Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

Факультет информационных технологий и программирования

Кафедра «Компьютерные технологии»

О.С. Ларионов

Определение сайтов-зеркал на основе анализа графа ссылок и дублирования контента

Бакалаврская работа

Научный руководитель: А.А. Романенко

Санкт-Петербург

2013

# Введение

В последнее время наблюдается значительный рост интернета. Уже сейчас он хранит в себе огромную коллекцию информации. Например, об этом свидетельствует объем «Машины времени интернета» [1]: более 10 петабайт. Количество сайтов в интернете, по данным аналитических компаний, превысило 620.5 млн. Однако, по данным исследователей, значительная их часть (30-45%) [2] являются дубликатами. Казалось бы, зачем создавать сайты с одним и тем же содержимым? Но у дубликатов есть как положительные, так и отрицательные стороны. С одной стороны, это поможет в случае, если доступ к одному хосту будет невозможен, например, по причине поломки оборудования. С другой стороны, дубликаты представляют проблему для поисковых систем: если интернет на треть заполнен идентичными сайтами, то индекс систем как минимум на треть содержит избыточную информацию.

В настоящей работе разработан алгоритм, позволяющий определить, являются ли сайты зеркалами, а так же вычислить, какой из предоставленных признаков схожести страниц является наиболее значимым.

# Глава 1. Основные понятия и обзор существующих методов

## Основные определения

Хотя понятие «Сайты-зеркала» доступно на интуитивном уровне, его весьма трудно формализовать. Зеркала бывают, как полностью идентичные, отличающиеся только хостом, на котором расположен сайт, так и структурно-идентичные, то есть такие, у которых адреса страниц совпадают (возможно, с небольшими изменениями, например, различными хостами или портами), а содержимое страниц сильно совпадает. «Сильно совпадает» значит, что контент отличается, например, только в верстке или дизайне. Если бы интернет не был такого размера, то обнаружить идентичные сайты было бы достаточно просто. Достаточно для каждого сайта создать документ, в котором сохранить все пути и страницы сайта. После этого несложно среди этих документов найти совпадающие. Однако, интернет слишком велик.

При обнаружении дубликатов мы встречаемся со следующими проблемами:

* Чаще всего доступен только список URL-адресов страниц, полученный веб-пауком, либо созданный прокси-сервером.
* Часть страниц из него может быть удалена, недоступна, 301 moved permanently, и так далее.
* Невозможно скачать и сохранить все страницы в памяти.

Это сильно усложняет задачу.

## Существующие алгоритмы

Значительная часть существующих алгоритмов являются вероятностными и действуют по приблизительно такому алгоритму

* Получают на вход список страниц (URL-адресов) с различных сайтов.
* Прореживают этот список по какому-либо принципу. Например, оставляют не более пятидесяти сайтов, либо сайты с количеством страниц, не меньшим ста.
* Запрашивают об этих сайтах различную информацию:
  + IP-адрес хоста
  + Миниатюру страницы
  + Список ссылок со страницы
  + …
* Составляют пары сайтов — кандидатов на зеркала
* Производят анализ пар на основе существующей информации

# Список используемых источников

1. Wayback Machine: Internet Archive.[http://web.archive.org/‎](http://web.archive.org)
2. *N. Shivakumar and H. Garcia-Molina.* Finding near replicas of documents on the web. **Proceedings of Workshop on Web Databases (WebDB’98)**, March 1998.