В связи с ростом интернета, появляется много сайтов-дубликатов. Согласно одному исследованию, сайтов, которые в той или иной степени являются дубликатами, примерно от 30 до 45 процентов. Такие сайты как полезны, так и вредны, информация об этом достаточно ценна, поэтому задача нахождения зеркал актуальна.

Для ее решения была разработана система для поиска зеркал. В качестве входных данных система получает большой список URL адресов с многих сайтов. Задача: обнаружить среди них сайты-зеркала. Система работает по следующему алгоритму:

1. Преобразовать список URLов в список пар хостов — кандидатов в зеркала.
2. Проверить эти пары хостов на дубликаты.

Для решения первой части в системе применяются так называемые компараторы. Это классы, реализующие интерфейс IComparator. В каждом компараторе заложен свой признак, и на основании этого признака компаратор генерирует список пар. Признаки могут быть как простые, например, одинаковый IP адрес, сходство URL адресов, так и сложные, например, связность страниц, одинаковые внешние ссылки, сходство деревьев сайтов. Вот пример компаратора на основе IP адреса. Он «считает», что если сайты находятся на одном IP адресе, то это сайты-кандидаты. Каждый из компараторов запускается на всем списке и генерирует свой список пар-кандидатов. Каждой паре компаратором назначается вес. Чем он больше, тем больше вероятность, что сайты в паре – это зеркала.

Далее на основе этих список строится единый. Для этого для каждой пары, которая есть хотя бы в одном из списков, подсчитывается суммарное значение ее веса в итоговом списке. Для этого берется вес пары в каждом из списков и домножается на значение влияния компаратора. Эти значения отражают точность и корректность компаратора при построении пар.

Итоговый список попадает на обработку PageChecker’у. PageChecker на основе названия хоста и информации о страницах, взятой из большого списка, генерирует несколько пар ссылок на страницы, которые требуется проверить, чтобы подтвердить факт дублирования.

Далее Чекер пытается получить эти страницы и сравнить сходство содержимого. Для этого применяется алгоритм w-шинглов. Если вкратце, то оба текста делятся на кусочки по w слов, называемые шинглами, или чешуйками. Для каждого текста составляется множество шинглов. После этого на основе формулы устанавливается степень сходства текстов. Значение формулы лежит в интервале от 0 до 1, где 0 соответствует полному различию, а 1 полному сходству. После проверки всех пар страниц, производится итоговый анализ на основе процента доступных страниц и их сходства, и подсчитывается результат.

Все компараторы можно запускать параллельно, проверку пар также можно запускать параллельно. Таким образом время работы алгоритма напрямую зависит от времени работы компараторов.