Сарсенова Айжан Зинетуллаевна
студентка
ФГБОУ ВПО «Саратовский государственный
университет им. Н.Г. Чернышевского»
г. Саратов, Саратовская область

АЛГОРИТМ ССЫЛОЧНОГО РАНЖИРОВАНИЯ

Аннотация*: в данной статье рассматриваются особенности реализации алгоритма ссылочного ранжирования PageRank. В работе приводится формула для подсчета ранга веб-страниц, а также анализируется работа и применимость алгоритма в реальной жизни.*

Ключевые слова*: ссылочное ранжирование, алгоритм PageRank, формула релевантности.*

Неотъемлемой составляющей каждого пользователя Всемирной паутины является поиск информации. Благодаря информационно-поисковым системам появилась такая возможность поиска, анализа и систематизации необходимой информации, причем так, что каждый пользователь может быстро найти релевантные и исчерпывающие сведения [1, с. 17]. Поисковая машина предоставляет список документов, в которых встречались ключевые слова запроса, выдавая результат в ранжированном виде.

Ранжированием называется упорядочивание результатов поиска по их релевантности [2, с. 71]. Под релевантностью подразумевается соответствие найденных документов изначальному запросу пользователя. У каждой поисковой машины есть своя «формула релевантности» для веб-страниц. Например, формула ранжирования по содержимому страницы состоит из трех характеристик: частоты слов, определяющей количество вхождений в документ слов, расположения поисковых слов в документе, то есть чем ближе слово расположено к началу страницы (возможно, даже в заголовке), тем выше ранг выданного результата, и расстояния между поисковыми словами.

Одним из самых популярных видов ранжирования является ссылочное ранжирование. Принцип ссылочного ранжирования основывается на довольно старой идее индекса цитируемости публикаций в научном мире для подсчета авторитета ученого, т. е. кого больше цитируют и на кого чаще ссылаются, тот авторитетен, следовательно, его работы важнее остальных [1, с. 71].

Эта идея была впервые применена основателями компании Google (Сергей Брин и Ларри Пейдж) и получила свое развитие в виде алгоритма PageRank, названный по фамилии одного из его изобретателей. Этот алгоритм присваивает каждой веб-странице ранг, который определяет ее «авторитетность» (важность, значимость) [2, с. 94]. «Авторитетность» веб-страницы вычисляется на основе других веб-страниц, которые ссылаются на нее, и общего количества ссылок, которые имеются на каждой из них.

То есть PageRank – это числовая величина, характеризующая «важность» веб-страницы. Чем больше ссылок на страницу, тем она «важнее». Кроме того, «вес» одной веб-страницы определяется весом ссылки, передаваемой другой веб-страницей. Таким образом, алгоритм PageRank – это метод вычисления веса страницы путем подсчета «важности» ссылок на нее. То есть рассчитывается вероятность того, что человек, который случайно переходит по ссылкам, дойдет до некоторой страницы. Чем больше веб-страниц, ссылающихся на данную, тем выше вероятность того, что пользователь случайно наткнется на эту страницу.

Если на некоторую веб-страницу ссылается очень «авторитетный ресурс», то и ранг самой страницы повышается. Однако, если этот «авторитетный» ресурс ссылается еще на тысячу других веб-страниц, то ранг этой страницы повышается незначительно [1, с. 72].

В связи с тем, что пользователь, конечно, может до бесконечности переходить по ссылкам и в таком случае перейти на все имеющиеся страницы, но чаще всего люди останавливаются на каком-то этапе. И, чтобы учесть данный факт, создателями алгоритма был введен коэффициент затухания (у каждой поисковой системы он равен определенному числу), означающий, что это вероятность того,

что пользователь продолжит кликать по имеющимся ссылкам на странице [2, с. 95].

Основная формула, называемая формулой «релевантности», по которой рассчитывается PageRank некоторой веб-страницы *A*, представлена в (1):

$$PR(A) = (1\text{-}d) + d * \left(\frac{PR(T_1)}{C(T_1)} + \ ... + \frac{PR(T_n)}{C(T_n)}\right), \quad (1)$$

где *PR(A)* – ранг веб-страницы *A*, d – коэффициент затухания, n – количество страниц, которые ссылаются на данную веб-страницу, T_i – i-тая ссылающаяся страница и C – общее количество ссылок на этой веб-странице. Причем нужно учесть, что невозможно вычислить ранг веб-страницы, пока неизвестны ранги ссылающихся на нее других веб-страниц, а эти ранги можно вычислить, только зная значения веб-страниц, которые ссылаются на них. Поэтому изначально всем веб-страницам присваивается произвольный PageRank (обычно начальный ранг ресурса равен 1, но на самом деле это не так важно).

После первого цикла подсчета ранга веб-страницы придется вернуться и пересчитать все ранги еще раз, так как PageRank остальных веб-страниц, ссылающиеся на веб-страницу *A*, уже изменится. И так придется сделать достаточное количество пересчетов, то есть так называемых итераций. В процессе разработки алгоритма PageRank создателям Google пришлось доказать *эргодическую теорему*, которая заключается в том, что процесс пересчета рангов веб-страниц в «конечном» итоге сойдется. Получается, что достаточно несколько раз пересчитать ранги всех веб-страниц из некоторого каталога ресурсов для того, чтобы их ранги зафиксировались и можно было в дальнейшем пользоваться ими для расчета релевантности в поисковых системах [1, с. 72].

Важность алгоритма PageRank трудно не оценить, так как он стал основной метрикой ранжирования результатов поиска в Google, что привело в свое время к резкому отрыву от конкурентов по качеству поиска. И в дальнейшем стал применяться в остальных крупнейших поисковых системах таких, как Yahoo, Yandex, Rambler и др. Следует отметить, что знание данного алгоритма также полезно веб-разработчикам в целях продвижения своего сайта.

Список литературы

1. Ашманов И.С. Продвижение сайта в поисковых системах [Текст] / И.С. Ашманов, А.А. Иванова. – М.: ООО «И.Д. Вильямс», 2007. – 304 с.

2. Сегеран, Т. Программируем коллективный разум [Текст] / Т. Сегеран; пер. с англ. А. Слинкина. – СПб.: Символ-Плюс, 2008. – 368 с.