УДК 515.878

Informatics and Mathematical Methods in Simulation
Vol. 7 (2017), No. 4, pp. 300-306

МОДЕЛЮВАННЯ ТА ОПТИМІЗАЦІЯ ДОСТУПУ ДО СТОРІНОК WEB-САЙТУ ДЛЯ РІЗНИХ ЗАКОНІВ РОЗПОДІЛУ ЙМОВІРНОСТЕЙ ЗВЕРТАННЯ ДО СТОРІНОК

М.І. Юськів, Г.Г. Цегелик

Львівський національний університет ім. Івана Франка,
вул. Університетська,1, Львів, 79000, Україна; e-mail: markiyan_yuskiv@live.com

У статті розглянуто підхід до математичного моделювання оптимального доступу користувачів до послідовно організованих сторінок Web-сайту. Знайдено вираз для математичного очікування загального часу, необхідного для пошуку сторінки, для різних законів розподілу ймовірності звертання до сторінок. Виведено співвідношення для знаходження параметрів, за яких математичне очікування досягає мінімуму. Запропоновано підходи для реалізації даного алгоритму у прикладних додатках. На основі побудованої математичної моделі оптимального доступу до сторінок інформаційного Web-сайту зі сторони користувача та отриманого співвідношення для знаходження параметрів, за яких математичне очікування досягає мінімуму, можна зробити висновок, що данне дослідження має потенціал для подальшого застосування його результатів на виробництві. Зокрема для побудови високопродуктивних модулів для реалізації пошуку у додатках з хмарною інфраструктурою.

Ключові слова: моделювання доступу до сторінок, Web-сайт, закони розподілу ймовірностей, мікросервісна архітектура, оптимізація доступу

Вступ

Мережа Інтернет є унікальним інформаційним ресурсом, через який кожен день проходять великі об’єми інформації. Основними складовими цього ресурсу є сервери та Web-сайти. Кожен сайт відповідно складається з набору сторінок. Сторінки сайту – атомарні одиниці, які між собою пов’язані логічними зв’язками. Відповідно, об’єктом дослідження є процес пошуку інформації на Web-сайті, де сторінки розбиті на блоки з використанням методу послідовного читання блоків користувачем та їхнього послідовного перегляду. Предметом дослідження виступає оптимальність вибраного методу та можливість його реалізації у промислових проектах.

Метою дослідження є побудова математичної моделі для заданого способу пошуку інформації на сайті та визначення ефективності пошуку інформації з використанням методу послідовного читання та перегляду блоків та сторінок у блоці для подальшої реалізації даного методу в рамках пошукової компоненти у додатку з мікросервісною архітектурою; визначення можливості подальшого використання результатів даного дослідження на виробництві.

Розглянуто організацію і перегляд сторінок сайту при заданих ймовірностях звертання до сторінок. За критерій ефективності пошуку сторінки приймається математичне очікування загального часу, необхідного для пошуку потрібної сторінки користувачем. Серед законів розподілу ймовірностей звертання до сторінок розглядаються рівномірний, бінарний, узагальнений та закон Зіпфа [1,2].

Основна частина

Припустимо, що інформація, яка міститься на Web-сайті, розміщена на N сторінках, розбитих на n блоків по m сторінок в кожному ($N = n \times m$). Пошук потрібної сторінки відбувається шляхом послідовного читання блоків користувачем і їх послідовного перегляду. Тоді математичне очікування загального часу, необхідного для пошуку сторінки на Web-сайті E виразиться наступною формулою:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} (ia + ((i-1)m + j)t)\, p_{(i-1)m+j},$$

де $a = b + dm$ – час читання блоку сторінок користувачем; b, d – деякі сталі; p_i – ймовірність звертання до i-ої сторінки; t – середній час перегляду однієї сторінки користувачем.

Знайдемо явний вираз для E як для рівномірного закону розподілу ймовірностей звертання до сторінок, так і для таких законів нерівномірного розподілу ймовірностей як [1,3]:

– "бінарного" закону: $p_i = \frac{1}{2^i}, \quad i = \overline{1, N-1}, \quad p_N = \frac{1}{2^{N-1}};$

– закону Зіпфа: $p_i = \frac{1}{iH_N}, \quad i = \overline{1, N}, \quad H_N = \sum_{k=1}^{N} \frac{1}{k};$

– узагальненого закону розподілу: $p_i = \frac{1}{i^c H_N^{(c)}}$, $i = \overline{1, N}$, $H_N = \sum_{k=1}^{N} \frac{1}{k^{(c)}}$, де c $(0<c<1)$.

Серед найпоширеніших методів оптимізації пошуку інформації на сайті, які використовуються у світовій практиці та згадуються у науковій періодиці, виділяють наступні види:

– пошук інформації з використанням теорії графів, що знайшло своє відображення у пошуку, реалізованому в мережі Facebook, та згадується у джерелі [1];

– розбиття сайту та сторінок на атомарні одиниці, що знаходять відображення у джерелах [4,5];

– використання експертних систем для повнотекстового пошуку з використанням доменів та попередніх результатів пошуку користувачем інформації на сайті [7,8];

– оптимізація пошукових запитів [8];

– кешування найпопулярніших запитів.

Продуктивність обчислювальних систем визначається ефективністю методів пошуку інформації на сайті. Оскільки в більшості систем опрацювання інформації типовими є випадки нерівномірного розподілу ймовірностей звертання до сторінки, то підхід використаний в даній статті, дає змогу не лише дослідити ефективність методу пошуку інформації для конкретного закону розподілу, а й мати залежності ефективності методу від зміни закону розподілу ймовірностей звертання до сторінок. Деякі часткові результати дослідження, одержані зарубіжними авторами, відображенні в джерелі [1]. Проте, як показує практика реальних проектів, не існує єдиного методу для покращення швидкості пошуку інформації на сторінці. Навпроти, поєднання покращення швидкодії алгоритмів, використання розподілених хмарних систем, оптимізація пошукових запитів та, при потребі, їх кешування на сьогоднішній день становлять основу підходів, які спільно використовуються для підвищення швидкодії пошуку інформації на сторінках Web-сайту.

Розглянемо докладно формальне визначення й властивості для різних законів розподілу ймовірностей звертання до сторінок Web-сайту.

У випадку рівномірного закону розподілу ймовірностей звертання до сторінок отримаємо наступний вираз для *E:*

$$E = \frac{1}{2}\big((n+1)a + (N+1)t\big),$$

або

$$E = \frac{1}{2}\left(\left(\frac{N}{m}+1\right)(b+dm) + (N+1)t\right).$$

Функція *E* досягає мінімуму при $m = \left(\frac{Nb}{d}\right)^{1/2}$. Тоді $n = \left(\frac{Nd}{b}\right)^{1/2}$.

У випадку, коли ймовірності звертання до сторінок задовольняють "бінарний" закон, отримаємо наступну формулу:

$$E = \left(\frac{2^m}{2^m - 1}a + 2t\right)\left(1 - 2^{-N}\right).$$

Якщо знехтувати величиною 2^{-N}, то з достатньо високою точністю можемо прийняти:

$$E = \left(\frac{2^m}{2^m - 1}a + 2t\right).$$

Для знаходження значення параметра *m,* при якому функція *E* досягає мінімуму, отримуємо рівняння: $2^m = 1 + \left(\frac{b}{d} + m\right)\ln 2.$

Якщо ймовірності звертання до сторінок розподілені за законом Зіпфа, то

$$E = \frac{1}{H_N}\left(\big((n+1)H_N - S_m(n)\big)(a+mt) + \left(\left(\frac{1}{n}S_m(n) - H_N + 1\right)N - mH_N\right)t\right),$$

де

$$S_m(n) = \sum_{k=1}^{n} H_{km}.$$

Використовуючи апроксимацію $S_m(n)$ функцією $\overline{S}_m(n)$,

$$\overline{S}_m(n) = n(H_N - 1) + \frac{1}{2}\ln n + C_1,$$

де

$$C_1 = \frac{1}{2}\ln 2\pi\,,$$

отримуємо:

$$E = \frac{1}{H_N}\left(\left(H_N + n - \frac{1}{2}\ln n - \frac{1}{2}\ln 2\pi\right)a + Nt\right).$$

Для знаходження значення параметра n, при якому функція E досягає мінімуму, отримуємо наступне рівняння: $(2n-1)n = \frac{Nd}{b}(2H_N - 2C_1 + 1 - \ln n)$.

У випадку, якщо розподіл ймовірностей звертання до сторінок задовольняє узагальнений закон розподілу, отримуємо:

$$E = \frac{1}{H_N^{(c)}}\left(\left((n+1)H_N^{(c)} - S_m^{(c)}(n)\right)\left(a + mt\right) + \left(H_N^{(c-1)} + mS_m^{(c)}(n) - NH_N^{(c)} - mH_N^{(c)}\right)t\right),$$

де

$$S_m^{(c)}(n) = \sum_{k=1}^{n} H_{km}^{(c)}.$$

Використовуючи апроксимацію $S_m^{(c)}(n)$ функцією $\overline{S}_m^{(c)}(n)$, де

$$\overline{S}_m^{(c)}(n) = nH_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}n + \frac{a^{(c)}(n)}{n^{1-c}}\right),$$

з достатньо високою точністю отримуємо

$$E = \frac{1}{H_N^{(c)}}\left(\left(H_N^{(c)} - \frac{N^{1-c}}{1-c}\right)\left(\frac{c-1}{2-c}n + \frac{a^{(c)}(n)}{n^{1-c}}\right)\right)\left(b + \frac{dN}{n}\right) + H_N^{(c-1)}t.$$

Якщо похідну від функції $a^{(c)}(n)$ за змінною n замінити скінченною різницею $a^{(c)}(n+1) - a^{(c)}(n)$, то для наближеного знаходження параметра n, при якому функція E досягає мінімуму, отримаємо рівняння:

$$n^{3-c} + (2-c)\left(n + \frac{2-c}{1-c}\frac{Nd}{b}\right)a^{(c)}(n) =$$

$$= (2-c)\frac{d}{b}N^c n^{1-c} + H_N^{(c)} + \frac{2-c}{1-c}n\left(n + \frac{Nd}{b}\right)\left(a^{(c)}(n+1) - a^{(c)}(n)\right).$$

У таблиці 1 наведено значення параметрів n, за яких математичне сподівання досягає мінімуму для рівномірного, бінарогого, закону Зіпфа та узагальненого (в якому $0 < c < 1$), законів розподілу ймовірностей звертання до записів, деяких значень b/d, t/d=0.1 та $N = 10^6$.

Використання результатів отриманого дослідження дає можливість на основі математичної моделі розрахувати математичне очікування для методу послідовного читання блоків користувачем і їх послідовного перегляду для різних законів розподілу ймовірностей звертання до сторінок. На основі вхідних даних ми можемо побачити, наскільки метод ефективний, та розробити подальші кроки для удосконалення алгоритмів пошуку інформації на Web-сайті.

Таблиця 1.

Значення параметрів *n*

b/d	Рівномірний	c = 0,2	c = 0,4	c = 0,6	c = 0,8	Закон Зіпфа	Бінарний
10	317,24	335,3	367,85	429,53	580,47	1025,5	297716,10
100	100,00	106,5	117,3	138,3	188,4	333,3	160730,1
1000	31,7	32,25	37,6	44,5	61,5	108,5	105792,4

До сильних сторін данного дослідження належить те, що без введення додаткових доробок виведений метод можна застосувати у реальних проектах. Для цього найкрщщим підходом буде розробка REST-сервіса як частини Web-додатку з мікросервісною архітектурою. Завданням цього сервіса є організація пошуку інформації на сайті. Перевагою такого підходу є можливість окремого тестування функціоналу пошуку, що, в свою чергу, не буде впливати на функціонал основної частиини Web-сайту та її тестування. Іншою перевагою саме такого підходу буде можливість використовувати розпаралелювання роботи та збільшення пропускної та обчислювальної спроможності через використання хмарних платформ, таких як Microsoft Azure, які дають багато можливостей для хостингу, тестування, діагностики та побудови метрик для REST-сервісів.

Висновки

Розглянуто актуальну проблему оптимізації пошуку інформації на Web-сайті з використанням методу послідовного перегляду.

Наукова новизна полягає у тому, що побудовано математичну модель оптимального доступу до сторінок інформаційного Web-сайту зі сторони користувача. За критерій оптимальності прийнято математичне очікування загального часу, необхідного для послідовного пошуку сторінки на Web-сайті. Математична модель враховує ймовірності звертання до сторінок, час читання блоку сторінок та час перегляду сторінок корсиувачем. Знайдено вирази для математичного очікування, залежні від різних законів розподілу ймовірності звертання до сторінок. Також виведено співвідношення для знаходження параметрів, за яких математичне очікування досягає мінімуму.

Практична цінність данного досліджегння у тому, що на основі отриманих даних можна побудувати програмний мікросервіс, який буде частиною додатку для оптимізації пошуку інформації на сайті.

Список літератури

1. Кнут, Д. Искусство программирования для ЭВМ. Т.3: Сортировка и поиск. – 2 изд. — М.: Издательский дом “Вильямс”, 2013. — 824 с.
2. Цегелик, Г.Г. Ефективність методу *r*-рівневого блочного пошуку для різних законів розподілу ймовірностей звертання до записів / Г.Г. Цегелик, А.В. Мельничанин. // Вісн. НУ “Львівська Політехніка”. Сер. інформ. Системи та мережі. — 2005. — № 549. — С. 184-192.
3. Цегелик, Г.Г. Математичне моделювання та оптимізація до інформації індексно-послідовних файлів баз данних / Г.Г. Цегелик, А.В. Мельничанин. // Волин. матем. вісн. Сер. прикл. матем. — 2009. — Вип. 6 (15) . — С. 179–196.

4. Юськів М.І., Цегелик Г.Г. Моделювання та ефективність доступу до послідовно організованих сторінок Web-сайту для різних законів розподілу ймовірностей звертання до сторінок – Інформатика та системні науки ІСН-2016: матеріали VII Всеукраїнської науково-практичної конференції за міжнародною участю, м. Полтава, 10–12 берез. 2016 р. — Полтава: ПУЕТ, 2016.
5. Юськів М.І., Цегелик Г.Г. Ефективність методу послідовного перегляду сторінок на Web-сайті для різних законів розподілу ймовірностей звертання до сторінок, м. Львів, 28-30 вересня 2016 р. — Львів: ФМІ, 2016.
6. Baeza-Yates R., Castilio C. Relating Web Structure and User Search Behavior. – Advances in Web mining and Web usage Analysis. Materials of 8th International Workshop. University of Philadelphia. — 2006. — Pp. 226–232.
7. Jansen B., Spink A. How are we searching the world wide web? A comparison of nine search engine transaction logs. Information Processing and Management: an International Journal - Special issue: Formal methods for information retrieval. — 2006. — Vol. 42, № 1. — Pp. 248–263.
8. Maharrey, B.K. Cloud Computing on Amazon's EC2 / Department of Computer Science. Auburn University. — 2010. — 6 p.
9. Ingwersen P., Jarvelin K. The Turn. Integration of information seeking and Retrieval in Context. — Dordrecht: Springer, 2005. — 417 p.
10. How the complexity of Google’s search ranking algorithms changes over time. [Електронний ресурс]. Режим доступу: https://medium.com/@nikhilbd/how-the-complexity-of-googles-search-ranking-algorithms-changes-over-time-6b0589a3c90f (Дата звернення 10.09.2017).

МОДЕЛИРОВАНИЕ И ОПТИМИЗАЦИЯ ДОСТУПА К СТРАНИЦАМ WEB-САЙТА ДЛЯ РАЗЛИЧНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ОБРАЩЕНИЯ К СТРАНИЦЕ

М.И. Юськив, Г.Г. Цегелик

Львовский национальний университет им.. Ивана Франка,
ул. Университетская,1, Львов, 79000, Украина; e-mail: markiyan_yuskiv@live.com

В статье рассмотрен подход к математическому моделированию оптимального доступа пользователей к последовательно организованных страниц Web-сайта. Найдено выражение для математического ожидания общего времени, необходимого для поиска страницы, для различных законов распределения вероятности обращения к страницам. Также выведено соотношение для нахождения параметров, при которых математическое ожидание достигает минимума. Предложены подходы для реализации данного алгоритма в прикладных приложениях. На основе построенной математической модели оптимального доступа к страницам информационного Web-сайта со стороны пользователя и полученного соотношения для нахождения параметров, при которых математическое ожидание достигает минимума можно сделать вывод, что данное исследование имеет потенциал для дальнейшего применения его результатов на производстве. В частности для построения высокопроизводительных модулей для реализации поиска в приложениях с облачной инфраструктурой.

Ключевые слова: моделирование доступа к страницам, Web-сайт, законы распределения вероятностей, микросервисная архитектура, оптимизация доступа

MODELING AND OPTIMIZATION OF ACCESS TO THE PAGES OF WEB-SITE FOR DIFFERENT DISTRIBUTION LAWS OF PROBABILITY OF ACCESSING TO THE PAGE

M.I. Yuskiv, G.G. Tsegelik

Lviv National University named after Ivan Franko,
1, Universitetskaya Str., Lviv, 79000, Ukraine; e-mail: markiyan_yuskiv@live.com

The article deals with mathematical modeling approach to optimal user access to pages organized sequentially Web-site. An expression is found for the expectation of the total time required to search the page for the various laws of probability distribution pages. Also, displayed for the value of the parameters under which expectation reaches a minimum. The approaches to implement the algorithm in applied applications. Based on the mathematical model of optimal access to information pages from the Web-site user side and got value for finding the parameters under which expectation reaches a minimum, we can conclude that the data the research has the potential for further application of the results at work. In particular for building high-performance modules to implement a search applications with using cloud.

Keywords: modelling of access, adherently organized pages, web-site, distribution laws probability, microservice arhitecture, optimization of access