Целых Александр Николаевич – Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге; e-mail: inf@tsure.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634310642; заместитель руководителя по информатике; кафедра прикладной информатики; д.т.н., профессор.

Чичерина Карина Сергеевна – e-mail: kchicherina@mail.ru; тел.: +79515275815; кафедра прикладной информатики; аспирантка.

Tselykh Alexander Nicolaevich – Taganrog Institute of Technology – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University"; e-mail: inf@tsure.ru; 44, Nekrasovsky, Taganrog, 347928, Russia; phone: +78634310642; deputy head of informatics; the department of applied information science; dr. of eng. sc.; professor.

Chicherina Karina Sergeevna – e-mail: kchicherina@mail.ru; phone: +79515275815; the department of applied information science; postgraduate student.

УДК 519.14

Э.М. Котов

МЕТОДЫ АНАЛИЗА ГИПЕРССЫЛОК ПРИ ИНФОРМАЦИОННОМ ПОИСКЕ В ГЛОБАЛЬНОЙ СЕТИ

Проведены исследования методов анализа гиперссылок, выделены два основных класса ранжирования на основе связей: методы глобального анализа – ранжирование, независящие от запроса, и методы локального анализа – ранжирование, зависящие от запроса. Дано описание и характеристика алгоритма PageRank в основу которого заложена модель случайного блуждания по веб-графу, которая используется для вычисления веса страницы (коэффициент PageRank) как вероятности ее достижимости и алгоритма HITS-поиска документов по заданной теме на базе гиперссылок, в основе которого лежит идентификация двух наборов страниц, которые могут быть важными: страницы-«авторитеты» и страницы-«концентраторы». Выявлены различия данных двух подходов к анализу гиперссылок.

Информационный поиск; ранжирование результатов поиска; анализ гиперссылок.

E.M. Kotov

METHODS OF THE ANALYSIS OF HYPERLINKS BY INFORMATION RETRIEVAL IN A GLOBAL NETWORK

This article describes a methods of the analysis of hyperlinks, two basic classes of ranging on the basis of communications are allocated: methods of the global analysis – ranging independent of inquiry and methods of the local analysis-ranging depending on inquiry. The description and the characteristic of algorithm PageRank in which basis is given the model of casual wandering under the web-graph who is used for calculation of weight of page (factor PageRank) is put in pawn as probabilities of its approachibility and algorithm HITS-search of documents in the set theme on the basis of hyperlinks in which basis identification of two sets of pages which can be important lays: pages "Hub" pages and pages "Authority" pages. Distinctions of the given two approaches to the analysis of hyperlinks are revealed.

Information retrieval; ranging of results of retrieval; the analysis of hyperlinks.

На сегодняшний день, наряду с возможностью иметь доступ к огромному объему информации, практически любого характера, возникает ряд проблем, связанных с извлечением адекватной информации из столь больших массивов, организацией поиска и классификации информационных ресурсов. Механизм, реали-

зованный в большинстве информационно-поисковых систем (ИПС), построенный на основе совпадения ключевых слов запроса и документа, оказывается малоэффективным, в связи с тем, что пользователь получает список формально релевантных документов, которые, в свою очередь, могут являться чрезмерно большими для просмотра.

Анализ гиперссылок значительно увеличивает релевантность результатов поиска, причем настолько, что все ведущие механизмы поиска в глобальной сети в той или иной степени используют различные методы анализа гиперссылок.

Статическая часть глобальной сети, состоящая из html-документов и гиперссылок между ними может быть представлена в виде направленного графа, в котором каждый узел является веб-страницей, а каждое направленное ребро – гиперссылкой. Совокупность таких узлов и направленных ребер называется веб-графом [1]. Ссылки в веб-графе не распределены случайным образом.

В основу методов анализа гиперссылок положено одно либо несколько из следующих допущений:

1) гиперссылка со страницы *A* на страницу *B* – это своего рода рекомендация автора страницы *A* – так называемое «допущение о рекомендательности»;

2) если страница *A* и страница *B* связаны гиперссылкой, они могут быть посвящены одной и той же теме, т.е. с большей вероятностью они относятся к одной и той же тематике нежели к разным – так называемое «допущение о тематической локальности»;

3) текст, связанный с анкерным тэгом (<*a*>) гиперссылки, описывает целевой документ, на который указывает гиперссылка – так называемое «допущение об анкерном описании».

Методы анализа гиперссылок используются для косвенной оценки качества документов. Схемы ранжирования на основе связей можно разделить на два класса:

- методы глобального анализа (ранжирование, не зависящее от запроса), например метод, использующий алгоритм PageRank.
- методы локального анализа (ранжирование, зависящее от запроса), например метод использующий алгоритм HITS.

Ранжирование, не зависящее от запроса, служит для определения «истинного» качества страниц. Алгоритм PageRank был предложен С. Брином и Л. Пейджем [2, 3] и использован для ранжирования в ИПС Google.

В основу данного алгоритма заложена модель случайного блуждания по веб-графу, которая используется для вычисления веса страницы (коэффициент PageRank) как вероятности ее достижимости.

Выполняется вычисление коэффициента PageRank каждой страницы, присваивая каждой ссылке на страницу весовой коэффициент, пропорциональный качеству страницы, содержащей гиперссылку. Чтобы определить качество ссылающейся страницы, используются ее коэффициенты PageRank рекурсивно, причем первоначальные значения PageRank задаются произвольно. В итоге страница имеет высокий коэффициент PageRank, если на нее ссылаются страницы с высоким коэффициентом PageRank.

Коэффициент PageRank для текущей веб-страницы можно определить по формуле:

$$PR(a) = \frac{d}{n} + (1-d) \cdot \sum_{(b,a) \in G} \frac{PR(b)}{C(b)}, \qquad (1)$$

где d – коэффициент настройки – константа, значение которой выбирается в пределах от 0,1 до 0,2; n – количество страниц (число узлов) в графе G; $C(b)$ – количество исходящих гиперссылок (число рёбер) со страницы b.

Для вычисления *PR(a)* требуется рекурсивная процедура, которая продолжается до достижения сходимости, и коэффициенты PageRank рассчитываются только один раз и не зависят от конкретных запросов, что в итоге позволяет эффективно отличить высококачественные страницы глобальной сети от низкокачественных.

Пусть есть *n* страниц *T={T1, T2, …, Tn}*, которые ссылаются на данный документ (веб-страницу *A*) , а *C(A)* – общее число ссылок с веб-страницы *A* на другие документы. Пусть *d* – это вероятность того, что пользователь, пересматривая какую-нибудь страницу из множества *T*, перейдет на страницу *A* по ссылке, а не другими средствами. Тогда вероятность продолжения перехода путем ручного введения адреса из случайной страницы будет составлять *1 – d,* а коэффициент PageRank для страницы *A* вычисляется по указанной выше формуле (1).

Во время обработки запроса показатель PageRank используется для ранжирования всех документов, соответствующих условиям запроса, как с учетом, так и без учета критерия ранжирования, зависящего от запроса.

Значения весов PageRank не зависят от запроса пользователя. Таким образом, вес PageRank является мерой статического качества веб-страницы, не зависящей от запросов пользователей. Предположим следующее утверждение – «результат ранжирования должен зависеть от запроса».

При ранжировании, зависящем от запроса, алгоритм присваивает показатель, который изменяет качество и адекватность выбранного множества страниц для данного запроса пользователя. Основная идея состоит в том, чтобы создать граф для каждого конкретного запроса, называемый графом соседей, и выполнить по этому графу анализ гиперссылок.

Алгоритм поиска документов по заданной теме на базе гиперссылок (hyperlink-induced topic search – HITS) был предложен Дж. Кляйнбергом. HITS является методом, в основе которого лежит идентификация двух наборов страниц, которые могут быть важными: страницы-«авторитеты» и страницы-«концентраторы» [4]. У страниц «авторитеты» и «концентраторы» есть взаимно укрепляющие отношения – хорошая страница-«концентратор» связывается со многими страницами-«авторитетами», и хорошая страница-«авторитет» связана со многими страницами-«концентраторами». Иными словами: хорошие авторитеты – страницы, которые содержат релевантную информацию (хорошие источники информации), и хорошие концентраторы – страницы, ссылающиеся на нужные страницы (хорошие источники ссылок). В результате метод HITS обеспечивает выбор из информационного пространства лучших «авторов» (первоисточников) и «посредников» (документов, от которых идут ссылки цитирования). То есть страница является хорошим посредником, если она содержит ссылки на ценные первоисточники, и наоборот, страница является хорошим первоисточником, если она упоминается хорошими посредниками.

На основе ранжированной выборки по запросу пользователя формируется стартовое множество *S* документов. Путем использования входящих и исходящих ссылок на документы из *S* строится расширенное множество *T* документов, находящихся на расстоянии одного ребра от стартовых узлов в веб-графе.

Простой учет количества входящих и исходящих ссылок на документы не является эффективным, поэтому далее следует итерационная процедура расчета показателей авторитетности и концентрации для всех узлов множества *T*. Это отражает следующий алгоритм:

1. Пусть *N* – множество узлов в графе соседей.
2. Для каждого узла *u* из *N a(v)* является весом авторитетности.
3. Для каждого узла *u* из *N h(v)* является весом концентрации.
4. Для всех узлов *u* из *N* все веса инициализируются значением 1.

5. Повторяется цикл до достижения сходимости.
6. Для всех узлов *u* из *N* рассчитывается, по формуле (2), вес авторитетности:

$$a\left(v\right)=\sum_{u_i\to v} h\left(u_i\right). \tag{2}$$

7. Для всех узлов *u* из *N* рассчитывается, по формуле (3), вес концентрации:

$$h\left(v\right)=\sum_{v\to u_i} a\left(u_i\right). \tag{3}$$

8. После каждой итерации выполняется нормализация весов.

Пусть *A* – квадратная матрица смежности подмножества веб-графа, в которой каждой странице соответствует одна строка и один столбец. Элемент A_{ij} равен единице, если существует гиперссылка со страницы *i* на страницу *j*, и нулю – в противном случае. Тогда формулы (2) и (3) запишем следующим образом (4):

$$\begin{aligned}\vec{h} &\leftarrow A\vec{a},\\ \vec{a} &\leftarrow A^T\vec{h},\end{aligned} \tag{4}$$

после взаимной подстановки получаем:

$$\begin{aligned}\vec{h} &\leftarrow AA^T\vec{h},\\ \vec{a} &\leftarrow A^TA\vec{h}.\end{aligned} \tag{5}$$

Введем λ_h – собственное значение матрицы AA^T и λ_a – собственное значение матрицы A^TA, и получим уравнения для собственных векторов:

$$\begin{aligned}\vec{h} &= \frac{1}{\lambda_h}AA^T\vec{h},\\ \vec{a} &= \frac{1}{\lambda_a}A^TA\vec{h}.\end{aligned} \tag{6}$$

Отметим следующие следствия: 1) результат итеративного вычисления стремится к стационарному значению, определенному структурой графа; 2) для вычисления результата можно применять любой метод вычисления собственного вектора стохастической матрицы.

Существует ряд ограничений модели HITS, рассматриваемых в работе [5]:

1. При использовании части веб-графа добавление ребер к нескольким узлам может сильно изменить конечный результат.
2. В большей степени подвержен манипулированию.
3. Взаимное усиление между хостами. Происходит, когда ряд документов относительно одного хоста указывает на единственный документ относительно второго хоста.
4. Динамически генерируемые ссылки.
5. Возможность попадания нерелевантных, но сильно связанных документов.
6. Нерелевантные узлы. Возникает ситуация, когда локальный подграф расширен, чтобы включать окружающие связи, и в результате страницы, не относящиеся к начальному вопросу, включены в граф, и как следствие происходит смещение темы.

Выделим существующие различия между рассмотренными методами PageRank и HITS:

1. Алгоритм PageRank вычисляет веса для всех веб-страниц (которые были проиндексированы) ещё до процедуры выполнения запроса пользователя. Алгоритм HITS применяется только к веб-страницам, полученным в результате выполнения определенного запроса пользователя.
2. Алгоритм HITS находит как «авторитеты», так и «концентраторы», PageRank – только «авторитеты».
3. Алгоритм PageRank требует нетривиальных вычислений, HITS – простой алгоритм, но очень затратный по времени вычисления.

Наиболее эффективными признаками для увеличения качества поиска являются признаки, основывающиеся на анализе ссылочной структуры веб-ресурсов, но в коллекциях, не обладающих данной структурой, можно получить улучшение качества поиска с использованием других признаков, подсчитываемых для целого документа или некоторых его атрибутов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Manning C.D., Raghavan P., Schütze H.* Introduction to information retrieval // Cambridge University Press. – 2008. – 544 p.
2. *Brin S., Page L.* The Anatomy of a Large-Scale Hypertextual Web Search Engine, Proc. Seventh World Wide Web Conf., Elsevier Science. – New York, 1998.
3. *Page L. et al.* The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies, Working Paper 1999-0120, Stanford Univ., Palo Alto, Calif., 1998.
4. *Kleinberg J.M.* Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46, 5, 1999. – P. 604-632.
5. *Bhart K., Henzinger M.* Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In Proceedings of ACM SIGIR'98 (Melbourne, Australia), 1998.

Статью рекомендовал к публикации д.т.н., профессор В.П. Карелин.

Котов Эдуард Михайлович – Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге; e-mail: kotov@tti.sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371743; кафедра прикладной информатики; старший преподаватель.

Kotov Eduard Michaylovich – Taganrog Institute of Technology – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University"; e-mail: kotov@tti.sfedu.ru; 44, Nekrasovsky, Taganrog, 347928, Russia; +78634371743; the department of applied information science; senior instructor.

УДК 519.14

А.В. Боженюк, Н.С. Опенько

ИССЛЕДОВАНИЕ И АНАЛИЗ МЕТОДОВ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ НЕЧЕТКОЙ ИНФОРМАЦИИ*

Описываются методы принятия решений на основе построения и обоснования механизмов нечеткого логического вывода. Данная задача является актуальной, потому что имеет широкое практическое применение. Рассматриваются основные схемы нечеткого вывода на основе методов Мамдани и Сугено, описываются их недостатки. Рассмотрен метод выбора решений на основе истинности правила modus ponens. Построена модель принятия решений на основе степени истинности правила modus ponens, использованная

* Работа поддержана РФФИ, проект № 11-01-00011а.