УДК 681.324:621.325

И.В. Ильина, И.И. Линник

Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков

МЕТОДЫ РАНЖИРОВАНИЯ ДОКУМЕНТОВ В ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

Проанализированы возможности алгоритмов ссылки ранжирования, которая применяется к коллекции документов, связанных гиперссылками. Рассмотрены основные методы, используемые для повышения ссылки ранжирования сайта.

***Ключевые слова**: PageRank, HITS, web-граф.*

Введение

Поиск в коллекциях документов является важной задачей. Об этом свидетельствует как большое количество поисковых систем, так и их постоянное развитие. Коллекции документов могут быть различных типов: блоги, новостные ленты, научные статьи или все множество веб-страниц. Поисковые системы, такие как Google или Yahoo, оперируют с последним типом коллекций. Принцип работы поисковой системы следующий: пользователь вводит запрос, после чего система возвращает те докумен-

ты из коллекции, которые наилучшим образом удовлетворяют запросу. Как правило, в традиционных поисковых системах ранжирование документов (определение релевантности документа запросу) производится на основе статистической информации о множестве слов в запросе и в документе [1].

Основная часть

В статье описаны основные методы выявления статических признаков ранжирования, используемых помимо векторной или вероятностной модели поиска для улучшения результатов поиска.

Если происходит поиск в Интернет, то расширение методов ранжирования за счет учета ссылочной структуры web может дать существенное улучшение результатов поиска. Статическая часть web, состоящая из HTML-документов и гиперссылок между ними может быть представлена в виде направленного графа, в котором каждый узел является web-страницей, а каждое направленное ребро – гиперссылкой. Совокупность таких узлов и направленных ребер называется web-графом.

Ссылки в web-графе не распределены случайным образом. Число ребер, входящих в узел, распределено скорее по степенному закону, а не по закону Пуассона, как было бы, если бы ссылки были расставлены случайным образом. Далее будет рассмотрено два интуитивных предположения, на основе которых базируется изложение методов анализа ссылок в web-графе [2].

1. Текст ссылки, указывающей на страницу ***B***, является хорошим описанием страницы ***B***.

2. Гиперссылка со страницы ***A*** на страницу ***B*** представляет собой признание авторитетности страницы ***B*** со стороны создателя страницы ***A***. Это не всегда так; например, многие ссылки со страницы на страницу внутри одного сайта обязаны своим появлением общему шаблону сайта. Например, большинство корпоративных web-сайтов имеют на каждой странице ссылки на уведомление об авторском праве. Ясно, что это не является свидетельством одобрения. Соответственно, алгоритмы анализа ссылки учитывают такие ссылки с меньшим весом.

1. Метод PageRank

Ниже рассмотрены методы вычисления весов и ранжирования, вытекающие исключительно из структуры ссылок. Первый метод присваивает каждому узлу web-графа вес в диапазоне от 0 до 1, известный как PageRank. Вес узла зависит от структуры ссылок. Процесс случайного перемещения по сети, начинается с какой-либо страницы (узла web -графа), и подчиняется следующим правилам. В каждый момент времени пользователь с страницы ***A*** переходит на случайную страницу, на которую со страницы ***B*** ведет гиперссылка. Идея алгоритма PageRank заключается в том, что узлы, часто посещаемые в результате случайного блуждания по сети, имеют большую важность, чем редко посещаемые узлы [1].

Если в узле ***A*** нет исходящий ссылок, то происходит телепортация (teleport) в случайный узел сети. Эта операция эквивалентна тому, что пользователь набирает URL-страницы в строке браузера.

Операция телепортации используется двояко:

1. Если узел не имеет исходящих ссылок, то пользователь вызывает операцию телепортации.

2. Если узел имеет исходящие ссылки, то телепортация происходит с вероятностью p, где αобычно равно 0,1, а с вероятностью $1-p$ переходит по случайной исходящей ссылке.

Применяя теорию марковских цепей, можно показать, что пользователь, следующий по описанному комбинированному алгоритму, находится в каждом узле н фиксированную долю времени , которая зависит от структуры веб-графа и от значения р н называется весом PageRank узла н.

Для подсчета PageRank можно использовать степенной метод (poweriteration). Он имитирует случайное блуждание: начав с определенного состояния и выполнив большое число шагов *t*, можно отследить частоты посещения каждого состояния. После большого количества шагов *t* эти частоты "устанавливаются", так что разница между вычисленными частотами опускается ниже заданного порогового значения. Эти частоты объявляются весами PageRank.

Значения весов PageRank не зависят от запроса пользователя. Таким образом, вес PageRank является мерой статического качества web-страницы, не зависящей от запросовпользователей. Ясно, что результат ранжирования должен зависеть от запроса [2].

2. Метод HITS

Метод HITS (Hyperlink-InducedTopicSearch), в котором по заданному запросу каждая web-страница получает два показателя: показатель портальности (hubscore) и показатель авторитетности (authorityscore). В этом методе вычисляется два ранжированных списка страниц, а не один. Метод основывается на разделении страниц на два основных типа: порталы (hub) и авторитетные источники (authority). Такое разделение особенно уместно при широком поиске (broad-topicsearch), когда запросы имеют вид "Я хочу знать о ХУПС". Авторитетным источником является официальный сайт университета. Порталами являются сайты, содержащие ссылки на авторитетные источники. Эти страницы не содержат в себе необходимую информацию, но содержат ссылки на нее, собранные заинтересованными людьми.Таким образом, показатель портальности страницы определяется как сумма показателей авторитетности страниц, на которые она ссылается. С другой стороны, если на страницу ссылаются хорошие порталы, то показатель ее авторитетности увеличивается [2].

Вывод

Рассмотрены основные методы, используемые для повышения ссылки ранжирования сайта. Умелое и взвешенное сочетание различных методов приводит к требуемому эффекту. Осторожность и соблюдение чувства меры относится к ссылочному ранжированию, как к одному из методов раскрутки web-ресурса.

Список литературы

1. Садовский А. Растолкованный PageRankили Все, что вы всегда хотели знать о PageRank [Электронный ресурс] / А. Садовский. – Режим доступа к ресурсу: http://digits.ru/ articles/promotion/pagerank.html.

2. Christopher D. Introduction to information retrieval / D. Christopher. – Cambridge Univ. Press. 2008. – 544 p.

Поступила в редколлегию 25.11.2012

Рецензент: д-р техн. наук, проф. И.В. Рубан, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

МЕТОД РАНЖИРУВАННЯ ДОКУМЕНТІВ В ІНФОРМАЦІЙНО-ПОШУКОВІЙ СИСТЕМІ

І.В. Ільїна, І.І. Линник

Стаття присвячена аналізу можливостей алгоритмів посилання ранжирування, яка застосовується до колекції документів, пов'язаних гіперпосиланнями. Розглянуті основні методи, які використовують для підвищення посилання ранжирування сайту.

Ключові слова:*PageRank, HITS, web-граф.*

RANKING METHOD DOCUMENTS IN THE INFORMATION-RETRIEVAL SYSTEM

I.V. Il’ina, I.I.Linnik

This article analyzes the possibilities of ranking algorithms links which is applied to the collections of documents related hyperlinks. The main methods used for the reference of raising site ranking.

Keywords: *PageRank, HITS, web-graph.*