Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Факультет информационных технологий и программирования
Кафедра «Компьютерные Технологии»

О. С. Ларионов

Обоснование темы бакалаврской работы «Определение сайтов-зеркал на основе анализа графа ссылок и дублирования контента»

Научный руководитель: ведущий программист Mail.Ru Group А. Романенко

Санкт-Петербург
2013

1 Актуальность

Интернет представляет собой огромную коллекцию информации. Объем "Машины времени интернета" превысил 10 петабайт, количество сайтов в интернете, по данным аналитических компаний, достигло 620,5 млн. что-нибудь про размер индекса Однако, не все они являются уникальными. Часть из них являются клонами, или зеркалами, друг друга. Знание, что один сайт является зеркалом другого, крайне полезно. Например, в поисковых системах оно может упростить индексирование сайтов (не просматривать по нескольку раз похожие страницы), улучшить выдачу по запросу (предлагать зеркала, если основной сайт недоступен; показывать только основной сайт и скрывать зеркала), помогать производить языковой анализ (если сайты-зеркала отличаются только языком). Существующие сейчас алгоритмы в основном являются закрытыми и/или запатентованными. Однако, и они не могут со стопроцентной точностью определить, являются ли сайты зеркалами.

2 Цель

Целью данной работы создание алгоритма для более точного определения сайтов-зеркал.

3 Задачи

1. Изучить существующие алгоритмы поиска сайтов-зеркал.
2. Определить свойства сайтов-зеркал.
3. Разработать метод обнаружения зеркал.
4. Реализовать, проверить и усовершенствовать метод, применив его для решения практических задач.

4 Обзор предметной области

4.1 Определение

Сперва уточним определение понятие "зеркало". Общепринятым является, что сайт B является зеркалом сайта A, если существует достаточно большое множество страниц на сайте A, что для любой страницы из этого множества существует *очень похожая* страница на сайте B. Степень похожести — весьма субъективная оценка, и устанавливается различными способами, например, выбором ключевых слов со страниц и их сравнение.

Можно заметить, что "зеркало" — это отношение эквивалентности. Действительно:

1. A является зеркалом самого себя.
2. если A является зеркалом B, то и B — зеркало A.
3. если A зеркало B, и B зеркало C, то A зеркало C.

4.2 Проблемы

При поиске зеркал мы сталкиваемся с несколькими проблемами:

- чаще всего доступен только список с URL-адресами страниц, полученный "веб-пауком сервером или каким-то другим образом;
- часть страниц может измениться, устареть или быть недоступна;

- получить/хранить страницы (сайты) целиком может оказаться затруднительно (например, не хватает оперативной памяти).
- ...

4.3 Известные решения

Большинство алгоритмов поиска зеркал действуют следующим образом:

1. получают на вход список страниц (URL-адресов) с различных сайтов;
2. оставляют в этом списке только сайты, в которых доступно достаточное количество страниц, например, более ста;
3. запрашивают различную информацию:
   - IP-адрес сервера;
   - миниатюру страницы;
   - список ссылок с этой страницы на другие;
   - ...
4. составляют пары сайтов-кандидатов на зеркала;
5. производят анализ пар на основе имеющейся информации.

Для анализа пар применяются следующие методы:

- Сравнение IP-адресов. Например, для $IPv4$, можно сравнивать целиком IP-адреса, либо какие-то части, например, первые три октета.
- Сравнение URL-адресов. Например, можно дописывать к префиксам ссылок различные суффиксы и сравнивать страницы, на которые ведут получившиеся ссылки.
- Сравнение списка ссылок с сайтов на другие. То есть можно оценивать, ведут ли ссылки с проверяемых сайтов друг на друга или на одни и те же сайты.
- Сравнение карты сайтов.
- Составление правил преобразования ссылок одного сайта в ссылки другого.
- ...

Далее, на основе результатов этого анализа, алгоритмы делают различные выводы. Например, если для каждого суффикса P на сайте A (http://A/P) есть похожий документ на сайте B (http://B/P) то сайты с некоторой вероятностью являются зеркалами. Проверка всех условий может занять длительное время, особенно если список сайтов велик. Поэтому стоит проверять не все, а выделить или придумать наиболее значимые. Наша цель состоит в том, чтобы предложить достаточные условия, после проверки которых можно было бы с высокой точностью ответить, являются ли сайты клонами.

5 Планируемые результаты

В ходе разработки предполагается реализовать алгоритм, который для заданного списка сайтов будет с высокой точностью и скоростью определять, какие сайты являются зеркалами друг друга.

Список литературы

[1] K. Bharat, A. Broder, *Mirror, mirror on the web: A study of host pairs with replicated content*, Computer Networks, 1999.

[2] K. Bharat, A. Broder, J. Dean, MR. Henzinger, *A comparison of techniques to find mirrored hosts on the WWW*, Journal of the American Society for Information Science, 2000.