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

Оглавление

[Введение 3](#_Toc357283824)

[Глава 1. Основные понятия и обзор методов 4](#_Toc357283825)

1.1. [Основные определения 4](#_Toc357283826)

[1.2. Существующие алгоритмы 5](#_Toc357283827)

[Глава 2. Разработка и реализация системы 7](#_Toc357283828)

[2.1. Постановка задачи 7](#_Toc357283829)

[2.2. Архитектура системы 7](#_Toc357283830)

[2.2. Подбор пар 8](#_Toc357283831)

[2.3. Алгоритмы подбора пар 9](#_Toc357283832)

[2.4. Проверка пар 10](#_Toc357283833)

[2.5. Проверка совпадения страниц 11](#_Toc357283834)

[2.6. Диаграмма классов системы 12](#_Toc357283835)

[Глава 3. Результаты тестирования 13](#_Toc357283836)

[Список используемых источников 14](#_Toc357283837)

# 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 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.

Таким образом, выявление таких сайтов является актуальной задачей. В настоящей работе разработана и реализована система, позволяющая обнаружить сайты, которые являются зеркалами, и проверить это.

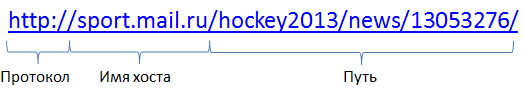
В главе 1 вводятся основные определения, производится постановка задачи, а также выполняется краткий обзор техник решения.

В главе 2 выполняется проектировка системы для решения поставленной задачи: разрабатывается общий концепт, приводятся используемые алгоритмы, описываются детали реализации системы.

В главе 3 приводятся результаты тестирования системы, и производится их анализ.

# Глава 1. Основные понятия и обзор методов

## Основные определения



**Рис. 1. Схема URL адреса**

Хотя понятие «Сайты-зеркала» доступно на интуитивном уровне, его весьма трудно формализовать. Зеркала бывают как полностью идентичные, отличающиеся только названием хоста, на котором расположен сайт, так и структурно-идентичные, то есть такие, у которых похожи схемы пути к страницам (возможно, с небольшими изменениями), а содержимое страниц по одинаковым путям 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. В таком случае можно было бы 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.
* Список может быть неактуален. Часть страниц из него может быть удалена, недоступна, 301 Moved Permanently, и так далее.
* Невозможно скачать и сохранить все страницы в памяти. Также трудно хранить в памяти данные по уже известным зеркалам.
* У некоторых сайтов-зеркал одинаковые страницы могут находиться по несовпадающим путям.

Это сильно усложняет задачу. Часть из этих проблем в этой работе удастся решить, часть еще предстоит.

## 1.2. Существующие алгоритмы

Существующие сейчас реализации для поиска зеркал являются закрытыми и/или запатентованными. Значительная часть описанных в работах алгоритмов являются вероятностными и действуют по такому принципу:

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

По используемой информации алгоритмы можно разделить на несколько категорий:

* Алгоритмы, основанные на IP адресе. Такие алгоритмы группируют входной список по IP-адресам и считают, что зеркалами являются те сайты, у которых IP-адрес одинаков.
* Алгоритмы, основанные на URL адресе. Такие алгоритмы для оценки сайтов используют различные части адреса (название хоста, схему пути до файла).
* Алгоритмы, основанные на переходах. Этот класс алгоритмов оценивает сайты, основываясь на переходах между страницами сайта, либо на ссылках на внешние ресурсы.
* Алгоритмы, основанные на деревьях сайтов. Основываясь на URL адресах страниц, эти алгоритмы строят деревья сайтов и сравнивают их.

Все эти алгоритмы несовершенны. Например, сейчас хостинг может выкладывать на одном и том же IP адресе много сайтов, благодаря технологии NAT. Алгоритмы, основанные на переходах, могут обмануть новостные сайты, так как часто они ссылаются на один источник или друг на друга (например, для повышения собственного веса в системе *PageRank*). Таким образом, процент верных ответов каждого из этих классов по отдельности невелик. Однако, эти алгоритмы достаточно просты в реализации и удобны в использовании. Поэтому на поверхности лежит идея группировки результатов этих алгоритмов. В работе [3] был проведен анализ алгоритмов из первых трех классов. В результате получилось, что лучшие результаты работы получаются тогда, когда используются алгоритмы из первого и третьего классов в совокупности. В этой работе произведена попытка улучшить результаты путем использования нескольких алгоритмов из разных классов в качестве критериев точности, а так же получение степени влияния различных алгоритмов на правильность итоговой оценки.

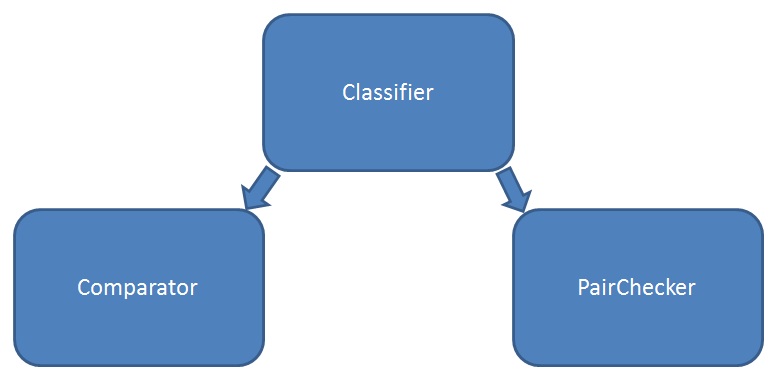
# Глава 2. Разработка и реализация системы

## 2.1. Постановка задачи

Таким образом, задачу поиска зеркал можно разбить на две: выделить из входного списка URL адресов пары сайтов, которые с большой вероятностью являются дубликатами, и проверить для каждой пары, действительно ли это сайты-зеркала.

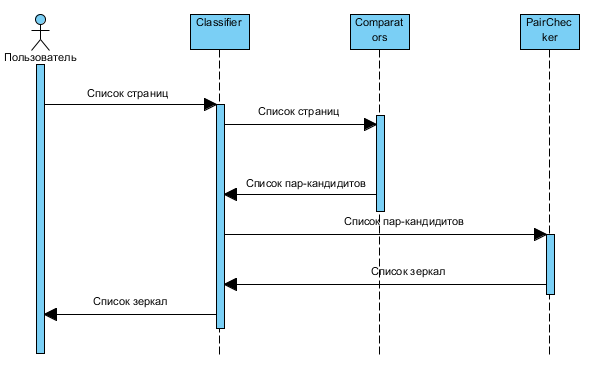
## 2.2. Архитектура системы

Система поиска состоит из трех модулей:

1. модуль, отвечающий за подбор пар-кандидатов;
2. модуль, проверяющий пары на зеркала;
3. основной модуль системы, с которым происходит общение пользователя. 

**Рис. 2. Компоненты системы**

Основной модуль (классификатор) получает на вход список сайтов с адресами страниц, затем с помощью компараторов (первого модуля) составляет список пар кандидатов, после же обрабатывается с помощью модуля проверки. После этого основной модуль выдает пользователю список с вероятными зеркалами.



**Рис. 3. Взаимодействие компонентов системы и пользователя**

## 2.2. Подбор пар

При подборе пар-кандидатов используется несколько компараторов. Компаратор это сущность, которая умеет создавать из большого списка адресов список пар по определенному признаку:

* По IP адресу сервера, на котором находится сайт
* По URL адресам страниц
* По связности сайтов
* …

Классификатор содержит в себе несколько компараторов. Каждому из них он передает список страниц, и от каждого компаратора получает в ответ свой список пар. Каждой паре в списке компаратор назначает вес. Чем больше этот вес, тем больше вероятность, что эти сайты являются зеркалами. Также у компараторов есть своя степень влияния на итоговый список. Она берется из предположения о том, что у каждого компаратора своя ценность. Например, информация от компаратора, который анализирует деревья сайтов может быть более важна, чем информация о том, что сайты находятся (или не находятся) на одном сервере.

Основываясь на всех списках от компараторов, а также влиянии компараторов, классификатор создает итоговый список взвешенных пар сайтов. Для этого он для каждой пары сайтов производит следующую операцию:

1. Находит ее во всех списках.
2. Вычисляет суммарный рейтинг пары на основе весов в каждом из списков и влиянии признаков.
3. Составляет итоговый список взвешенных пар.

После этого список фильтруется, чтобы убрать пары с малым весом.

## 2.3. Алгоритмы подбора пар

В текущей системе реализованы следующие компараторы для подбора пар:

1. Компаратор на основе IP адресов. Этот компаратор работает по следующему алгоритму. Для каждого сайта во входном списке вычисляется IP адрес, на котором находится сайт (с помощью запроса к DNS серверу). После этого для каждого из полученных IP адресов составляется список находящихся на нем сайтов. Далее алгоритм составляет пары кандидатов из сайтов, находящихся на одном IP-адресе. Чтобы уменьшить вероятность ложных срабатываний, считается, что если количество сайтов на одном IP адресе больше порогового значения, то этот IP адрес игнорируется. Этой оптимизацией можно отбросить сервера хостингов. В качестве примера таких сайтов можно привести сайты http://debian.com и http://debian.org.

В качестве веса пары берется число, обратное количеству сайтов на этом IP адресе.

1. Компаратор на основе URL адресов. Для определения кандидатов на основе URL адресов страниц сайта используется модификация алгоритма шинглов [4]. Анализ происходит следующим образом. Каждый URL адрес из списка разделяется на части по символу “/”. Например, адрес

www.google.com/calendar/events/Saint-Petersburg-2012/

разбивается следующим образом:

www.google.com calendar events Saint-Petersburg-2012

После этого все числа заменяются символом “\*” для удобства анализа.

www.google.com calendar events Saint-Petersburg-\*

Во многих случаях числа используются как идентификатор чего-либо в коллекции. Эта оптимизация поможет выделить значимую часть из таких адресов как, например, http://news.mail.ru/politics/13238342/ или http://news.mail.ru/politics/13236325/. Обе эти ссылки приведутся к виду http://news.mail.ru/politics/\*/.

Далее несколько подряд идущих частей группируется в терм, и к нему добавляется номер первой части. Например, если группировать в терм по 2 части, то из указанных выше частей получатся следующие термы:

* www.google.com\_calendar\_1
* calendar\_events\_2
* events\_Saint-Petersburg-\*\_3

Позиции помогут обнаружить адреса, сходные по структуре, а не только по частям адреса.

После преобразования всех адресов в термы, для каждого сайта собирается множество самых частых из полученных термов. Далее высчитывается оценку похожести хостов . Для каждого терма, который встречается на нескольких сайтах, вычисляется его вес в оценке по следующей формуле:

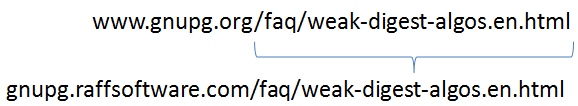
,

где — количество хостов, на которых встретился этот терм. Этот вес добавляется в оценку. После обработки всех термов общая оценка нормализуется по формуле:

где и — количество самых частых термов host1 и host2 соответственно. Нормализация делается, так как у сайтов с большим числом адресов получается много термов, и, соответственно, оценка сходства получается очень большой. Знаменатель дроби помогает сделать ее пропорциональной размеру сайта. Нормализованная оценка похожести выступает в качестве веса пары.

## 2.4. Проверка пар

После того, как классификатор обработал данные компараторов и составил итоговый список пар для проверки, этот список передается системе проверки. Ее задача подтвердить, либо опровергнуть, что пара сайтов это действительно зеркала. Для каждой пары сайтов система выберет несколько страниц с этих сайтов и проверит их на сходство. Для выбора пар страниц система может использовать любую стратегию. В общем случае, когда ничего, кроме адресов страниц, о сайтах неизвестно, можно выбирать произвольный адрес страницы на одном из сайтов и добавлять путь к файлу на первом сайте к имени хоста другого сайта.



**Рис. 4. Преобразование ссылок**

Создав несколько пар страниц, система получает их содержимое и измеряет их сходство.

## 2.5. Проверка совпадения страниц

Как уже было сказано, для проверки двух страниц на совпадение можно использовать несколько стратегий.

1. Можно сравнивать на точное совпадение. Для этого, например, можно вычислять хэш-функцию. Однако, эта проверка слишком строга, так как в реальных ситуациях сайты часто могут отличаться только версткой или картинками, оставаясь сходными по содержанию.
2. Другая идея состоит в том, чтобы брать только текста страницы, без различных тегов, ссылок, картинок, рекламы и так далее. Это позволит убрать со страницы большинство динамического содержимого, которое не относится к сути страницы и может мешать работе системы. В данной работе для такой проверки страниц используется алгоритм шинглов. Он работает следующим образом. Несколько подряд идущих слов в тексте объединяются в кортежи, называемые чешуйками (shingles). При этом выборка происходит внахлест. Например, для текста «а роза упала на лапу азора» можно получить три 4-шингла (4 — это количество слов в кортеже):

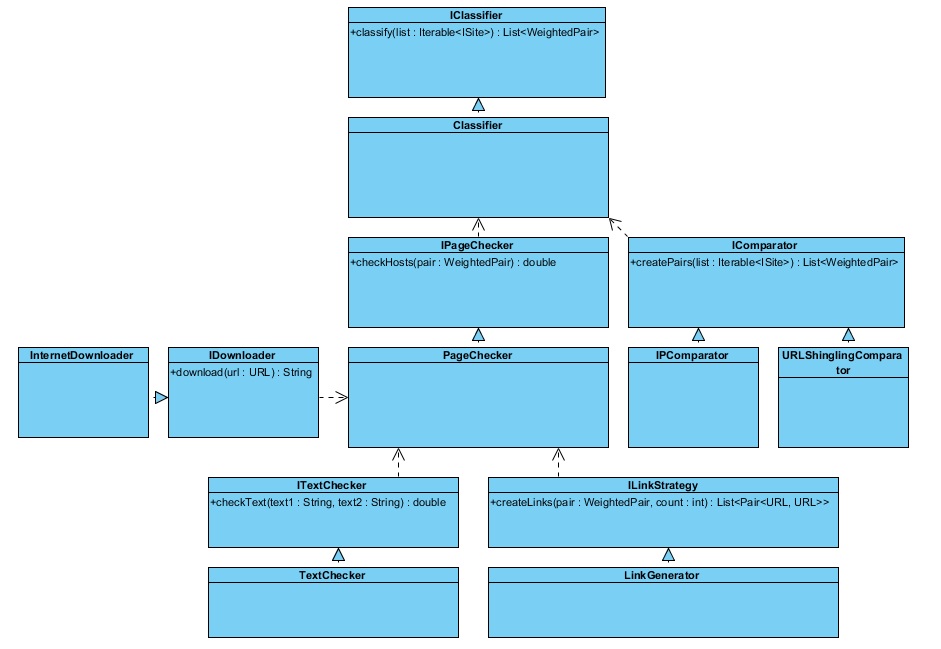
* (а, роза, упала, на)
* (роза, упала, на, лапу)
* (упала, на, лапу, азора)

Для каждой страницы page1 и page2 строятся множества шинглов и , которые встретились в тексте. Для оценки сходства используется коэффициент Жаккарда:

Нетрудно заметить, что эта оценка лежит в интервале , при этом 1 соответствует идентичности страниц, а 0 полному различию.

Проверив предлагаемые пары страниц на сходство, можно сделать итоговое предположение о том, действительно ли сайты являются зеркалами. Если более 50 процентов страниц оказывались похожими более чем на 0.5, то сайты признаются зеркалами.

## Диаграмма классов системы



**Рис. 3. Диаграмма классов системы**

На этой диаграмме можно увидеть иерархию классов, представленных в системе. Для проверки список с URL-адресами передается в Classifier, который делегирует составление списков классам, реализующим интерфейс IComparator. После этого он соединяет полученные списки пар в один и передает каждую его пару на обработку классу PageChecker. Получив итоговые результаты, Classifier выдает итоговый список пар, которые с наибольшей вероятностью являются зеркалами.

Как можно видеть из диаграммы классов, в большинстве случаев в программе используются интерфейсы, а не конкретные реализации классов. Благодаря этому систему достаточно просто расширять, добавлять новые компараторы, изобретать свои схемы подбора ссылок и проверки текста, даже переделывать принцип построения итогового списка. Также в такой системе несложно покрывать unit-тестами каждый компонент программы в отдельности и использовать фреймворки для внедрения зависимостей (dependency injection) для облегчения совместного взаимодействия компонентов. В каждом из компонентов хранится текущее состояние (stateless-компоненты), таким образом, система доступна для многопоточного использования.

# Глава 3. Результаты тестирования

## Тестирование на небольшом списке

Сначала система была испытана на небольшом списке из 32 тысяч адресов с сайтов:

* "www.php.net", "fi1.php.net", "jp1.php.net"
* "cran.r-project.org", "cran.gis-lab.info"
* "arxiv.org", "uk.arxiv.org"
* "www.wikileaks.org", "www.wikileaks.no"
* "www.gnupg.org", "gnupg.raffsoftware.com", "gnupg.parentinginformed.com"

Существующих компараторов оказалось достаточно, чтобы обнаружить все зеркала. Для каждой из пар зеркал при проверке суммарное сходство оказалось более 90 процентов. Проблемы возникли только при обнаружении пары (www.wikileaks.org; www.wikileaks.no), так как www.wikileaks.no является неполным зеркалом. При анализе этой пары сходство страниц оказалось около 50 процентов.

# 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.
3. *K. Bharat, A. Broder, J. Dean, M. Henzinger.* A Comparison of Techniques to Find Mirrored Hosts on the WWW. **Fourth ACM Conference on Digital Libraries**, 1999.
4. *A. Broder, S. Glassman, M. Manasse, G. Zweig.* Syntactic Clustering of the Web. <http://www.std.org/~msm/common/clustering.html>
5. *K. Bharat, A.Broder.* Mirror, mirror on the Web: a study of host pairs with replicated content.