В связи с ростом интернета, появляется много сайтов-дубликатов. Такие сайты как полезны, так и вредны, информация об этом достаточно ценна, поэтому задача нахождения зеркал актуальна.

Для ее решения в данной работе была разработана система для поиска зеркал. Первая проблема, с которой мы встречаемся при поиске: это как найти пары сайтов, которые могут являться зеркалами, работать быстро и требовать не слишком много информации о сайтах. Вторая проблема: как проверить найденные пары на дубликаты.

Первая проблема в данной работе решается при помощи компараторов. В качестве входных данных для них выступает список сайтов с адресами страниц. Каждый компаратор оценивает эти сайты по какому-либо признаку. В данной работе используются три признака.

* IP адреса сайтов
* Схемы путей URL адресов страниц
* Граф зеркал

Каждый компаратор, оценив по своему признаку, генерирует свой список пар кандидатов. Каждой паре компаратор назначает вес. Получив все списки, система создает общий список. Для каждой пары в нем определяется общий вес, который зависит от веса пары в каждом списке и влиянии компараторов.

После этого необходимо решить вторую проблему, а именно проверить эти пары кандидаты. Для этого используется модуль проверки сайтов, или чекер. Система передает ему список пар, для каждой пары чекер составляет список пар страниц, которые нужно скачать и проверить на совпадение. Проверяется только текст страниц, без разметки, графики и так далее.

Для проверки совпадения чекер использует алгоритм шинглов. Для этого оба текста делятся на кортежи по несколько элементов, строятся множества шинглов для каждого текста, а дальше по несложной формуле коэффициента Жаккарда находится степень сходства.

После проверки всех пар страниц система на основании доступности страниц и их сходства вычисляет для каждой пары сайтов кандидатов суммарный коэффициент сходства.

Следующей проблемой является возможность сохранения полученных данных и их дальнейшего использования. Так как сайтов много, информацию о них невозможно всегда хранить в памяти. Для решения этой проблемы применяется графо-ориентированная база данных.

База представляет собой граф, где вершинами являются сайты, а ребрами обозначается, что сайты являются зеркалами. По этой базе можно удобно отвечать на запросы пользователя, являются ли сайты зеркалами, и если да, то каков их коэффициент сходства.

Итого система состоит из четырех модулей:

* Основной модуль
* Модуль компараторов
* Модуль проверки сайтов
* База зеркал

Каждый из модулей представляет собой независимое приложение. Модули можно запустить на различных компьютерах, таким решается проблема горизонтального масштабирования системы.

Было проведено два тестирования системы. При первом тестировании использовался список из 40 тысяч адресов страниц с 15 сайтов. Все зеркала были обнаружены за 20 секунд, причем бОльшую часть времени занял процесс получения страниц из интернета. При втором тестировании проводился анализ системы на сайтах со схожей структурой путей. Но и тут систему не удалось обмануть, так как содержимое страниц оказывалось различным.

Таким образом, система является работающей и может использоваться в промышленных задачах.