УДК 004.891.2

Конюшенко О. В.[1], аспірант.

Реалізація системи персоналізованого пошуку на основі аналізу діяльності членів віртуальної спільноти

[1] Київський національний університет імені Тараса Шевченка,03680, м. Київ, пр-т. Глушкова 4д, e-mail: okonyushenko@gmail.com

O. V. Konyushenko[1], PhD student.

Implementation of personalized search system based on the analysis of the activity of the members of the virtual community

[1] Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova st., 4d, e-mail: okonyushenko@gmail.com

У статті детально представлені математичні моделі, на основі яких функціонує розроблена система персоналізованого пошуку, яка дозволяла б здійснювати пошук документів тематичного порталу з урахуванням результатів кооперативної діяльності віртуального співтовариства порталу по оцінці документів, охарактеризовані основні аспекти її роботи, представлена архітектура.

Ключові слова: персоналізований пошук, віртуальні спільноти, генетичні алгоритми, тематичні портали

Detailed mathematical models presented in this article formed a strong background for designed system of personalized search that should allow to search documents in thematic portal based on the results of the cooperative activity of an portal virtual community on the evaluation of documents; in article described main aspects of its work and architecture.

User level of interest in information section of documents, determined by informational portrait of this person. Informational portrait of the user based on vector: elements are concepts with weight which indicating level of interest in concept of the user.

User interface was expanded with module which collect and analyze user actions concerning resources evaluation. To make an initial relevant determination of finding certain resource, system uses the functionality of basic search engine with subsequent change based on alorithms of personalization and consideration of the performance of virtual communities.

Keywords: personalized search, virtual communities, genetic algorithms, thematic portals

Статтю представив д. ф.-м. н., проф. Анісімов А. В.

Вступ

У зв'язку із збільшенням об'єму даних, розміщених в Internet і корпоративних базах даних, актуальною стає проблема підвищення якості пошукових і навігаційних систем [1]. Наразі більшість пошукових систем можуть дізнаватися про інформаційну потребу користувача тільки за допомогою ключових слів, які задані в запиті. При цьому, система не має в своєму розпорядженні відомостей про контекст, в якому користувач вживає ці слова. Та і самому користувачеві буває важко підібрати ключові слова, які точно описують тематику, що його цікавить. Пошукова система надає користувачеві результати, відсортовані за релевантністю до запиту. При цьому враховується текст документів або структура посилань між ними, але не враховується ряд інших факторів, які могли б покращити якість пошуку [2,3], зокрема:

- коло інтересів і індивідуальні особливості користувача, який веде пошук;
- результати кооперативної діяльності віртуального співтовариства по оцінці якості ресурсів.

Перший з цих факторів можна охарактеризувати як проблему впровадження персоналізованого пошуку, другий – поширення впливу результатів діяльності одного користувача (зокрема, оцінювання якості ресурсів) на результати пошуку інших, що в перспективі може призвести до підвищення ефективності пошуку.

Щодня пошукові системи отримують запити від користувачів, які вибирають результати, що їх зацікавили. Цю інформацію пошукові системи збирають, систематизують і видають у вигляді цифр статистики: популярності запитів, інтересів, рейтингів тощо. Зрозуміло, ці цифри загальні і не відносяться до конкретної людини. Але при впровадженні засобів однозначної ідентифікації користувача дістаємо можливість збору інформації про конкретного користувача. Спеціалізована обробка цієї інформації дозволяє дізнатися багато цікавого про користувача, що може стати привабливим при формуванні індивідуальних результатів видачі на запит. Сама процедура пошуку значно спрощується, оскільки пошукова машина вже «знає» багато чого про його потреби і інтереси. Це і називають персоналізацією пошуку [4].

Актуальність досліджень у цій галузі підтверджується як діяльністю багатьох наукових спільнот у цьому напрямку, так і впровадженням систем, які реалізують дані ідеї, великими корпораціями, що представлені на ринку пошуку інформації. Зокрема, відносно недавно компанія Google представила свій сервіс персоналізованого пошуку Вікіпошук (SearchWiki) [5]. Але сервіс має ряд обмежень, спричинених природою Internet. З огляду на це, було прийнято рішення про проведення досліджень в цій галузі на базі окремого веб-порталу.

Мета роботи полягає у розробці основних принципів функціонування та впровадженні системи персоналізованого пошуку, яка дозволяла б здійснювати пошук документів тематичного порталу з урахуванням результатів кооперативної діяльності віртуального співтовариства порталу по оцінці якості документів.

Побудова «інформаційного портрета користувача»

Інформаційним портретом користувача (ІПК) називатимемо набір параметрів і їх значень, що описують сферу інтересів користувача та галузі знань, які його цікавлять [6].

Зазвичай, під ІПК розуміється вектор, елементами якого є поняття з вказівкою ваги, що характеризує міру зацікавленості ним користувача. Такий портрет можна складати методом попереднього анкетування користувача або враховуючи активність користувача при роботі з інформаційними ресурсами. Перший спосіб вважається дуже трудомістким, таким що вимагає з одного боку складання тестових наборів, які охоплюють різні галузі знань, а з іншого боку, великих зусиль і тимчасових витрат при проходженні користувачем цих тестів. Проте, тести можна використовувати для первинного наближеного формування інформаційного портрета. Другий підхід не вимагає від користувача великої кількості додаткових дій і потенційно забезпечує більшу точність.

Для побудови ІПК традиційно використовують персональні програмні агенти або персоніфікацію пошукової системи [7]. У першому випадку система будує портрет на основі документів, що переглядаються користувачем, з різних джерел. У другому, система відстежує документи, відібрані користувачем в результатах, виданих по запиту конкретною пошуковою системою.

У обох випадках в основі можуть лежати одні і ті ж алгоритми побудови ІПК. Відмінність в тому, що персональний агент будує вектори документів нальоту, а персоніфікатор пошукової системи може використовувати вектори документів, визначені раніше при індексуванні.

Зазвичай застосовують спосіб побудови ІПК, при якому система класифікує вибрані користувачем документи відповідно до деякої онтологічної структури [8]. Категорії онтології, в які попало більше документів, складають ІПК. Документ співвідноситься з категорією на основі скалярного добутку їх n-вимірних векторів. Реалізації відрізняються способом побудови n-вимірного простору і методом обчислення скалярного добутку. Як координати зазвичай використовують ключові слова і їх ваги.

Можна вважати, що задача формування ІПК зводиться до задачі автоматичної категоризації текстів. Проте для навчання більшості систем категоризації потрібна вже готова структура категорій з великою навчальною вибіркою по кожній категорії. Такі структури дуже загальні і не завжди адекватні інтересам користувачів. У зв'язку з цим більший інтерес становлять алгоритми, які ефективно вирішують дану задачу в умовах малих навчальних вибірок, коли навчання починається з одного-двох документів, і система класифікації перенавчається при додаванні нового документа.

Системи побудови ІПК також розрізняються способом отримання оцінки користувачем документа. Це може відбуватися в автоматичному або ручному режимі. Наприклад,

якщо користувач переглядає документ довше певного часу, то вважається, що користувач оцінив документ позитивно. У ручному режимі користувач явно вказує, які документи він вважає відповідними запиту. В будь-якому випадку, система оцінювання має бути організована так, щоб забезпечувати достатню точність при невеликих зусиллях з боку користувача.

З огляду на вищесказане, у нашій системі ІПК реалізується вектором, елементами якого є поняття з вказівкою ваги, що характеризує міру зацікавленості поняттям користувача. У пошуковій історії зберігаються всі запити користувача та позначаються всі документи які він оцінив при цьому запиті, оскільки припускається, що тільки у разі відповідності документа його інформаційним потребам він прочитає його та залишить оцінку в системі.

Міра зацікавленості користувача *u* в категорії *i* визначається як

$$c_{ui} = \frac{\sum_j \left(w_j^u \times w_j^i\right)}{\sqrt{\sum_j \left(w_j^u\right)^2} \times \sqrt{\sum_j \left(w_j^i\right)^2}}, \text{де}$$

w^i – інформаційний портрет розділу *i*, який формується наступним чином

$$w_j^i = \begin{cases} w_j^u, \text{якщо поняття } j \\ \text{відноситься до розділу } i \\ 0, \text{інакше} \end{cases}$$

w^u – інформаційний портрет користувача *u*.

Компоненти ІПК можуть змінюватись ґрунтуючись на пошуковому запиті, оцінюванні ресурсу при певному пошуковому запиті. Тобто $w_j = w_j + \frac{K}{n_l} * damp(q^s)$ за умови, що термін l запиту q^s відповідає терміну w_j, де n_l – число розділів, до яких можна віднести термін q_l^s,

$$K = \begin{cases} k1, \textit{якщо термін } q_l^s \textit{ відповідає терміну } w_j \\ k2, \textit{якщо термін } q_l^s \textit{ відповідає терміну } w_j \\ \textit{та була отримана оцінка ресурсу} \end{cases}$$

Далі отриманий вектор w^u нормується $w_j^u = \frac{w_j^u}{\sqrt{\sum_l w_l^u}}$. Функція *damp(q)* дозволяє здійснити пріоритизацію більш нових пошукових історій над більш старими без виключення останніх. Загальний вигляд цієї функції $damp(q^s = \frac{k}{p(s)})$, де k – певний коефіцієнт, s – значення яке характеризує давність запиту (порядковий номер, дата), *p(s)*– певна монотонно незростаюча функція, яка характеризує давність повідомлення.

Також система передбачає автоматичне додавання до інформаційних портретів нових понять. При аналізі пошукових запитів і оцінених ресурсів можуть виявлятися поняття, які не входять до інформаційного портрету, але, як виявляється, певним чином характеризують той чи інший розділ. Наприклад, є певне поняття *t*, якого немає в інформаційних портретах, але воно вживається в ряді запитів *{q1,..., qm }*, після видачі яких були оцінені ті чи інші ресурси. Якщо певні характеристики таких дій, зокрема кількість та частота, будуть перевищувати певне порогове значення, система може прийняти рішення про включення цього поняття в інформаційні портрети. Розділ або розділи, до яких буде віднесене це поняття будуть визначатися розділами документів, оцінених при роботі з запитами, до якого це поняття входило.

Побудова оцінок цінності документів та компетентності користувачів

Одне з головних завдань у нашій пошуковій системі є обробка інформації про якість того чи іншого ресурсу веб-порталу з врахуванням інформаційних потреб конкретного користувача.

Релевантність документа запиту у даній системі буде визначатися за формулою *R=R1+R2*, де *R1* – релевантність, отримана за допомогою базової пошукової системи, *R2* – оцінка якості документа, яка визначається розробленими у цій роботі методами, що будуть розглядатися далі.

Величина *R2* в свою чергу визначатиметься наступним чином:

$$R2 = k\frac{\sum_{i=1}^{n} R_i}{n}$$

де n – кількість розділів, до яких віднесений даний документ, R_i – оцінка якості документу відносно розділу *i*, *k* – коефіцієнт.

Передбачається, що якість документа для користувача у певному розділі залежить від власне цінності ресурсу відносно цього розділу та зацікавленості у розділі користувача. Пов'язуються ці величини функцією:

$$R_i = f(c_i, v_i)$$

де c_i– міра зацікавленості користувача у розділі i, v_i– цінність матеріалу відносно розділу i. Функція, яка використовується при обчисленні цієї величини задається наступним чином:

$$f(a,b) = \begin{cases} a*b, b \geq 0 \\ (2-a)*b, b < 0 \end{cases}$$

Враховуючи, що в даній системі зацікавленість користувача може бути представлена числом з інтервалу $c \in [0..1]$, а оцінка цінності ресурсу може бути від'ємною, наведена формула дозволяє коректно обробляти всі випадки.

Міра зацікавленості користувача у розділі визначається інформаційним портретом користувача. Інший фактор, який використовується при оцінці якості ресурсу – цінність матеріалу відносно певного розділу визначається наступним чином

$$v_i = \frac{\sum_{j=1..n} mark * a_{ij}}{n}$$

де a_{ij}– авторитет користувача j у розділі i, *mark* – оцінка користувача ресурсу. Варто зазначити, що користувач може оцінювати певний ресурс наступними оцінками: «*very bad*», «*bad*», «*normal*», «*good*», «*very good*», які відображаються в системі відповідно значеннями *-1, -0.5, 0, 0.5, 1*.

Авторитет користувача j у розділі i визначається наступним чином.

Початкове значення $a_{ij}=0$.

Після нових оцінювань матеріалів авторитети користувачів перераховуються за наступними формулами:

$$\Delta a_{ui} = K \frac{\sum_{j=1..n, j \neq u} d * f(c_{ij}, a_{ij})}{(n-1)}$$

$$\text{де } d = \begin{cases} 1, \text{ якщо } mark_j = mark_u \\ 0.5, \text{якщо } \left| mark_j - mark_u \right| = 0.5 \\ -\left| mark_j - mark_u \right|, \text{інакше} \end{cases}$$

Δa_{ui} – зміна авторитету користувача u в розділі i, K – коефіцієнт.

Загальна архітектура функціонування системи

У ході реалізації системи пошуку інформації тематичного веб-порталу було прийнято рішення вести її розробку на базі вже розробленої пошукової системи, яка має надавати наступні сервіси: індексування ресурсів порталу, пошук і визначення базових оцінок релевантності документів. Потім ця система повинна розширитися модулями представлення додаткових елементів інтерфейсу, які дозволяли б проводити роботу з розширеним функціоналом базової системи, модулями прийому та обробки наданої користувачем інформації.

З огляду на все вищезазначене було сформовано ряд вимог до базової пошукової системи:

- повинна бути з відкритим кодом (open-source), це необхідно як для додавання нових модулів до системи, так і для можливості певної адаптації її коду до вимог системи, яка розробляється;
- має мати достатньо розвинуті можливості у сфері пошуку та ранжування ресурсів;
- повинна мати індексатор, який добре працює як з текстами написаними латиницею, так і кирилицею.

Однією з пошукових систем, яка задовольняє цим умовам, є система PhpDig [9] , яка і була обрана як базова в нашій реалізації.

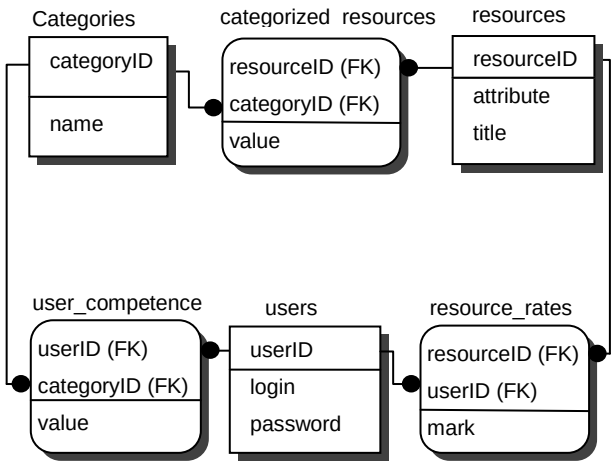

Рис.1. ER-модель частини БД, яка зберігає необхідні дані для проведення оцінювань

На рисунку 1 представлена ER-модель частини БД, яка зберігає дані для проведення оцінювань. Оцінку результатів діяльності віртуальної спільноти порталу система проводить за трьома основними сутностями: користувач (*users*), ресурс (*resources*), розділ (*Categories*). Таблиця *user_competence* зберігає дані про авторитет користувача в певній категорії, *categorized_resources* – цінність, яку мають ресурси в певних розділах, *resource_rates* – дані про оцінювання користувачами ресурсів.

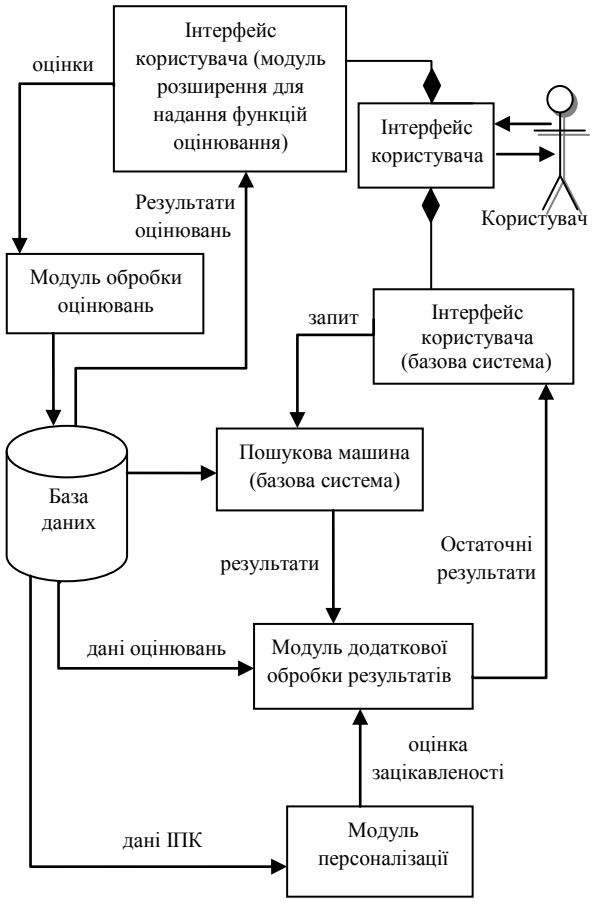

Рис.2. Загальна схема функціонування системи

На рисунку 2 представлена загальна схема функціонування системи. Інтерфейс користувача був розширений модулем обробки дій користувача щодо оцінювання ресурсів. Для здійснення початкового пошуку визначення релевантості певного ресурсу запиту система використовує функціональні можливості базової пошукової системи з подальшою їх зміною на основі алгоритмів персоналізації та врахування результатів діяльності віртуальних співтовариств, описаних вище.

Висновки

У роботі описано розв'язання актуальної наукової задачі розробки основних принципів функціонування та впровадження системи персоналізованого пошуку, яка дозволяла б здійснювати пошук документів тематичного порталу з урахуванням результатів кооперативної діяльності віртуального співтовариства порталу по оцінці документів. Детально представлені математичні моделі, на основі яких функціонує розроблена система, охарактеризовані основні аспекти її роботи, представлена архітектура.

Можна сформулювати основні подальші напрямки досліджень та розвитку та удосконалення системи. Актуальною задачею подальших досліджень при збереженні основних засад роботи може бути удосконалення математичних моделей, що лежать в основі функціонування системи.

Список використаних джерел

1. *Глибовець М. М.* Веб сервіси оброблення документів / Глибовець М. М., Жигмановський А. А. Заболотний Р. І., Захоженко П. О. – Київ: Національний університет "Києво-Могилянська академія". 2012. – 212 с.
2. *Глибовец Н. Н.* Создание динамической системы распространения контента с использованием протокола BitTorrent / Глибовец Н. Н., Мельник В. Е., Сидоренко М. О. // Компьютерная математика. – 2012. – №2. – С. 76-85.

References

1. GLIBOVEC M. M., ZHIGMANOVS'KIJ A. A., ZABOLOTNIJ R. I. and ZAHOZHENKO P. O. (2012) *Veb servisi obroblennja dokumentiv.* Kyiv: Nacional'nij universitet "Kievo-Mogiljans'ka akademija".
2. GLIBOVEC N. N., MEL'NIK B. E. and SIDORENKO M. O. (2012) *Sozdanie dinamicheskoj sistemy rasprostranenija kontenta s ispol'zovaniem protokola BitTorrent.* Komp'juternaja matematika. 2.p. 76-85.

3. *Глибовець М. М.* Алгоритм та онлайн-застосування пошуку осередків зацікавленості за обраною предметною областю / Глибовець М. М., Сидоренко М. О. // Вісник Київського національного університету імені Тараса Шевченка. Серія: фізико-математичні науки. – 2012. – № 3. – С. 133-140.
4. *Ландэ Д. В.* Поиск знаний в Internet. Профессиональная работа / Ландэ Д. В. – Москва: Вильямс, 2005. – 272 с.
5. *SearchWiki* — персонализированный поиск от *Google*– 2008. – Режим доступу: http://internetno.net/category/anonsi/searchwiki/
6. *Широков А.В.* Разработка модели информационного портрета пользователя для персонифицированного поиска. – Режим доступу: http://*page-adviser.ru/papers/pdf/shirokov.pdf*
7. *Qiu F., Cho J.* Automatic identification of user interest for personalized search. / Qiu F., Cho J. //: WWW *'06*: Proceedings of the 15th international conference on World Wide Web, 22 – 26 May 2006, New York, USA : ACM Press. – New York. – 2006. – P.727-736
8. *Trajkova J.* Improving Ontology-Based User Profiles / J. Trajkova, S. Gauch – Режим доступу: http://eolo.cps.unizar.es/docencia/MasterUPV/Articulos/Improving%20Ontology-Based%20User%20Profiles.pdf
9. Режим доступу: http://www.phpdig.net/

3. GLIBOVEC M. M. and SIDORENKO M. O. (2012) Algoritm ta onlajn-zastosuvannja poshuku oseredkiv zacikavlenosti za obranoju predmetnoju oblastju. *Visnyk Kyivskoho natsionalnoho universitetu imeni Tarasa Scevchenka. Seriya fizyko-matematychni nauky*. 3. p. 133-140.
4. LANDJE D. V. (2005) *Poisk znanij v Internet. Professional'naja rabota*. Moskva: Vil'jams. Internetno (2008)
5. *SearchWiki* — personalized search from *Google* – Available from: http://internetno.net/category/anonsi/searchwiki/
6. SHIROKOV A.V. *Razrabotka modeli informacionnogo portreta pol'zovatelja dlja personificirovannogo poiska.* – Available from: http://*page-adviser.ru/papers/pdf/shirokov.pdf*
7. QIU, F. and CHO J. (2006) Automatic identification of user interest for personalized search. In *WWW '*06*: Proceedings of the 15th international conference on World Wide Web*, Monday 22nd to Friday 26th May 2006. New York: ACM Press. pp.727-736
8. TRAJKOVA, J. and GAUCH, S. Improving Ontology-Based User Profiles. – Available from: http://eolo.cps.unizar.es/docencia/MasterUPV/Articulos/Improving%20Ontology-Based%20User%20Profiles.pdf
9. Available from: http://www.phpdig.net/

Надійшла до редколегії 29.04.14