УДК 004.67

ПОБУДОВА ПЕРСОНАЛЬНО–ОРІЄНТОВАНИХ МОДЕЛЕЙ РАНЖУВАННЯ РЕЗУЛЬТАТІВ ПОШУКУ ІНФОРМАЦІЇ

В.В. Зосімов

Миколаївський національний університет ім. В.О. Сухомлинського,

zosimovvv@gmail.com

В статті описано побудову моделі ранжування видачі пошукових систем на основі оцінок користувачів щодо якості та зручності використання веб–ресурсу. Такі моделі ранжування є суб'єктивними для кожного користувача. Це досягається за рахунок того, що при побудові моделей основну роль грає ступінь довіри користувача тим оцінкам, на основі яких вона будується. Для визначення коефіцієнта загасання ступеня довіри до оцінок на кожному рівні, застосовувався УІА МГУА, який вже показав високу точність в ході експериментів з побудови моделей ранжування видачі пошукових систем Google і Яндекс.
Ключові слова: пошук інформації, ранжування, індуктивне моделювання, метод групового урахування аргументів, ітераційний алгоритм.

The article describes the building of the search results ranking model based on the user evaluation of web resource quality and usability. Such ranking models are subjective for each user. This is achieved due to the fact that the main influence on the model has the user`s trust rank to the expert`s evaluation on which basis it is built. To determine the trust rank, the generalized iterative algorithm of the group method of data handling was applied, which has already showed high accuracy in the experiments on building Google and Yandex ranking models.
Keywords: information search, ranking, inductive modeling, group method of data handling, iterative algorithm.

В статье описано построение модели ранжирования выдачи поисковых систем на основе оценок пользователей относительно качества и удобства использования веб–ресурса. Такие модели ранжирования являются субъективными для каждого пользователя. Это достигается за счет того, что при построении моделей основную роль играет степень доверия пользователя к тем оценкам, на основе которых она строится. Для определения коэффициента затухания степени доверия для оценок на каждом уровне, применялся ОИА МГУА, который уже показал высокую точность в ходе экспериментов по построению моделей ранжирования выдачи поисковых систем Google и Яндекс.
Ключевые слова: поиск информации, ранжирование, индуктивное моделирование, метод группового учета аргументов, итерационный алгоритм.

Вступ

На сьогоднішній день реклама в інтернеті є найбільш ефективним способом просування бізнесу. Це призвело до того, що пошукові системи перестали бути інструментом отримання інформації, а перетворилися в рекламні майданчики, які видають користувачеві не ті сторінки, які найбільш релевантні його інформаційним потребам, а ті, в просування яких вкладено більше коштів. Пошуковим системам вигідне штучне зниження якості видачі органічного пошуку, так як на його фоні більш доречно виглядає контекстна реклама, яка в більшості випадків мало релевантна пошуковому запиту. На користь того, що пошуковим системам не вигідна висока якість органічного пошуку, говорить той факт,

що найбільша пошукова система Google в 2009 році [1] запустила технологію «Вікіпошук». Вона дозволяла користувачам налаштовувати результати пошукової видачі. Користувачам надавалась можливість сортування та видалення результатів пошуку. Також була реалізована глобальна система оцінок веб–ресурсів. Однак ця технологія пропрацювала не більш півроку і була відключена з посиланням на низьку затребуваність серед користувачів. Хоча очевидно, що персоналізація пошукової видачі зробить більшість сучасних методів штучного просування веб–ресурсів, неефективними.

Алгоритми ранжування пошукових систем враховують велику кількість факторів, але основну вагу має рейтинг сторінок (Page Rank – PR), який обчислюється на основі аналізу кількості та якості зовнішніх посилань на сторінку [2]. Такі методи оцінки є об'єктивними для всіх користувачів, але їх легко сфальсифікувати при наявності певного рекламного бюджету. З цього випливає, що вони орієнтовані на задоволення потреб рекламодавців, а не користувачів.

Цією проблемою займався Бріц Р.А. [3]. У його роботі представлена теоретична модель ранжування, заснована на статистиці відвідувань веб–ресурсів і часу перегляду документів. Однак в ній не враховується думка користувачів про веб-ресурсі, а також рейтинг довіри між користувачами. Також методи ранжування, засновані на врахуванні думки користувачів, широко застосовуються при ранжуванні товарів в інтернет–магазинах. Однак в них за основу береться кількість оцінок та їх цифровий вираз і не враховується рейтинг довіри між користувачами.

Виходячи зі сказаного, актуальним завданням є розробка методів персоналізації результатів роботи пошукових систем шляхом надання користувачу інструментів управління пошуковою видачею, а також розробки нових моделей ранжування, заснованих на суб'єктивних уявленнях користувача.

1. Ранжування на основі оцінок користувачів

Для побудови персоналізованої моделі ранжування, заснованої на оцінках користувачів, недостатньо числових значень оцінок. Необхідно визначити ступінь довіри користувача до тих оцінок, на основі яких будується модель ранжування. Для цього в модель необхідно ввести коефіцієнт довіри (*Trust Rank – TR*), який буде визначати, наскільки думка користувача збігається з даною оцінкою. Далі, для уникнення плутанини в термінах, будемо називати всіх учасників, крім поточного користувача, експертами.

При цьому необхідно виділити кілька рівнів зв'язку користувача з експертами:

1. На основі оцінок, виставлених особисто користувачем.

2. На основі оцінок, виставлених експертами, що мають спільні оцінки з поточним користувачем (експерти першого рівня).

3. На основі оцінок експертів, що мають спільні оцінки експертів першого рівня (експерти другого рівня)

4. На основі оцінок експертів третього рівня

По мірі віддалення від першого рівня, коефіцієнт довіри повинен спадати. На першому рівні значення дорівнюватиме 1. Будемо вважати що з кожним рівнем він зменшується удвічі. Через низьке значення коефіцієнта довіри на рівнях нижче четвертого, їх значення не будуть враховуватися.

Загальний рейтинг веб-ресурсу будемо вважати як суму середнього арифметичного всіх персоналізованих оцінок. Значення розраховується відповідно до моделі, побудованої за допомогою УІА МГУА на основі ряду персональних даних користувача на кожному рівні зв'язків.

Для побудови моделі персоналізованої оцінки було обрано ряд суб'єктивних ознак x_n, які можуть прямо або побічно впливати на оцінку відвідувача, де n – порядковий номер ознаки всередині моделі:

$x_1 = (r_1+Q_1)$, де r – оцінка експерта, Q – коефіцієнт корегування оцінки експерта;

x_2 – коефіцієнт затухання рівня довіри;

x_3 – активність експерта, що виражається в загальній кількості оцінок експерта;

x_4 – різниця у віці, що на відміну від абсолютного значення повних років, є більш суб'єктивним показником відносно користувача. Збіг віку – 0,9, кожен рік різниці – (– 0,01);

x_5 – дохід. Різниця в тисячах гривень;

x_6 – відповідь на питання 3+3*3=? Вірна відповідь – 0.9, невірна – 0.1;

x_7 – стать. Збігається з користувачем – 0.9, ні – 0.1;

x_8 – наявність домашніх тварин. Збігається з користувачем – 0.9, ні – 0.1

x_9 – освіта. Різниця між освітою користувача та експерта, виражена в числовому значенні. Шкала оцінювання освіти від 1 до 5, де 1 – середня освіта, і 6 – вчений ступінь;

x_{10} – фах освіти, гуманітарна, технічна. Збігається з користувачем – 0.9, ні – 0.1;

x_{11} – сімейний стан. Збігається з користувачем – 0.9, ні – 0.1;

x_{12} – наявність автотранспортного засобу. Збігається з користувачем – 0.9, ні – 0.1;

x_{13} – заняття спортом. Співпадає з користувачем – 0.9, ні – 0.1.

Важливо відзначити, що при переході на другий рівень зв'язків користувачі, які не мають спільних оцінок з основним користувачем, вважаються експертами, а експерти, які знаходяться на рівень вище, вважаються поточними користувачами.

2. Збір вхідних даних

Для отримання вхідних даних були обрані 80 перших веб-ресурсів з пошукової видачі Google за запитом встановлення металопластикових вікон. Експерти оцінювали якість представленої на веб-ресурсі інформації та зручність використання веб–ресурсу.

Для максимальної реалістичності побудованої моделі експерти і веб-ресурси були розбиті на 4 групи. Кожна група експертів належала до свого рівня довіри і оцінювала свою групу веб-ресурсів. Для забезпечення зв'язку оцінок експертів на кожному рівні з думкою поточного користувача, до них додавалося кілька експертів з верхнього рівня, які вже мали оціночний зв'язок з поточним користувачем. Коефіцієнт корегування оцінки *Q* для кожного користувача обчислювався на основі середнього арифметичного всіх різниць оцінок.

В таблицях 1-7 представлені дані, отримані в ході експерименту.

Позначення, що використовуються в таблицях:

- U_0 – поточний користувач;
- $U_{1–35}$ – експерти;
- $S_{1–80}$ – веб–ресурси, що оцінювались;
- $Q–>U_n$ – значення *Q*, розраховані для користувача з номером *n* відносно кожного експерта, представленого в таблиці.

Таблиця 1

Оцінки експертів першого рівня

Сайт / Кор.	S_1	S_2	…	S_9	S_{10}	…	S_{19}	S_{20}	$Q–>U_0$
U_0	10	10	…	9	9	…	3	5	
U_1	10	10	…	9	8	…	4	4	–0.25
U_2	8	8	…	7	7	…	3		1.00
U_3	9	9	…			…	3	3	0.07
U_4	10	10	…	8	8	…		3	–0.12
U_5	9	10	…	10		…	3		0.00
U_6	9	10	…		10	…	4		0.45
U_7	10	10	…	9	9	…	3	5	0.00
U_8	8	8	…		10	…		4	0.54
U_9	7	9	…	8	8	…	4		0.53
U_{10}	7	9	…	10	8	…			0.15
U_{11}	9		…		9	…	4	4	0.43
U_{12}	8		…	9		…		4	0.09
U_{13}	10	10	…	9	9	…	3	5	0.00
U_{14}	7	7	…	8	8	…	2	4	0.50

Кожен експерт першого рівня виконує роль користувача на другому рівні. Відповідно коефіцієнт *Q* був розрахований для кожного з них, відносно кожного експерту другого рівня. За тим же принципом були розраховані *Q* коефіцієнти для третього та четвертого рівнів.

Таблиця 2

Оцінки експертів другого рівня

Сайт / Кор.	S_{21}	S_{22}	…	S_{28}	S_{29}	…	S_{39}	S_{40}
U_1	8	8	…	3		…	7	
U_3	9	8	…	2		…	5	8
U_5	10	8	…	2		…	6	7
U_{10}	10	10	…	3	10	…	6	7
U_{11}	8	7	…	3	10	…		8
U_{12}	8	7	…	3		…	6	8
U_{14}	8	9	…	3		…	6	8
U_{15}		8	…	5		…	7	
U_{16}		8	…	3	10	…	8	6
U_{18}	9	9	…		9	…	6	5
U_{19}	9	9	…		9	…		
U_{20}	9		…	4	9	…	6	5

Таблиця 3

Оцінки експертів третього рівня

Сайт / Кор.	S_{41}	S_{42}	…	S_{52}	S_{53}	…	S_{59}	S_{60}
U_1	8	7	…	10		…	4	
U_5		6	…	10	8	…	5	5
U_{12}	7	6	…	9		…		
U_{15}		8	…	10	8	…	5	
U_{17}	9	7	…	8	6	…	3	
U_{19}	8	8	…	8	7	…		
U_{21}		6	…			…	3	5
U_{22}	7	7	…	9	9	…	4	4
U_{23}	9		…	10		…	5	
U_{24}		8	…	9	7	…	3	
U_{25}	9	8	…		8	…	4	
U_{26}		7	…	8		…	4	
U_{27}	8	7	…	8	9	…		
U_{28}			…	10	9	…		

Таблиця 4

Q коефіцієнт для другого рівня

	Q–>U_1	Q–>U_3	Q–>U_5	Q–>U_{10}	Q–>U_{11}	Q–>U_{12}	Q–>U_{14}
U_{15}	– 0.22	–0.58	–0.25	–0.08	–0.64	–0.36	–0.71
U_{16}	–0.33	–0.57	–0.14	–0.06	–0.15	–0.38	–0.69
U_{17}	0.33	0.38	0.29	0.38	0.23	–0.07	–0.21
U_{18}	–0.33	0.13	0.50	0.56	0.38	0.19	–0.13
U_{19}	–0.43	0.11	0.33	0.58	0.17	0.08	–0.20
U_{20}	–0.13	0.00	0.20	0.09	–0.22	–0.08	–0.30
U_{21}	–0.60	–0.50	–0.22	0.40	0.00	–0.40	–0.33

Таблиця 5

Q коефіцієнт для третього рівня

	Q–>U_1	Q–>U_5	Q–>U_{12}	Q–>U_{15}	Q–>U_{17}	Q–>U_{19}	Q–>U_{21}
U_{22}	0.64	0.19	–0.08	0.80	–0.08	–0.21	0.00
U_{23}	–0.75	–1.00	–1.38	–0.33	–1.30	–1.44	–2.63
U_{24}	0.36	0.07	–0.17	0.64	–0.29	–0.43	–0.40
U_{25}	–0.82	–1.17	–1.44	–0.27	–1.27	–1.45	–1.75
U_{26}	0.73	0.50	0.50	0.90	0.22	0.11	–0.57
U_{27}	0.45	0.17	0.40	0.55	–0.15	–0.15	–0.33
U_{28}	–0.11	–0.18	–0.50	0.20	–0.67	–1.17	0.00

Таблиця 6

Оцінки експертів четвертого рівня

Сайт / Кор.	S_{61}	S_{62}	…	S_{70}	S_{71}	…	S_{79}	S_{80}
U_1	7	5	…	2		…		
U_{17}	8	5	…	3		…		
U_{19}	7		…	3	6	…	7	9
U_{22}	7	5	…	3	6	…		9
U_{24}	6	4	…	4	7	…	6	
U_{27}	7	3	…	3		…	8	8
U_{28}	8	4	…			…	8	9
U_{29}	9	5	…	3	5	…		
U_{30}	8		…			…	8	
U_{31}	8	4	…	3	6	…		
U_{32}	9		…	4	6	…	8	8
U_{33}	7	5	…	4	5	…		
U_{34}	7		…	2	5	…	8	
U_{35}	8	3	…	3	4	…		

Таблиця 7

Q коефіцієнт для четвертого рівня

	Q–>U_1	Q–>U_{17}	Q–>U_{19}	Q–>U_{22}	Q–>U_{24}	Q–>U_{27}	Q–>U_{28}
U_{29}	–0.10	0.45	0.36	0.07	–0.07	–0.64	–0.75
U_{30}	0.60	1.00	0.54	0.54	0.33	–0.08	–0.27
U_{31}	0.40	0.83	0.27	0.43	0.50	–0.17	–0.20
U_{32}	–0.50	0.00	–0.13	–0.13	–0.12	–0.73	–0.50
U_{33}	–0.64	–0.31	–0.54	–0.44	–0.44	–1.23	–1.20
U_{34}	–0.50	0.18	0.00	0.00	0.13	–0.42	–0.60
U_{35}	0.29	0.56	0.44	0.73	0.50	–0.22	0.17

3. Побудова моделі коефіцієнта довіри оцінкам користувачів

Коефіцієнти довіри користувача були розраховані на основі моделі, побудованої за допомогою УІА МГУА, яка вже показала високу точність в експериментах з побудови моделей Google та Yandex [4].

Будь-який алгоритм МГУА вирішує завдання дискретної оптимізації для побудови моделі оптимальної складності за мінімумом заданого зовнішнього критерію на основі розподілу вибірки даних.

Ітераційні алгоритми МГУА вирішують таке завдання, послідовно наближаючись до мінімального значення критерію, використовуючи процедуру мережевого типу, засновану на аналогії з біологічним відбором живих організмів: ускладнення моделей на кожному ряді відбувається через попарне "перетинання" F кращих моделей попереднього ряду. Процес ускладнення припиняється після того, як критерій починає збільшуватися.

Комбінаторна оптимізація структур частинних моделей у комбінованому алгоритмі дає алгоритм під повною назвою «комбінований ітераційно-комбінаторний алгоритм» КІКА. Його подальше узагальнення, що враховує як КІКА та всі його частинні випадки, а також різні варіанти режимів роботи, що реалізуються через інтерфейс користувача, називаються «Узагальнений ітераційний алгоритм» або УІА МГУА.

Формально в загальному випадку для ряду r можна визначити УІА МГУА так [5]:

1) вхідною матрицею є $X_{r+1} = (y_1^r, ..., y_F^r, x_1, \ldots, x_m)$ з ряду $r+1$, де $x_1, \ldots, x_m$ є вхідними аргументами $y_1^r, ..., y_F^r$ є проміжними аргументами ряду r.

2) застосовуються оператори переходу виду

$$\begin{aligned} y_l^{r+1} &= f(y_i^r, y_j^r),\ l = 1,2,\ldots, C_F^2, \quad i, j = \overline{1, F}, \\ y_l^{r+1} &= f(y_i^r, x_j), l = 1,2,\ldots, Fm, i = \overline{1, F},\ j = \overline{1, m} \end{aligned} \tag{1}$$

може бути застосовано у ряді $r+1$ щоб побудувати лінійні, білінійні та квадратичні часткові описи:

$$z = f(u,v) = a_0 + a_1 u + a_2 v;$$
$$z = f(u,v) = a_0 + a_1 u + a_2 v + a_3 uv;$$
$$z = f(u,v) = a_0 + a_1 u + a_2 v + a_3 uv + a_4 u^2 + a_5 v^2. \quad (3)$$

3) для кожного опису за допомогою комбінаторного алгоритму знаходиться оптимальна структура, наприклад, для лінійного часткового опису, вираз має вид:

$$f(u,v) = a_0 d_1 + a_1 d_2 u + a_2 d_3 v, \quad (4)$$

де $d_k, k = 1,2,3$, $d_k = \{0, 1\}$ елементи двійкового структурного вектора $d = (d_1\, d_2\, d_3)$, що приймають значення 1 від 0 (включення чи невключення відповідного аргумента). Потім краща модель буде записана у вигляді $f(u,v,d_{opt})$, де

$$d_{opt} = \arg\min_{l=\overline{1,q}} CR_l, \; q = 2^p - 1, \; f_{opt}(u,v) = f(u,v,d_{opt}). \quad (5)$$

4) алгоритм зупиняється при виконанні умови $CR^r > CR^{r-1}$, де CR^r, CR^{r-1} значення критерію для кращих моделей (r–1)-го та r-го рядів відповідно. Якщо умова тримається, алгоритм зупиняється, в іншому випадку відбувається перехід до наступного ряду.

Рівняння (6) відображає модель, побудовану з використанням УІА МГУА:

$$TR = 0.00721 + 0.04271x_1 - 0.00853x_2 - 0.0215x_3 + 0.02511x_4 + \\ + 0.0153x_7 + 0.03341x_9 + 0.0742x_{10} + 0.1482x_1^2 - 0.01994x_1 x_2 \quad (6)$$

Точність моделі 93%.

Згідно з цією моделлю, інформативними виявились наступні ознаки:

$x_1 = (r_1 + Q_1)$;

x_2 – коефіцієнт затухання рівня довіри;

x_3 – активність експерта;

x_4 – різниця у віці;

x_7 – стать;

x_9 – освіта;

x_{10} – фах освіти – гуманітарна, технічна.

Змінні x_4, x_7, x_9, x_{10} є базовими і найбільш значущими соціальними характеристиками людини, і саме вони мають найбільший вплив на формування думки користувача.

Точність ранжування на основі *TR* була перевірена на другій групі вебресурсів, яку користувач раніше не оцінював. Таблиця 8 містить результати ра-

нжування списку веб–ресурсів різними методами: ранжування Google, ручне ранжування користувача (експерта) і автоматичне на основі *TR*.

Таблиця 8

Результати ранжування

Ранжування Google	Ручне ранжування користувача	Автоматичне ранжування на основі *TR*
7	1	1
12	2	3
3	3	2
4	4	9
18	5	5
10	6	4
19	7	7
1	8	8
9	9	6
11	10	10
2	11	13
20	12	12
5	13	19
14	14	14
15	15	16
6	16	17
13	17	15
17	18	18
16	19	11
8	20	20

З таблиці 8 видно, що результати ранжування на основі *TR* набагато ближче до ручного ранжування, ніж значення пошукової системи Google.

Висновки

Описана методика побудови суб'єктивної моделі ранжування веб-ресурсів на основі оцінок користувачів показала високу ефективність. Це говорить про перспективність розвитку даного напрямку.

Наступним етапом можлива побудова моделей ранжування, заснованих на великій кількості факторів, по аналогії з сучасними пошуковими системами,

проте в їх основі замість алгоритму Page Rank, заснованого на посиланнях, застосовувати показник *TR*, заснований на оцінках.

Отриману модель ранжування набагато складніше сфальсифікувати, тому що вона заснована на суб'єктивних оцінках користувача. Для кожного користувача модель ранжування буде унікальною. І для її фальсифікації необхідно буде розробляти методи визначення побажань кожного користувача, що набагато складніше, ніж штучне збільшення кількості посилань на свій веб-ресурс з авторитетних джерел.

Література

1. SearchWiki – https://ukraine.googleblog.com/2009/05/

2. Page Rank – https://ru.wikipedia.org/wiki/PageRank

3. Брицов Р.А. Ранжирование информации на основе оценок и поведения пользователей. T-Comm: Телекоммуникации и транспорт. – 2016. – Том 10. – №1. – С. 62-

4. Zosimov V., Stepashko V., Bulgakova O., Inductive Building of Search Results Ranking Models to Enhance the Relevance of Text Information Retrieval Proceedings. International Workshop on Database and Expert Systems Applications, DEXA Volume 2016, pp. 291–295.

5. Stepashko V., Bulgakova O., Zosimov V., Construction and Research of the Generalized Iterative GMDH Algorithm with Active Neurons. In: N. Shakhovska, V. Stepashko (eds). Advances in Intelligent Systems and Computing. Springer, Cham, vol 689, 2018, pp. 492–510.