In this article established the conditions of the existence of an extremal element for the problem of finding the distance between two sets of linearly normalized space, the conditions of the unity of an extremal element for its equivalent problem, the properties of the function of the distance and formulas for finding an extremal element for the problem of finding the distance between two closed spheres of this space.

Key words: *the linear normed space, the distance between sets, the extremal element, the conditions of the existence and the unity, the properties of the function of the distance.*

Отримано: 29.09.2020

УДК 519.87+519.176
DOI: 10.32626/2308-5878.2020-21.99-114

Н. А. Гук, д-р фіз.-мат. наук, професор,
С. В. Диханов, аспірант,
І. О. Долотов, магістрант

Дніпровський національний університет
імені Олеся Гончара, м. Дніпро

АНАЛІЗ СТРУКТУРИ САЙТУ З ВИКОРИСТАННЯМ ПОНЯТТЯ МОДУЛЯРНОСТІ

У роботі здійснюється аналіз структури веб-сайту, який має ієрархічну організацію розділів. Ієрархічна структура передбачає розбиття всієї інформації на окремі категорії за темами. Гіпертекстову модель веб-сайту зображено математичною моделлю у вигляді орієнтованого незваженого веб-графу, вершинами якого є web-сторінки, а ребрами — гіперпосилання між ними. Висувається гіпотеза, що сторінка, яка посилається на іншу, має з нею тематичну схожість, а групи пов'язаних між собою сторінок утворюють кластер.

З використанням локальної інформації про гіперпосилання між сторінками сайту здійснюється кластеризація сторінок. Для оцінки якості кластеризації використовується функціонал модулярності, який характеризує різницю між долею ребер у середині кластеру при заданому розбитті та долею ребер, якщо б вони були сгенеровані в графі випадковим чином. Випадковий граф обирається у якості нульової моделі.

Для максимізації значень функціоналу модулярності застосовується Лувенський метод. Розроблено жадібну схему алгоритму, яка зводить задачу до послідовності локальних задач оптимізації. Пропонується здійснювати відбір пар «вершина — кластер», з'єднання яких призводить до збільшення значення функціоналу модулярності. Для довільної вершини гра-

фу відшукується цільовий кластер виходячи з аналізу списків суміжності вершини.

Алгоритм було реалізовано у вигляді прикладного програмного забезпечення, побудованого із використанням принципів функціонального програмування, та застосовано для аналізу структури сайту інтернет-магазину. Досліджено залежність значення функціоналу від кількості кластерів розбиття та параметрів ітераційного процесу.

Здійснено аналіз контенту сторінок веб-сайту всередині кластера, який виявив їх тематичну схожість. Для більшості кластерів є можливим формування семантичного опису. Виконано порівняння результатів кластеризації з експертним розбиттям, обчислено значення точності та повноти розбиття на кластери.

Ключові слова: *веб-сайт, веб-граф, ієрархічна структура, гіперпосилання, функціонал модулярності, кластеризація, тематична зв'язність сторінок, Лувенський метод, жадібний алгоритм.*

Вступ. Сьогодні веб-ресурси стають ефективним інформаційним інструментом для просування компаній на ринку та відіграють важливу роль у розвитку бізнесу. Грамотний підхід до визначення цілей функціонування ресурсу та проектування його структури забезпечує успішне просування сайту в мережі Інтернет та сприяє розширенню цільової аудиторії.

Аналіз функціонування веб-сайту є актуальною задачею, оскільки дозволяє виявити помилки в організації структури, оцінити потенційні можливості та перспективи його розвитку, своєчасно виконати оптимізацію веб-ресурсу, що сприяє підвищенню його ефективності. Результати такого аналізу можна використовувати при проведенні процедури реінжинірингу сайтів.

Підґрунтям для проведення аналізу веб-сайтів є необхідність якісної оптимізації веб-ресурсів та визначення перспектив його розвитку, планування маркетингової або рекламної кампанії підприємства, створення нового або оновлення старого проекту. При здійсненні аналізу сайту важливо визначити ціль дослідження та обрати методи для проведення аналізу.

Структура веб-сайту представляє собою логічну схему розміщення сторінок ресурсу, із використанням якої легко бачити, які сторінки та в якій ієрархії будуть розміщені на сайті, а також визначити шляхи до категорій, підкатегорій, карток товарів.

Аналіз структури сайту дозволяє виявляти помилки в логічній організації веб-ресурсу, визначати, чи якісно налаштовано внутрішні зв'язки між сторінками ресурсів, чи зручно для користувачів знаходити необхідну інформацію. Подібний аналіз також є найважливішим техні-

чним інструментом з точки зору SEO (Search Engine Optimization) при виконанні заходів пошукової оптимізації.

Правильна структура веб-сайту суттєво впливає на ранжування. Швидкий обхід веб-сайту пошуковим роботом забезпечує швидку індексацію сторінок сайту пошуковими системами. Важливим елементом аналізу структури є відшукання посилань, які ведуть на неіснуючі сторінки, а також визначення сторінок, що містять інформацію, але на них не існує жодного посилання з інших сторінок ресурсу. Такі помилки заважають швидкої індексації ресурсу пошуковим роботом та можуть бути усунені при перелінкуванні веб-сайту.

Крім того, веб-сайт із логічною структурою розподіляє внутрішню посилальну вагу по всіх сторінках в залежності від їх важливості, як для користувача, так і для просування ресурсу у мережі Інтернет.

Тому розробка методів та інструментів для проведення аналізу структури веб-сайту є актуальною задачею, розв'язання якої дозволить виявити помилки в логічної організації ресурсу та надати рекомендації щодо їх усунення.

Аналіз літературних джерел. Сайти, розташовані в мережі Інтернет, мають різноманітну структуру, яка визначається обраною моделлю. Для корпоративних сайтів, інтернет-магазинів, каталогів продукції найчастіше використовується тематична організація структури, яка має рівні ієрархії та виконує класифікацію великого обсягу інформації. Така структура передбачає розбиття всієї інформації на окремі категорії за темами. З використанням такої організації матеріалу користувач має змогу легко та швидко відшукувати на сайті необхідну для нього інформацію.

Аналіз структур існуючих сайтів найчастіше здійснюється вебмастером або веб-аналітиком за допомогою візуального оцінювання. Структура сайту, яку побудовано з використанням одного зі спеціальних сканерів, наприклад, Website auditor от SEO Powersuite, ScreamingFrog, дозволяє оцінити обсяг тематичних розділів, проаналізувати внутрішні гіперпосилання та наочно представити, які сторінки веб-ресурсу є найбільш популярними при внутрішньому перелінкуванні, а які не мають вихідних зв'язків з іншими сторінками. Однак якщо ресурс має великий обсяг і складну ієрархію розділів, візуальний аналіз є трудомістким процесом та може стати неефективним з точки зору виявлення помилок.

Сьогодні в літературі для аналізу зв'язності веб-простору та структури окремих сайтів широко застосовується представлення сайту у вигляді веб-графу [1].

Практично кожен сайт організовано як набір веб-сторінок, навігація по яких здійснюється за допомогою гіперпосилань. Якщо зобразити сайт у вигляді графа, то вершинами графа будуть web-сторінки, а дугами — гіперпосилання [2]. Запропонована модель зображення сайту є логічною та відображає його структуру [3].

Для дослідження структури веб-простору широко застосовуються теоретико-графові методи, які ґрунтуються на відшуканні компонент зв'язності, пошуку найкоротших шляхів, побудові основних дерев, спектральні методи аналізу графів [4], а також спеціальні методи для аналізу веб-простору [5].

В роботі [6] запропоновано використовувати графічну модель для аналізу структури контент-орієнтованих сайтів з метою оптимізації навігації для користувачів. Дослідження моделі у вигляді графу дозволяє здійснювати автоматизований аналіз веб-сайту з метою виявлення проблемних місць у його структурі та можливих шляхах її оптимізації. Крім того, з використанням графів виконується візуалізація структури веб-сайтів, що спільно з накопиченими статистичними даними веб-аналітики (кількість унікальних відвідувань, показник відмов, кількість переходів з однієї сторінки на інші) робить можливим їх наочний аналіз та обробку.

Часто, для аналізу структури веб-простору або веб-сайту застосовується кластеризація, з використанням якої здійснюється розбиття множини вершин графу на підмножини відповідно до деяких ознак.

В роботі [7] запропоновано метод побудови семантичних кластерів у гіпертекстовій структурі веб-сайту на основі даних статистики переходів користувачів між сторінками сайту.

Кластеризацію графу можна також здійснити шляхом мінімізації (максимізації) деякого функціоналу якості. Найчастіше для побудови такого функціоналу використовується поняття модулярності графу [8], на її максимізації ґрунтується багато алгоритмів виділення кластерів. Значення модулярності показує, наскільки отримане розбиття є якісним у тому сенсі, що вершини всередині кластеру мають більшу високу щільність зав'язків одна з одною, ніж з вершинами інших кластерів.

У роботі для аналізу структури веб-сайту пропонується використовувати локальну інформацію про веб-граф, а саме посилання між сторінками. Вважається, що сторінка, яка посилається на іншу, має з нею тематичну схожість, тобто можна припустити, що вони є тематично локалізованими. На основі інформації про посилання необхідно виконати кластеризацію сторінок веб-ресурсу шляхом максимізації функціоналу модулярності. Отримане розбиття необхідно проаналізувати шляхом перевірки тематичної схожості сторінок веб-сайту, які опинились у одному кластері.

Постановка задачі. Розглядається веб-сайт інтернет-магазину з ієрархічною структурою, в якій виділено головну сторінку, інформаційні сторінки про компанію та способи торгівлі, сторінки з тематичними статтями, каталог товарів із зазначеними категоріями, підкатегоріями та карточками товарів. Передбачається, що існує певна градація інформації на окремі категорії у відповідності до тематики. Для товарів магазину передбачена можливість фільтрації за певними ознаками. Навігацію на

веб-сайті організовано таким чином, що користувач має можливість, як з головної, так і з будь-якої сторінки сайту, перейти в будь-яку категорію, підкатегорію та на конкретну сторінку з товаром. Сторінки сайту є html-документами та мають власні URL адреси, які є стандартизованим способом запису адреси в мережі Інтернет. Сторінки сайту містять внутрішні гіперпосилання, за допомогою яких здійснюється перехід до інших сторінок цього сайту, таким чином будується внутрішня структура сайту.

Використовуючи інформацію про гіперпосилання необхідно виконати кластеризацію сторінок ресурсу та здійснити перевірку тематичної схожості сторінок, які потрапили до одного кластеру.

Гіпертекстову модель веб-сайту можна зобразити математичною моделлю у вигляді орієнтованого незваженого графа $G = (V, E)$, де V — множина вершин, елементи якої відповідають сторінкам сайту, E — множина ребер графу, елементи якої відповідають гіперпосиланням між сторінками. Граф зображується матрицею суміжності $\left[A_{ij} \right]$.

Задачу кластеризації сторінок веб-сайту сформулюємо у такий спосіб: необхідно побудувати розбиття множини V вершин веб-графу на кластери C_k, $C_k \subset C$, $k = \overline{1, K}$, $\forall i, j; i, j = \overline{1, K}$; $i \neq j : C_i \cap C_j = \varnothing$, $V = \bigcup_{k=1}^{K} C_k$ за допомогою відображення $\varphi : V \to C$, для якого виконується:

$$\varphi^* = \arg\max_{\varphi \in \Phi} Q(\varphi),$$

де $Q(\varphi)$ — функціонал якості розбиття.

Метод розв'язання задачі. Припустимо, що веб-граф має кластерну структуру, в якій щільність зв'язків між вершинами всередині кластеру вище, ніж щільність зв'язків в іншої частині графу. Крім того, передбачається, що щільність зв'язків всередині кластерів у такому графі значно вища, ніж щільність зв'язків у довільному графі, який не має структури.

Однією з можливих метрик якості розбиття є значення модулярності [9]. Модулярність характеризує різницю між долею ребер усередині кластера при заданому розбитті та долею ребер, якщо б вони були генеровані в графі випадковим чином.

Випадковий граф обирається у якості нульової моделі, передбачається, що він зберігає степені вершин вихідного графа, а ймовірність того, що кінці ребер потраплять у вершину i дорівнює:

$$d_i / (2m - 1),$$

де d_i — степінь вершини i; m — загальна кількість ребер в графі.

Тоді ймовірність існування ребра між парою вершин $i, j \in V^2$ дорівнює $d_i d_j / 2m$. Для великих значень m відніманням 1 в знаменнику можна знехтувати.

Тоді функціонал модулярності можна зобразити у такій спосіб:

$$Q = \frac{1}{2m} \sum_{i,j \in V^2} \left[A_{ij} - \alpha \frac{d_i d_j}{2m} \right] \delta(i, j), \qquad (1)$$

де A_{ij} — елемент матриці суміжності графа; α — параметр для регулювання числа кластерів у розбитті, який залежить від розміру графа та зазвичай обирається з інтервалу [0;2]; дельта-функція

$$\delta(i, j) = \begin{cases} 1, \text{ якщо вершини } i,\ j \text{ належат одному кластеру,} \\ 0, \text{ інакше.} \end{cases}$$

Алгоритм, запропонований у роботі [8], передбачає побудову розбиття графу на кластери таким чином, щоб значення Q було максимальним.

Функціонал (1) не є неперервним, тому задача оптимізації відноситься до класу задач дискретної оптимізації. Пошук глобального максимуму функціоналу (1) є NP-повною задачею у сильному сенсі [10] тому алгоритми повного перебору не є ефективними, кількість варіантів розбиття m вузлів по C_k кластерам зростає експоненційно зі збільшенням числа k.

Однак існує ряд евристичних прийомів, які дозволяють знайти значення модулярності близьке до максимального за поліноміальний час. Для розв'язання задач кластеризації розробляються жадібні алгоритми, які зводять задачу до послідовності локальних задач оптимізації. Для їх розв'язання здійснюється вибір пар «кластер-кластер» [11, 12] або «вершина — кластер» [13], з'єднання яких призводить до збільшення значення модулярності.

Ефективним підходом до розв'язання задач кластеризації вебграфів є Лувенський метод максимізації модулярності (Louvain Modularity Maximization), який відноситься до жадібних евристик [13].

Для побудови алгоритмічної схеми визначається деяке початкове розбиття множини вершин графу на кластери, як правило, кожна вершина асоціюється з власним кластером. На першому етапі роботи алгоритму вершини перебираються у деякому заданому порядку, для довільної вершини u почергово розглядаються всі кластери крім того, до якого ця вершина зараз належить. Цільовий кластер C_k відшукується серед тих кластерів, у яких є вершини суміжні з вершиною u. Серед розглянутих кластерів обирається той, переміщення вершини у який призводить до максимального збільшення значення приростання модулярності.

Тоді задачу пошуку цільового кластеру можна сформулювати як локальну задачу оптимізації у такий спосіб:

$$C_{loc} = \arg\max_{C_k} \Delta Q\left(u, C_k\right), \tag{2}$$

де прирощення модулярності при переміщенні вершини до кластеру обчислюється:

$$\Delta Q\left(u, C_k\right) = Q_{C_k \cup u} - Q_{C_k / u},\ k = \overline{1, K-1},$$

K — кількість кластерів.

Оскільки $\delta\left(i, j\right)$ дорівнює 1 тільки у тому випадку, коли вершини i та j знаходяться в одному кластері, значення модулярності можна отримати шляхом підсумовування значень модулярності по кожному кластеру окремо

$$Q = \frac{1}{2m}\sum_{k=1}^{K}\left(\sum_{i,j\in C_k} A_{ij} - \frac{\alpha}{2m}\sum_{i,j\in C_k} d_i d_j\right).$$

Вводячи позначення $\sum_{i,j\in C_k} d_i d_j = \sum_{i\in C_k} d_i \sum_{j\in C_k} d_j = \left(\sum_{i\in C_k} d_i\right)^2$, обчислимо прирощення модулярності, яке викликано переміщенням вершини u в кластер C_k:

$$\Delta Q\left(u, C_k\right) = \left[\frac{\sum_{i,j\in C_k} A_{ij} + d_{u\to C_k}}{2m} - \alpha\left(\frac{\sum_{i\in C_k} d_i + d_u}{2m}\right)^2\right] -$$

$$-\left[\frac{\sum_{i,j\in C_k} A_{ij}}{2m} - \alpha\left(\frac{\sum_{i\in C_k} d_i}{2m}\right)^2 - \alpha\left(\frac{d_u}{2m}\right)^2\right] = \frac{d_{u\to C_k}}{2m} - \alpha\frac{\sum_{i\in C_k} d_i d_u}{2m^2},$$

де $\sum_{i\in C_k} d_i$ — загальна кількість ребер, які зв'язують вершину i кластеру C_k з вершинами графу; $d_{u\to C_k}$ — кількість ребер, які зв'язують вершину u з вершинами кластера C_k; d_u — степінь вершини u.

Якщо знайдено максимальне та додатне значення прирощення модулярності, то вершина u переміщується до кластеру C_k. Якщо всі значення прирощення модулярності від'ємні, то вершина залишається у своєму кластері. Процес повторюється послідовно для всіх

вершин доти, доки вдається досягти збільшення значення модулярності при переміщенні вершини у кластер.

На другому етапі алгоритму граф, який поділено на кластери на попередньому етапі, перетворюється у метаграф G'. У метаграфі вершинами стають кластери, які отримано в результаті виконання процедури першого етапу.

Кратні ребра між кожною парою вершин у метаграфі G' замінюються на одне ребро, якому у відповідність ставлять вагу. Вага дорівнює кількості ребер, що з'єднують вершини відповідних кластерів. Локальна оптимізація значення прирощення модулярності та побудова метаграфа з'єднуються в одну ітерацію. Такі ітерації виконуються, поки існує можливість збільшити значення модулярності.

Запропонований підхід реалізовано у вигляді алгоритму.

Алгоритм кластеризації.

Вхід: граф $G = (V, E)$, який подано матрицею суміжності.

Вихід: множина кластерів $\{C_k\}$, $k = \overline{1, K}$.

0. Для графу $G = (V, E)$ виконати:

Перший етап:

1. Задати розбиття множини вершин графу V на кластери $\left\{C_k^{(t)}\right\}$, $k = \overline{1, K}$.
2. Задати значення параметру α.
3. Визначити порядок обходу вершин.
4. Повторювати для $u \in V$: знайти кластер для переміщення вершини u із розв'язання задачі (2)

$$C_{loc} = \arg\max_{C_k} \Delta Q(u, C_k).$$

5. Якщо $\Delta Q_{u \to C_k} > 0$, то перенести вершину u в кластер $C_k^{(t+1)}$ оновити інформацію о кластерах: $C_k^{(t+1)} = C_k^{(t)} \cup u$, $C_j^{(t+1)} = C_j^{(t)} / u$.
6. Якщо жодна з вершин не переміщується в новий кластер, то перейти до шагу 7.
7. Для отриманого розбиття $\left\{C_k^{(t+1)}\right\}$, $k = \overline{1, K}$ на кластери обчислити значення модулярності (1), перейти до другого етапу.

Другий етап:

1. Побудувати метаграф $G' = (V', E')$, де $V' = \{C_k\}$
2. Якщо $|V'| = |V|$, то вихід, інакше перейти до шагу 0.
3. Вихід.

Обчислювальна складність алгоритму дорівнює $O(m \log m)$.

Наведений алгоритм було реалізовано у вигляді прикладного програмного забезпечення, побудованого із застосуванням принципів функціонального програмування. Застосування такого підходу передбачає референційну прозорість усіх функції у програмі. Функції працюють лише з переданими ним змінними, що дозволяє виключити помилки при зверненні до пам'яті комп'ютера. Крім того, при використанні такого підходу порядок виконання операцій не є важливим, що дозволяє легко виконувати обчислення паралельно.

При обробці графів великої вимірності це дозволяє суттєво скоротити час на обчислення.

Аналіз отриманих результатів. Тестування розробленого програмного забезпечення було виконано на прикладі графа, наведеного у роботі [14], кластеризацію якого виконано шляхом побудови блочно-діагональних матриць сильної зв'язності. На рис. 1 наведено вихідний граф з результатами розбиття на кластери [14], які відокремлено пунктирною лінією, та результат роботи розробленого програмного забезпечення.

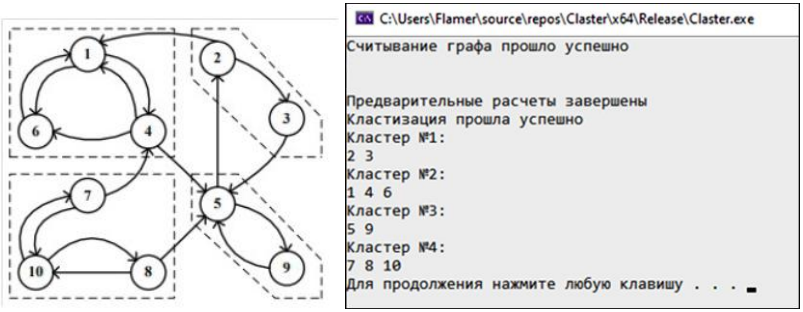

Рис. 1. *Результат розбиття на кластери із використанням матриць сильної зв'язності [14] та розробленого алгоритму*

За результатами застосування розробленого програмного забезпечення отримано таке ж саме розбиття на кластери, як і у роботі [14]. Кластери об'єднують вершини, які мають більшу кількість зв'язків одна з одною, ніж з вершинами інших кластерів. Значення функціоналу (1) на кожної ітерації виконання алгоритму поступово збільшувалось.

З літературних джерел [11] відомо, що застосування функціоналу модулярності в якості метрики кластеризації призводить до проблеми роздільної здатності, невеличкі кластери не відокремлюються. В разі, коли розглядався граф з [14], у якого потужність множин $|V| = 10$ та $|E| = 18$, вказана проблема не виявилась, однак зі збільшенням розмірів графу функціонал модулярності потребує модифікації шляхом підбору параметра масштабу α.

Далі описаний алгоритм було застосовано для аналізу структури сайту інтернет-магазину http://semena-dnepr.org.ua.

Із застосуванням спеціального програмного забезпечення, яке реалізує обхід графу та виконує процедуру краулінгу, було побудовано веб-граф сайту. Усі вершини графа є унікальними, тобто одна html-сторінка відповідає одній вершині, в побудованому графі $|V| = 486$, $|E| = 12096$. За результатами кластеризації отримано 33 кластера, на рис. 2 наведено залежність значення функціоналу (1) від номеру ітерації алгоритму та залежність значення модулярності від кількості кластерів на етапах виконання алгоритму (рис. 3). З аналізу залежностей можна бачити, що значення функціоналу дорівнює 0,368 та стабілізується після виконання 6 ітерацій алгоритму.

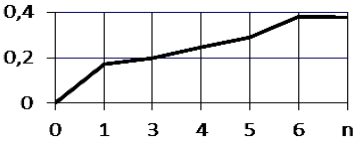

Рис. 2. *Залежність значення модулярності від номеру ітерації алгоритму*

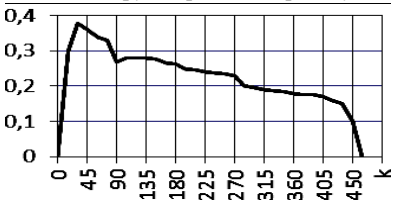

Рис. 3. *Залежність значення модулярності від кількості кластерів*

Аналіз контенту сторінок сайту, які опинились у одному кластері, дозволяє виявити тематичну схожість елементів всередині кластеру. В табл. 1, табл. 2 наведено приклади формування кластерів розбиття та їх характеристики.

Таблиця 1

Опис кластеру групи товарів Кабачки

Характеристика кластеру	**Номер вершини в графі**	**Гіперпосилання**	**Семантичний опис кластеру**
Номер кластеру: 5 Кількість елементів в кластері: 16	33	http://semena-dnepr.org.ua/product-category/semena-ovoshhej/kabachki/	Сторінка категорії товарів Кабачки
	154	http://semena-dnepr.org.ua/product/nasinnja-kraini-cukkini-ajeronavt-2-g/	Сторінки товарів, які знаходяться у категорії Кабачки
	155	http://semena-dnepr.org.ua/product/nasinnja-kraini-bjanka-2-g/	

Продовження таблиці 1

	156	http://semena-dnepr.org.ua/product/kab-gribovskij-10-g/	
	157	http://semena-dnepr.org.ua/product/nasinnja-kraini-gribovskij-2-g/	
	158	http://semena-dnepr.org.ua/product/cukkini-zolotinka-10-g/	
	167	http://semena-dnepr.org.ua/product/nasinnja-kraini-cukkini-zolotinka-2g/	
	168	http://semena-dnepr.org.ua/product/iskander-f1-5-sht/	
	169	http://semena-dnepr.org.ua/product/kavili-f1-5-sht/	
	170	http://semena-dnepr.org.ua/product/nasinnja-kraini-cukesha-2-g/	
	171	http://semena-dnepr.org.ua/product/nasinnja-kraini-patisson-belyj-2-g/	
	172	http://semena-dnepr.org.ua/product/nasinnja-kraini-patisson-oranzhevyj-2-g/	
	178	http://semena-dnepr.org.ua/product/nasinnja-kraini-svekla-cilindra-3g-kabachok-bjanka-2g/	
	179	http://semena-dnepr.org.ua/product/kabachok-kustovoj-10-g/	
	180	http://semena-dnepr.org.ua/product/kabachok-kustovoj-2-g/	
	186	http://semena-dnepr.org.ua/product/nasinnja-kraini-ogurec-kustovoj-0-5g-kabachok-gribovskij-2-g/	

Таблиця 2

Опис кластеру групи товарів Кавуни

Характеристика кластеру	**Номер вершини в графі**	**Гіперпосилання**	**Семантичний опис кластеру**
Номер кластеру 10 Кількість елементів в кластері: 9	46	http://semena-dnepr.org.ua/product-category/semena-ovoshhej/arbuzy/	Сторінка категорії товарів Кавуни
	195	http://semena-dnepr.org.ua/product/nasinnja-kraini-ogonek-5-g/	Сторінки товарів, які знаходяться у категорії Кавуни
	210	http://semena-dnepr.org.ua/product/arbuz-ogonek-500-g/	
	226	http://semena-dnepr.org.ua/product/nasinnja-kraini-astrahanskij-1-g/	

Продовження таблиці 2

	227	http://semena-dnepr.org.ua/product/nasinnja-kraini-bingo-1-g/	
	228	http://semena-dnepr.org.ua/blog/vyrashhivanie-rassady-ovoshhnyh-kultur/	
	232	http://semena-dnepr.org.ua/product/nasinnja-kraini-krimson-svit-10-sht/	
	233	http://semena-dnepr.org.ua/product/nasinnja-kraini-melitopolskij-1-g/	
	239	http://semena-dnepr.org.ua/product/nasinnja-kraini-ogonek-1-g/	

Для більшості кластерів, отриманих в результаті розбиття, виявилось можливим сформулювати семантичний опис, який відповідає або назві категорії товарів, або пошуковим запитам за конкретними критеріями та ознаками товарів.

Однак у розбитті існують і такі кластери, для яких не вдалось сформулювати семантичний опис. Найчастіше це великі за кількістю елементів кластери. Так до кластеру №1 належить близько 100 вершин, частину з яких можна віднести до сторінок навігації сайту, сторінок сортування товарів за певними критеріями, сторінок із контактною інформацією та сторінок форуму із запитаннями. Зазначені сторінки за будовою сайту мають велику кількість гіперпосилань одна на одну, що пояснює їх привласнення до одного кластеру.

Також до цього кластеру було віднесено сторінки товарів, які за семантичними ознаками повинні були опинитися у окремих кластерах. Наприклад, це сторінки товарів з продажу насіння кукурудзи та саджанців винограду. Це пояснюється тім, що кількість сторінок з товарами, що належать до цих груп, є замалою. Так на сайті у категорії Насіння кукурудзи наявними є лише три сторінки із товарами, з яких міг би складатися окремий кластер за семантичним описом. Таким чином, було виявлено суттєвий недолік алгоритму по відокремленню невеличких кластерів.

Оскільки застосування поняття модулярності у якості метрики розбиття дозволяє отримати розв'язок близький до оптимального, в роботі для оцінювання якості кластеризації було виконано порівняння отриманого розбиття з експертним.

Експертне розбиття було побудовано виходячи зі структури інтернет-магазину, кластерам відповідали категорії товарів, а сторінки з товарами в певних категоріях ставали елементами кластеру. Результат кластеризації із застосуванням запропонованого підходу було оцінено відносно запропонованого експертного розбиття. Використовувалися стандартні для подібного класу задач метрики: точність та повнота, які обчислювались наступним чином [15]:

$$P = \frac{T}{T+F}\ ;\ R = \frac{T}{T+FN}$$

де T — кількість кластерів експертного розбиття, відокремлених алгоритмом; F — кількість кластерів, що є відсутніми у експертному розбитті, але відокремлені алгоритмом; FN — кількість кластерів експертного розбиття, що є відсутніми у розбитті, яке отримано із застосуванням алгоритму.

Будемо вважати, що кластер C_k експертного розбиття знайдено алгоритмом, якщо серед знайдених алгоритмом кластерів існує такий, який складається більше ніж з половини вершин кластеру C_k та менше ніж з половини вершин будь-якого з інших кластерів експертного розбиття. За результатами роботи алгоритму було досягнуто значення точності кластерізації — 0,83, значення повноти — 0,7.

За аналогічними формулами розраховується точність та повнота в середині кластеру, значення Т дорівнює максимальної кількості вершин експертного кластеру, які виділено алгоритмом у окремий кластер під час розбиття, F та FN — число вершин, доданих та виключених алгоритмом з експертного розбиття відповідно. У табл. 2 наведено відповідні значення метричних характеристик для окремих кластерів. Середні значення точності та повноти по усім кластерам розбиття дорівнюють 0,8 та 0,76 відповідно.

Таблиця 3

Значення метрик якості для досліджуваного графу

Номер кластеру у розбитті	**Кількість елементів у кластері**	**Значення точності у кластері**	**Значення повноти у кластері**
1	103	0, 58	0,62
5	16	0,87	0,91
10	10	0,81	0,76
18	5	0,67	0,63
31	12	0,83	0,86

З аналізу табл. 3 можна бачити, що для великих та маленьких кластерів спостерігається менша точність та повнота, що обумовлено нездатністю алгоритму відокремлювати невеличкі кластери та привласнювати їх елементи до великих за кількістю елементів кластерів.

Висновки. У роботі здійснюється аналіз структури веб-сайту, який має ієрархічну організацію розділів. З аналізу публікацій за темою роботи сформульовано гіпотезу щодо тематичної зв'язності сторінок сайту, які посилаються одна на одну, а також обрано математичну модель та метод розв'язання задачі.

Гіпертекстову модель веб-сайту зображено математичною моделлю у вигляді орієнтованого незваженого веб-графу, вершинами

якого є web-сторінки, а ребрами — гіперпосилання між ними. Для аналізу структури веб-сайту здійснюється кластеризація веб-сторінок. В якості метрики для оцінки результату розбиття вершин на кластери використовується функціонал модулярності. Вершини веб-графу, щільність зв'язків між якими значно вища, ніж щільність зв'язків між вершинами у довільному графі, який не має структури, відносяться до одного кластеру. Граф, у якому ребра розташовані випадковим чином обирається у якості нульової моделі.

Для максимізації значень функціоналу модулярності застосовується Лувенський метод. Розроблено жадібну схему алгоритму, яка зводить задачу до послідовності локальних задач оптимізації. У якості евристики пропонується здійснювати відбір пар «вершина — кластер», з'єднання яких призводить до збільшення значення функціоналу модулярності.

Алгоритм реалізовано у вигляді прикладного програмного забезпечення, побудованого із застосуванням принципів функціонального програмування. Здійснено аналіз структури веб-сайту інтернет-магазину. Досліджено залежність значення функціоналу модулярності від кількості кластерів розбиття та параметрів ітераційного процесу.

Здійснено аналіз тематичної зв'язності сторінок веб-сайту всередині кластеру. Для більшості кластерів сформульовано семантичний опис. Виконано порівняння результатів кластеризації з експертним розбиттям, обчислено значення точності та повноти розбиття на кластери.

Список використаних джерел:

1. Ольшевский А. И. Описание способов представления web-сайтов в виде фреймовой модели для реализации функциональных операций в Интернет-клиентских системах. *Искусственный интеллект*. 2008. № 1. С. 110-116.
2. Chandresh P. S., Suman S., Suman K., Saurabh L. Analysis to Visualize a Web Graph. *International Journal of Computer Science Issues*. 2012. Vol. 9. Issue 3. № 2. P. 247-253.
3. Tint S., Aung M. Web graph clustering using hyperlink structure. *Advanced Computational Intelligence: An International Journal (ACII)*. 2014. Vol. 1. № 2. P. 17-24.
4. Ying X., Wu X., Barbara D. Spectrum based fraud detection in social networks. *27 th Intern. Conf. on Data Engineering*: ICDE'2011 (Hannover, Germany, April 11-16, 2011): Proc. Wash.: IEEE. 2011. P. 912-923.
5. Flake G. W., Lawrence S. R., Giles C. L., Coetzee F. M. Self-Organization and Identification of Web Communities. *IEEE Computer*. 2002. Vol. 35 (3). P. 66-71.
6. Салин В. С., Папшев С. В., Сытник А. А. Графовая модель веб-сайта как основа для анализа его структуры. *Телематика'2012*: тр. XIX всерос. науч.-метод. конф., г. Санкт-Петербург. 2012. С. 190-191.
7. Салин В. С., Папшев С. В. Семантическая сегментация веб-гипертекста на основе дискретных математических моделей. *Компьютерная лингвистика и вычислительные онтологии*: сб. науч. ст. XVIII Объединенной

конф. «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г. ун-т ИТМО. 2015. С. 119-129.
8. Newman M. E. J., Girvan M. Finding and evaluating community structure in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*. 2004. Vol. 69. № 2. Article ID 026113.
9. Girvan M., Newman M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA 99*. 2002. P. 7821-7826.
10. Brandes U., Delling D., Gaertler M., Goerke R., Hoefer M., Nikoloski Z., Wagner D. Maximizing modularity is hard. URL: https://arxiv.org/pdf/-physics/0608255v2.pdf.
11. Clauset A., Newman M. E. J., Moore C. Finding community structure in very large networks. *Phys. Rev. E*. 2004. Vol. 70, № 6. Article ID 066111.
12. Newman M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*. 2004. Vol. 69. Article ID 066133.
13. Blondel V., Guil-laume J., Lambiotte R., Lefebvre E. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008. № 10. P. 10008-10020.
14. Шестопалова О. Е., Кизин П. В. Модификация матричного алгоритма разбиения графов для решения задач декомпозиции. *Вестник Полоцкого Государственного Университета. Серия С.* 2011. С. 26-32.
15. Olson D. L., Delen D. Advanced Data Mining Techniques. Springer, 2008. 180 p.

ANALYSIS OF THE SITE STRUCTURE USING THE CONCEPT OF MODULARITY

The analysis of the structure of the website, which has a hierarchical organization of sections, is carried out. The hierarchical structure the division of all information into separate categories by topic is involved. The hypertext model of a website is represented by a mathematical model in the form of an oriented unweighted web graph. Web pages are vertices of a graph, and hyperlinks between them are edges of a graph. A hypothesis is put forward about the thematic coherence of pages that link to each other. Groups of related pages are thought to form a cluster.

Using local information about hyperlinks between site pages, site pages are clustered. As a clustering quality metric the modularity functional is used. Modularity characterizes the difference between the fraction of edges within a cluster at a given partition and the fraction of edges if they were generated in the graph at random. A random graph as the zero model is chosen.

The Louvain method to maximize the values of the modularity functional is used. A greedy scheme of the algorithm, which reduces the problem to a sequence of local optimization problems, is developed. It is proposed to select vertex-cluster pairs, the connection of which leads to an increase in the value of the modularity functional. For an arbitrary vertex of the graph, the target cluster is found based on the analysis of the lists of adjacency of the vertex.

Using the principles of functional programming application software that implements the algorithm is developed. The software to analyze the structure of the online store site is used. The dependence of the value of the modularity functional on the number of partition clusters and the parameters of the iterative process is investigated.

Analysis of the content of the website pages within the cluster, which revealed their thematic similarity, was performed. For most clusters the formation of a semantic description is possible. The results of clustering are compared with the expert partition. The values of accuracy and completeness of division into clusters are calculated.

Key words: *website, web-graph, hierarchical structure, hyperlinks, modularity functional, clustering, thematic coherence of pages, Louvain method, greedy algorithm.*

Отримано: 18.09.2020

УДК 519.6
DOI: 10.32626/2308-5878.2020-21.114-126

М. Р. Петрик, д-р фіз.-мат. наук,
Д. М. Михалик, канд. техн. наук,
І. В. Гоянюк, аспірант

Тернопільський національний технічний університет
імені Івана Пулюя, м. Тернопіль

ВИСОКОПРОДУКТИВНІ ОБЧИСЛЕННЯ ДЛЯ МОДЕЛЮВАННЯ ФІЛЬТРАЦІЙНОГО МАСОПЕРЕНОСУ В СЕРЕДОВИЩІ МІКРОПОРИСТИХ ЧАСТИНОК З УРАХУВАННЯМ ЗВОРОТНІХ ЗВ'ЯЗКІВ

Методами інтеґральних перетворень Лапласа і Фур'є побудований високшвидкісний точний аналітичний розв'язок крайової задачі фідьтраційного масопереносу, що включає два взаємозв'язаних типи переносу: на мікрорівні — в мікропорах вологовмістких частинок та макрорівні — в системі макропор міжчастинкового простору в обмеженому середовищі мікропористих частинок. Шляхом розв'язання оберненої задачі з використанням експериментальних концентраційних розподілів в системі Microsoft Visual C++ розраховані профілі приведених коефіцієнтів консолідації для частинок та системи макропор і виконана перевірка моделі на адекватність.

Процеси фільтраційного масо переносу є важливими технологічними операціями при розділенні сумішей, екстрагуванні рідин із різних біологічних матеріалів в багатьох галузях промисловості. Тому дослідження методології математичного моделювання з використанням методів інтеґральних перетворень Фур'є і Лапласа і побудови високошвидкісного та точного аналітичного розв'язку дозволять реалізовувати високопродуктивні обчислення з ефективним розпаралелюванням обчислювального процесу для багатоядерних комп'ютерів, що є досить необхідним в переробній, хімічній, фармакології та інших галузях індустрії. Таким чином, це забезпечення виконання ефективних процедур перевірки моделі на