УДК 004.4'414
DOI: 10.31673/2412-9070.2022.033133

С. С. КОРОТКОВ[1], асистент кафедри;
А. О. БАРАБАШ[2], аспірант,
[1] Державний університет телекомунікацій, Київ
[2] Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сикорського»

ПРОБЛЕМА СЕМАНТИЧНИХ ПРОТИРІЧ У ВЕЛИКИХ ОБСЯГАХ ДАНИХ

Розглянуто проблему усунення надмірності семантично близької текстової інформації на основі латентно-семантичного аналізу та одного з алгоритмів нечіткого висновку. Дано опис латентно-семантичного аналізу як способу виявлення семантичної близькості документів. Сформульовано варіант правил нечіткого висновку щодо розв'язання завдання усунення надмірності семантично близької інформації. Запропоновано оцінити ступінь впливу модуля усунення протиріч на оперативність функціонування інформаційних систем.

Ключові слова: латентно-семантичний аналіз; усунення надмірності; правила нечіткого висновку; семантично близька текстова інформація.

Вступ

Обсяг цифрового всесвіту кожні два роки розширюється вдвічі. У 2020 році він становив 44 зетабайти. Такий великий обсяг даних передбачає високі ризики та вимоги.

У світовій практиці найефективніше можливості великих даних використовуються у секторі фінансів, інфокомунікаційних технологій, продажу та логістики.

Для більш якісного та надійного зберігання і використання великих обсягів даних пропонується застосовувати латентно-семантичний аналіз як спосіб виявлення семантичної близькості документів.

Основна частина

Розглянемо проблему усунення надмірності семантично близької текстової інформації на основі латентно-семантичного аналізу та одного з алгоритмів нечіткого висновку.

Суть пропонованого підходу до усунення надмірності семантично близької текстової інформації полягає в такому:

♦ оцінити ступінь збігу тексту. Якщо немає повного збігу тексту, переходимо до кроку 2. Інакше виконується усунення дубльованих даних;

♦ визначити семантично близьку текстову інформацію за допомогою одного з методів семантичного аналізу, наприклад латентно-семантичного;

♦ ухвалити рішення щодо можливості усунення семантично близької інформації на основі одного із алгоритмів нечіткого висновку, наприклад алгоритму Мамдані.

Семантично близька інформація на транспорті може надходити до баз даних із різних корпоративних інформаційних систем.

Як основні складові вмісту текстових документів, що порівнюються, можуть виступати такі інформаційні ознаки: час, місце і дія. Крім того, під час порівняння документів природно зважати на ступінь (наприклад, відсоток) збігу тексту.

Потрібно зауважити, що визначення складу основних інформаційних ознак текстових документів, порівнюваних за їхньою семантичною близькістю, є найважливішим і нетривіальним завданням, яке досі не дістало вичерпного розв'язку.

Використання алгоритму латентного семантичного аналізу. Латентно-семантичний аналіз (ЛСА) є методом оброблення інформації природною мовою, що аналізує взаємозв'язок між колекцією документів і термами, котрі зустрічаються в них, зіставляє деякі фактори (тематики) всіх документів і термів. В основу методу ЛСА покладено принципи факторного аналізу, зокрема виявлення латентних (прихованих) зв'язків явищ або об'єктів, що вивчаються. За допомогою факторного аналізу можна виявити приховані змінні фактори, які відповідають за наявність лінійних статистичних зв'язків кореляцій між спостережуваними змінними.

Головні цілі використання ЛСА – виявлення семантичних зв'язків між термами та латентними залежностями всередині безлічі текстових документів, розподіл (класифікація) документів на групи, розширення пошукових запитів та розв'язання деяких інших завдань.

Метод ЛСА призначено також для вилучення контекстно-залежних значень слів за допомогою статистичного оброблення великих наборів текстових даних. Його можна застосовувати в процесі пошуку та індексації інформації, у завданнях фільтрації, а також для виявлення взаємозв'язку слів за контекстами.

Метод ЛСА використовує сингулярне розкладання вихідної матриці A «терми на документи». У результаті маємо три матриці U, S та V. Підсумок розкладання записується як добуток: $A = USV$.

Далі потрібно знизити ранг вихідної матриці k. Вихідна матриця містить так звані шуми (напри-

клад, випадковий збіг внутрішніх характеристик у двох документах). Зниження рангу дає змогу послабити вплив «шумів», а також зменшити трудомісткість і час оброблення вихідної матриці, що доцільно для великих матриць. Надмірне зниження рангу вихідної матриці здатне призвести до втрати значної інформації, а отже, ми можемо отримати незадовільні зв'язки між об'єктами.

Також зниження рангу спричинює скорочення кількості стовпців і рядків у складових матрицях U, S і V. У підсумку дістаємо скорочені матриці U_k, S_k і V_k, а результат зниження рангу матриці записують як добуток: $X = U_k S_k V_k$.

До того ж під час зниження рангу потрібно вибрати таке оптимальне значення k, щоб здобути більш точні результати.

У нашому разі за допомогою методу ЛСА оцінюється семантична близькість порівнюваних документів за допомогою вибраних інформаційних ознак (наприклад, час, місце та дія).

Після здобуття кількісних оцінок семантичної близькості порівнюваних документів переходимо до етапу нечіткого висновку з метою ухвалення рішення щодо усунення надмірності інформації.

Усунення протиріч. Для вірогідного та ефективного аналізу і подання результатів, якісного прогнозування та ухвалення вірних рішень потрібно усунути протиріччя. Практика використання сучасних інформаційних систем і баз даних показує, що суперечності дуже часто є в текстових, числових і графічних даних, які зберігаються. До основних причин виникнення протиріч належать:

- старіння збереженої інформації та потреба в її своєчасному оновленні;
- людський фактор, що зумовлює помилки введення нових та редагування наявних даних;
- інтегрування даних, що надходять із різних джерел.

Усі помилки можна поділити на кілька категорій, але так чи інакше вони пов'язані з людським фактором, а саме з помилками введення даних. Існує кілька окремих баз даних, інформація в які надходить від співробітників різних відомств, адже бази розрізнені, а іноді їми керують різні СКБД. Отже, виникає безліч протиріч під час комплексних запитів і спроб пов'язати помилкові дані з баз різних служб, відомств, підсистем. Щоб підвищити вірогідність даних та скоротити кількість помилок ручного введення, потрібно задіяти дані додаткових баз, алгоритми пошуку та усунення протиріч, даючи змогу своєчасно виявляти протиріччя, усувати або не допускати їх виникнення.

Пошук та усунення протиріч — одне з найважливіших завдань забезпечення ефективного використання даних та оперативної роботи інформаційних систем. Суперечності (конфлікти) у базі даних можна поділити на три основні групи:

1) *конфлікти іменування* — використання одного і того самого імені для різнотипних речей або кількох імен для одного і того самого об'єкта;

2) *структурні* (одні з найчастіших конфліктів) — застосування моделей, ключів або політик, що різняться за структурою, для подібних або тих самих об'єктів;

3) *семантичні конфлікти* — дані чи знання суперечливі по суті, наприклад, їх значення суперечать різним системам сприйняття світу.

Для виявлення та усунення протиріч застосовуються різні підходи та інструментальні засоби — загалом для виявлення конфліктів іменування та структурних конфліктів або для запобігання конфліктам завдяки введенню більш детальних описів зв'язків між об'єктами, для заборони певних дій чи інших заходів, що потребують великої попередньої підготовки та ручної роботи. Більшість інструментальних засобів припускають роботу з єдиною базою даних, що є результатом тривалої роботи з інтеграції баз. У разі об'єднання даних на рівні сховища засіб аналізу має вміти працювати з різнорідними даними та виконувати пошук протиріч у складних умовах різнорідності типів даних.

Для встановлення семантичних протиріч доцільно використовувати метод латентно-семантичного аналізу, ефективність застосування якого цілком доведено. Ухвалення рішення щодо можливості і варіант вирішення протиріччя семантично близької інформації можна розв'язати на основі методів нечіткого виведення з використанням алгоритму Мамдані або нейронних мереж.

Вочевидь, що екосистема додаткових елементів вплине на оперативність роботи підсистем. У цьому разі роботу алгоритмів виявлення та усунення протиріч можна розглядати як розігрів в інформаційних системах критичного значення. Щоб оцінити оперативність функціонування цих систем, пропонуються такі підходи до впровадження функції усунення протиріч:

- спочатку запит виконується у чистому вигляді без участі модуля усунення протиріч. Якщо в результаті запиту ми дістаємо не більш як один факт, то результат передається на верхній рівень. У разі отримання більше одного факту є наявність протиріччя, а результат передається в модуль усунення протиріччя;
- усі здобуті результати передаються відразу в модуль усунення протиріччя незалежно від кількості отриманих фактів.

Очевидно, що застосування різних підходів по-різному позначатиметься на оперативності функціонування інформаційної системи. Також ступінь впливу модуля розігріву залежатиме від кількості протиріч.

Висновки

Усунення надмірності у великих обсягах даних на основі латентно-семантичного аналізу та нечіткого висновку, на наш погляд, може дати змогу помітно зменшити обсяг даних, що зберігаються. Подальші дослідження передбачається продовжити в напрямах: практичної реалізації запропонованого підходу, оцінювання вірогідності та ефективності одержуваних рішень, обґрунтування складу основних інформаційних ознак семантичної близькості сегментів даних, що порівнюються, аргументування вибору найкращого алгоритму нечіткого висновку.

Список використаної літератури

1. ***Шрам Г.*** *Оптимальне використання ресурсів пам'яті // Журнал мережних рішень LAN. 2011. № 3.*

2. ***Щербінін А.*** *Рішення щодо дедуплікації даних // Storage News. 2008. № 2. С. 2–7.*

3. ***Хорошилов А. А.*** *Методи автоматичного встановлення смислової близькості документів на основі їх концептуального аналізу // Праці 15-ї всерос. наук. конф. «Електронні бібліотеки: перспективні методи та технології, електронні колекції». Ярославль, 14–17 жовтня 2013 р. Секція 6.*

4. ***Штовба С. Д.*** *Введення в теорію нечітких множин та нечітку логіку. Вінниця: Вид-во Вінницького держ. техн. ун-ту, 2001. 198 с.*

5. ***Інтелектуальне*** *інформаційно-керівне середовище для організації перевезень та транспортного обслуговування // Праці 2-ї наук.-техн. конф. «Інтелектуальні системи управління залізничним транспортом». М., 15–16 листопада 2012 р. С. 66–72.*

6. ***Кураленок І. Є., Некрестьянов І. С.*** *Автоматична класифікація документів на основі латентно-семантичного аналізу // Праці 1-ї всерос. наук.-метод. конф. «Електронні бібліотеки: перспективні методи та технології, електронні колекції». СПб, 1999. С. 89–96.*

7. ***Landauer T., Foltz P., Laham D.*** *An introduction to Latent Semantic Analysys // Discourse Processes, 1998. 25. P. 259–284.*

8. ***Хомоненко А. Д., Краснов С. А.*** *Застосування методів латентно-семантичного аналізу для автоматичної рубрикації документів // Вісті ПГУПС. 2012. №2 (31). С. 124–132.*

9. ***Агєєв М. С., Добров Б. В., Лукашевич Н. В.*** *Автоматична рубрикація текстів: методи та проблеми // Навч. записки Казан. держ. ун-ту. Фізико-математичні науки. 2008. Т. 150. Кн. 4. С. 25-40.*

10. ***Войцеховський С. В., Хомоненко А. Д.*** *Виявлення шкідливих програмних впливів на основі нечіткого висновку // Проблеми інформаційної безпеки. Комп'ютерні системи. 2011. № 3. С. 81–91.*

S. S. Korotkov, A. O. Barabash

THE PROBLEM OF SEMANTIC CONTRADICTIONS IN LARGE VOLUMES OF DATA

The problem of eliminating the redundancy of semantically close textual information based on latent semantic analysis and one of the fuzzy inference algorithms is considered. A description of latent semantic analysis as a method of detecting the semantic proximity of documents is given. A variant of fuzzy inference rules for solving the task of eliminating the redundancy of semantically close information is formulated. It is proposed to evaluate the degree of impact of the contradiction elimination module on the operational efficiency of information systems.

In order to assess the efficiency of the functioning of such systems, two approaches to the implementation of the function of eliminating contradictions are proposed:

1. Initially, the request is executed in its pure form without the participation of the contradiction elimination module. If as a result of the query we receive no more than one fact, then the result is passed to the upper level. When receiving more than one fact, there is a contradiction, the result is transferred to the contradiction elimination module.

2. All obtained results are transferred immediately to the contradiction elimination module, regardless of the number of obtained facts.

It is obvious that the application of different approaches will affect the operational efficiency of the information system in different ways. In the first approach, the degree of influence of the heating module will depend on the number of contradictions. The article proposes to evaluate the degree of influence of the module of elimination of contradictions on the efficiency of the functioning of information systems.

Redundancy in large volumes of data based on latent semantic analysis and fuzzy inference can, in our opinion, significantly reduce the amount of data stored.

Keywords: latent semantic analysis; elimination of redundancy; fuzzy inference rules; semantically close text information.