

ЕЛЕКТРОННА БІБЛІОТЕКА – СЕРЕДОВИЩЕ АДАПТИВНОГО АГРЕГУВАННЯ ІНФОРМАЦІЇ

An Electronic Library As An Environment Of Adaptive Aggregation Of Informarion
Dmitry Lande, Olga Barkova

Д.В. Ланде, О.В. Баркова

У статті запропоновано узагальнену схему функціонування мережі електронних бібліотек, яка базується на феномені конвергенції двох гілок функціонування бібліотеки – обслуговування читачів і комплектування. Розглянуто окремі параметри мережі електронних бібліотек. Надано оцінку інтенсивності поповнення фонду електронної бібліотеки у складі пірингової бібліотечної мережі.

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.

Ключові слова: електронна бібліотека, математична модель, комплектування, пірингова мережа, мережа масового обслуговування

Keywords: electronic library, peer-to-peer network, math model, gathering of collections, queueing network

Метою функціонування електронної бібліотеки є максимально повне та оперативне обслуговування користувача. Для цього повинна бути максимально повною власна база даних електронної бібліотеки, налагоджений міжбібліотечний зв'язок (обмін) з найкращими електронними бібліотеками, залучені зовнішні ресурси пошуку електронних видань, задіяні кращі методики та експерти по пошуку в мережевому середовищі. Адаптивне агрегування документальної інформації полягає у тому, що комплектування інформації здійснюється на основі мережевих ресурсів, які збираються у фонд (базу даних) електронної бібліотеки у відповідності з потребами користувачів, виражених у запитах, тобто адаптуються під їх інформаційні потреби [1].

При 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 як клієнтом, так і сервером. На відміну від мереж з архітектурою «клієнт-сервер», така організація мережі дозволяє зберігати працездатність мережі при довільній кількості та поєднанні вузлів. У багатьох випадках P2P є накладеними мережами, що 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 цій централізованій архітектурі ключова роль відводиться серверам, які визначають мережу незалежно від наявності клієнтів, тобто при падінні цих серверів мережа стає неробочою. Очевидно, що зростання кількості клієнтів мережі типу

«клієнт/сервер» приводить до зростання навантажень на серверну частину, внаслідок чого вона може виявитися перенавантаженою. Децентралізована пірінговая мережа, навпаки, стає продуктивнішою при збільшенні кількості вузлів, підключених до неї. Дійсно, кожен вузол додає в мережу P2P свої ресурси (дисковий простір і обчислювальні можливості), в результаті сумарні ресурси мережі збільшуються.

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

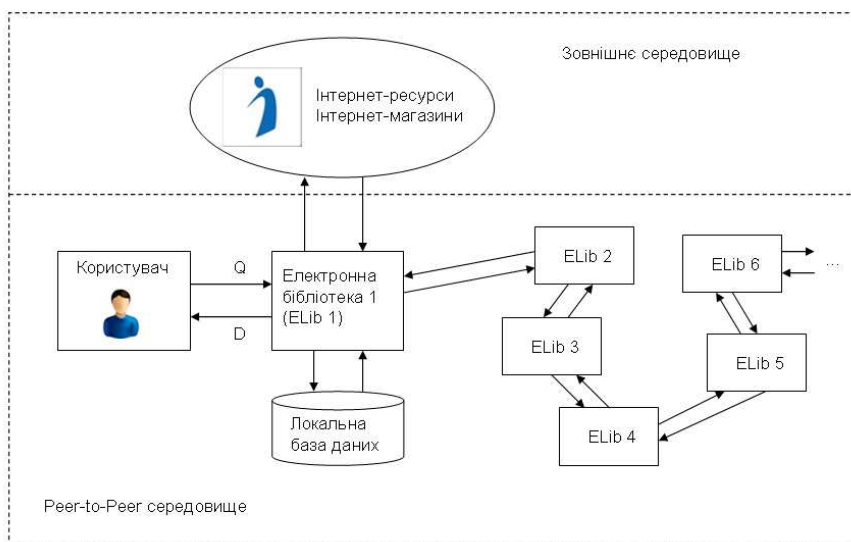


Рис. 1. Функціональна схема, що поєднує обслуговування користувача і комплектування

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

Розглянемо деякі алгоритми пошуку в пірінгових мережах, які можуть застосовуватися в мережах електронних бібліотек.

У більшості пірінгових мереж, орієнтованих на обмін файлами, використовується 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 по ключах** зводиться до знаходження ID вузла, на якому зберігається ключ ресурсу [3]. При цьому запит користувача проходить певний маршрут і досягає вузла, на якому знаходиться ключ, що співпадає з запитом. Далі цей вузол пересилає першому вузлу адреси всіх вузлів, що мають ресурс, відповідний ключу, за яким проводиться пошук.

Метод широкого первинного пошуку (Breadth First Search, BFS) в мережі P2P розмірності N реалізується таким чином. Вузол q генерує запит, який адресується до всіх сусідів (найближчих за деякими критеріями вузлів). Коли вузол p отримує запит, виконується пошук в 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 доставляється по мережі вузлу, що запрошує (рис. 2).

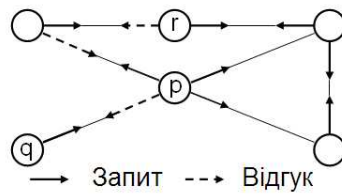


Рис. 2. Метод BFS

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 дозволяє уникнути перевантаження всієї мережі повідомленнями полягає в приписуванні кожному запиту параметра часу життя (Time-to-live, TTL). Параметр TTL визначає максимальне число переходів, по яких можна пересилати запит. При типовому пошуку початкове значення для TTL складає 5-7 і зменшується кожного разу, коли запит пересилається. Коли 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 не потрібно глобальної інформації про стан контенту мережі; вузол може отримувати локальні рішення так швидко, як це буде потрібно. З іншого боку, цей метод ймовірносний. Тому деякі великі сегменти мережі можуть виявитися недосяжними.

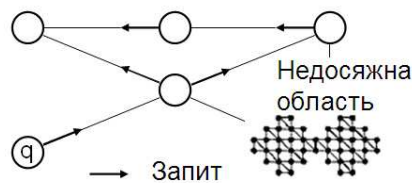


Рис. 3. Метод RBFS

Інтелектуальний пошуковий механізм (Intelligent Search Mechanism, ISM) (рис. 4) забезпечує поліпшення швидкості і ефективності пошуку інформації за рахунок мінімізації витрат на зв'язок, тобто на число повідомлень, що 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 найбільш відповідають даному запиту.

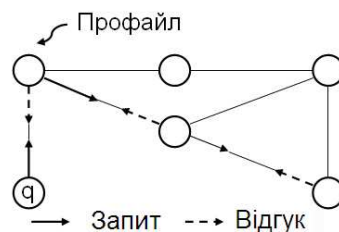


Рис. 4. Метод ISM

Інтелектуальний пошуковий механізм складається з двох компонент – профайлу (profile) і способу його ранжирування, так званого рангу релевантності. Кожен вузол

мережі буде 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, в якому зберігається по T запитів для N вузлів. Як тільки стек заповнюється, відбувається заміна «того запиту, який не використовувався найдовше» (Least Recently Used, LRU) для збереження останніх запитів. Функція «ранг релевантності» (Relevance Rank, RR) використовується вузлом P_i , 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_i порівнює q зі всіма запитами в 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, задаючий вагу запитів. У цій формулі Q - множина запитів, на які була відповідь у вузла P_i ; $S(P_i, q_j)$ - кількість результатів, що повернулися вузлом P_i за запитом q_j ; оцінка Sim розраховується за правилом, розглянутим 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 $q_j = (q_j^{(1)}, q_j^{(2)}, \dots, q_j^{(n)})$, де n - розмір словника, а $q_j^{(i)}$ дорівнює одиниці, якщо запит q_j містить слово з номером i , а якщо не містить, то дорівнює нулю. Операція $q_j \cdot q$ 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 дозволяє 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 буде дорівнювати одиниці, якщо у відгуку існує такий документ d , що $Sim(q, d) > L$, інакше $R(q)$ буде дорівнювати нулю.

У випадку сценарію 1, якщо вже 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 задовільнений, 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 загальноприйнятого прийому аналітичного спрощення, обґрунтовується, що дані апроксимації у ряді випадків показують непротиворічність даних розподілів реальним даним, а умови стаціонарності, ординарності і відсутності післядії задовільняють реальним бібліотечним потокам.

Позначимо:

λ_i – інтенсивність потоку заявок в електронній бібліотеці i ;

μ_i – інтенсивність обслуговування заявок в електронній бібліотеці i ;

γ_i – інтенсивність поповнення фонду електронної бібліотеки i .

На справді, 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 $\exists j: R_j(q) = 1, j = 0, \dots, M$, де M – множина тих вузлів, куди пересилається запит у відповідності з вибраним алгоритмом пошуку (індекс $j = 0$ відповідає зовнішньому середовищу, яке розглядається в моделі як узагальнений вузол).

Серед параметрів, які відповідають всій мережі масового обслуговування назовемо наступні: кількість систем масового обслуговування (M), що входять до мережі, матрицю ймовірності переходів $P = \|p_i^j\|$, інтенсивності вхідних потоків (завантаження даних із зовнішнього середовища), середній час обслуговування запитів в узлах мережі масового обслуговування ($\bar{T}_{o1}, \dots, \bar{T}_{oM}$), середній час обробки запиту в мережі \bar{T}_π (час між входом заявки в мережу і виходом з неї).

Розглянемо частковий випадок мережі електронних бібліотек, коли запити користувачів обслуговуються послідовно в кожній з бібліотек. Нехай відомо: середній час обслуговування запитів ($\bar{T}_{o1}, \dots, \bar{T}_{oM}$) та оцінки ймовірностей вдалого ($R_i(q) = 1$) обслуговування запитів в цих електронних бібліотеках (отримані емпірично) (p_1, \dots, p_M). Відповідно, позначимо q_1, \dots, q_M – ймовірності невдалого обслуговування запиту ($R_i(q) = 0$). Тоді середній час перебування запиту в мережі електронних бібліотек (обробки запиту) складатиме:

$$\begin{aligned} \bar{T}_\pi &= \bar{T}_{o1} + (1 - p_1)(\bar{T}_{o2} + (1 - p_2)(\bar{T}_{o3} + \dots + (1 - p_{M-1})\bar{T}_{oM}))) \dots = \\ &= \bar{T}_{o1} + q_1(\bar{T}_{o2} + q_2(\bar{T}_{o3} + \dots + q_{M-1}\bar{T}_{oM}))) \dots = \\ &= \bar{T}_{o1} + q_1\bar{T}_{o2} + q_1q_2\bar{T}_{o3} + \dots + q_1q_2 \dots q_{M-1}\bar{T}_{oM} = \\ &= \sum_{i=1}^N \bar{T}_{oi} \prod_{j=1}^N q_{j-1}. \end{aligned}$$

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

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

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

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

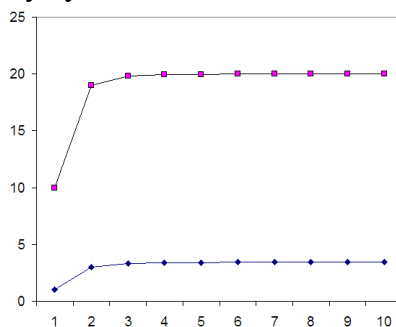


Рис. 5. Залежності середнього часу перебування запиту в мережі від кількості переходів (♦ – прямий порядок, ■ – зворотний порядок)

Наведемо формальний вираз для визначення інтенсивності γ_i 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 $j = 2, \dots, M$. Ймовірність того, що запит не задоволено у жодному з цих вузлів, дорівнює: $q_2 q_3 \dots q_M$. Відповідно, ймовірність того, що вона знайшлась щонайменш у одному вузлі з $j = 2, \dots, M$ дорівнює $1 - q_2 q_3 \dots q_M$. Звідси інтенсивність поповнення фонду електронної бібліотеки визначається формулою:

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

або у загальному вигляді:

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

Таким чином, у статті запропоновано узагальнену схему функціонування мережі електронних бібліотек, яка базується на феномені конвергенції двох гілок функціонування бібліотеки – обслуговування читачів і комплектування. Розглянуто найважливіші параметри мережі електронних бібліотек, як 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.

Література:

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. D. Zeinalipour-Yazti, Vana Kalogeraki, Dimitrios Gunopulos. Information Retrieval in Peer-to-Peer Networks // IEEE CiSE Magazine, Special Issue on Web Engineering, 2004. – p. 1-13. URL: www.cs.ucr.edu/~csyiazti/papers/cise2003/cise2003.pdf
4. V. Kalogeraki, D. Gunopulos and D. Zeinalipour-Yazti, "A Local Search Mechanism for Peer-to-Peer Networks". Proc. of CIKM'02, McLean VA, USA, 2002.
5. Уолрэнд Дж. Введение в теорию сетей массового обслуживания. – М.: Мир, 1993. – 336 с.