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

[тыц] Для ее решения была разработана система для поиска зеркал. В качестве входных данных система получает большой список URL адресов страниц с многих сайтов. Задача: обнаружить среди них сайты, которые являются зеркалами. Алгоритм, по которому работает система, состоит из двух частей:

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

[тыц] Для решения первой части в системе применяются компараторы. Каждый компаратор основывается на каком-либо признаке сходства сайтов-дубликатов, и на основании этого признака компаратор генерирует список пар-кандидатов. Признаки могут быть как простые, например, одинаковый IP адрес, сходство URL адресов, так и сложные, например, связность страниц, одинаковые внешние ссылки, сходство деревьев сайтов. Для примера, компаратор на основе IP адреса «считает», что если сайты находятся на одном IP адресе, то это сайты-кандидаты. Каждой паре компаратором назначается вес. Чем он больше, тем больше вероятность, что сайты в паре – это зеркала.

[тыц] Далее на основе этих списков строится единый. Для этого для каждой пары, которая есть хотя бы в одном из списков, подсчитывается суммарное значение ее веса. Для этого берется вес пары в каждом из списков и домножается на значение влияния компаратора. Это влияние отражает точность и корректность компаратора при построении пар. Например, информация о том, что сайты имеют похожие деревья, может быть более значима, чем информация о том, что IP адреса одинаковы или различны. Итоговый список фильтруется, чтобы отсеять пары с малым весом.

[тыц] Далее список обрабатывается PageChecker’ом. Входным значением для него является взвешенная пара сайтов. 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.

[тыц] Далее Чекер пытается получить эти страницы и проверить сходство содержимого. Для этого применяется алгоритм шинглов. Шинглы, или чешуйки, это упорядоченные кортежи из w слов. Для каждого текста вычисляется множество шинглов, которые присутствуют в нем. 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. Например, можно считать, что, если все страницы доступны, и суммарное сходство составляет более 50 процентов, то сайты являются зеркалами.

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

[тыц] Вопросы?