

УДК 519.765:519.767:004.93

**КЛАСТЕРНИЙ АНАЛІЗ ТЕКСТОВИХ ДОКУМЕНТІВ
В ПРОСТОРІ СЕМАНТИЧНИХ КОНЦЕПТІВ**

Б.М. Павлишенко

Львівський національний університет імені Івана Франка

e-mail:pavlsh@yahoo.com

В сучасних інформаційних системах часто використовують комбіновані алгоритми інтелектуального аналізу даних, зокрема, текстових масивів [1,2,3]. Кластеризація текстових документів є однією із складових такого аналізу. Для представлення текстових документів часто використовують модель векторного простору [3,4]. В цій моделі кожний документ відображається як вектор в багатомірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми із словників текстових масивів. Текстовий масив можна представити у вигляді матриці частот слів (термів) в документах. Колонки в такій матриці визначають документи, а рядки – частоти лексем в цих документах. Мірою відстані між двома документами може бути кут між векторами цих документів в утвореному векторному просторі. Такий підхід має також ряд проблем, зокрема, розмірність аналізованого простору є великою, оскільки зумовлена розміром словника. Одним із шляхів вирішення цієї проблеми є використання латентного семантичного аналізу [4,5,6]. Суть такого аналізу полягає в сингулярному розкладі матриці ознак типу “терми–документи” і аналізі текстових масивів у новому векторному просторі меншої розмірності. Базис цього простору є побудований на лінійних комбінаціях квантитативних характеристик лексем словника. Такий новий векторний простір часто називають простором концептів. Розмірність нового простору визначається кількістю найбільших сингулярних чисел – елементів діагональної матриці сингулярного розкладу. Документи також можуть бути квантитативно близькими не тільки за частотами окремих лексем, а також за характеристиками заданих лексемних об’єднань, зокрема семантичних полів [7,8]. Розмірність матриці ознак «семантичні поля–документи» є суттєво меншою в порівнянні із матрицею ознак для лексем словника текстових масивів. Семантичні поля формуються на основі експертного аналізу, одні і ті ж лексеми можуть одночасно належати різним семантичним полям. Сингулярна декомпозиція матриці семантичних ознак дасть можливість аналізувати текстові масиви в ще меншому векторному просторі. Визначити ефективність такої декомпозиції можна аналізуючи утворення кластерної структури в новому семантичному просторі концептів для класифікованих за деякою ознакою текстових документів. Такою ознакою може бути, наприклад, спільний стиль або автор.

Розглянемо модель, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Нехай існує деякий словник лексем, які зустрічаються в текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \{ w_i \mid i = 1, 2, \dots, N_w \} \quad (1)$$

Сукупність текстових документів опишемо такою множиною

$$D = \{ d_j \mid j = 1, 2, \dots, N_d \} \quad (2)$$

Введемо множину семантичних полів

$$S = \{ s_k \mid k = 1, 2, \dots, N_s \} \quad (3)$$

Під семантичним полем розуміють таку множину лексем, які об’єднані деяким спільним поняттям [7,8]. Прикладом семантичних полів може бути - поле руху, поле комунікації, поле сприйняття та інші. Документ d_j з множини текстових документів D

можна представити як впорядковану множину слів, порядок елементів якої відповідає порядку слів в цьому документі

$$T_j^d = \{t_{lj} \mid l = 1, 2, \dots, N_j^t\} \quad (4)$$

Введемо відображення лексемного складу словника W на множину семантичних полів S за допомогою деякого оператора U_{ws}

$$U_{ws} : w_i \rightarrow s_k, \quad i = 1, 2, \dots, N_w; k = 1, 2, \dots, N_s \quad (8)$$

Оператор U_{ws} задамо таблицею, яка визначається експертним лексикографічним аналізом [7,8]. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \xrightarrow{U_{ws}} s_k, i = 1, 2, \dots, N_w \right\} \quad (9)$$

Введемо мультимножину образів відображення U_{ws} семантичних полів для окремого документа d_j

$$S_j^d = \{n_{kj}^{sd}(s_k) \mid k = 1, 2, \dots, N_s\} \quad (10)$$

де n_{kj}^{sd} - кількість лексем семантичного поля s_k в лексемному складі документа d_j

$$n_{kj}^{sd} = \sum_{l=1}^{N_j^t} f_s(t_{lj}, s_k), \quad \text{де } f_s(t_{lj}, s_k) = \begin{cases} 1, & t_{lj} \in W_k^s \\ 0, & t_{lj} \notin W_k^s \end{cases} \quad (11)$$

Введемо матрицю семантичних ознак типу “частоти_семантичних_полів-документи”

$$M_{sd} = (p_{kj}^{sd})_{k=1, j=1}^{N_s, N_d} \quad (12)$$

де p_{kj}^{sd} - частота семантичного поля s_k в лексемному складі документа d_j , яку обрахуємо за формулою

$$p_{kj}^{sd} = \frac{n_{kj}^{sd}}{N_j^t} \quad (13)$$

Вектор

$$V_j^s = (p_{1j}^{sd}, p_{2j}^{sd}, \dots, p_{N_s j}^{sd}) \quad (14)$$

відображає документ d_j в N_s -мірному семантичному просторі текстових документів. Запропонована модель дає можливість визначити матрицю частотних семантичних ознак типу “частоти_семантичних_полів-документи” і ввести новий базис для текстових характеристик. В семантичному базисі можуть спостерігатись якісно нові групування текстових документів.

Розглянемо утворення простору семантичних концептів внаслідок сингулярного розкладу матриці частот семантичних полів. Нехай існує матриця типу “частоти_семантичних_полів-документи” M_{sd} , яка описується формулою (12). Вектор V_j^s (14) відображає документ d_j в N_s -мірному просторі текстових документів. Добуток двох векторів $(V_p^s)^T V_q^s$ визначає кількісну міру близькості цих векторів у N_s -мірному семантичному просторі текстових документів. Відповідно добуток матриць $(M_{sd})^T M_{sd}$ містить скалярні добутки векторів $(V_p^s)^T V_q^s$ всіх документів і відображає їхні кореляції в просторі семантичних векторів. Нехай існує сингулярна декомпозиція матриці M_{sd}

$$M_{sd} = U_{sd} \Sigma_{sd} Y_{sd}^T \quad (17)$$

Тоді добуток матриць $(M_{sd})^T M_{sd}$ можна розглянути у вигляді

$$(M_{sd})^T M_{sd} = (U_{sd} \Sigma_{sd} Y_{sd}^T)^T (U_{sd} \Sigma_{sd} Y_{sd}^T) = Y_{sd} \Sigma_{sd}^T \Sigma_{sd} Y_{sd}^T \quad (18)$$

У відповідності до теорії сингулярного розкладу матриць [5,6] діагональна матриця Σ_{sd} містить сингулярні числа в порядку їх спадання. Якщо взяти K найбільших сингулярних чисел матриці Σ_{sd} і відповідно K сингулярних векторів матриць U_{sd} і Y_{sd} то отримаємо K -рангову апроксимацію матриці M_{sd} :

$$(M_{sd})_K = (U_{sd})_K (\Sigma_{sd})_K (Y_{sd})_K^T \quad (19)$$

Матриця $(Y_{sd})_K$ відображає зв'язок між векторами документів $\mathcal{E}_j^s$ у новому комбінованому K -мірному семантичному просторі, який будемо називати простором семантичних концептів. Зв'язок між вектором V_j^s документу в початковому семантичному просторі та вектором $\mathcal{E}_j^s$ у просторі семантичних концептів можна описати так

$$\begin{aligned} V_j^s &= (U_{sd})_K (\Sigma_{sd})_K \mathcal{E}_j^s \\ \mathcal{E}_j^s &= (\Sigma_{sd})_K^{-1} (U_{sd})_K^T V_j^s \end{aligned} \quad (20)$$

Отже, ранг апроксимації матриці M_{sd} , який визначається числом K , також визначає розмірність простору семантичних концептів. Очевидно, що число K , може бути суттєво меншим за розмірність N_s початкового семантичного простору. Це зменшує розмірність задачі аналізу подібності текстових документів у семантичному векторному просторі.

Для групування текстових документів в просторі семантичних концептів ми використали алгоритм ієрархічної агломеративної кластеризації. Завдання кластеризації полягало в побудові відображення множини документів на множину кластерів. Для обрахунку міри міжкластерної міри близькості в просторі семантичних концептів ми використовували евклідову відстань. Враховуючи те, що кластери можуть складатись з декількох об'єктів існують різні методи формування та об'єднання кластерів на основі відстаней між об'єктами всередині кластера. В наших дослідженнях ми використовували метод Варда. В цьому методі обраховують квадрати евклідових відстаней від окремих документів до центру кожного кластера. Далі ці відстані сумують. В новий кластер об'єднується ті кластери при об'єднанні яких виходить найменший приріст суми квадратів відстаней. Графічним зображенням результату ієрархічної кластеризації є дендрограма, на якій відображається процес агломеративного об'єднання кластерів. По осі абсцис відкладають номери кластерів, а по осі ординат відстані між кластерами. При певних значеннях відстаней починається об'єднання кластерів. З ростом порогової міжкластерної відстані кластери об'єднуються аж до повного злиття кластерів в один кластер. Для отримання інформативної кластерної структури вибирається деякий поріг міжкластерної відстані при якому утворюється оптимальна з точки зору аналізу текстових масивів кластерна структура.

Для аналізу ефективності розглянутих алгоритмів кластеризації взято текстову вибірку 155 художніх творів англійської класики 4 відомих авторів (Ч.Дікенс, Дж.Лондон, В.Скотт, М.Твен). Для утворення семантичного простору сформовано 15 семантичних полів в які входить близько 5000 неозначених форм дієслова. Для кожного документа були розраховані частотні словники, на основі яких розраховані частотні спектри семантичних полів документів. Отже, кожний документ розглядається як вектор

в 15-мірному початковому семантичному просторі. Розраховано дендрограми ієрархічної кластеризації при різній розмірності простору семантичних концептів. Вибраний ранг апроксимації матриці семантичних ознак суттєво впливає на формування кластерної структури. Проаналізовано класифікацію текстових документів за авторами. Вибрано таку порогову міжкластерну відстань, при якій утворюється кількість кластерів рівна кількості авторів текстів у досліджуваній вибірці. В аналізованому випадку це чотири кластери. Як впливає із отриманих даних, тексти деяких авторів є відсутні в одних кластерах і максимально зконцентровані в інших. Такий нерівномірний розподіл текстів за авторами в кластерах свідчить про те, що кластерна структура документів в просторі семантичних концептів відображає класифікаційну структуру документів за авторами. В дослідженнях для тестової вибірки знайдена оптимальна розмірність простору семантичних концептів $K=5$. У випадку кластеризації документів в просторі семантичних концептів більшої розмірності ($K>5$), розподіл документів одного і того ж автора по кластерах може якісно відрізнятися, однак спостерігаються домінантні кластери для документів деяких авторів. При низькій розмірності $K \in \{1,2,3\}$ кластери текстів з домінуючими авторами зникають і розподіл за авторами по кластерах стає більш рівномірним.

Отже, формування простору семантичних полів дає можливість отримувати новий структурний поділ документів за семантичними ознаками. Сингулярний розклад матриці семантичних ознак типу “частоти_семантичних_полів-документи” дає можливість аналізувати текстові документи в новому просторі семантичних концептів. Ієрархічна кластеризація документів в такому просторі відображає класифікаційну структуру документів за різними ознаками, зокрема, за авторством текстів. Розмірність простору семантичних концептів визначається рангом апроксимації матриці семантичних ознак при сингулярному розкладі і може бути суттєво меншою за розмірність простору семантичних полів. У випадку дослідження авторства текстів вибір розмірності простору семантичних концептів зумовлений рівнем відображення класифікаційного поділу документів за авторами в кластерній структурі, що визначається наявністю домінуючих кластерів для документів окремих авторів.

Література

1. Ким Д.О., Мьюллер Ч.У., Клекка У.Р. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215с.: ил
2. Жамбю М. Иерархический кластер-анализ и соответствия: пер. с фр. – М.: Финансы и статистика, 1988. – 342 с: ил.
3. Брасегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. Анализ данных и процессов: учеб. пособие.-СПб.:БХВ-Петербург,2009.-512с.:ил.
4. Pantel Patrick, Turney Peter D. From Frequency to Meaning: Vector Space Models of Semantics[Електронний ресурс]//arXiv:1003.1141, 2010, <http://arxiv.org/abs/1003.1141>
5. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis //Journal of the American Society for Information Science Volume 41, Issue 6, pages 391–407, 1990
6. Andri Mirzal Clustering and Latent Semantic Indexing Aspects of the Singular Value Decomposition [Електронний ресурс]// arXiv:1011.4104v2, 2011, <http://arxiv.org/abs/1011.4104v2>
7. Вердиева З.Н. Семантические поля в современном английском языке. – М.: Высшая школа, 1986. – 120с.
8. Левицкий В.В., Стернин И.А. Экспериментальные методы в семасиологии. – Воронеж: Изд-во ВГУ, 1989. – 192 с.