Задача 2 — Поисковый робот

Кашин Андрей

23 марта 2013 г.

1 Алгоритм

Работа происходит в несколько потоков. Каждый поток крутится в цикле, и на каждой итерации берет URL из общей очереди, скачивает страничку по этому URL'у, парсит её, и для каждой найденной ссылки проверяет, есть ли она в хэш-таблице уже добавленных ссылок, если её еще там нет, то он добавляет её туда а также добавляет эту ссылку в общую очередь Таким образом каждая ссылка будет добавлена в очередь не более одного раза, а значит посещена не более одного раза.

2 Теоретическое обоснование

Были произведены некоторые замеры времени работы отдельных компонент алгоритма. Оценивались времена скачки страницы и времена её обработки (парсинга). Время парсинга было принебрежимо мало, а время скачки, наоборот, было очень большим. Таким образом был сделан упор на ускорения скачки страниц за счет скачки в несколько потоков. Ясно, что это единственное узкое место(время скачки приблизительно 300ms, а время разбора страницы — 1ms), расходы на синхронизацию ничтожно малы в этом случае. Максимизация скорости работы робота заключалась в максимизации скорости скачки, а следовательно в использовании пропускной способности канала на максимум, путем паралельной скачки.

3 Результаты тестовых запусков

Тестовые запуски показывают, что время скачки возрастает, при увеличении количества потоков, используемых для скачки. Это еще раз подтверждает, что это критичное место для робота. Время работы программы с ключами

$$crawler\ yandex.ru\ 2\ 100\ /tmp/yandex$$

в 16 потоков, равно $T = 150s$, $Size = 5.184mb$.