УДК 004.891

д.т.н., проф. **Лєнков С.В.** (ВІКНУ)
к.т.н., доц. **Джулій В.М.** (ХмНУ)
Романовська І.А. (ХмНУ)

МЕТОД ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ВИДАЧІ РЕЗУЛЬТАТУ ПОШУКУ В ІНФОРМАЦІЙНОМУ СЕРЕДОВИЩІ

У статті проведено аналіз ранжування сторінок у сучасних інформаційних пошукових системах. Розглянуто одну з основних складових ранжування – критерій PageRank. За останні роки розвитку інформаційних технологій він істотно вплинув на створення списку авторитетності сторінок. Завдання цього критерію полягає у визначенні так званої «ваги» сторінки, яка вираховується за відповідною формулою: пріоритетним є врахування зв'язків між сторінками – їхніх посилань. Чим більше посилань є на сторінку, тим більша її «вага» і вище вона стоїть у видачі результатів пошукової системи. Саме ці зв'язки є головною складовою формули обрахунку, вони поділяються на внутрішні та зовнішні. Проблема виникає з внутрішніми зв'язками, тому що вони дають посилання одне на одного, створюючи при цьому цикл посилань, що штучно збільшує «вагу» сайту з кожним кроком перевірки. Для вирішення проблем циклічності та штучного збільшення авторитетності пропонується визначення альтернативного показника TruePageRank, який відрізняється такими особливостями: посилання, які виходять з певної сторінки, не враховуються; в процесі обчислення TruePageRank вихідна структура зв'язків між сторінками перетворюється в антициклічний граф, в якому посилання, що утворюють циклічні зв'язки, не враховуються і увага приділяється прямим посиланням. При цьому об'єктивне значення авторитетності не втрачається, тому що знімається лише штучне збільшення значення. Для нового критерію потрібні ті самі значення, тобто не потрібно шукати нову інформацію.

TruePageRank – ще один з методів обробки вихідних даних, що дозволяє отримати альтернативний показник авторитетності сторінки, який досить об'єктивно відображає її інформаційну цінність.

Ключові слова: пошук, ранжування, PageRank, посилання, сайт, інформація.

Вступ. Кількість інформації в мережі Інтернет зростає з кожним днем все швидше, але, на жаль, цього не можна сказати про її якість. Користувач у пошуках потрібної інформації може провести все життя, якщо тільки випадково не знайде шуканий матеріал, єдиний вихід для нього – скористатися пошуковими системами, які зберігають інформацію про адреси і зміст веб-сторінок. Пошукові машини намагаються вирішити проблему – серед сотень однотипних документів вибрати кращий, за цим усім стоїть багато алгоритмів та функцій, які наявні у кожній пошуковій системі, і в кожній вони відрізняються. Але у всі вони характеризуються критерієм ранжування. У наш час використовуються текстові та довідкові критерії ранжування сторінок при пошуку. Перші визначають доречність («релевантність») документа виходячи з наявності слів запиту в тексті і заголовках сторінки. Однак наявність великої кількості документів може знецінити витончені механізми розрахунку релевантності, що базуються тільки на вмісті сторінки. Це і відбулося, коли власники сайтів зрозуміли, яку вигоду вони отримують від цільових відвідувачів, яких безкоштовно надають пошукові системи. Якість пошуку зіпсувалася, кількість документів зросла – «релевантний» документ стало дуже легко створити.

Постановка завдання. З метою поліпшення якості пошуку частину роботи з визначення «хороших», «важливих» документів побічно поклали на веб-майстрів Мережі. Розміщуючи посилання на зовнішній сайт, розробник ніби рекомендує його відвідувачам – саме цю особливість Інтернету вирішили використовувати для поліпшення якості пошуку. Підвищена вагомість документа визначається, таким чином, з урахуванням посилань ззовні на сайт, що містить цей документ.

Однак при врахуванні цитованості ресурсу виникають деякі труднощі. Кількість зовнішніх посилань на сайт не підходить для подання цитованості – з появою безкоштовних

хостингів кількість посилань дуже легко збільшити. Але важливість таких посилань нікчемна порівняно з посиланнями з відомих ресурсів. PageRank і є таким параметром важливості, він показує цитованість сторінки. Щодо цього критерію свого часу було висловлено багато різних і часто суперечливих думок, які заперечують вагомість PageRank. Розробники і промоутери сайтів прагнуть максимізувати його значення для створюваних інформаційних ресурсів.

Досить поширеним явищем в Інтернеті є ситуації, коли в структурі посилань сайту зустрічаються циклічні зв'язки, в тому числі і зворотні циклічні зв'язки. Наявність таких зв'язків не тільки ускладнює процес розрахунку показника авторитетності, а й істотно знижує його роль як індикатора семантичної значущості певної сторінки. При розрахунку значення показника PageRank частина ваги, переданої сторінкою A сторінці B під час однієї ітерації, повертається назад сторінці A під час наступної ітерації. Так виникає проблема циклічності. Під проблемою циклічності мається на увазі наявність у структурі посилань так званих штучних циклічних зв'язків. Крім них існують також об'єктивні (природні) циклічні зв'язки, але вони меншою мірою впливають на проблему обчислення авторитетності за допомогою PageRank. Штучні циклічні зв'язки передбачають навмисне, тобто штучне збільшення значення показника авторитетності PageRank з метою його підвищення. Якщо брати до уваги ступінь важливості вхідних посилань на сторінки сайту, то при розрахунку авторитетності більшу роль мають відігравати прямі зв'язки, ніж циклічні.

Виклад основного матеріалу дослідження. Сучасні пошукові системи Інтернету в процесі обробки запитів користувачів застосовують досить складні алгоритми ранжування документів (точніше, посилань на ці документи), які видаються як результати пошуку. Деякі фахівці говорять про кілька десятків і навіть про сотні критеріїв упорядкування, використовуваних у різних поєднаннях [1].

Для початку розглянемо сам процес ранжування документів. Ранжування – послідовне розміщення чогось, у нашому випадку – це розміщення шуканих сторінок у видачі пошукової системи. Фактори, що впливають на ранжування, поділяються на статичні, динамічні і власні. Статичні фактори не залежать від запиту до пошукової системи. Динамічні фактори залежать від запиту і поділяються на внутрішні (організація документа) і зовнішні (ранжування посилань). Власні належать до самої системи. Зупинимося на статичних факторах. Хорошим прикладом є критерій PageRank [2]. Порядок здійснення ранжування має такий вигляд:

- Знайти всі сторінки, що відповідають ключовим словами пошуку.
- Сортувати відповідно до «сторінкових факторів», таких як ключові слова.
- Врахувати текст посилань на сторінки.
- Відкоригувати результати з даними PageRank.

PageRank – це система, що базується на алгоритмі, який дозволяє визначати цифрову «вагу» кожного елемента мережі пов'язаних документів для з'ясування важливості кожного елемента у відношенні до іншого. Значення показника PageRank враховується в процесі розрахунку підсумкової релевантності документа при формуванні видачі пошуковою системою у відповідь на запит користувача [2].

Формула розрахунку PageRank має такий вигляд:

$$PR(A) = (1 - d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(T n)), \quad (1)$$

де $PR(A)$ – «вага» PageRank сторінки A (та «вага», яку ми хочемо визначити);

d – коефіцієнт затухання, що відображає, яку частку «ваги» може передати сторінка-донор на сторінку-акцептор. Зазвичай його приймають рівним 0,85, це означає, що сторінка може передати 85% «ваги»;

$PR(Ti)$ – PageRank сторінки, яка вказує на сторінку A ;

$C(Ti)$ – кількість посилань на сторінку;

PR(Tn)/C(T n) A – означає, що дія виконується для кожної сторінки, яка вказує на сторінку A .

У публікаціях найчастіше згадуються такі методи обчислення PageRank [2]:

- ітераційний;
- матричний;
- функціональний.

Ітераційний метод розрахунку PageRank

Цей метод найбільш часто використовується в практиці демонстраційних обчислень. Являє собою метод розв'язання системи рівнянь, що визначають PageRank сторінок:

1) вибирається геометрія сайту, розставляються посилання, формується система рівнянь шляхом запису формули (1) для кожної сторінки;

2) задаються початкові значення PageRank для кожної сторінки, вони можуть бути будь-якими;

3) розраховується новий набір значень PageRank за рівнянням виду (1), виходячи з наявного поточного набору значень;

4) розраховується середній PageRank по всьому набору сторінок, PageRank кожної сторінки ділиться на отриману величину; в результаті середній PageRank стає рівним одиниці;

5) якщо набір значень PageRank змінився порівняно з вихідним набором кроку 3, повертаємося до кроку 3; якщо цього не відбулося, то розрахунок закінчується.

Матричний метод розрахунку PageRank

Нижче подано матрицю зв'язків (рис. 1), матрицю можна помножити на вектор значень PageRank m-го кроку ітерації, отриманий вектор помножити на d, додати одиничний вектор, помножений на $(1-d)$, і отримати наступне наближення вектора PageRank з номером $m+1$, який потрібно пронумерувати (щоб сума проекцій вектора PageRank дорівнювала N). При навичках роботи з математичними програмами цей спосіб може бути більш зручним.

	1	2	3	4
1	0	1/3	1/3	1/3
2	0	0	1/2	1/2
3	0	0	0	1
4	1	0	0	0

Рис. 1. Матриця зв'язків

На рисунку 1 сторінка 1 посилається на 2, 3, 4; сторінка 2 – на 3 і 4; сторінка 3 – на 4, а 4 – на 1. Наведена матриця містить значення $Mij = 1/Cj > i$, тобто значення в кожному осередку розділене на загальну кількість посилань Cj на сторінці j.

Функціональний метод розрахунку PageRank

Суть цього методу полягає в тому, щоб розрахувати стабільні значення PageRank, не застосовуючи ітераційних методів.

Зауважимо, що у функціональному методі міститься не одна, а дві якісних відмінності від ітераційних методів:

- до розгляду залучаються кількісно невизначені значення PageRank зовнішніх (у відношенні до розглянутого сайту) сторінок.
- використовується неітераційний метод розв'язання системи лінійних алгебраїчних рівнянь, що визначає PageRank для сторінок розглянутого сайту. Залучення до розгляду «вхідного» потоку авторитетності являє собою спробу врахування впливу зовнішніх сайтів на PageRank конкретного сайту, розглянутого ізольовано від мережі. При цьому «вхідний», зовнішній, PageRank конкретної сторінки розглядається як інтегрований потік – він символізує не одне, а певну кількість зовнішніх посилань. У загальному випадку, природно,

потрібно враховувати «вхідний» PageRank для всіх сторінок розглянутого сайту, а не тільки для головної сторінки.

Деякі особливості дозволяють зробити висновок про те, що функціональний метод обчислення PageRank – це гарне теоретичне рішення, придатне для наочних ілюстрацій у простих випадках, коли розглядаються окремі сайти, але не більше. Для аналізу великих фрагментів мережі Інтернет цей метод не придатний [3].

Типовими структурами зв'язків усередині сайту є:

- ієрархічна (рис. 2);
- циклічна (рис. 3);
- «широке зв'язування» (рис. 4).

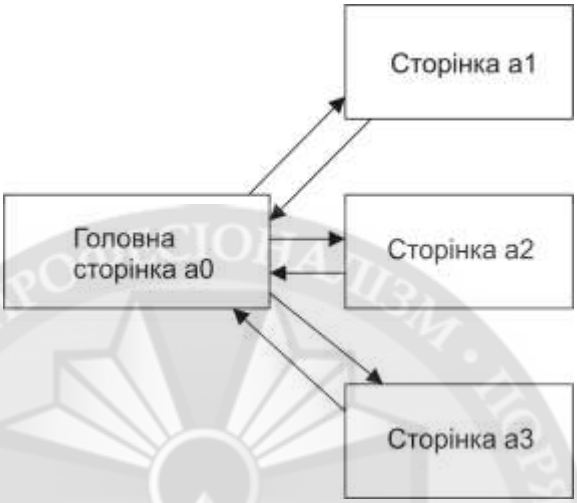

Рис. 2. Ієрархічна структура

Рис. 3. Циклічна структура

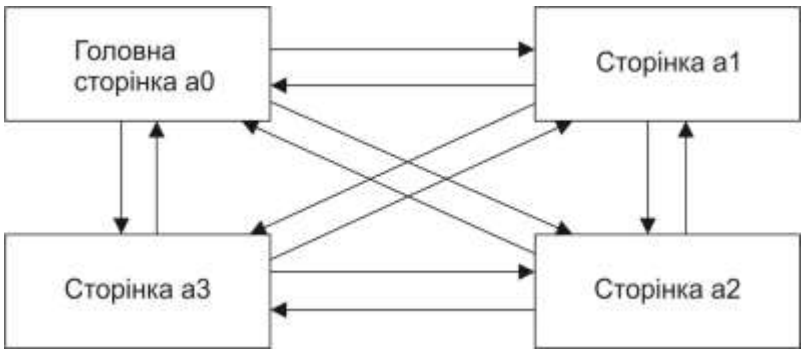

Рис. 4. Структура «широке зв'язування»

Вже досить часто стали зустрічатися проблеми циклічності, а саме так званої штучної циклічності. Цей фактор не рідко враховується при створенні сайту з метою його просування. Наявність природних циклічних зв'язків є наслідком прагнення оптимізувати структуру сайту і використовувати існуючий у мережі готовий інформаційний матеріал. Такі об'єктивні циклічні зв'язки становлять невід'ємну частину структури сайту. Якщо брати до уваги ступінь важливості вхідних посилань на сторінки сайту,то при розрахунку авторитетності більшу роль мають відігравати прямі зв'язки, ніж циклічні. Розглянемо приклад: нехай два сайти мають приблизно однакові суми значень показників авторитетності сторінок ($PRA = PRB$) (рис. 5, рис. 6).

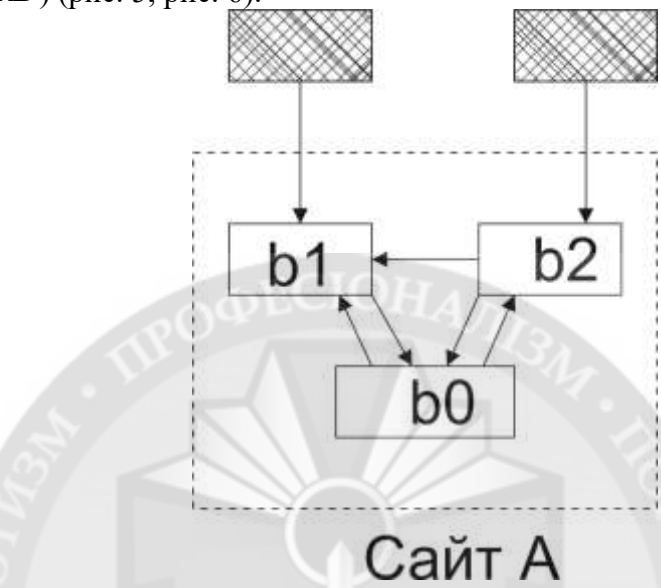

Рис. 5. Зв'язки сайту A

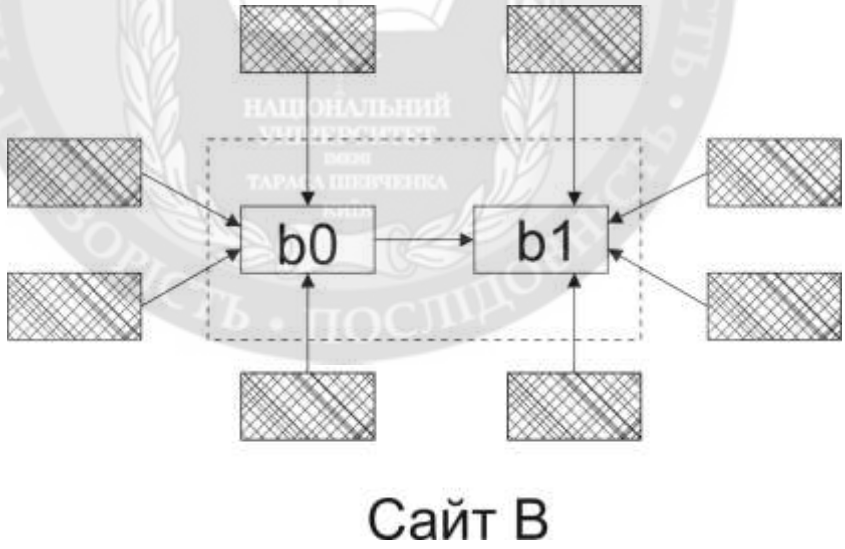

Рис. 6. Зв'язки сайту B

Величина значення авторитетності сайту A досягається завдяки циклічним зв'язкам між сторінками, з урахуванням того, що прямих зв'язків (як внутрішніх, так зовнішніх) мало. Величина значення авторитетності сайту B , навпаки, досягається завдяки прямим зв'язкам. Прямі вхідні посилання (зовнішні – з інших сайтів) семантично більшою мірою підкреслюють авторитетність сторінок, а отже, сайту в цілому. Авторитетність сайту

визначається кількістю вхідних посилань на цей сайт, а не внутрішньою структурою посилань. Отже, значення авторитетності сайту B має перевищувати значення авторитетності сайту A. Алгоритм обчислення авторитетності PageRank не має на увазі вирішення цього протиріччя ($PRA = PRB$).

Для вирішення проблем циклічності і штучного збільшення значення авторитетності можна розробити альтернативний показник TruePageRank, який відрізняється такими особливостями:

- вихідні посилання з певної сторінки не враховуються;
- у процесі обчислення TruePageRank вихідна структура зв'язків між сторінками перетворюється в ациклічний граф, у якому посилання, що утворюють циклічні зв'язки, не враховуються. Ця особливість приводить до того, що в інформаційних ресурсах Інтернету виключаються з розгляду численні внутрішні посилання, при цьому відбувається відмова як від штучних циклічних зв'язків, так і від об'єктивних, і увага приділяється прямим посиланням, при цьому об'єктивне значення авторитетності не губиться, тому що знімається лише зміна значення;
- у результаті викреслювання вихідних посилань у структурі зв'язків утворюється термінальна ситуація, яка характеризується відсутністю вихідних посилань, на конкретний сайт з безлічі сторінок і означає закінчення обчислення значення TruePageRank на певному етапі процедури.

Загальна формула для розрахунку показника авторитетності має такий вигляд:

$$NPRb = (1 - d) + d(NPRi(A[b])/Ci + ... + NPRn(A[b])/Cn), \quad (1)$$

де $NPRb$ – значення показника авторитетності сторінки;

d – коефіцієнт загасання (ймовірність того, що користувач, який прийшов на сторінку, перейде по одному з посилань на цій сторінці, а не припинить подорож по мережі, зазвичай встановлюється рівною 0,85);

$NPRi$ – значення показника авторитетності i-ї сторінки, що посилається на сторінку b;

Ci – загальна кількість посилань на i-ту сторінку;

$NPRn$ – значення показника авторитетності n сторінки, що посилається на сторінку b;

Cn – загальна кількість посилань на n сторінку;

$A[b]$ – поточна матриця зв'язків між сторінками, що відповідає певному етапу процедури.

З метою послідовного аналізу алгоритму обчислення TruePageRank розглянемо приклад розрахунку показника авторитетності однієї зі сторінок сайту, що містить таку структуру посилань (рис. 7):

Рис. 7. Структура посилань сайту

Структура посилань, зображена на рис. 7 містить зворотні й циклічні зв'язки. Метою розрахунку авторитетності сторінки за допомогою алгоритму TruePageRank є запобігання впливу зворотних і циклічних зв'язків на шуканий результат і запобігання штучному збільшенню показника авторитетності сторінки.

Висновки.У статті розглянуто методи роботи ранжування сторінок та їхній вплив на авторитетність сторінки і результати пошуку. Також наведено формальний опис критерію PageRank, формулу його розрахунку. Виділено циклічну структуру зв'язків та недолік в роботі з циклічними посиланнями, а саме врахування з однаковим коефіцієнтом як прямих, так і циклічних зв'язків. Це призводить до неправильного визначення «ваги» сайту і дає можливість штучно збільшувати авторитетність сторінки. У зв'язку з цим запропоновано новий метод ранжування – TruePageRank та його модифіковану формулу. Подаються алгоритм роботи та рисунки у вигляді блок-схем Планується розробити вдосконалений TruePageRank, який буде відкидати циклічні зв'язки та враховувати тільки прямі посилання із сайтів, що значно змінить показники результатів авторитетності та не допустить штучного підвищення «ваги» PageRank.

ЛІТЕРАТУРА:

1. Муляр І.В. Адаптивний метод впорядкування інформаційного забезпечення з використанням пошукової функції // Зб. наук. праць Військового інституту Київського НУ ім. Тараса Шевченко. – К.: ВІКНУ, 2011. – Вип. №33. – С. 192-196.

2. Райдингс К., Садовский А. Растолкованный PageRank, или Все, что вы всегда хотели знать о PageRank [Электронный ресурс] / К. Райдингс, А. Садовский. – Режим доступа: http://digits.ru/ articles/promotion/ pagerank.html (дата звернення 10.09.2016).

3. Сбітнєв А.І., Алгоритм пошуку семантично подібних документів / А.І Сбітнєв, С.В. Лєнков, В.М Джулій., І.В Муляр., Л.В. Охрамович // Науково-практичний журнал «Сучасні інформаційні технології у сфері безпеки та оборони». – Київ, 2013. - №2(17). – С. 24 – 29.

4. Трофименко Е. PageRank: иерархия и обмен ссылками [Электронный ресурс] / Е. Трофименко. – Режим доступа: http://promosite.ru/articles/pagerank-exchange.php (дата звернення 12.09.2016).

REFERENCES:

1. Mulyar I.V. Adaptyvnyy metod vporyadkuvannya informatsiynoho zabezpechennya z vykorystannyam poshukovoyi funktsiyi // Zb. nauk. prats' Viys'kovoho instytutu Kyyivs'koho NU im. Tarasa Shevchenko. – K.: VIKNU, 2011. – Vyp. – №33. – S. 192-196

2. Raydynhs K., Sadovskyy A. Rastolkovannыy PageRank, yly Vse, chto vы vsehda khotely znat' o PageRank [Эlektronnыy resurs] / K. Raydynhs, A. Sadovskyy. – Rezhym dostupa: http://digits.ru/ articles/promotion/ pagerank.html (data zvernennya 10.09.2016).

3. Sbitnyev A.I., Alhorytm poshuku semantychno podibnykh dokumentiv / A.I Sbitnyev, S.V. Lyenkov, V.M Dzhuliy., I.V Mulyar., L.V. Okhramovych // Naukovo-praktychnyy zhurnal «Suchasni informatsiyni tekhnolohiyi u sferi bezpeky ta oborony». – Kyyiv, 2013. – №2(17). – S. 24 – 29.

4. Trofimenko E. PageRank: ierarhija i obmen ssylkami [PageRank: hierarchy and link exchange]. Available at: http://promosite.ru/articles/pagerank-exchange.php (Accessed 12 September 2016).

Без рецензії.

д.т.н., проф Ленков С.В., к.т.н., доц. Джулий В.Н., Романовская И.А.

МЕТОД ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ВЫДАЧИ РЕЗУЛЬТАТА ПОИСКА В ИНФОРМАЦИОННОЙ СРЕДЕ

В статье проведен анализ ранжирование страниц в современных информационных поисковых системах. Рассмотрены одни из основных составляющих ранжирование - критерий PageRank. За последние годы развития информационных технологий он существенно повлиял на

создание списка авторитетности страниц. Задача этого критерия заключается в определении так называемого «веса» страницы, который рассчитывается по соответствующей формуле: приоритетным является учет связей между страницами - их ссылок. Чем больше ссылок является на страницу, тем больше её «вес» и выше она стоит в выдаче результатов поисковой системы. Именно эти связи являются главной составляющей формулы расчета, они делятся на внутренние и внешние. Проблема возникает с внутренними связями, так как они дают ссылки друг на друга, создавая при этом цикл ссылок, искусственно увеличивает «вес» сайта с каждым шагом проверки. Для решения проблем цикличности и искусственного увеличения авторитетности предлагается определение альтернативного показателя TruePageRank, который отличается следующими особенностями: ссылки, которые выходят с определенной страницы, не учитываются; в процессе вычисления TruePageRank исходная структура связей между страницами превращается в антициклических граф, в котором ссылки, образуют циклические связи, не учитываются и внимание уделяется прямой ссылке. При этом объективное значение авторитетности не теряется, потому что снимается только искусственное увеличение значения. Для нового критерия нужны те же значения, то есть не нужно искать новую информацию.

TruePageRank - еще один из методов обработки исходных данных, позволяет получить альтернативный показатель авторитетности страницы, который достаточно объективно отражает её информационную ценность.

Ключевые слова: поиск, ранжирования, PageRank, ссылки, сайт, информация.

Prof. Lenkov S.V., Ph.D. Dzhuliy V.N., Romanovska I.A.

METHOD OF INCREASING THE EFFICIENCY OF THE ISSUANCE OF A SEARCH RESULT IN THE INFORMATION ENVIRONMENT

This article provides an analysis of ranking pages in modern informative search systems. Considered one of the main components of the ranking - the criterion PageRank. In recent years, IT development was significantly influenced the creation of the list of authority pages. The purpose of this criterion is to determine the so-called "weight" of the page, which is calculated by the appropriate formula: the priority is the consideration of the links between pages - their links. The more links there are on a page, the higher its "weight" and it is higher in issuarance of search system's results. These links are the main components of the calculation formula, they are divided into internal and external. There is a problem with the internal connections, because they provide a link at each other, creating a cycle of links that artificially increases the "weight" of the site with each step of verification. To resolve the cyclical problems and artificial increase of credibility is proposed the determining of alternitive index TruePageRank, which is characterized by the following features: links, which come from a particular page do not count; in the process of calculating PageRank the original structure of links between pages is converted into an anti-cyclical graph in which the links are forming cyclic connections, is not considered and focuses on the direct link. The objective determination of credibility isn't lost because only the artificial increase in value is removed. The new criteria requires the same values, that is not necessary to search for new information.

TruePageRank - is another method of processing of output data, allowing you to get the alternative index of page's authority, which is quite objectively reflects its information value.

Keywords: search, ranking, PageRank, link, website, information.