УДК 004.738.52

РАНЖИРОВАНИЕ WEB-СТРАНИЦ С ИСПОЛЬЗОВАНИЕМ ВЗАИМНОЙ ИНФОРМАЦИИ МЕЖДУ ГИПЕРССЫЛКАМИ

Р.М. Алгулиев, Р.М. Алыгулиев

Институт информационных технологий Национальной академии наук Азербайджана, г. Баку

Для повышения эффективности ранжирования web-страниц предложены три модификации алгоритма PageRank. Особенность первой из них состоит в измерении степени независимости гиперссылок, на основе которой определяется вес гиперссылки. Вторая и третья модификации, учитывающие тематическую близость web-страниц, представляют собой усовершенствованные варианты алгоритмов WPR и Topic-Centric, соответственно.

ВВЕДЕНИЕ

Появление в конце прошлого столетия World Wide Web (WWW) сделало Интернет одним из основных источников информации. Сегодня WWW — это динамично изменяющаяся среда, а представленные в ней информационные ресурсы крайне разнородны. Распределенный характер WWW сильно затрудняет поиск нужной информации среды разнообразных материалов, охватывающих самые разные сферы человеческой деятельности. При поиске нужной информации в Web поисковым машинам приходится охватывать огромное число связанных гиперссылками страниц. По мере увеличения web-ресурсов и рассредоточения их источников поиск нужной информации в Интернете становится еще более трудоемким. При этом следует разрабатывать такие технологии и подходы, которые отвечали бы увеличивающимся потребностям пользователей. Один из таких подходов — это усовершенствование существующих технологий информационного поиска.

Поисковые машины по их методам индексирования делятся на два поколения. В поисковых машинах первого поколения, разработанных на ранних стадиях создания Web, для ранжирования web-страниц была использована частота слов или мера подобия, т. е. были использованы традиционные методы поиска документов. В отличие от обычных текстовых документов web-страницы имеют ряд специфических особенностей. В гипертекстовой среде носителями информации, кроме web-страниц, являются и гиперссылки. Поэтому для улучшения поисковой точности поисковые машины второго поколения как дополнительный источник информации используют гиперссылки. Поисковые машины, следуя по гиперссылкам, посещают огромное web-пространство, в результате которого собирают дополнительную информацию о web-страницах. Потом эта информация ими используется при ранжировании web-страниц. Поэтому в последние годы многие исследования были посвящены анализам гиперссылок.

На сегодня разработаны некоторые алгоритмы для решения упомянутой проблемы. Среди них наиболее популярны алгоритмы PageRank [1, 2], HITS (Hypertext Induced Topic Search) [3] и SALSA (Stochastic Approach for Link Structure Analysis) [4]. Известные поисковые машины — Google, Yahoo и др. — в той или иной степени используют эти алгоритмы. Алгоритмы PageRank и HITS ранг web-страниц вычисляют итеративно. Алгоритм HITS с помощью обычных методов информационного поиска сначала идентифицирует web-страницы, релевантные запросу пользователя, а затем упорядочивает их таким образом, чтобы самые релевантные web-страницы были представлены в верхней части списка. Другой алгоритм PageRank ранжирует целые связанные гиперссылками web-страницы, затем среди них выбирает страницы, релевантные запросу пользователя, сохраняя при этом их ранги. Алгоритм SALSA является комбинацией алгоритмов PageRank и HITS.

Настоящая статья посвящена усовершенствованию алгоритма PageRank, где предлагаются три модификации.

1. КРАТКИЙ ОБЗОР АЛГОРИТМОВ PAGERANK И HITS

Основные обозначения:

$G(W, E)$ — ориентированный web-граф;
W — множество web-страниц;
E — множество гиперссылок;
$u \to v$— гиперссылка из страницы u в страницу v;
$B(u) = \{v : v \to u\}$ — множество страниц, которые ссылаются на страницу u (Backwards links);
$F(u) = \{v : v \to u\}$ — множество страниц, на которые ссылается страница u (Forwards links);
n — общее число страниц в web-графе $G(W, E)$;
$PR(u)$ — PageRank страницы u;
$A(u)$ — ранг страницы u как «авторитет»;
$H(u)$ — ранг страницы u как «концентратор»;
$|U|$ — мощность множества U;
$\mathrm{sim}(u, v)$ — мера близости страниц u и v, которая определяется метрикой косинуса;
$\omega(v \to u)$ — вес гиперссылки $u \to v$;
$r(q, t)$— степень релевантности запроса q тематике T.

Для извлечения информации из структуры гиперссылок и перекрестных ссылок, отслеживания дефектов их структуры, анализа связей между ссылками и объектами ссылок (Web Structure Mining) широко применяется алгоритм PageRank. Это статический алгоритм, предназначенный для оценки качества страниц, не зависящий от каких-либо запросов, т. е. с его помощью вычисляется «глобальная значимость» страниц. Суть алгоритма заключается в следующем. Представьте себе случайного пользователя, перемещающегося по Web. Пусть пользователь посещает страницу v. На каждом шаге пользователь либо «перепрыгивает» на другую страницу в Web, выбранную случайным образом, либо он следует по гиперссылке на текущей странице, при этом не возвращаясь и не посещая одну и ту же страницу дважды. Если через $(1 - d)$ обозначить вероятность случайного прыжка, то вероятность перехода по ссылке будет d. Таким образом, показатель PageRank страницы u можно вычислить по следующей рекурсивной формуле:

$$RP(u) = \frac{(1-d)}{n} + d \sum_{v \in B(u)} \frac{1}{|F(v)|} RP(v). \qquad (1)$$

В правой части $1/n$ соответствует тому, что среди n страниц каждая страница выбирается с одинаковой вероятностью. Здесь также предполагается, что исходящие ссылки на странице v выбираются с одинаковой вероятностью, равной $1/|F(v)|$. Для сходимости процесса (1) вероятность d (ее называют коэффициентом демпфирования [1, 2]) выбирается из интервала $d \in [0{,}8;\ 1]$. Из алгоритма (1) видно, что чем больше ссылок на страницу, тем она становится «важнее».

Алгоритм HITS [3], как и алгоритм PageRank, основан на анализе web-структуры, но в отличие от него у каждой страницы выделяются две роли: роль «авторитета» (authority) и роль «концентратора» (hub). Алгоритм HITS, анализируя входящие и исходящие гиперссылки, ранжирует web-страницы. По этому алгоритму страница, на которую ссылаются другие страницы, называется «авторитетом», а страница, которая ссылается на другие страницы, называется «концентратором»:

$$A(u) = \sum_{v \in B(u)} H(v), \qquad (2)$$

$$H(u) = \sum_{v \in F(u)} A(v). \qquad (3)$$

Цель алгоритма HITS заключается в поиске наиболее качественных «авторитетов» и наиболее качественных «концентраторов». Из формул (2) и (3) видно, что для каждой страницы алгоритм HITS вычисляет два ранга: ранг $A(u)$, показывающий качество страницы как «авторитета», и ранг $H(u)$, показывающий качество страницы как «концентратора». Как и в алгоритме PageRank, в первом приближении рангам страниц присваивается произвольное ненулевое значение и затем производится итерационный процесс, состоящий из последовательного применения операций (2) и (3).

Несмотря на то, что поисковые машины второго поколения достигли более высокой точности, чем машины первого поколения, в дальнейшем их эффективность снизилась, были выявлены уязвимости перед недобросовестными методами манипулирования рейтингом (спамдексингом). Причина состояла в том, что в первоначальных вариантах алгоритмов степень значимости страниц определялась числом входящих гиперссылок. Другими словами, при вычислении показателя значимости страниц эти алгоритмы не учитывали тематической близости страниц. В настоящее время перспективно направление, связанное с комбинированным учетом информации о гиперссылочной связности web-страниц и результатов контентного анализа этих страниц.

Для улучшения качества ранжирования предложены некоторые модифицированные варианты алгоритмов PageRank и HITS. Далее приводятся некоторые модифицированные варианты алгоритма PageRank.

Например, при ранжировании web-страниц алгоритм Topic-Centric [5] учитывает меру близости страниц:

$$PR(u) = \frac{(1-d)}{n} + d \sum_{v \in B(u)} \frac{\mathrm{sim}(u, v)}{\sum_{x \in F(v)} \mathrm{sim}(x, v)} PR(v). \quad (4)$$

Согласно определению, $\mathrm{sim}(u, v) = 1$ соответствует максимальной близости страниц u и v, а $\mathrm{sim}(u, v) = 0$ — их полному различию.

В алгоритме TSPR (Topic-Sensitive PageRank) [6] web-страницы сначала группируются по тематике, а потом на каждом тематическом разделе вычисляется ранг страницы. Тематические разделы отбираются из верхнего уровня ODP (Open Directory Project). Пусть U_k означает множество URL-адресатов страниц в тематическом разделе T_k. При вычислении PageRank вектора для тематического раздела T_k алгоритм TSPR предполагает, что случайный пользователь двигается (следует по гиперссылкам или «прыгает») только по страницам из множества U_k, т. е. не выходит за рамки тематики. Тогда показатель PageRank страницы u на тематическом разделе T_k будет определен так:

$$PR_k(u) = d \sum_{v \in B(u)} \frac{PR_k(v)}{|F(v)|} + \begin{cases} (1-d)/|U_k|, \text{ если } u \in U_k, \\ 0, \text{ если } u \notin U_k. \end{cases}$$

Тогда зависящий от запроса показатель значимости (query sensitive importance score) страницы u

$$S_q(u) = \sum_k PR_k(u) r(q, T_k).$$

Подчеркнем, что результат поиска ранжируется с учетом этих счетов.

В модели Intelligent Surfer [7] для вычисления ранга страницы u, зависящего от запроса q, алгоритм PageRank преобразуется к следующему виду:

$$PR_q(u) = (1-d) \frac{r(q, u)}{\sum_{x \in W} r(q, x)} + d \sum_{v \in B(u)} \frac{r(q, u)}{\sum_{y \in F(v)} r(q, y)} PR_q(v).$$

В моделях Topical PageRank и Topical HITS [8] каждой странице u сопоставляется два вектора: контент-вектор $C(u)$ и авторитет-вектор $A(u)$.

Контент-вектор C_u: $[C(u^1), ..., C(u^i), ..., C(u^m)]$ является вероятностью распределения контента страницы u, где компонента $C(u^i)$ представляет собой относительный вклад i-го тематического раздела в контент страницы u.

Авторитет-вектор A_u: $[A(u^1), ..., A(u^i), ..., A(u^m)]$ определяет степень значимости страницы u по тематическим разделам, где компонента $A(u^i)$ определяет степень значимости страницы u относительно i-го тематического раздела (как известно, в алгоритме HITS, кроме авторитет-вектора определяется и концентратор-вектор).

Степень значимости $A(u^i)$ страницы u на i-м тематическом разделе вычисляется таким образом:

$$A(u^i) = d(1-\alpha) \sum_{v \in B(u)} \frac{1}{|F(v)|} A(v^i) + d\alpha \sum_{v \in B(u)} \frac{1}{|F(v)|} C(v^i) \sum_{k \in T} A(v^k) + \frac{(1-d)}{n} C(u^i) \sum_{v \in G} \sum_{k \in T} A(v^k), \quad (5)$$

где $i \in T = \{1, 2, ..., m\}$, α — вероятность перехода пользователя на i-й тематический раздел на странице.

Пусть $A(v) = \sum_{k \in T} A(v^k)$ и $\sum_{v \in G} A(v) = 1$, тогда формула (5) примет более компактный вид:

$$A(u^i) = d \sum_{v \in B(u)} \frac{(1-\alpha) A(v^i) + \alpha C(v^i) A(v)}{|F(v)|} + \frac{(1-d)}{n} C(u^i).$$

Когда на странице тематические переходы скрыты, т. е. при $\alpha = 0$, данная модель сводится к оригинальному PageRank-алгоритму.

В Topical HITS-алгоритме зависимость между векторами $A(u)$ и $H(u)$ задается следующими соотношениями:

$$A(u^i) = \sum_{v \in B(u)} \frac{(1-\alpha) H(v^i) + \alpha C(v^i) H(v)}{|F(v)|},$$

$$H(v^i) = \sum_{v \in F(u)} \frac{(1-\alpha) A(u^i) + \alpha C(u^i) A(u)}{|B(u)|},$$

где $H(v) = \sum_{k \in T} H(v^k)$ и $A(u) = \sum_{k \in T} A(u^k)$.

В случае, когда страница не разбита на тематические разделы, т. е. если на странице тематические детали скрыты, эта модель сводится к нормализованному алгоритму HITS или упрощенному варианту алгоритма SALSA [4]:

$$A(u) = \sum_{v \in B(u)} \frac{1}{|F(v)|} H(v),$$

$$H(v) = \sum_{v \in F(u)} \frac{1}{|B(v)|} A(u).$$

В работе [9] была предложена модель WPSS (Web Page Scoring Systems), обобщающая все вышеупомянутые модели. Прежде чем произвести

ранжирование, алгоритм WICER (Weighted Inter-Cluster Edge Rank) [10] предлагает кластеризацию страниц по тематике. Для уточнения весов гиперссылок в алгоритме WLRank (Weighted Links Rank) [11] предлагается учитывать различные атрибуты. А именно, полагается, что атрибуты — tag, anchor text и др. — дают гиперссылкам дополнительные веса, в результате чего улучшается точность поисковых машин.

Заметим, что во всех перечисленных работах главная цель — найти web-страницы с максимальным рангом, отвечающим потребностям пользователей. В этом контексте и работа [12] не исключение. Для нахождения web-страниц с максимальным рангом в ней, на основе теории потока в сетях, предложена оптимизационная модель ранжирования web-страниц, которая была сведена к задаче линейного программирования.

Изложенный краткий обзор позволяет сделать вывод, что исследование проблемы ранжирования web-страниц является перспективным направлением разработки технологий информационного поиска.

2. ТРИ МОДИФИКАЦИИ АЛГОРИТМА PAGERANK

Первая модификация. Как было отмечено, вероятность нахождения страницы с максимальным рангом непосредственно связана с выбором гиперссылок. В данной модификации для определения вероятности выбора гиперссылки используется точечная взаимная информация (Pointwise Mutual Information — PMI) [13]. При определении вероятности выбора гиперссылки точечной взаимной информацией учитывается, что если web-страницы указываются (цитируются) одной и той же страницей, то каждая гиперссылка (ссылочная страница) содержит долю информации о других гиперссылках (ссылочных страницах).

Пусть $F(v)$ — множество страниц, на которые ссылается страница v. Тогда, следуя работе [13], точечная взаимная информация между гиперссылками $v \to u_i$ и $v \to u_j$ определяется формулой:

$$PMI(v \to u_i,\ v \to u_j) = \log_2\left(\frac{p(u_i, u_j)}{p(u_i)p(u_j)}\right), \qquad (6)$$

где $p(u_i)$— вероятность цитирования страницы u_i (вероятность следования пользователя по гиперссылке $v \to u_i$), а $p(u_i, u_j)$ — вероятность коцитирования страниц u_i и u_j (совместная вероятность следования пользователя по гиперссылкам $v \to u_i$ и $v \to u_j$).

В дальнейшем для простоты записи вместо обозначения $PMI(v \to u_i,\ v \to u_j)$ будем применять обозначение $PMI(u_i, u_j)$.

Интуитивно, точечной взаимной информацией между гиперссылками $v \to u$ и $v \to u_j$ измеряется количество информации по отношению друг к другу.

Из определения (6) следует, что если вероятности выбора гиперссылок $v \to u_i$ и $v \to u_j$ независимы, то нет приращения информации. Это означает, что если гиперссылки независимы, то гиперссылка $v \to u_i$ не содержит никакой информации о гиперссылке $v \to u_j$, и наоборот. Следовательно, их точечная взаимная информация равна нулю, т. е. информация о гиперссылке $v \to u_i$ не дает никакой информации о гиперссылке $v \to u_j$ (и наоборот). Действительно, если гиперссылки $v \to u_i$ и $v \to u_j$ независимы, то $p(u_i, u_j) = p(u_i)p(u_j)$ и, следовательно,

$$PMI(u_i, u_j) = \log_2\left(\frac{p(u_i, u_j)}{p(u_i)p(u_j)}\right) = \log_2 1 = 0.$$

Для определения вероятности выбора гиперссылки $v \to u_i$ вычисляется суммарная точечная взаимная информация $PMI(u_i)$, которая определяется между ней и остальными исходящими гиперссылками страницы v, и общая точечная взаимная информация $PMI(F(v))$, которая вычисляется между всевозможными парами гиперссылок, исходящими из страницы v.

Суммарную точечную взаимную информацию гиперссылки $v \to u_i$ будем определять формулой:

$$PMI(u_i) = \sum_{\substack{u_j \in F(v) \\ u_j \neq u_i}} PMI(u_i, u_j). \qquad (7)$$

Общая точечная взаимная информация $PMI(F(v))$ получается суммированием формулы (7):

$$PMI(F(v)) = \sum_{u_i \in F(v)} PMI(u_i) = \sum_{u_i \in F(v)} \sum_{\substack{u_j \in F(v) \\ u_j \neq u_i}} PMI(u_i, u_j). \qquad (8)$$

Тогда доля ранга $PR(v)$, которая распределяется по страницам $u_i \in F(v)$, т. е. вес гиперссылки $v \to u_i$ будет вычисляться отношением формул (7) и (8):

$$\omega(v \to u_i) = \frac{PMI(u_i)}{PMI(F(v))} = \frac{\sum\limits_{\substack{u_j \in F(v) \\ u_j \neq u_i}} PMI(u_i, u_j)}{\sum\limits_{u_i \in F(v)} \sum\limits_{\substack{u_j \in F(v) \\ u_j \neq u_i}} PMI(u_i, u_j)}, \quad u_i \in F(v). \qquad (9)$$

Легко видеть, что если вероятности выбора гиперссылок $v \to u_i$ и $v \to u_j$ независимы, то приходим к неопределенности. Во избежание неопределенности формула (9) преобразуется к виду:

$$\omega(v \to u_i) = \frac{1 + \sum\limits_{\substack{u_j \in F(v) \\ u_j \neq u_i}} PMI(u_i, u_j)}{|F(v)| \sum\limits_{u_i \in F(v)} \sum\limits_{\substack{u_j \in F(v) \\ u_j \neq u_i}} PMI(u_i, u_j)},$$

$$u_i \in F(v). \qquad (10)$$

Таким образом, оригинальный PageRank-алгоритм имеет вид:

$$PR(u) = \frac{(1-d)}{n} + d \sum_{v \in B(u)} \omega(v \to u) PR(v). \qquad (11)$$

Из формулы (11), как следствие, легко можно получить оригинальный PageRank-алгоритм. Действительно, если предположить, что выбор гиперссылок независим, тогда из формулы (10) вытекает, что вероятность выбора каждой гиперссылки равна величине $\omega(v \to u_i) = 1/|F(v)|$, которая совпадает с вероятностью выбора гиперссылки в оригинальном PageRank-алгоритме, выраженным формулой (1).

Теперь переходим к вычислению вероятности $p(u_i, u_j)$, которую можно определить так:

$$p(u_i, u_j) = p(u_j|u_i)p(u_i). \qquad (12)$$

Поскольку каждая страница представляется как «мешок слов», то условная вероятность

$$p(u_j|u_i) = \sum_{k=1}^{K} p(u_j|u_i, w_k)p(w_k|u_i),$$

где K означает общее число слов в наборе страниц $\{v\} \cup F(v)$.

При допущении независимости появления слова w_k в страницах u_i и u_j последнюю формулу можно выразить так:

$$p(u_j|u_i) = \sum_{k=1}^{K} p(u_j|u_i, w_k)p(w_k|u_i) \approx \sum_{k=1}^{K} p(u_j|w_k)p(w_k|u_i),$$

и формула (12) принимает вид:

$$p(u_i, u_j) = p(u_i) \sum_{k=1}^{K} p(u_j|w_k)p(w_k|u_i). \qquad (13)$$

Пусть f_{ik} — число появления слова w_k в странице u_i. Тогда вероятность появления слова w_k в странице u_i

$$p(w_k|u_i) = f_{ik} / \sum_{s=1}^{K} f_{is}. \qquad (14)$$

Условная вероятность $p(u_j|w_k)$ определяется согласно формуле Байеса:

$$p(u_j|w_k) = \frac{p(w_k|u_j)p(u_j)}{p(w_k)}, \qquad (15)$$

где $p(w_k) = \sum\limits_{i=1}^{|F(v)|} p(w_k|u_i)p(u_i)$.

С учетом формул (13)—(15)

$$p(u_i, u_j) = \frac{p(u_i)p(u_j)}{\sum\limits_{s=1}^{K} f_{is} \sum\limits_{s=1}^{K} f_{js}} \sum_{k=1}^{K} \frac{f_{ik}f_{jk}}{p(w_k)}.$$

Подставляя последнее выражение в формулу (6), для точечной взаимной информации получим следующую формулу:

$$PMI(u_i, u_j) = \log_2 \left(\frac{1}{\sum\limits_{s=1}^{K} f_{is} \sum\limits_{s=1}^{K} f_{js}} \sum_{k=1}^{K} \frac{f_{ik}f_{jk}}{p(w_k)} \right).$$

Вторая модификация. В работе [14] показано, что при вычислении PageRank-вектора следует учитывать не только входящие гиперссылки, но и исходящие. Этот алгоритм, как и все перечисленные модификации алгоритма PageRank, позволяет избегать равномерного распределения ранга страницы между ссылочными страницами. Здесь для каждой гиперссылки вычисляются два веса: $\omega^+(v \to u)[0, 1]$ и $\omega^-(v \to u)$.

Вес $\omega^+(v \to u)$ определяется числом входящих гиперссылок страницы u и ссылочных страниц страницы v:

$$\omega^+(v \to u) = \frac{|B(u)|}{\sum\limits_{x \in F(v)} |B(x)|}, \qquad (16)$$

а вес $\omega^-(v \to u)$ — числом исходящих гиперссылок страницы u и всех страниц множества $F(v)$:

$$\omega^-(v \to u) = \frac{|F(u)|}{\sum\limits_{x \in F(v)} |F(x)|}. \qquad (17)$$

С учетом формул (16) и (17) оригинальный PageRank-алгоритм модифицируется так [14]:

$$PR(u) = \frac{(1-d)}{n} + d \sum_{v \in B(u)} \omega^+(v \to u)\omega^-(v \to u) PR(v).$$

Как можно заметить из формул (16) и (17), в этой модификации не учитывается тематическая близость страниц. Для учета тематической близос-

ти страниц нами предлагаются следующие определения весов (16) и (17):

$$\omega^{+}(v \to u) = \frac{\sum\limits_{x \in B(u)} \mathrm{sim}(u, x)}{\sum\limits_{y \in F(v)} \sum\limits_{z \in B(y)} \mathrm{sim}(y, z)}, \qquad (18)$$

$$\omega^{-}(v \to u) = \frac{\sum\limits_{x \in F(u)} \mathrm{sim}(u, x)}{\sum\limits_{y \in F(v)} \sum\limits_{z \in F(y)} \mathrm{sim}(y, z)}. \qquad (19)$$

Если предположить, что для любой гиперссылки $(x \to y)$ мера подобия между страницами x и y принимает одно и то же значение, то из нашего определения, как следствие, получается результат работы [14]. Действительно, пусть для любой гиперссылки $(x \to y)$ $\mathrm{sim}(x, y) = a = \mathrm{const}$, тогда из формул (18) и (19) получаются соответствующие формулы (16) и (17).

Третья модификация. Предлагаемая модификация является усовершенствованным вариантом алгоритма Topic-Centric [5]. В отличие от алгоритма Topic-Centric в нашем варианте вероятность выбора гиперссылки не только зависит от степени близости страниц v и $u \in F(v)$, она также зависит от степени близости страниц множества $F(v)$:

$$PR(u) = \frac{(1-d)}{n} + d \sum_{v \in B(u)} \left(\frac{\lambda \mathrm{sim}(u, v) + (1-\lambda) \sum\limits_{\substack{x \in F(v) \\ x \neq u}} \mathrm{sim}(u, x)}{\lambda \sum\limits_{x \in F(v)} \mathrm{sim}(x, v) + (1-\lambda) \sum\limits_{x \in F(v)} \sum\limits_{\substack{y \in F(v) \\ y \neq x}} \mathrm{sim}(x, y)} \right) \times PR(v), \qquad (20)$$

где $0 \leqslant \lambda \leqslant 1$.

Если влияние мер близости страниц множества $F(v)$ свести к нулю, т. е. если в формуле (20) положить $\lambda = 1$, то получается формула (4).

Можно предложить и другую модификацию:

$$PR(u) = \frac{(1-d)}{n} + d \sum_{v \in B(u)} \left(\frac{\mathrm{sim}(u, v) \sum\limits_{\substack{x \in F(v) \\ x \neq u}} \mathrm{sim}(u, x)}{\sum\limits_{x \in F(v)} \mathrm{sim}(x, v) \cdot \sum\limits_{x \in F(v)} \sum\limits_{\substack{y \in F(v) \\ y \neq x}} \mathrm{sim}(x, y)} \right) PR(v).$$

ЗАКЛЮЧЕНИЕ

Исследования последних лет показали, что точность алгоритмов анализа гиперссылок — HITS, PageRank и др. — непосредственно зависит от выбора гиперссылки. Другими словами, степень точности ранжирования результатов поиска непосредственно зависит от вероятности выбора гиперссылки. В оригинальном PageRank-алгоритме и в некоторых других его модификациях вероятности выбора гиперссылок считались равными, т. е. гиперссылки выбирались с одинаковой вероятностью. Проведенные эксперименты показали, что такой подход не гарантирует нахождение релевантных страниц, отвечающих потребностям пользователей. Одна из главных причин заключается в том, что в традиционных алгоритмах ранг страницы определяется числом гиперссылок. Дальнейшие исследования подтверждают, что без контентного анализа документов невозможно решить проблему эффективности поисковых машин. Эффективность поиска оценивается степенью релевантности отобранных документов к запросу пользователя. Для повышения эффективности поиска в работе предложены три модификации алгоритма PageRank. Каждая модификация при определении веса гиперссылки учитывает тематическую близость страниц и их соседей, связанных гиперссылками. Предложенные в данной статье модификации алгоритма PageRank представляют собой усовершенствованные варианты результатов работ [1, 2, 5, 14].

ЛИТЕРАТУРА

1. *Brin S., Page L.* The anatomy of a large-scale hyper-textual Web search engine // Computer Networks and ISDN systems. — 1998. — Vol. 30, — N 1—7. — P. 107—117.
2. *Berkhin P.* A survey on PageRank computing // Internet Mathematics. — 2005 — 2006. — Vol. 2, N 1. — P. 73—120.
3. *Kleinberg J.M.* Authoritative sources in a hyperlinked environment // Journal of the ACM. — 1999. — Vol. 46, N 5. — P. 604—632.
4. *Lempel R., Moran S.* SALSA: the stochastic approach for link-structure analysis // ACM Trans. on Information Systems. — 2001. — Vol. 19, N 2. — P. 131—160.
5. *Ingongngam P., Rungsawang A.* Topic-centric algorithm: a novel approach to Web link analysis // Proc. of the 18th Intern. Conf. on Advanced Information Networking and Applications (AINA'04). — Fukuoka, Japan, 2004. — Vol. 2. — P. 299— 301.
6. *Haveliwala T.H.* Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search // IEEE Trans. on Knowledge and Data Eng. — 2003. — Vol. 15, N 4. — P. 784—796.
7. *Richardson M., Domingos P.* The intelligent surfer: probabilistic combination of link and content information in PageRank // Advances in Neural Information Processing Systems. MIT Press. — 2002. — Vol. 14. — P. 1441—1448.
8. *Nie L., Davison B. D., Qi X.* Topical link analysis for Web search // Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — Seattle, USA, 2006. — P. 91—98.
9. *Diligenti M., Gori M., Maggini M.* A unified probabilistic framework for Web page scoring systems // IEEE Trans. on Knowledge and Data Engineering. — 2004. — Vol. 16, N 1. — P. 4—16.
10. *Padmanabhan D., Desikan P., Srivastava J.* WICER: a weighted inter-cluster edge ranking for clustered graphs // Proc. of the 2005 IEEE/WIC/ACM Intern. Conf. on Web Intelligence (WI'2005). — Compiegne, France, 2005. — P. 522—528.
11. *Baeza-Yates R., Davis E.* Web page ranking using link attributes // Proc. of 13th World Wide Web Conference (WWW13). — New York, USA, 2004. — P. 328—329.
12. *Алыгулиев Р.М.* Оптимизационная модель ранжирования Web-страниц // Системы управления и информационные технологии. — 2006. — № 3(25). — С. 4—7.
13. *Efron M.* Using cocitation information to estimate political orientation in Web documents // Knowledge and Information Systems. — 2006. — Vol. 9, N 4. — P. 492—511.
14. *Xing W., Ghorbani A.* Weighted PageRank algorithm // Proc. of the Second Annual Conf. Communication Networks and Services Research (CNSR'04). — Fredericton, Canada, 2004. — P. 305—314.

e-mail: rasim@science.az; a.ramiz@science.az

Статья представлена к публикации членом редколлегии В.Л. Эпштейном. ❑