Семинарское задание №6
Crawling Order

Мосиенко Константин Викторович

2019

Вам предстоит несколько раз смоделировать обход simple википедии с целью сравнить различные политики обхода, которые в данном случае будут просто выбирать, какой конкретно урл скачать следующим. Для этого понадобится скачанная википедия и предрассчитанный по ней pagerank. Для каждой политики предлагается построить график зависимости $WC(t)$, с предположением, что мы скачиваем одну страницу в секунду. Предлагается рассмотреть следующие политики:

- Сверх читерная. Данная политика обходит страницы в порядке убывания $PR(p)$, сверх читерность её в том, что она знает обо всех урлах сразу, поэтому тут даже ничего не надо моделировать. Просто сортируем все $PR(p)$ по убыванию и рисуем график частичных сумм.
- Читерная. Данная политика просто знает наперёд $PR(p)$ всех страниц и из всех страниц в очереди предлагает скачать ту, у которой он выше.
- Обход в ширину.
- Выбор случайного урла из очереди.
- Приоритезация по количеству входящих ссылок.
- (Дополнительный балл) Используйте для приоритезации OPIC. Обратите внимание, что для его работы необходимо скачивать каждую страницу по много раз. С какими ещё проблемами вы столкнулись?

Постройте все функции на одном графике.