УДК 004.55

Н.Г. Аксак[1], С.А. Коргут[2], Н.Г. Стрельцова[3]
[1]ХНУРЕ, м. Харків, Україна, axak@kture.kharkov.ua;
[2]ХНУРЕ, м. Харків, Україна, korguts@gmail.com;
[3]ХНУРЕ, м. Харків, Україна, lexy@email.ua

РАСШИРЕННЫЙ BROWSERANK НА ОСНОВЕ УЧЕТА ПОВЕДЕНЧЕСКИХ ФАКТОРОВ ПОЛЬЗОВАТЕЛЕЙ

В работе предлагается модификация метода BrowseRank, отличающаяся от существующего введением качественного коэффициета, учитывающего дополнительную информацию о поведении пользователей, что позволяет более точно определять важность веб-страницы. Приведены результаты экспериментального моделирования, подтверждающие эффективность предложенного подхода.

PAGERANK, BROWSERANK, ДАННЫЕ О ПОВЕДЕНИИ ПОЛЬЗОВАТЕЛЕЙ, ПОВЕДЕНЧЕСКИЙ ГРАФ, ЦЕПИ МАРКОВА, Q-МАТРИЦЫ

Введение

Вес или важность страницы, отражающая ее «ценность» для посетителей Интернет, является ключевым фактором в алгоритмах веб-поиска, потому что для современных поисковых систем сканирование, индексирование и ранжирование документов во многом зависит именно от этого фактора [1].

Вес страницы в большинстве современных поисковых систем вычисляется при помощи графа переходов PageRank, либо основанном на нем BrowseRank, и известен как процесс анализа вебссылок и данных о поведении пользователя [4,10]. Большинство подобных алгоритмов предполагают, что если большое количество значимых страниц ссылается на документ – то он, по всей вероятности, тоже важен, и значимость вычисляется по графу переходов BrowseRank, который учитывает некоторые данные о поведении пользователя.

Из-за стремительного роста объема информации в Интернет, точный расчет веса веб-страницы становится критическим и является серьезной проблемой для поисковых систем [2]. Кроме того, проблемой для поисковых систем является ссылочный спам [7] – создание веб-страниц с целью вводить в заблуждение поисковые системы. Это страницы, в основном созданные в коммерческих целях, использующие различные методы для достижения более высокого рейтинга на страницах поисковых результатов.

Поэтому критичными являются точность определения важности страницы и стойкость метода вычисления к ссылочному спаму. В работе предлагается модификация метода BrowseRank для решения указанных проблем.

1. Цель и задачи исследования

Целью работы является анализ методов вычисления важности Web-документа для улучшения работы поисковых систем.

Для постижения поставленной цели необходимо решить следующие задачи:

– исследовать наиболее популярные методы вычисления значимости веб-страниц;

– исследовать влияние доступной информации о поведении пользователя для повышения релевантности документов.

2. Обзор существующих методов вычисления важности страницы

PageRank (PR) представляет собой метод вычисления веса страницы путём подсчёта важности ссылок на неё [5, 9, 12, 14], который применяется к коллекции веб-документов, связанных гиперссылками и назначает каждому из них некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов, и представляется числом от 0 до 10. Чем больше ссылок на страницу, тем она становится «значительнее» [3]. Кроме того, вес страницы A определяется весом ссылки, передаваемой страницей B:

$$W_{PR}(A) = (1-d) + d\left(\frac{W_{PR}(X_1)}{C(X_1)} + \ldots + \frac{W_{PR}(X_n)}{C(X_n)}\right),$$

где $W_{PR}(A)$ – вес PageRank анализируемой страницы A; d — “damping factor” или коэффициент затухания в диапазоне от 0 до 1, который обычно устанавливают равным 0.85-0.9, выражает вероятность дальнейшего путешествия пользователя по ссылкам; $W_{PR}(X_i)$ — вес PageRank страницы X_i, указывающей на страницу A; $C(X_i)$ — число ссылок со страницы X_i.

Однако PR это только один из показателей, которые используются современными поисковыми системами для определения релевантности и важности страницы. PR отражает приблизительное качество страницы, но не связан с ее тематической релевантностью и не учитывает поведенческие факторы. Основной недостаток данного метода – его неустойчивость перед спамом (искусственной накруткой).

BrowseRank – метод корпорации Microsoft позволяющий вести учет поведения пользователей при ранжировании документов в результатах поиска [11].

Основное отличие BrowseRank от PageRank заключается в структуре графа. Если в технологии

PageRank узлами графа служат документы HTML, а ребрами – ссылки, то в технологии BrowseRank в качестве ребер выступают «клики» или количество переходов. Помимо этого, сохраняется вся мета-информация о длительности сессии в рамках того или иного ресурса.

Метод BrowseRank использует цепи Маркова на графе переходов для вычисления важности страницы, по сути, имитируя случайный «веб-серфинг» пользователя по ссылкам сайтов с учетом времени посещения страницы и типом перехода (по ссылке, по вводу URL).

Перед PageRank у этого метода есть следующие преимущества: большая устойчивость к ссылочному спаму и определение продолжительности сессии, что позволяет более надежно оценить, насколько документ важен для посетителя.

Данные о поведении пользователя в BrowseRank представлены в тройках, состоящих из <URL, Time, Type>. Здесь URL содержит адрес веб-страницы, которую посещал пользователь, Time означает время посещения страницы, а Type указывает, как пользователь попал на данную страницу путем ввода URL в браузере (INPUT), либо посредством перехода по ссылке (CLICK). Записи сортируются в хронологическом порядке.

На их основе строится поведенческий граф:

$$G = < V, W, T, \delta >,$$

где $V = \{v_i\}$ – множество вершин; $W = \{w_{ij}\}$ – множество весов; $T = \{t_i\}$ – время посещения; $\delta = \{\delta_i\}$ – вероятность распределения; $(i, j = 1,..., N)$ – количество веб-страниц в графе переходов.

Модель. В модели BrowseRank для представления случайного веб-серфинга на графе переходов используется Марковская цепь с непрерывным временем [11].

X_t – страница, которую пользователь посетил во время $t, (t > 0)$. Процесс $X = \{X_t, t \ge 0\}$ является процессом Маркова с непрерывным временем. $P_{ij}(t)$ – это вероятность перехода со страницы i на j для интервала времени t в этом процессе. Существует стационарное распределение вероятностей π, которое не зависит от t и связано с $P(t) = [p_{ij}(t)]_{N \times N'}$ такое, что для каждого $t > 0$

$$\pi = \pi P(t).$$

Каждое *i*-е вхождение распределения π обозначает отношение времени нахождения пользователя на *i*-ой странице к общему времени, проведенному на всех страницах, т.е. распределение является показателем важности страницы.

Алгоритм. Распределение вероятностей, служащее показателем веса страницы, вычисляется при помощи матрицы интенсивностей переходов. Матрица интенсивностей переходов определяется как производная $P(t)$ при *t*, стремящемся к 0, т.е. $Q = P'(0)$, матрица $Q = (Q_{ij})_{N \times N}$ далее – Q-матрица [6].

При конечном пространстве состояний соответствие между Q-матрицей и $P(t)$ является однозначным и $-\infty < q_{ij} < 0$; $\sum_j q_{ij} \to 0$. Соответственно Q процесс является Марковским процессом с непрерывным временем, то есть процесс перехода по ссылкам $X = \{X_t, t \ge 0\}$ является Q-процессом из-за конечности пространства состояний. Q-процесс относится к модели с вложенной Марковской цепью. Так называемая вложенная Марковская цепь – это Марковский процесс с дискретным временем, представленный матрицей вероятностей переходов с нулевыми значениями во всех диагональных позициях и $-\frac{q_{ij}}{q_{ii}}$ во всех остальных позициях, где все параметры $q_{ij}, i, j = 1,..., N$ имеют те же значения, как и раньше.

X представлен как процесс, а Y как вложенная Марковская цепь, полученная из Q-матрицы. А $\pi = (\pi_1,..., \pi_N)$ и $\bar{\pi} = (\bar{\pi}_1,..., \bar{\pi}_N)$ обозначают стационарную вероятность распределений процессов X и Y:

$$\pi_i = \frac{\frac{\bar{\pi}_i}{q_{ii}}}{\sum_{j=1}^{N} \frac{\bar{\pi}_i}{q_{jj}}}. \qquad (1)$$

Вычисление q_{ii}**.** Для Q-процесса время просмотра страницы t_i для i-ой вершины графа регулируется экспоненциальным распределением с параметром q_{ii}

$$P(t_i > t) = \exp(q_{ii} t).$$

Это означает, что определяется на основе большого числа значений времени просмотра страниц, взятых из данных о поведении пользователя. Эта задача не является тривиальной, потому что данные о пользователе обычно содержат погрешности, и для решения этой проблемы используется дополнительная модель погрешностей для получения объективных значений параметра .

Для страницы i используются значения времени посещения $Z_1, Z_2,..., Z_{mi}$ имеющие такое же распределение, как случайная переменная Z. Предполагается, что эта Z является комбинацией реально проведенного на странице времени t_i и погрешностей , т. е.:

$$Z = u + t_i.$$

Погрешность u регулируется распределением Хи-квадрат как $x^{(k)}$, его среднее значение будет k, а дисперсия – $2k$. Среднее значение и дисперсия Z равны μ и δ^2 соответственно. Поскольку u и t_i не зависят друг от друга, получается:

$$\mu = E(Z) = E(u + t_i) = k - \frac{1}{q_{ij}},$$

$$\delta^2 = Var(Z) = Var(u + t_i) = 2k + \frac{1}{q_{ii}^2}.$$

Выборочное среднее $\overline{Z} = \frac{1}{m_i}\sum_{l=1}^{m_i} Z_l$ и выборочная дисперсия $S^2 = \frac{1}{m_i - 1}\sum_{l=1}^{m_i}(Z_l - \overline{Z})^2$ являются объективными оценками для μ и δ^2.

Далее q_{ii} вычисляется через решение следующей задачи оптимизации:

$$\min_{q_{ii}}((\overline{Z} + \frac{1}{q_{ii}}) - \frac{1}{2}(S^2 - \frac{1}{q_{ii}^2})),$$

$$q_{ii} < 0.$$

К недостаткам метода BrowseRank относится то, что он имеет свои ограничения как модель для определения важности страницы, поскольку учитывает не все доступные данные о поведении пользователя, которые могут послужить надежной информацией для определения важности страницы. К таким данным относятся количество страниц, просмотренных на сайте, общее время посещения сайта и отдельных страниц, возвраты пользователя в результаты поиска после просмотра сайта.

3. Расширенный BrowseRank

Несмотря на свои очевидные преимущества, BrowseRank обладает следующим недостатком - он моделирует процесс блуждания по ссылочному графу с учетом данных о поведении пользователя, ограничиваясь только временем посещения страницы. Не учитываются другие доступные данные о поведении пользователя, которые могут улучшить показатель качества и, следовательно, важности веб-страницы.

Для решения этой проблемы предлагается использовать дополнительные данные о поведении пользователя, такие как глубина просмотра и возврат в результаты поиска.

Под глубиной просмотра будем понимать количество просмотренных страниц сайта за сессию. Если посетители сайта просматривают всего 1-2 страницы, это может свидетельствовать о том, что ресурс им неинтересен (они попали на него случайно). Другая возможная причина — сайт имеет слишком сложную и неудобную навигацию, и пользователи просто не могут найти нужную им информацию. В любом случае глубина просмотра отражает степень релевантности/качества ресурса.

Возврат в результаты поиска после посещения страницы, как правило, говорит о ее нерелевантности/некачественности. Соответственно, данный показатель может быть использован как надежный источник информации для вычисления качества страницы.

Как правило, в зависимости от целей, поставленных перед сайтом, и сложности решаемых задач, сайты можно отнести к разным категориям (табл. 1).

Таблица 1

Назначение сайтов и их категории

Категории	Назначение сайтов
1	интернет-магазины, аукционы
2	информационные справочники
3	тематические, городские порталы, сайты новостей
4	блоги, форумы
5	корпоративные сайты, сайты организаций
6	общий доступ к файлам

Таким образом, граф сёрфинга пользователя, сгенерированного из данных о поведении пользователя принимает следующий вид:

$$G = < V, W, T, TR, RS, \delta >,$$

где – $V = \{v_i\}$ множество вершин; $W = \{w_i\}$ – множество весов; $T = \{t_i\}$ – время посещения; $TR = \{tr_i\}$ – глубина просмотра; $RS = \{rs_i\}$ – возврат в результаты поиска; $\delta = \{\delta_i\}$ – вероятность распределения; $(i, j = 1,..., N)$ — количество веб-страниц в графе переходов.

Для сбора информации о переходах и времени нахождения на страницах для большого числа пользователей строится граф переходов пользователей (рис. 1, 2). Каждое ребро графа представляет URL из данных о поведении пользователя и связанные с ним метаданные. Каждое направленное ребро представляет переход между двумя вершинами, отражающих количество переходов в качестве веса страницы. Другими словами, граф переходов пользователей является взвешенным графом с ребрами, содержащими метаданные и вершинами, содержащими вес веб-страницы.

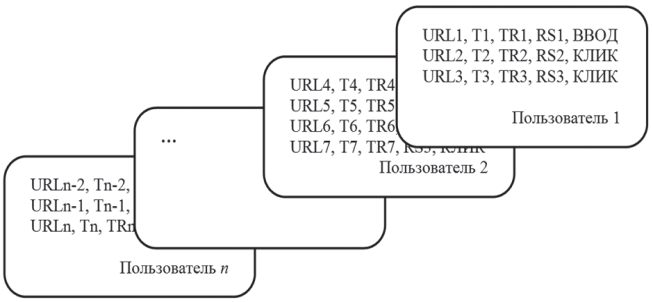

Рис. 1. Информация о поведении пользователя в расширенном BrowseRank

Механизм извлечения данных о переходах пользователя осуществляется следующим образом:

Новая сессия инициируется в случае 30-минутной и более паузы с момента предыдущей активности, либо в случае ввода названия сайта в адресную строку.

В рамках каждой сессии создаются пары url из соседних записей. Пара url означает, что переход был осуществлен при помощи ссылки. На основе этих данных вычисляется количество переходов рамках одного веб-сайта (глубина просмотра TR) и показатель возврата в результаты поиска RS.

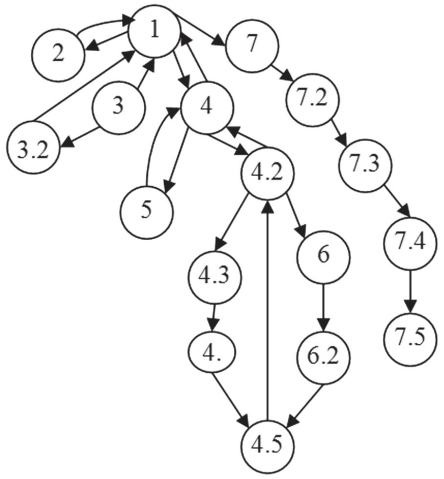

Рис. 2. Граф переходов в расширенном BrowseRank

В каждой сессии, сегментированной по типу перехода, первый url введен непосредственно пользователем. Такие url считаются «доверительными» и называются «зеленым» трафиком. Обрабатывая данные о поведении пользователей, переходы на указанные url считаются следствием случайного распределения. Нормализация на частоту посещения этих документов дает начальные вероятности посещения соответствующих страниц.

Для каждой пары url продолжительность сессии первого url вычисляется простой разностью дат. Если url был последним в сессии, возможны два варианта. Для сессий, сегментированных по времени, продолжительность просмотра последнего url рассчитывается на основании данных о просмотрах других страниц. Для сессий, сегментированных по типу, время просмотра последнего url рассчитывается исходя из времени начала следующей сессии.

4. Вычисление вероятности переходов для вложенной Марковской цепи

Вероятность переходов во вложенной Марковской цепи описывает «чистые», без погрешностей, переходы пользователя в графе. Их оценка может быть основана на наблюдаемых у пользователя переходах между веб-страницами, на «зеленом» трафике (когда url введен непосредственно пользователем) и на директ-трафике (трафик с закладок). Директ-трафик указывает о том, что пользователи переходят из закладок по памяти, что свидетельствует о качестве сайта. Для объединения этих типов данных при вычислении вероятности переходов используется следующий метод. Вначале берется граф переходов $G =< V, W, T, TR, RS, \delta >$. Затем к нему добавляется псевдо-вершина $(N+1)$ и два вида ребер:

– ребра от последней страницы каждой сессии к псевдо-вершине, связанные с количеством переходов на нее в качестве ее веса, для последнего типа данных (трафик с закладок) вес усиливается введением качественного коэффициента γ;

– ребра от псевдо вершины к первой странице каждой сессии, связанные с вероятностью случайного распределения. Новый граф обозначается как

$$\overline{G} =< \overline{V}, \overline{W}, \overline{T}, \overline{TR}, \overline{RS}, \bar{\delta} >,$$

где $|\overline{V}| = (N+1), \bar{\delta} =< \bar{\delta}_1, ..., \bar{\delta}_N, 0 >$.

Модель рассматривается как случайные переходы по графу $\overline{G}$.

Для учета дополнительных данных о поведении пользователя при расчете вероятности переходов в графе серфинга пользователя введем качественный коэффициент γ:

$$-\frac{a_{ij}}{a_{ii}} = \begin{cases} a\dfrac{\gamma\overline{w}_{ij}}{\sum_{k=1}^{N+1}\gamma\overline{w}_{ik}} + (1-a)\delta_i, i \in V, j \in \overline{V} \\ \delta_j, i = N+1, j \in V \end{cases}.$$

Другими словами, когда пользователь переходит по ссылкам графа, он переходит по ребрам с вероятностью a или начинает с новой страницы с вероятностью $(1-a)$.

Для расчета качественного коэффициента для категорий сайтов 1-5 применятся формула:

$$\gamma = 1 - 0.5 \times (\frac{1}{T+TR} + RS), \quad (2)$$

а для категории 6:

$$\gamma = 0.5 \times (\frac{1}{T+TR} - RS + 1), \quad (3)$$

где T – общее время пребывания пользователя на сайте в течение одной сессии $(T > 0)$; TR – глубина просмотра выражает количество страниц сайта, которые посетил пользователь в течение одной сессии $(TR \geq 1)$; RS – возврат в результаты поиска, принимает значения 0, если возврат не произошел, 1, если возврат имел место быть.

В табл. 2 приведены значения качественного коэффициента γ для примера, изображенного на рис. 2.

Таблица 2

Примеры данных о поведении пользователя с учетом дополнительных параметров

url	Время, мин.	Глубина просмотра	Возврат в результаты поиска	Коэффициент γ(2)	Коэффициент γ(3)
site1	2,5	1	0	0,857142	0,214286
site2	0,1	1	1	0,499998	0,250002
site3	0,33	2	1	0,285407	0,571888
site4	8,1	5	1	0,462121	0,306818
site5	0,3	1	1	0,115384	0,826923
site6	1,4	2	1	0,352941	0,470588
site7	20,2	5	0	0,980158	0,029762

На рис. 3-5 приведены диаграммы, которые отображают значение качественного коэффициента в зависимости от различных значений показателей времени, глубины просмотра и возврата в результаты поиска.

Как видно из рисунков, введение дополнительных сведений о поведении пользователя позволяет точнее оценить релевантность web-ресурса.

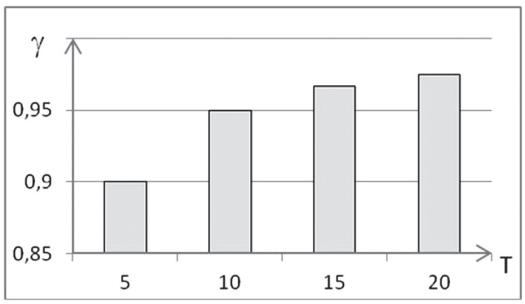

Рис. 3. Значение коэффициента γ в зависимости от значения времени пребывания пользователя на сайте для категорий 1-5

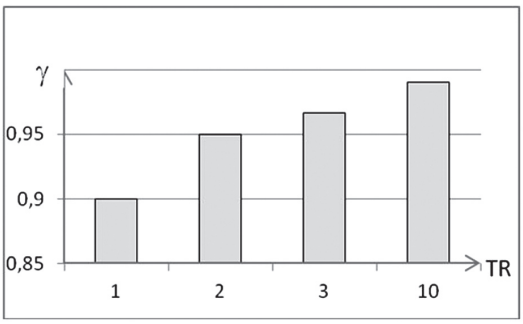

Рис. 4. Значение коэффициента в зависимости от глубины просмотра для категорий 1-5

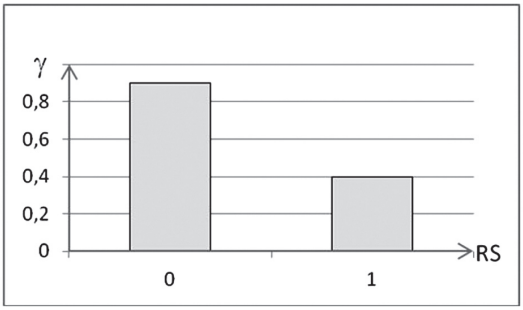

Рис. 5. Значение коэффициента в зависимости от возврата в результаты поиска для категорий 1-5

Таким образом, для реализации модификации метода BrowseRank необходимо следующее.

Входными данными являются время посещения страницы, тип перехода и информация о поведении пользователя, выходные данные — значение важности страницы π .

1. Создать граф переходов пользователя на основе данных о его поведении.

2. Вычислить качественный коэффициент γ по формулам (2) или (3).

3. Вычислить q_{ij} для всех страниц.

4. Вычислить матрицы вероятностей переходов для вложенной Марковской цепи и получить стационарное распределение вероятностей.

5. Вычислить стационарного распределения вероятностей -процесса по соотношению (1).

Выводы

В ходе проведенных исследований обнаружено, что наиболее популярные методы вычисления важности страницы имеют недостатки, связанные с информацией о посещении веб-ресурса, что влечет к недостаточно точному вычислению ее значимости.

Проведен анализ на влияние дополнительных параметров о поведении пользователя (глубина просмотра и показатель возврата в результаты поиска), который показал, что введение качественного коэффициента в метод BrowseRank, позволяет более точно определять важность веб-страницы, эффективно бороться с поисковым спамом и выдавать пользователям поисковых систем более релевантные результаты. Приведены результаты экспериментального моделирования, подтверждающие эффективность предложенного подхода.

Список литературы: 1. *B. Amento, L. Terveen,* and *W. Hill.* Does authority mean quality? Predicting expert quality ratings of web documents. In SIGIR ' 00. ACM, 2000. **2**. *R. Baeza-Yates* and *B. Ribeiro-Neto.* Modern Information Retrieval. Addison Wesley, May 1999. **3**. *M. Bianchini, M. Gori,* and *F. Scarselli.* Inside pagerank. ACM Trans. Interet Technol., 5(1):92–128, 2005. **4**. *P. Boldi, M. Santini,* and *S. Vigna*. Pagerank as a function of the damping factor. In WWW ' 05. ACM, 2005. **5**. *S. Brin* and *L. Page*. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998. **6**. *G. H. Golub* and *C. F. V. Loan*. Matrix computations (3rd ed.). Johns Hopkins University Press, Baltimore, MD, USA, 1996. **7**. Z. *Gyongyi* and *H. Garcia-Molina*. Web spam taxonomy. In AIRWeb '05. 2005. **8**. *Z. Gyongyi, H. Garcia-Molina,* and *J. Pedersen*. Combating web spam with trustrank. In VLDB '04, pages 576–587. VLDB Endowment, 2004. **9**. *T. Haveliwala*. Ecient computation of pageRank. Technical Report, Stanford University, 1999. **10**. *T. Haveliwala* and *S. Kamvar*. The second eigenvalue of the google matrix. Technical Report, Stanford University, 2003. **11**. *Yuting Liu, Bin Gao, Tie-Yan Liu*. BrowseRank: Letting Web Users Vote for Page Importance. SIGIR 2008 (Best Student Paper Award). **12**. *Sergey Brin, Lawrence Page*. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Science Department, Stanford University, Stanford, 1996. **13**. *Z. Gyuŋgyi, H. Garcia-Molina, J. Pedersen*: Combating Web Spam with TrustRank. **14**. *Matthew Richardson, Amit Prakash, Eric Brill*. Beyond PageRank: Machine Learning for Static Ranking. — 2006.

Поступила до редколегії 25.01.2013

УДК 004.55

Розширений browserank на основі обліку поведінкових факторів користувачів / Н.Г. Аксак, С.А. Коргут, Н.Г. Стрельцова // Біоніка інтелекту: наук.-техн. журнал. – 2013. – № 1 (80). – С. 99-103.

У роботі пропонується удосконалений метод BrowseRank, що дозволяє більш точно визначати важливість веб-сторінки за рахунок урахування додаткової інформації про поведінку користувачів.

Іл. 5. Табл. 2. Бібліогр.: 14 найм.

UDK 004.55

Advanced browserank: algorithm accounting behavioral factors of users / N.G. Axak, S.A. Korgut, N.G. Streltsova // Bionics of Intelligense: Sci. Mag. – 2013. – № 1 (80). – P. 99-103.

The paper proposes an improved method BrowseRank, allowing more accurately determine the importance of web pages by taking into account additional information on user behavior.

Fig. 5. Tab.2. Ref.: 14 items.