УДК 004.9 doi: 10.15421/321909

Н.А. Гук, С.В. Диханов, С.Ф. Сірик
Дніпровський національний університет імені Олеся Гончара

АНАЛІЗ СТРУКТУРИ САЙТА ЗА ДОПОМОГОЮ WEB-ГРАФА

У роботі запропоновано методику аналізу структури веб-сайту з використанням даних про гіпертекстові зв'язки між сторінками. Розроблено математичну модель веб-сайту у вигляді веб-графа. Для аналізу структури веб-графу використовується метод кластеризації k -середніх, у якості кластерів обрано типи сторінок сайту. Апробацію запропонованого підходу виконано на прикладі існуючого інтернет-магазину. В результаті аналізу отримано розбиття сторінок сайту на кластери, які відповідають ієрархічним елементам структури: категоріям товарів, підкатегоріям, сторінкам товарів.

Ключові слова: веб-сайт, веб-граф, веб-аналітика, матриця суміжності, кластеризація, метод k - середніх.

В работе предложена методика анализа структуры веб-сайта на основе данных о гипертекстовых связях между страницами. Разработана математическая модель веб-сайта в виде веб-графа. Для анализа структуры веб-графа используется метод кластеризации k -средних, в качестве кластеров выбраны типы страниц сайта. Апробацию предложенного подхода выполнено на примере существующего интернет-магазина. В результате анализа получено разбиение страниц сайта на кластеры, которые соответствуют иерархическим элементам структуры: категориям товаров, подкатегориям, страницам товаров.

Ключевые слова: веб-сайт, веб-граф, веб-аналитика, матрица смежности, кластеризация, метод k - средних.

A technique for analyzing the structure of a website based on data on hypertext links between pages is proposed. An analysis method based on the topology of links between pages was selected. The mathematical model of the website in the form of a web graph is developed. Structural relationships between pages are represented by binary values in the graph adjacency matrix. The problem of clustering is formulated. To analyze the structure of the web graph the clustering method k-means is used. A metric for determining the distance between cluster elements has been introduced. Assessment of the complexity of the algorithm is performed. Website pages correspond to hierarchical units of the structure. The structure distinguishes between pages of categories and subcategories of goods, pages of goods, and thematic articles. Types of site pages are selected as clusters. Typical pages for each cluster are selected as centroids.

An iterative algorithm for constructing a web graph has been developed. The queue is selected as the data structure for storing local information when crawling pages. Testing of the proposed approach is carried out on the example of an existing online store. A division of the site pages into clusters was obtained as a result of the analysis. A division is corresponded to hierarchical elements of the structure: product categories, subcategories, product pages.

The type of pages that are poorly identified by the algorithm is revealed. Using the results of clustering, you can improve the site structure during reengineering. Application of the developed methodology will improve the indexing of the site in the search engine.

Keywords: website, web graph, web analytics, adjacency matrix, clustering, k - means.

Вступ. У сучасному світі існує велика кількість інформації, яку можна представити у вигляді об'єктів та відносин між ними. Прикладом таких об'єктів можуть служити ресурси, розташовані у мережі Інтернет. Веб-сайти представляють собою набір HTML-документів (сторінок), які мають певну ієрархію та пов'язані між собою.

Аналіз структури веб-сайту дозволяє вирішити ряд завдань, таких як забезпечення високого рейтингу сайту в пошуковій системі, ефективне просування ресурсу, зручність та адаптивність ресурсу для користувача.

Відомо, що світові пошукові системи Google, Yahoo!, Bing та інші формують у пошуковій видачі групи близьких за семантикою веб-документів, релевантних запиту користувача. Сайти потрапляють у пошукову видачу відповідно до свого рейтингу, який враховує кількість та якість зовнішніх посилань на сайт, релевантність тексту до пошукового запиту, та визначається за методикою вебометрії [1]. Перебування на перших сходинках гарантує залучення значної кількості відвідувачів сайту та дозволяє збільшити продаж товарів та послуг підприємства.

Аналіз структури сайту також є необхідним, коли при розробці сучасних веб-сайтів використовується механізм побудови адаптивної навігації. Під час перебування користувача на сайті виконується аналіз його поведінки, фіксуються переходи по сторінках веб-сайту, а для подальшого перегляду формуються списки веб-сторінок, близьких за змістом до вже відвіданих користувачем сторінок [2]. Крім того, при адмініструванні веб-сайтів дослідження структури необхідно для аналізу зв'язаності груп сторінок, релевантних близьким за змістом запитам користувачів. Сформовані групи веб-сторінок використовуються при реінжинірингу існуючих сайтів з метою підвищення їх зручності для навігації користувачів [3].

Для розв’язання сформульованих задач широко застосовується кластеризація веб-сторінок за тематикою або ієрархією, коли схожі сторінки об'єднуються в групи. З кожним днем розміри сайтів збільшуються, зростає їх структурна складність, тому розробка методів для аналізу структур веб-сайтів з метою покрашення їх рейтингу та зручності для користувачів є актуальною задачею.

Огляд літератури. Для здійснення процедури кластеризації сторінок веб-сайту існують різні підходи. Методи, які ґрунтуються на автоматичної обробці вмісту текстів веб-сторінок, дозволяють визначати тематику сторінки за найбільш характерними словами і використовують статистичний аналіз тексту за кількістю входжень ключових слів [4]. При аналізі текстів також широко застосовується векторна модель зображення тексту, коли текст описується характеристичним вектором у лексичному просторі з використанням частотного спектру слів. Побудовані вектори інтерпретуються як

точки в багатовимірному просторі, кластеризація документів виконується з використанням метрики відстані між векторами. Однак при великих обсягах текстів такі методи вимагають багато часу на обробку. У разі, коли на сторінках розміщуються як текстові, так і графічні матеріали схожі за змістом, лінгвістичні методи аналізу гіпертекстової інформації застосувати неможливо.

У роботі [5] запропоновано ідею комбінованого методу класифікації веб-документів, який враховує зміст тексту сторінки і гіперпосилання, що ведуть зі сторінки, що розглядається, на пов'язані з нею сторінки.

Найбільш широко для аналізу структури сайтів застосовуються методи, засновані виключно на аналізі топології зв'язків між сторінками веб-ресурсу. Кожен сайт являє собою набір веб-сторінок. Сторінка є html-документом та має власну URL адресу, яка являє собою стандартизований спосіб запису адреси в мережі Інтернет. Сторінки сайту містять внутрішні гіперпосилання, за допомогою яких здійснюється перехід до іншої сторінки цього сайту, і зовнішні гіперпосилання, що визначають шлях до сторінок іншого сайту. Внутрішні посилання формують внутрішню структуру сайту, зовнішні посилання визначають взаємозв'язок сайту з іншими сайтами в мережі Інтернет. Вважається, що, якщо дві сторінки пов'язані гіперпосиланнями, то сторінка, що посилається, рекомендує для читання цитовану сторінку, а вказаний зв'язок носить характер рекомендації. По-друге, сторінка, яка посилається на іншу, має з нею тематичну схожість, тобто можна вважати, що вони є тематично локалізованими.

В якості моделей зображення сайтів широко використовуються модель у вигляді графа, модель мережі Петрі, модель теорії автоматів, решітчаста модель.

Найбільш дослідженою і популярною для аналізу структури сайту є модель у вигляді орієнтованого графа, така модель представлення сайту логічна і найкраще відображає його структуру. В роботі для аналізу структури сайту – пошуку зв'язків між різними сторінками ресурсу – застосовується веб-граф [6], а в якості методу аналізу структури веб-сайту використовуються методи кластерного аналізу графів.

Постановка задачі та математична модель. Визначимо гіпертекстову модель веб-сайту H як набір, що складається з двох множин:

$$H = \{P, L\}, \tag{1}$$

де $P = \{p_1, ..., p_n\}$ – множина сторінок сайту;

$L = \{l \mid \exists\, p_1, p_2 \in P : p_1, p_2 \in l(p_1, p_2)\}$ – множина гіперпосилань між сторінками.

Структурі гіпертекстової моделі веб-сайту відповідає математична модель у вигляді орієнтованого незваженого графа $G = (A, E)$, у якому $A = P$, $E = L$. У побудованому графі A – множина вершин, елементи якої описують сторінки сайту, E – множина ребер графу, елементи якої відповідають гіперпосиланням між сторінками. Для аналізу структури веб-сайту необхідно виконати кластеризацію сторінок.

Сформулюємо задачу кластеризації у такий спосіб. Розглянемо граф $G=(A,E)$. Множину всіх непустих підмножин множини A позначимо C. Розбиттям множини A вершин графа G на кластери будемо називати таке відображення $\varphi: A \to C$, для якого виконується:

$$\mathrm{E}(\varphi)=\{C_j\}_{j=\overline{1,k}} \subset C, \forall i, j\,(1\le i,1\le k): C_i \cap C_j = \varnothing,\ A=\bigcup_{i=1}^{k} C_i$$

Елементи множини значень $\mathrm{E}(\varphi)$ відображення φ будемо називати кластерами. Множину всіх φ можливих для графа G, будемо позначати, як $\Phi=\Phi_G$. З використанням деякої оцінки розбиття $Q(\varphi)\in R$ задачу кластеризації множини вершин A графа G можна зобразити у такий спосіб:

$$\varphi^{*}=\arg\max_{\varphi\in\Phi} Q(\varphi). \qquad (2)$$

Метод розв'язання задачі. Існуючі методи кластеризації умовно можна розділити на три групи. До графових методів кластеризації відносять алгоритм найкоротшого незамкнутого шляху, очевидний алгоритм, алгоритм побудови мінімального остовного дерева, алгоритми обчислення модулярності графу. До статистичних методів відносяться ЕМ-алгоритм, метод k-середніх та інші, до третьої групи методів – методів ієрархічної кластеризації – слід віднести агломеративну та розділову кластеризацію.

Для аналізу графових моделей також існують пакети прикладних програм, в яких реалізовані основні методи кластеризації та зручні засоби для візуалізації побудованих рішень.

В роботі для кластеризації веб-графа використовується метод k-середніх. Ітеративний алгоритм методу здійснює групування сторінок сайту по фіксованій кількості кластерів та полягає в наступному: випадковим чином обирається k векторів, які визначаються як центроїди (найбільш типові представники) кластерів. Потім кластери $C_1, C_2, ..., C_k$ наповнюються. Для кожного з векторів, які залишилися, деяким чином визначається близькість до центроїду відповідного кластера. У роботі близькість визначається як нормований скалярний добуток:

$$Sim\left(x, c^{j}\right)=\frac{\sum_{k=1}^{K} x_k c_k^{j}}{|x|\left|c^{j}\right|}, \qquad (3)$$

де x – вектор, c^{j} – центроїд кластера C_j, K – кількість кластерів.

Після визначення відстані вектор приписується до того кластеру, до якого він є найбільш близьким. Вектори групуються і перенумеровуються відповідно приналежності до кластерів. Потім для кожного з нових кластерів наново обирається центроїд c^{j}, координати якого визначаються у такий спосіб:

$$c_k^i = \frac{1}{|c_i|} \sum_{x \in C_i} x_k \,. \tag{4}$$

Після цього знову здійснюється процеси наповнення кластерів, обчислення нових центроїдів, поки процес формування кластерів не стабілізується. У якості критерію зупинки алгоритму обирається відсутність переходу об'єктів з кластера в кластер на двох сусідніх ітераціях.

Алгоритм k - середніх максимізує функцію якості кластеризації Q:

$$Q(C_1,...,C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} Sim\left(x, c^j\right). \tag{5}$$

Складність побудованого алгоритму визначається як $O(KN)$, де N – число вершин веб-графу, K – число кластерів. За допомогою описаного алгоритму можливо знайти лише локальну оптимальну конфігурацію кластерів, вигляд якої суттєво залежить від початкового вибору центроїдів.

До недоліків сформульованого алгоритму також слід віднести неможливість знаходження глобального оптимуму. Крім того, сформульований алгоритм для організації роботи в якості вхідного параметра вимагає вказати число кластерів. Для сформульованої задачі існує апріорна інформація про ієрархічну структуру сайту, яку можна використати для завдання кількості кластерів.

Алгоритм побудови веб-графа.

Граф, що відображає структуру веб-сайту складається з множини вершин $A = \{A_i\}$, $i = \overline{1, N}$. Головною сторінкою сайту завжди є A_1. Кожній вершині A_i можна поставити у відповідність множину вихідних ребер $E_i = \{e_{ij}\}$, де дуга e_{ij} пов'язує вершини A_i та A_j і відповідає гіперпосиланню з веб-сторінки A_i на веб-сторінку A_j. Таким чином, структуру сайту можна зобразити у вигляді орієнтованого графу, вершинами якого є сторінки A_i, а дугами – гіперпосилання e_{ij}. При побудові графа можливі випадки, коли модель буде вироджуватися в односпрямований граф (відношення між вершинами є антисиметричним), а також у дерево.

Граф структури сайту може бути представлений математично у вигляді матриці суміжності. Вхідними даними рекурсивного алгоритму формування графу є посилання на головну сторінку веб-сайту і бажана кількість сканованих рівнів. Можливість завдання глибини сканування сайту зроблено для того, щоб обмежити час роботи програми для дуже великих сайтів. На виході у програми створюються 2 файли: в одному файлі містяться адреси всіх знайдених сторінок, із зазначенням рівня кожної сторінки, в другому файлі міститься список гіперпосилань у вигляді: початкова сторінка, кінцева сторінка. За необхідністю можливо додавати вагу посилання, враховуючи, наприклад, кількість посилань з однієї сторінки на іншу.

Алгоритм починає роботу з пошуку головної сторінки сайту A_1, для сторінки відшукуються гіперпосилання на інші сторінки сайту e_{ij}, за необхідністю може бути збережено інформацію про вміст сторінки.

Пошук гіперпосилань виконується циклічно. Для збереження сторінок вебсайту використовується відповідна структура даних – черга. Кожне знайдене посилання перевіряється на приналежність до даного сервера. Якщо сторінка належить серверу, то перевіряється умова, не створена чи вже сторінка A_j, на яку посилається сторінка A_i. Якщо сторінки A_j ще не існує, то A_j створюється і розташовується у черзі Q для пізнішої обробки. Якщо сторінка A_j вже існує, то для поточної сторінки A_i, що розглядається, додається посилання e_{ij} і пошук триває. Після того, як всі посилання e_{ij} для сторінки A_i знайдені, з черги Q обирається нова сторінка A_j, для якої виконується чергова ітерація циклу алгоритму. Алгоритм закінчує свою роботу, коли черга Q порожня. За результатами роботи алгоритму можливо сформувати матрицю суміжності, яка відображає взаємозв'язок сторінок сайту.

Аналіз результатів. Запропонований підхід застосовано для аналізу структури існуючого протягом тривалого періоду сайту підприємства – інтернет-магазину.

Для побудови веб-графу розроблено програмне забезпечення, вхідними даними для якого являються посилання на головну сторінку та глибина сканування (кількість рівнів структури, що досліджується). Під рівнями структури мається на увазі ієрархія «категорія-підкатегорія-сторінка товару». Окремим елементом структури є сторінки, на яких розташовано інформаційні матеріали про товари, зазначені сторінки можуть відповідати, як категоріям, так й сторінкам товарів.

Посилання на головну сторінку ідентифікується за доменним ім'ям. Шляхом сканування веб-ресурсу за протоколом http збирається інформація стосовно гіперпосилань між сторінками. За результатами сканування будується граф, вершинами якого є адреси відшуканих сторінок із вказанням ієрархічного рівня сторінки, а дугами – гіперпосилання між сторінками. Побудований граф є орієнтованим.

У табл. 1 наведено фрагмент матриці суміжності веб-графа, великими літерами позначено сторінки ресурсу, на перетині рядків та стовпців бінарна одиниця позначає наявність зв'язку між сторінками. У наведеному фрагменті існують вершини, які відповідають сторінкам різних типів – категоріям товарів, підкатегоріям, сторінкам товарів, статтям з інформацією про товари.

Для аналізу побудованої структури було застосовано метод k- середніх, кількість кластерів вважалась відомою, дорівнювала 4 та відповідала кількості рівнів структури. Критерієм закінчення ітераційного процесу обрано стабілізацію процесу формування кластерів, яка спостерігається при переході

до 5 ітерації. У табл. 2 наведено метричні характеристики, які відображають відстані кожного з векторів сторінок до центрів кожного з кластерів.

Таблиця 1

Фрагмент матриці суміжності веб-графа

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	0	0	0	1	0	0	1	0	0	0	0	1	1	1	0	1
B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
D	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
E	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
F	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
H	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
I	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0
J	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
K	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
L	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1
M	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
N	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0

Таблиця 2

Метричні характеристики кластерів

	1	2	3	4	Результат кластерізації
	Відстань до центру кластера				(номер кластеру)
A	2.236	2.04	1.542	1.959	3
B	0.0	2.005	1.777	1.984	1
C	2.449	2.272	1.777	1.76	4
D	2.0	1.778	1.119	1.016	4
E	2.236	2.04	1.334	1.858	3
F	2.236	0.855	1.552	1.992	2
G	2.236	2.04	1.369	1.426	3
H	2.0	1.078	1.187	1.723	2
I	2.0	2.272	1.59	2.214	3
J	2.0	1.778	0.918	1.723	3
K	2.236	0.671	2.1	2.39	2
L	2.449	2.272	1.799	1.646	4
M	2.449	2.005	1.498	1.586	3
N	2.449	2.272	1.786	0.95	4
O	2.0	1.778	0.633	1.646	3
P	2.0	2.272	1.803	0.622	4

Для наведеного у табл. 1 фрагменту даних в результаті кластеризації побудовано набір з 4 кластерів. Аналіз розбиття на кластери показав, що об’єкти кожного з кластерів мають спільні ознаки, отримане розбиття

згрупувало сторінки веб-сайту за типами: кластер 1 відповідає сторінці, яка визначає категорію товару, до кластеру 2 потрапили сторінки підкатегорій, до кластеру 3 – сторінки окремих товарів, до кластеру 4 – сторінки, на яких розміщено тематичні матеріали у вигляді статей.

Однак за результатами аналізу повного веб-графу інтернет-магазину виявлено, що ні всі сторінки відповідають типовим кластерам. Найбільш погано виконується ідентифікація сторінок, на яких розташовано статті про товари, що пов'язано з їх узагальнюючим змістом та інформаційною спрямованістю. Сторінки із статтями посилаються на сторінки товарів з різних категорій. Усунення виявленої невідповідності призведе до більш ефективної індексації сторінок сайту та дозволить суттєво покращити структуру сайту для зручності його використання. За результатами кластеризації можливо виконати реінжиніринг ресурсу та налаштувати логічну структуру сайту, що підвищить помітність ресурсу для пошукових систем та зручність для користувачеві.

Висновки. У роботі для аналізу структури веб-сайту побудовано модель сайту у вигляді графу, розроблено алгоритм автоматичної генерації матриці суміжності веб-графу. Для аналізу зв'язків сторінок сайту застосовано метод кластеризації k-середніх. За допомогою розробленого програмного забезпечення виконано кластеризацію сторінок за типом. Отримані результати кластеризації можна застосовувати для покращення структури сайту при реінжинірингу.

Бібліографічні посилання

1. Webometrics Ranking of World Universities: [Електронний ресурс]. – Режим доступу: http://www.webometrics.info
2. **Hollink, V.** Navigation behavior models for link structure optimization [Text] / V. Hollink, M. Van Someren and B. J.Wielinga // User Modelling and User-Adapted Interaction. – 2007. – Vol. 17, iss. 4. – Pp. 339–377.
3. **Салин, В.С.** Об одном подходе к реинжинирингу гипертекстовых структур [Текст] / В.С. Салин,С.В. Папшев // Математические методы в технике и технологиях – ММТТ-26: сб. тр. XXVI междунар. науч. конф. – 2013. – С. 118–120.
4. **Крейнес, М.Г.** Модели текстов и текстовых коллекций для поиска и анализа информации [Текст] / М.Г. Крейнес // Матем. модел. эколого-экономич. систем: экономика ТРУДЫ МФТИ. – 2017. – Том 9, № 3. – С. 132–142
5. **Calado, P.** Combining linkbased and content-based methods for web document classification [Text] / P. Calado, M. Cristo, E. Moura et al. // CIKM'03, Nov. – 2003. – Pp. 3–8.
6. **Печников, А.А.** Разработка инструментов для вебометрических исследований гиперссылок научных сайтов [Текст] / А.А. Печников, Н.Б. Луговая, Ю.В. Чуйко, И.Э. Косинец // Вычислительные технологии. – 2009. – Т. 14, № 5. – С. 66–78.

Надійшла до редколегії 24.10.2019.