УДК 519.87+519.176 doi: 10.15421/322204

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

АНАЛІЗ ВМІСТУ ВЕБСТОРІНОК ІЗ ЗАСТОСУВАННЯМ ЧАСТОТНОЇ МОДЕЛІ ТЕКСТУ

Розглядається задача кластерізації сторінок вебсайту на основі аналізу їх текстів методом k-середніх. Для зображення тексту застосовано векторну модель та метрику TF-IDF. Для вилучення текстів застосовано структурний підхід до аналізу HTML документів на основі тегів. Результати кластерізації можуть бути застосовані для покращення логічної будови сайту та формулювання рекомендацій для реінжинірингу.

Ключові слова: вебсайт, векторна модель, метрика TF-IDF, кластеризація, метод k-середніх, реінжиніринг.

S.V. Dykhanov, N.A. Guk
Oles Honchar Dnipro National University

WEB PAGE CONTENT ANALYSIS USING A FREQUENCY TEXT MODEL

The problem of automatic clustering of a collection of texts by topic has a wide practical application in various areas of natural language processing. Examples of such applications are recommender systems and intelligent assistants, determining user profiles in social networks, analyzing the emotional colouring of texts, clustering the abstracts of scientific articles and books, analyzing the semantic core of websites, and identifying artificial texts.

In this paper, we consider the problem of analyzing the texts of web pages. To represent web pages in the work, a vector model of text representation is used. We will use the TF-IDF metric to vectorize web page texts.

To reduce the dimension of the vector space, pre-processing was applied to the texts of the collection. The procedures of lemmatization and cleaning of texts by removing stop words were performed. To extract texts from web pages, we used a structural approach to parsing HTML documents based on tags. An automatic procedure for parsing a web page in order to extract text is considered as an automaton with a finite number of states. The movement and change of states of the automaton takes place taking into account the context of the page.

The clustering of documents in the collection was performed using the k-means method. To determine the number of clusters k, a threshold value is set, which is calculated as the ratio of the average distance between clusters to the average distance between documents within a cluster.

Computational experiments were carried out for the website of an online store. The clustering results obtained using the notion of graph modularity have been improved. The analysis of the quality of the resulting of clustering in relation to the expert partition was carried out using the accuracy and recall metrics. The results of clustering can be used to improve the logical structure of the site and formulate recommendations for reengineering.

Keywords: website, vector model, TF-IDF metric, clustering, k-means method, reengineering.

С.В. Дыханов, Н.А. Гук
Днипровский национальный университет имени Олеся Гончара

АНАЛИЗ СОДЕРЖАНИЯ ВЕБСТРАНИЦ С ПРИМЕНЕНИЕМ ЧАСТОТНОЙ МОДЕЛИ ТЕКСТА

Рассматривается задача кластеризации страниц вебсайта на основе анализа текстов страниц методом k-средних. Для представления текста используется векторная модель и метрика TF-IDF. Для извлечения текстов предлагается структурный подход к анализу HTML документов на основе тегов. Результаты кластеризации могут быть применены для улучшения логического строения сайта и формулировки рекомендаций по реинжинирингу.

Ключевые слова: вебсайт, векторная модель, метрика TF-IDF, кластеризация, метод k-средних, реинжиниринг.

Вступ. Задачі автоматичної кластеризації колекції текстів за тематикою мають широке практичне застосування у різних областях обробки природної мови. Яскравими прикладами таких застосувань є розробка рекомендаційних систем та інтелектуальних асистентів, задачі визначення профілів користувачів у соціальних мережах, задачі аналізу емоційного фарбування текстів, задачі кластеризації анотацій наукових статей та книжок, задачі аналізу семантичного ядра вебсайтів та визначення релевантності ключових слів пошуковим запитам, задачі ідентифікації штучних текстів.

Незважаючи на актуальність та широке практичне застосування задача кластеризації текстів досі викликає багато труднощів під час розв'язання.

В даній роботі розглядається задача аналізу текстів вебсторінок. Відомо, що для аналізу структури вебсайтів широко застосовуються методи кластеризації, за допомогою яких визначають групи схожих сторінок. В залежності від обраної цільової функції задачі кластеризації кластери можуть утворюватись сторінками, які мають значну кількість посилань одна на одну, сторінками, яки є схожими за структурою та стилем оформлення, сторінками, які є затребуваними у певних груп користувачів. Однак, подальший аналіз вмісту визначених кластерів показує, що зазвичай в них знаходяться елементи, які у порівнянні з експертним розбиттям повинні належати до інших кластерів. Тому потребує розробки підхід щодо аналізу схожості вебсторінок всередині кластерів, який спирається на визначення їх близькості за певними ознаками. Враховуючи контентну орієнтацію вебсайтів саме аналіз текстів сторінок дозволить вдосконалити вміст кластеру та відокремити сторінки, які опинились у кластері випадково.

Аналіз підходів до зображення та кластеризації текстів. Підходи до аналізу текстів та визначенню їх близькості передбачають вибір моделі представлення тексту, вибір ознак, за якими буде здійснюватися порівняння, та вибір методу кластеризації.

До аналізу текстів можуть застосовуватись моделі та методи, які враховують семантику природних мов [1].

У роботі [2] розвиваються підходи по формуванню наборів семантично пов'язаних слів на основі концептів, за допомогою яких можливо розуміти сенс багатозначних слів по їх оточенню, визначати близькість текстів за відсутності

спільних слів, подолати проблему надлишковості. У [3] розглядається застосування контекстних векторів ознак. При застосуванні методу латентно-семантичного аналізу [4] використовується поняття контекстно-залежного змісту слів, передбачається, що між окремими словами та контекстом, у якому вони вживаються, існує зв'язок. Сукупність контекстів, у яких певне слово вживається, визначає подібність значень слів та множин слів, що у свою чергу дозволяє визначити асоціативну та семантичну близькість і враховувати кореляцію між термами та документами.

Однак застосування таких підходів передбачає витратні процедури побудови концептів онтології, семантичних фреймів для зображення речень тексту та семантичної мережі тексту, розробку процедур їх порівняння із застосуванням алгоритмів теорії графів [5], нейронних мереж глибокого навчання [6].

Більш простим у використанні та практичної реалізації є зображення текстів у вигляді частотних векторних моделей з подальшою їх кластерізацією шляхом обчислення мери близькості векторів у векторному просторі із застосуванням відповідних метрик. Використання векторного зображення текстів вимагає значно менших витрат, ніж врахування семантики мов, ручне складання баз знань або онтологій. В залежності від вигляду та розміру текстів можна застосовувати різні методи визначення ознак, широко вживаними є частотне зображення текстів на основі «bag of words» та «word embeddings» [7].

Векторна модель тексту на основі «bag of words» враховує лише частоту застосування слів у тексті або колекції текстів, при цьому не враховується порядок слів у тексті, граматичні та синтаксичні конструкції. Для кожного тексту будується вектор, який має вимірність словника. Для покращення якості моделі можуть застосовуватись методи редукції простору ознак, методи видалення шуму.

Модель «word embeddings» враховує контекстну близькість слів. Гіпотезою моделі враховується, що слова, які зустрічаються у тексті поряд з однаковими словами та, відповідно, мають однаковий сенс, у векторному зображенні повинні мати схожі вектори. Для визначення мери їх близькості застосовуються косинусна або евклідова відстань. Серед недоліків моделі можна зазначити складність отримання контекстних векторів для слів, що рідко зустрічаються у текстах колекції, неможливість врахування семантичної та синтаксичної неоднозначності понять, неможливість відображення ієрархічної природи мови.

Серед методів кластеризації, що застосуються для векторних моделей тексту, широко застосовуються як класичні методи – k-means, k-median, метод найближчого сусіда, так і нейромережеві та нейронечітки підходи.

Аналіз літературних джерел показав, що застосування частотної моделі тексту є прийнятним для колекцій, які складаються з невеликої кількості текстів та мають обмежений словник. Але додаткового дослідження потребують

особливості застосування зазначених моделей та методів до задач аналізу схожості текстів вебсторінок.

Постановка задачі. Розглядається задача автоматичного розподілення текстів вебсторінок з множини $T=\{t_1,t_2,...,t_n\}$ на кластери з множини $C=\{C_1,C_2,...,C_k\}$, $C_i\subset C,\ i=\overline{1,k}\ \forall i,j;i,j=\overline{1,k};i\neq j: C_i\cap C_j=\varnothing$. Для зображення текстів застосується векторне представлення у вигляді матриці M вимірності $m\times n$, де кожен *i*-ий рядок матриці $m_i=\{m_{i1},m_{i2},...,m_{ik}\}$ складається з *k* ознак тексту. Необхідно знайти такий розподіл текстів по кластерах, щоб кожний кластер складався з близьких за метрикою векторів, а вектори, що належать до різних кластерів суттєво відрізнялись. Враховуючи особливості формування вебсторінок необхідно також розробити метод вилучення тексту з вебсторінки, обрати метод вилучення ознак тексту та метод кластеризації.

Математична модель зображення тексту вебсторінки. Для зображення текстів вебсторінок використовується векторна модель подання тексту (vector space model). Векторна модель дозволяє виконати зображення текстів у вигляді векторів у спільному для колекції текстів D векторному просторі.

Текст $d\in D$ у векторній моделі розглядається як невпорядкована множина термів T. Під векторизацією тексту мається на увазі розбиття тексту на унікальні слова (словосполучення або n-грами) та подальше кодування. Вимірність вектору тексту дорівнює кількості різних термів у всій колекції текстів D, і є однаковою для кодування всіх документів.

Для векторізації текстів вебсторінок будемо застосовувати метрику TF-IDF:

$$M=tf(t,d)\times idf(t,D).$$

Частота входження терму t в документ d обчислюється у такій спосіб:

$$tf(t,d)=\frac{n_t}{\sum_k n_k},$$

де n_t – число входжень терму t в тексті d, K – загальна кількість термів в тексті.

Інверсія частоти, з якою певний терм зустрічається у текстах колекції, обчислюється у такий спосіб:

$$idf(t,D)=\log\frac{|D|}{\left|\{d_i\in D|t\in d_i\}\right|},$$

де $|D|$ – число текстів у колекції; $\left|\{d_i\in D|t\in d_i\}\right|$ число текстів з колекції D, в яких зустрічається терм t.

Для зменшення вимірності векторного простору до текстів колекції застосовано попередню обробку, яка складалась з виконання процедур лематизації та очищення текстів шляхом видалення «стоп-слів».

Лематизація зводить словоформи у тексті до нормальної форми. В українській мові нормальними формами вважаються наступні морфологічні зображення: для іменників – називний відмінок, однина; для прикметників – називний відмінок, однина, чоловічий рід; для дієслів, дієприкметників, дієприслівників – дієслово в інфінітиві недосконалого виду. До «стоп-слів» відносять слова, які часто зустрічаються у текстах колекції, але не несуть змістовного навантаження. У роботі застосовується словник «стоп-слів» та видаляються всі слова довжиною до трьох символів. Метрика TF-IDF обчислюється для нормалізованого тексту. За рахунок скорочення загального числа термів зменшується вимірність векторного простору.

Особливості процесу вилучення тексту вебсторінки. Особливістю задачі, що розглядається, є необхідність вилучення текстів з вебсторінок. Зазвичай вебдокументи в мережі Інтернет зображуються в форматі Hyper Text Markup Language 5 (HTML5) та містять не лише текст, а і іншу інформацію – CSS стилі, JavaScript код, елементи навігації та оформлення сторінок. На основі HTML сторінок браузер будує структурну модель сайту – Document Object Model (DOM). Зазначена модель є деревом, у вершинах якого зберігаються посилання та атрибути елемента, а нащадками такого елемента є вкладені в нього HTML-теги та тексти з інформацією (контентом).

Для вилучення текстів з вебсторінок у роботі застосовано структурний підхід до аналізу HTML документів на основі тегів. Передбачається, що у HTML-документах імена тегів та атрибутів є стандартизованими.

Автоматична процедура аналізу вебсторінки з метою вилучення тексту розглядається як автомат зі скінченою кількістю станів. Рух та зміна станів автомату відбувається з урахуванням контексту сторінки. В залежності від позиції на вебсторінці визначаються стани: з зовні тегу, в середині тегу, в середині коментаря, в середині JavaScript коду. Слід зазначити, що в середині JavaScript коду можуть також знаходитися HTML-теги, але вони не входять до структури побудованого DOM-дерева. Оскільки для аналізу необхідно вилучити лише текстові данні з вебсторінки, то було відокремлено назви елементів та їх нащадків, які не переглядаються та не належать до побудованої структури. До таких віднесено елементи стильового оформлення, посилання, елементи коментарів, сховані елементи та елементи, що розташуються за межами бачення браузеру.

Шляхом переходів між станами та з врахуванням змісту стану з документу відокремлюється структурована інформація. На основі отриманих даних можливо побудувати ієрархію, яка відтворює структуру вкладення тегів у вигляді дерева об’єктів.

Метод кластеризації. Кластерізацію документів колекції виконано із застосуванням методу *k*-середніх. За допомогою цього методу відбувається розбиття множини елементів векторного простору на заздалегідь відоме число кластерів *k*. Неієрархічний метод *k*-середніх дозволяє отримати розподіл n елементів на *k* кластерів, так щоб кожний елемент з колекції текстів належав до окремого кластеру з найближчим до нього середнім зна-

ченням. Метод базується на мінімізації суми квадратів евклідових відстаней між кожним елементом з колекції текстів та центром його кластера, тобто функції:

$$\sum_{i=1}^{N} r\left(d_i, m_j\left(d_i\right)\right)^2,$$

де r – метрика, d_i – i-ий документ колекції, а $m_j\left(d_i\right)$ – центр кластера, якому на j-ій ітерації приписаний елемент d_i.

У процесі виконання алгоритму мінімізується середньоквадратична відстань на елементах кожного кластера. На кожній ітерації відбувається обчислення центру мас для кожного кластера, отриманого на попередньому кроці, потім вектори розбиваються на кластери знову відповідно до того, який з нових центрів виявився ближчим за обраною метрикою. Алгоритм завершується, коли на якийсь ітерації не відбувається зміни кластерів.

Наведемо опис алгоритму кластеризації методом k-середніх. Маємо масив векторних зображень текстів вигляді матриці, кожен рядок матриці відповідає певному тексту та має ряд ознак. Відповідно значень ознак об'єкт розташовується у багатовимірному просторі.

1. Визначається k – кількість кластерів, що необхідно утворити;
2. Обирається k об’єктів, метрики TF-IDF яких є віддаленими одна від одної. Обрані об’єкти вважаються центрами кластерів;
3. Кожний об’єкт привласнюється до одного з k кластерів – того, відстань до якого найкоротша;
4. Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об'єктів, що входять у цей кластер;
5. Ітеративне повторюються кроки 3-4 алгоритму, поки центри кластерів не стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об'єкти), дисперсія всередині кластера буде мінімізована, а між кластерами – максимізована.

Для визначення кількості кластерів k встановлюється значення порогу, яке розраховується як відношення середньої відстані між кластерами до середньої відстані між документами в середині кластеру. Тоді на кожної ітерації розподіл документів для поточної кількості кластерів буде розраховуватись до тих пір, поки відстань між поточною кількістю кластерів не стане нижче значення кластерізаційного порогу.

Для виконання програмної реалізації моделі представлення вебсторінок та алгоритму кластеризації було застосовано мову програмування Python та спеціалізовані бібліотеки. Для зображення векторної моделі представлення текстів з вебсторінок застосовано клас CountVectorizer з бібліотеки Scikit-learn. Для врахування вагомості слів у текстах використано процедуру TfidfVectorizer. Для виконання кластеризації попередньо оброблених текстів використовується клас KMeans з бібліотеки Scikit-learn.

Аналіз результатів. Для проведення обчислювальних експериментів було застосовано результати кластеризації, які попередньо було отримано та описано у роботі [8]. Наведений у роботі підхід до кластеризації текстових документів було застосовано для аналізу вмісту кластерів, які утворились для сайту інтернет-магазину http://semena-dnepr.org.ua із застосуванням поняття модулярності. Сайт інтернет-магазину має близько 500 унікальних вебсторінок, які шляхом кластерзації з використанням поняття модулярності було розподілено на 33 кластери.

Аналіз якості отриманого розбиття по відношенню до експертного розбиття здійснюється з використанням метрик точності та повноти [9] виявив суттєвий недолік алгоритму модулярності. Невеличкі кластери не відокремлювались, а привласнювались до більш великих кластерів.

У таблиці 1 наведено значення метрик якості розбиття для досліджуваного вебсайту. З аналізу таблиці можна бачити, що гірші значення точності спостерігаються для великих та маленьких кластерів.

Так до кластеру №1 належить близько 100 вебсторінок, за результатами перегляду їх вмісту виявилось, що частина з них належить до сторінок навігації сайту, сторінок сортування товарів за певними критеріями, сторінок із контактною інформацією та сторінок форуму із запитаннями, також до кластеру було віднесено окремі сторінки товарів, які за семантичними ознаками повинні були утворити окремі кластери.

Таблиця 1

Значення метрик якості для розбиття із застосуванням поняття модулярності

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

Тому з метою вдосконалення вмісту вже утворених із застосуванням поняття модулярності кластерів було застосовано розроблений підхід щодо кластеризації текстів вебсторінок. В результаті було отримано розподіл 103 елементів з кластеру 1 попереднього розбиття [8] на 5 кластерів, результати наведено у табл. 2.

Середні значення точності та повноти по усім кластерам розбиття дорівнюють 0.938 та 0.912 відповідно.

Таким чином, із застосуванням запропонованого підходу вдосконалено вміст кластерів, усунено недоліки попереднього розбиття із застосуванням поняття модулярності та отримано розв’язок, який має кращі показники якості (наближається до оптимального).

Таблиця 2

Значення метрик якості для розбиття елементів 1-го кластеру з врахуванням вмісту тексту сторінок

Номер кластеру у розбитті	Кількість елементів у кластері	Значення точності у кластері	Значення повноти у кластері	Семантичний опис кластеру
1	1	1	1	Головна сторінка сайту
2	4	0.97	0.93	Сторінки з контактною інформацією
3	35	0.86	0.79	Сторінки навігації сайтом
4	34	0.88	0.91	Сторінки сортування товарів
5	21	0.93	0.86	Сторінки форуму з публікаціями
6	3	1	1	Сторінки товарів з категорії «Кукурудза»
7	5	0.93	0.9	Сторінки товарів з категорії «Саджанці»

Висновки. Для зображення текстів вебсторінок використано векторну модель із врахуванням частоти вживаності слів у тексті, значення якої розраховується за допомогою метрики TF-IDF. Для вилучення текстів з вебсторінок у роботі розроблено структурний підхід до аналізу HTML документів на основі тегів.

Розроблено алгоритми кластеризації вебсторінок з врахуванням вмісту їх текстів методом k-середніх. Для оцінки якості розбиття використано метрики точності, повноти, які можна застосувати за наявності експертного розбиття сторінок вебресурсу на кластери.

Запропонований підхід застосовано для аналізу груп сторінок окремого вебсайту, розподіл яких по кластерам отримано з використанням поняття модулярності. Оцінено результат вдосконалення вмісту кластерів. Результати кластерізації можуть бути застосовані для покращення логічної будови сайту та формулювання рекомендацій для реінжинірингу.

Бібліографічні посилання

1. Wang J., Peng J., Liu O. A classification approach for less popular webpages based on latent semantic analysis and rough set model. *Expert Systems with Applications*. 2015. Vol. 42, is 1. Pages 642-648.
2. Lakhzoum D., Izaute M, Ferrand L. Semantic similarity and associated abstractness norms for 630 French word pairs. *Behavior Research Methods.* 2021. 53. Pp. 1166–1178
3. Assylbekov Z., Takhanov R. Context Vectors are Reflections of Word Vectors in Half the Dimensions. *Journal of Artificial Intelligence Research*. 2019. 66. Pp. 225-242.
4. Gefen D., Endicott J. E., Fresneda J. E., Miller J., Larsen K. R. (2017). A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community. *Communications of the Association for Information Systems*. 2017. Vol. 41. Pp 450-496
5. E. Castillo, O. Cervantes, D. Vilariño Text Analysis Using Different Graph-Based Representations. *Computacion y Sistemas*. 2018. 21(4). Pp. 581-599
6. Dubey, G., Sharma, P. (2022). A Neural Network Based Approach for Text-Level Sentiment Analysis Using Sentiment Lexicons. *Artificial Intelligence and Speech Technology. AIST 2021. Communications in Computer and Information Science*. 2022. vol 1546.
7. Peng Jin, Yue Zhang, Xingyuan Chen, and Yunqing Xia Bag-of-embeddings for text classification. *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). AAAI Press*. 2016. Pp. 2824-2830.
8. Гук Н.А., Диханов С.В., Долотов І.О. Аналіз структури сайту з використанням поняття модулярності. *Математичне та комп'ютерне моделювання. Серія: Фізико-математичні наук*. 2020. Вип. 21. С. 99-114.
9. Olson D. L., Delen D. Advanced Data Mining Techniques. Springer, 2008. 180 p

Надійшла до редколегії 29.08.2022.