных естественных ссылок и 1850 платных ссылок, отобранных вручную методом случайной выборки (точность алгоритма оценивалась только по естественным ссылкам). Мы смогли идентифицировать ссылки, принадлежащие одному сервису обмена ссылками, и получили таким образом подборку платных ссылок, которую использовали в качестве второго тестового образца.

Из 300 миллионов ссылок, присутствующих в нашем графе, алгоритм идентифицировал как платные 50 миллионов ссылок (17 %).

Таблица 1. Результаты идентификации платных ссылок.

Тестовая выборка	Точность	Эффективность
1. 3550 ссылок	95 %	93 %
2. 140 000 ссылок	-	96%

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru

4. ЗАКЛЮЧЕНИЕ

Идентификация платных ссылок позволяет оценивать релевантность ссылок для коммерческих и некоммерческих запросов независимо друг от друга. В первом случае, платные ссылки учитываются при вычислении коммерческого рейтинга, а во втором — игнорируются. Это делает формулу вычисления рейтингов более эффективной и повышает качество поиска, нейтрализуя влияние чрезмерной оптимизации на некоммерческие поисковые запросы и делая результаты поиска более разнообразными.

Данный алгоритм может быть усовершенствован за счет применения аналогов microHITS к блокам ссылок в рамках механизма сегментации страниц Яндекса [7].

5. БЛАГОДАРНОСТИ

Хотим поблагодарить Сергея Певцова, Илью Сегаловича, Аркадия Борковского и Сергея Волкова за полезные замечания по данному вопросу.

6. ИСТОЧНИКИ

[1] Kleinberg, J. (1997). Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5): 604–632.
[2] T. H. Haveliwala. Topic-sensitive pagerank. In Proc. 11th International WWW Conference, pages 517-526, 2002.
[3] Lafferty J., Zhai, C. Document language models, query models, and risk minimization for IR. In Proceedings of SIGIR-2001, pp 111-119.
[4] K. Bharat and M.R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, Proc. 21St Annual International ACM SIGIR, pp.104–111, 1998.
[5] B. Wu and B. Davison. Undue influence: Eliminating the impact of link plagiarism on web search rankings. Technical report, LeHigh University, 2005.
[6] Yasuhito Asano, Yu Tezuka, Takao Nishizeki. Improvement of HITS algorithms for spam links. APWeb/WAIM 2007, LNCS 4505, pp 479-490, 2007.
[7] S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction. ACM 1-58113-348-0/01/0005, 2001.

Авторы:

Кирилл Николаев
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
kvn@yandex-team.ru

Екатерина Зудина
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
zudina@yandex-team.ru

Андрей Горшков
Яндекс
Москва, ул. Самокатная, д.1
7-495-739-70-00
gorshkov@yandex-team.ru