УДК 027+[004.77+004.82]

Дмитро Ланде,
зав. відділу Інституту проблем реєстрації інформації НАН України, д-р техн. наук

Ольга Баркова,
заст. директора ТОВ «Спеціалізований центр БАЛІ»

Електронна бібліотека як середовище адаптивного агрегування інформації

Автори статті пропонують узагальнену схему функціонування мережі електронних бібліотек, яка ґрунтується на феномені конвергенції двох напрямів діяльності бібліотеки – обслуговування користувачів і формування фонду. Розглядаються окремі параметри мережі електронних бібліотек, дається оцінка інтенсивності поповнення фонду електронної бібліотеки у складі піринговоі бібліотечної мережі.

К л ю ч о в і с л о в а: електронна бібліотека, математична модель, комплектування, пірингова мережа, мережа масового обслуговування.

The generalized schema of operation of an electronic libraries network is proposed, which is based on a phenomena of the confluence of the two main functions of library – servicing customers and gathering collections. Some parameters of electronic libraries network have been examined. The estimate of intensity of collections augment of an electronic library as a part of a peer-to-peer network has been performed.

K e y w o r d s: electronic library, math model, gathering of collections, mass services network.

У глобальному інформаційному просторі, який ґрунтується на комп'ютерних мережах і цифрових технологіях, електронні бібліотеки стають ключовими інформаційними системами, що здійснюють накопичення і систематизацію документних ресурсів, інформаційне обслуговування користувачів шляхом організації доступу до першоджерел, систематизованого знання та надання комплексу сучасних мережевих інформаційних сервісів.

Бібліотека як інформаційна система подолала шлях від традиційної до автоматизованої, а згодом і до електронної бібліотеки з цифровим поданням об'єктів збереження (документних ресурсів). Її інформаційно-пошукові засоби змінювалися від паперових каталогів і картотек до електронних, які формуються засобами автоматизованих бібліотечно-інформаційних систем (АБІС) і надають доступ до «зовнішніх», відносно до цих АБІС, об'єктів-ресурсів у цифрових медіа форматах, до систем повнотекстового пошуку для текстових ресурсів. Що характерно, останні забезпечують пошук не тільки документів, але й інформації в інформаційному змісті документа або масиву документів. При цьому атрибутивний пошук за метаданими залишається актуальним, особливо для нетекстових об'єктів електронних бібліотек.

Техніко-технологічне середовище сучасної електронної бібліотеки на сьогодні є сукупністю внутрішньої інформаційно-технологічної інфраструктури автоматизованої бібліотеки з технічними засобами збереження цифрових об'єктів та зовнішнього інформаційно-комунікаційного середовища, яке розвивається на базі глобальних комп'ютерних мереж. При цьому, розвиток комп'ютерних мереж і засобів автоматизації бібліотек створює сприятливі умови для забезпечення електронних бібліотек необхідним комплексом інструментальних засобів.

З іншого боку, у зв'язку з еволюційними змінами техніко-технологічного середовища функціонування електронних документних ресурсів, зокрема, з розвитком інформаційних мереж та технологій інформаційного пошуку, постає необхідність розроблення нових мережевих моделей електронних бібліотек та відповідних технологій використання розподілених електронних документних ресурсів. Розвиток комп'ютерних мереж, насамперед, вимагає певного адаптування інфраструктур та технологій електронних бібліотек до сучасних методів й технологій мережевого середовища, зокрема таких, як хмарні обчислювання та пірінгові мережі.

У статті розкриваються нові концептуальні підходи до розбудови розподіленої електронної

бібліотеки із застосовуванням пірингової мережі (Peer-to-peer, P2P) та використання у такій мережі технології адаптивного агрегування.

Головною метою діяльності електронної бібліотеки є максимально повне та оперативне обслуговування користувачів. Щоб досягти її, потрібно володіти повною власною базою даних (тобто фондом) електронної бібліотеки, міжбібліотечним зв'язком (обміном) з найкращими електронними бібліотеками, мати змогу залучати зовнішні ресурси, користуватися послугами кваліфікованих експертів, задіювати кращі методики пошуку в мережевому середовищі. Адаптивне агрегування документної інформації полягає у тому, що комплектування інформації здійснюється на основі мережевих ресурсів, які збираються у фонд (базу даних) електронної бібліотеки відповідно до інтересів користувачів, виражених у запитах, тобто адаптуються під їх інформаційні потреби [1].

Аналогічний підхід використовується в АБІС у типовому модулі оперативного каталогізування у процесі книговидачі, коли документ, який ще не відображений в електронному каталозі, обробляється під час його повернення користувачем до бібліотеки. Таким чином здійснюється ретрокаталогізація найбільш запитуваної частини бібліотечного фонду.

Під час реалізації електронних мереж доцільно застосовувати пірингові мережі (Peer-to-peer, P2P – рівний з рівним), які засновані на рівноправ'ї учасників. У таких мережах відсутні виділені сервери, а кожен вузол (peer) є і клієнтом, і сервером. На відміну від мереж з архітектурою «клієнт/сервер», така організація мережі дає змогу зберігати працездатність мережі при довільній кількості та поєднанні вузлів. У багатьох випадках P2P є накладеними мережами, які використовують існуючі транспортні протоколи стека TCP/IP, – TCP або UDP. Слід зазначити, що на практиці пірингові мережі складаються з вузлів, кожен з яких взаємодіє лише з деякою підмножиною інших вузлів мережі (через обмеженість ресурсів) [2].

Архітектура пірингових мереж принципово відрізняється від традиційної централізованої архітектури «клієнт/сервер», де мережа залежить від центральних вузлів (серверів), які забезпечують підключені до мережі термінали (клієнти) необхідними сервісами. У централізованій архітектурі ключова роль відводиться серверам, які визначають мережу незалежно від наявності клієнтів, тобто при падінні цих серверів мережа стає непрацездатною. Цілком очевидно, що зростання кількості клієнтів мережі, на зразок «клієнт/сервер», призводить до зростання навантажень на серверну частину, внаслідок чого вона може виявитися перенавантаженою. Децентралізована пірингова мережа, навпаки, стає продуктивнішою у разі збільшення кількості вузлів, підключених до неї. Дійсно, кожен вузол додає у мережу P2P свої ресурси (дисковий простір і обчислювальні можливості). В результаті сумарні ресурси мережі збільшуються.

Проаналізуємо модель мережі електронних бібліотек, яку вважатимемо пірінговою мережею, до того ж сполученою з глобальною мережею Інтернет, яка розглядається як зовнішнє середовище (рис. 1). Особливість цієї архітектури полягає у конвергенції двох основних гілок функціонування бібліотеки – обслуговування читачів (користувачів) і комплектування (поповнення власного фонду – локальної бази даних).

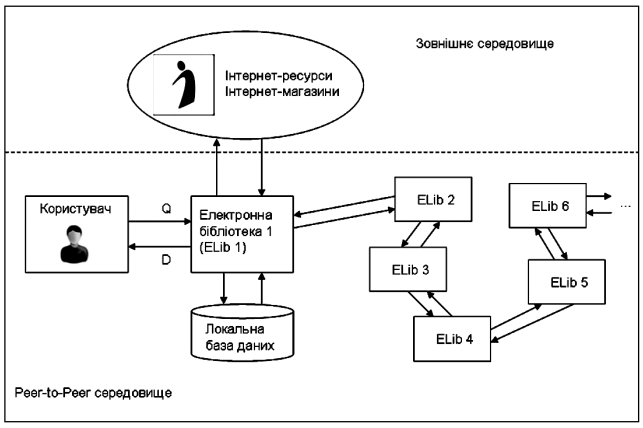

Рис. 1. Функціональна схема, що поєднує обслуговування користувача і комплектування

На рис. 1 користувач передає до електронної бібліотеки 1 (ELib 1) запит Q, і отримує відгук – документ D.

А тепер розглянемо деякі алгоритми пошуку в пірінгових мережах, які можуть застосовуватися у мережах електронних бібліотек.

У більшості пірингових мереж, орієнтованих на обмін файлами, використовуються два види об'єктів, яким приписуються відповідні ідентифікатори (ID): вузли і ресурси (наприклад, файли), що характеризуються ключами (Key), тобто мережа може бути представлена двовимірною матрицею розмірності MN, де M – кількість вузлів, N – кількість ресурсів. *Алгоритм пошуку ресурсів за ключами* зводиться до знаходження ID вузла, на якому зберігається ключ ресурсу [3]. При цьому запит користувача проходить певний маршрут і досягає вузла, де розміщено ключ, котрий збігається із запитом. Далі цей вузол пересилає першому вузлу

адреси всіх вузлів, що мають ресурс, відповідний ключу, за яким проводився пошук.

Метод широкого первинного пошуку (Breadth First Search, BFS) у мережі P2P розмірності реалізується таким чином. Вузол q генерує запит, який адресується до всіх сусідів (найближчих за деякими критеріями вузлів). Коли вузол p отримує запит, виконується пошук у його локальному індексі. Якщо деякий вузол r приймає запит (Query) і обробляє його, то він генерує повідомлення-відгук (QueryHit), щоб повернути результат. Повідомлення QueryHit включає інформацію про релевантні документи, яка доправляється мережею вузла, котрий робить запити (рис. 2).

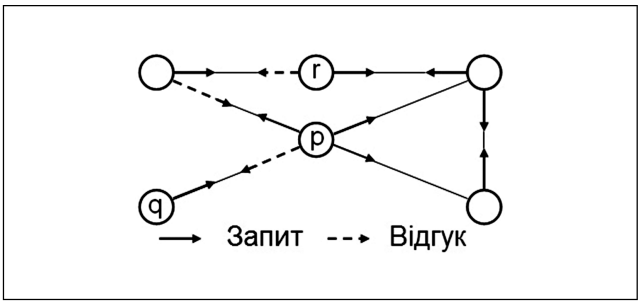

Рис. 2. Метод BFS

Коли вузол q отримує QueryHits від більш ніж одного вузла, він може завантажити файл з найбільш доступного ресурсу. Повідомлення QueryHit повертаються тим же шляхом, що і первинний запит.

У BFS кожен запит викликає надмірне навантаження мережі, оскільки він передається по всіх зв'язках. При цьому вузол з низькою пропускною спроможністю може стати вузьким місцем. Одним з методів, що дає змогу уникнути перевантаження всієї мережі повідомленнями, полягає у приписуванні кожному запиту параметра часу життя (Time-to-live, TTL). Параметр TTL визначає максимальне число переходів, якими можна пересилати запит. У разі типового пошуку початкове значення для TTL становить 5–7 і зменшується кожного разу, коли запит пересилається. Якщо TTL стає рівним 0, повідомлення більше не передається. BFS гарантує високий рівень якості пошуку за рахунок великого числа переданих повідомлень.

Метод випадкового широкого первинного пошуку (Random Breadth First Search, RBFS) був запропонований як поліпшення BFS [4]. У методі RBFS (рис. 3) вузол q пересилає пошукове розпорядження тільки частині вузлів мережі, вибраної у випадковому порядку. Яка саме частина вузлів – це параметр методу RBFS. Перевага RBFS полягає в тому, що тут немає потреби у глобальній інформації про стан контенту мережі; вузол може отримувати локальні рішення так швидко, як це буде потрібно. З іншого боку, цей метод ймовірнісний. Тому деякі великі сегменти мережі можуть виявитися недосяжними.

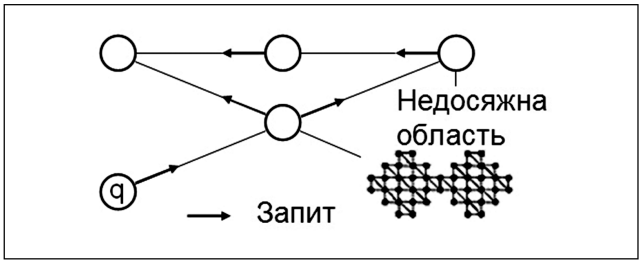

Рис. 3. Метод RBFS

Інтелектуальний пошуковий механізм (Intelligent Search Mechanism, ISM) (рис. 4) забезпечує швидкість і ефективність пошуку інформації за рахунок мінімізації витрат на зв'язок, тобто на число повідомлень, котрі передаються між вузлами, та мінімізації кількості вузлів, які опитуються для кожного пошукового запиту [4]. Щоб досягти цього, для кожного запиту оцінюються лише ті вузли, які найбільше відповідають даному запиту.

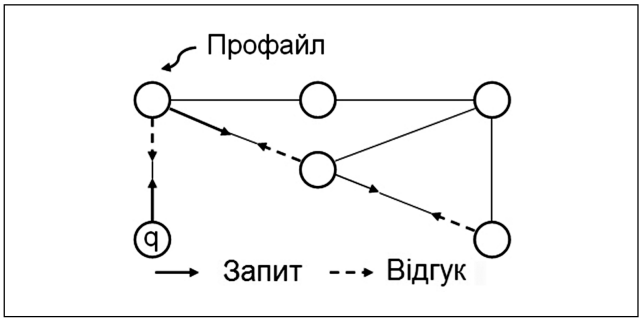

Рис. 4. Метод ISM

Інтелектуальний пошуковий механізм складається з двох компонентів – профайлу (profile) і способу його ранжирування, так званого рангу релевантності. Кожен вузол мережі будує інформаційний профайл для кожного з сусідніх вузлів. Профайл містить останні відповіді кожного з вузлів. За допомогою рангу релевантності здійснюється ранжирування профайлів вузлів для вибору тих сусідніх, які даватимуть найбільш релевантні документи за запитом. Механізм профайлів використовується для того, щоб зберігати останні запити, а також кількісні характеристики результатів пошуку.

При реалізації моделі ISM використовується єдиний стек запитів, у якому зберігається по T запитів для N вузлів. Як тільки стек заповнюється, відбувається заміна «того запиту, який не використ-

товувався найдовше» (Least Recently Used, LRU) для збереження останніх запитів. Функція «ранг релевантності» (Relevance Rank, RR) використовується вузлом P_l, щоб виконувати оперативну класифікацію його сусідів для визначення тих, які слід опитувати першими за запитом q. Для обчислення рангу релевантності RR кожного вузла P_i, P_l порівнює q зі всіма запитами у структурі профайлу, для якого відомий список відповідей на попередні запити, і обчислює:

$$RR(P_i, q) = \sum_{j \in Q} Sim(q_j, q)^{\alpha} \cdot S(P_i, q_j),$$

де α – параметр, який задає вагу запитів. У цій формулі Q – множина запитів, на які була відповідь у вузла P_i; S (P_i, q_j) – кількість результатів, повернутих вузлом P_i за запитом q_j; оцінка Sim розраховується за правилом, яке розглядається у векторно-просторовій моделі пошуку:

$$Sim(q_j, q) = \frac{q_j \cdot q}{|q_j||q|}.$$

У цьому випадку запит розглядається як n-вимірний вектор у просторі слів $q_j=(q_j^{(1)}, q_j^{(2)},..., q_j^{(n)}$, де n – розмір словника, а $q_j^{(1)}$ дорівнює одиниці, якщо запит q_j містить слово з номером i, або дорівнює нулю, якщо він його не містить. Операція $q_j^{(1)} \cdot q$ відповідає скалярному добутку відповідних векторів, а $|q_j||q|$ – добутку норм цих векторів.

Ранг релевантності RR забезпечує вищий ранг вузла, який повертає більше результатів. Крім того, використовується параметр α, який дає змогу збільшувати вагу запитів, найбільш подібних початковому.

Передбачається можливість двох основних сценаріїв формування відгуку системи: електронна бібліотека надає читачеві:

1) перший релевантний документ і зупиняє обслуговування;

2) максимально повну добірку, що відповідає запиту.

Якщо позначити через L поріг відповідності відгуку електронної бібліотеки запиту користувача q, то відгук електронної бібліотеки $R(q)$ на запит q дорівнюватиме одиниці, якщо у відгуку наявний такий документ d, що $Sim\ (q,d)>L$, інакше $R(q)$ буде дорівнювати нулю.

У випадку сценарію 1, якщо після звернення до локальної бази даних $R(q)$=1, то запит користувача задовольняється, інакше йде звернення до мережі електронних бібліотек. Якщо у будь-якій з них здійснюється $R(q)$=1, то також вважається, що запит виконаний, інакше йде звернення до зовнішнього середовища, зокрема Інтернет-ресурсів.

Для сценарію 3 пошук не переривається навіть у випадку $R(q)$=1 у будь-якому вузлі. Обмеження – фізичний час або TTL.

Розглянемо мережу електронних бібліотек як мережу масового обслуговування [5], для якої розподіл вхідних потоків (запитів) приймається пуасонівським, а розподіл часу обслуговування – експоненційним. Вибір даних ймовірнісних законів як апроксимацій реальних випадкових розподілів, крім загальноприйнятного прийому аналітичного спрощення, обгрунтовується: дані апроксимації у низці випадків показують суперечність цих розподілів реальним даним, а умови стаціонарності, ординарності і відсутності післядії задовольняють реальні бібліотечні потоки.

Позначимо:

λ_i – інтенсивність потоку заявок в електронній бібліотеці;

μ_i – інтенсивність обслуговування заявок в електронній бібліотеці;

γ_i– інтенсивність поповнення фонду електронної бібліотеки.

Насправді, γ_i відповідає інтенсивність обслуговування запитів в електронній бібліотеці i з незадовільним результатом, коли необхідний документ знаходиться у інших джерелах, після чого автоматично заноситься до фонду електронної бібліотеки i, тобто $R_1(q)$=0, але при цьому $\exists j : R_j(q) = 1,\ j = 0,..., M$, де – множина тих вузлів, куди пересилається запит відповідно до обраного алгоритму пошуку (індекс $j=0$ відповідає зовнішньому середовищу, яке розглядається в моделі як узагальнений вузол).

Серед параметрів, які відповідають усій мережі масового обслуговування, назвемо такі: кількість систем масового обслуговування, (M), що входять до мережі; матрицю ймовірності переходів P=$||p^j_i||$; інтенсивність вхідних потоків (завантаження даних із зовнішнього середовища); середній час обслуговування запитів у вузлах мережі масового обслуговування ($\overline{T}_{O1},..., \overline{T}_{OM},$), середній час обробки запиту в мережі $\overline{T}_{\pi}$ (час між входом заявки в мережу і виходом з неї).

Розглянемо окремий випадок мережі електронних бібліотек, коли запити користувачів обслуговуються послідовно в кожній з бібліотек. Нехай відомо: середній час обслуговування запитів ($\overline{T}_{O1},..., \overline{T}_{OM},$) та оцінки ймовірностей вдалого ($R_i(q)$=1) обслуговування запитів у цих електрон-

них бібліотеках (отримані емпірично) ($p_1,...,p_M$). Відповідно, позначимо $q_1,...,q_M$ – ймовірності невдалого обслуговування запиту ($R_i(q)$=0). Тоді середній час перебування запиту в мережі електронних бібліотек (обробки запиту) становитиме:

$$\overline{T}_\pi = \overline{T}_{o1} + (1-p_1)(\overline{T}_{o2} + (1-p_2)(\overline{T}_{o3} + ... + (1-p_{M-1})\overline{T}_{oM}))...) =$$

$$= \overline{T}_{o1} + q_1(\overline{T}_{o2} + q_2(\overline{T}_{o3} + ... + q_{M-1}\overline{T}_{oM}))...) =$$

$$= \overline{T}_{o1} + q_1\overline{T}_{o2} + q_1q_2\overline{T}_{o3} + ... + q_1q_2...q_{M-1}\overline{T}_{oM} =$$

$$= \sum_{i=1}^{N} \overline{T}_{oi} \prod_{j=1}^{N} q_{j-1}.$$

Отже, задача мінімізації середнього часу перебування запиту в мережі електронних бібліотек може бути записана у формальному вигляді:

$$\overline{T}_\pi = \sum_{i=1}^{N} \overline{T}_{oi} \prod_{j=1}^{N} q_{j-1} \to \min$$

при $0 < q_i < 1,\ \forall i = 1,...,M$
та допущенні $q_0 = 1,\ \overline{T}_{oi} << \overline{T}_{o0}, \forall i = 1,...,M.$

Таким чином, задача зводиться до знаходження оптимального ранжирування вузлів мережі – електронних бібліотек за двома параметрами – середнім часом обробки запиту і ймовірністю отримання позитивного (або негативного) результату. Це задача нелінійного програмування, яку при обмеженій кількості електронних бібліотек можна розв'язати навіть методом прямого перебору.

Як окремий випадок розглянемо мережу з десяти вузлів, середній час обробки запиту в яких становить відповідно 1, 2, 3,…, 10 с, а ймовірності задоволення запиту відповідно, – 0,1, 0,2, 0,3, …, 1. На рис. 5 наведено залежності середнього часу перебування запиту в мережі від кількості переходів, коли звернення надходять від першого до десятого вузла послідовно, а також у зворотному порядку обходу вузлів.

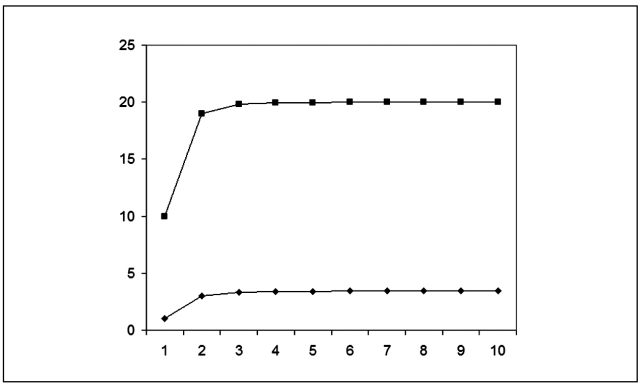

Рис. 5. Залежності середнього часу перебування запиту в мережі від кількості переходів (◆ – прямий порядок, ■ – зворотний порядок)

Наведемо формальний вираз для визначення інтенсивності поповнення фонду електронної бібліотеки i (не заперечуючи загальності, можна вважати i=1).

Справедливе співвідношення $\gamma_1 = P_s\lambda_1$, де P_s – ймовірність події, що запит не було задоволено у першій електронній бібліотеці, але було задоволено у деякому вузлі j, де j=2,..., M. Ймовірність того, що запит не задоволено у жодному з цих вузлів, дорівнює: $q_2q_3\cdots q_M$. Відповідно, ймовірність того, що вона знайшлась щонайменш в одному вузлі з j=2,..., M дорівнює $1-q_2q_3\cdots q_M$. Звідси інтенсивність поповнення фонду електронної бібліотеки визначається формулою:

$$\gamma_1 = \lambda_1 q_1 (1 - q_2 q_3 \cdots q_M)$$

або у загальному вигляді:

$$\gamma_j = \lambda_j q_j (1 - \prod_{i=1, i\neq j}^{M} q_i) \quad .$$

Таким чином, у статті запропоновано узагальнену схему функціонуванння мережі електронних бібліотек, яка ґрунтується на феномені конвергенції двох гілок функціонування бібліотеки – обслуговування читачів і комплектування. Розглянуто найважливіші параметри мережі електронних бібліотек, зокрема середній час перебування запиту користувача (читача) в цій мережі, надано оцінку інтенсивності поповнення фонду електронної бібліотеки у випадку побудови міжбібліотечних зв'язків і технології обігу інформації за принципом пірингових мереж.

Висновок

Сучасний стан розвитку мережевих систем і технологій зумовлює необхідність розбудови електронних бібліотек як розподілених систем, а також застосування технологій розподіленого пошуку. На думку авторів, найбільшу ефективність процесів розподіленого пошуку зараз демонструють піринговi мережі. Тому електронну бібліотеку слід створювати як систему з розподіленими інформаційними ресурсами та формувати системи взаємопов'язаних електронних бібліотек корпоративного, галузевого та національного рівнів. Ефективність функціонування таких систем залежить від технологічності формування їх ресурсів та організаційно-технологічної взаємодії суб'єктів інформаційної інфраструктури.

Список використаних джерел

1. *Додонов А. Г., Ландэ Д. В.* Методы и средства мониторинга, адаптивного агрегирования и обобщения информационных потоков // Информационные технологии и безопасность. Проблемы научного и правового обеспечения кибербезопасности в современном мире. Материалы международной научной конференции ИТБ-2011. – К.: ИПРИ НАН Украины, 2011. – С. 6–9.

2. *Ландэ Д. В., Снарский А. А., Безсуднов И. В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М. : Либроком (Editorial URSS), 2009. – 264 с.

3. *Уолрэнд Дж.* Введение в теорию сетей массового обслуживания. – М. : Мир, 1993. – 336 с.

4. *Kalogeraki V., Gunopulos D. Zeinalipour-Yazti D.* A Local Search Mechanism for Peer-to-Peer Networks // Proc. of CIKM'02, McLean VA, USA, 2002.

5. *Zeinalipour-Yazti D., Kalogeraki Vana Gunopulos Dimitrios.* Information Retrieval in Peer-to-Peer Networks [Elektronic resours] // IEEE CiSE Magazine, Special Issue on Web Engineering, 2004. – P. 1–13. – URL : www.cs.ucr.edu/~csyiazti/papers/ cise2003/cise2003.pdf