УДК 519.765:519.767:004.93

Б. Павлишенко
Львівський національний університет імені Івана Франка

СЕМАНТИЧНА КЛАСТЕРИЗАЦІЯ ТЕКСТОВИХ ДОКУМЕНТІВ МЕТОДОМ *K*–СЕРЕДНІХ

Запропоновано алгоритм кластеризації текстових документів методом k–середніх у векторному просторі частотних характеристик семантичних полів. Показано ефективність семантичного кластерного аналізу при вивченні класифікацій текстових документів, зокрема за авторством.

Ключові слова: інтелектуальний аналіз текстів, кластерний аналіз, метод k–середніх, семантичні поля.

The algorithm of clusterization of text documents by k–means method in the vector space of frequencies characteristics of semantic fields has been proposed. The effectiveness of semantic cluster analysis for investigation of text documents classifications, particularly authorship has been shown.

Key words: text mining, cluster analysis, k–means method, semantic fields.

Вступ

Кластеризація текстових документів є одним із складових методів інтелектуального аналізу текстових даних. Під кластеризацією розуміють розбиття даних за певними групами, які називають кластерами, в яких дані згруповано за деякими спільними характеристиками. Завдання кластеризації полягає у побудові відображення множини вхідних даних на множину кластерів [1–4]. Важливим етапом кластеризації є формування векторного простору текстових документів, у якому аналізують кластери. За поширеною векторною моделлю документи відображають як вектори у багатомірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми із словників текстових масивів [4,5]. Текстовий масив можна представити у вигляді матриці ознак слів (термів) та документів. Такими ознаками можуть бути текстові частоти лексем. У матриці ознак колонки визначають документи, а рядки – частоти лексем у цих документах. Кожна колонка матриці ознак є вектором частот лексем для певного документа. 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. Документи також можуть бути квантитативно близькими не тільки за частотами окремих лексем, але й за характеристиками заданих лексемних об'єднань, зокрема семантичних полів [6,7]. Розмірність матриці ознак «семантичні_поля–документи» є істотно меншою порівняно з матрицею ознак для лексем словника текстових масивів. Семантичні поля формуються на основі експертного аналізу, одні і ті самі лексеми можуть одночасно належати до різних семантичних полів. Перспективним, на нашу думку, є аналіз відображення в семантичній кластерній структурі документів класифікаційної структури документів за рядом ознак, зокрема за авторством текстів.

Постановка задачі

Для аналізу ефективності семантичної кластеризації текстових документів розглянемо формування семантичного простору, утвореного частотними характеристиками семантичних полів. Далі розглянемо кластеризацію методом к-середніх у семантичному просторі текстових документів. На прикладі вибірки текстових документів реалізуємо ітераційну кластеризацію текстів методом *k*–середніх і проаналізуємо утворену в семантичному просторі кластерну структуру.

Формування векторного семантичного простору текстових документів

Кластерний аналіз текстових документів реалізують у певному векторному просторі. Проаналізуємо формування семантичного векторного простору, утвореного частотними характеристиками семантичних полів словникового складу текстових масивів. Розглянемо модель, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Нехай існує деякий словник лексем, які зустрічаються у текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \{ w_i \mid i = 1,2..., N_w \} \tag{1}$$

Сукупність текстових документів опишемо такою множиною

$$D = \{ d_j \mid j = 1,2..., N_d \} \tag{2}$$

Введемо множину семантичних полів

$$S = \{ s_k \mid k = 1,2..., N_s \} \tag{3}$$

Під семантичним полем розуміють таку множину лексем, які об'єднані деяким спільним поняттям [6,7]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та інші. Документ d_j з множини текстових документів D можна подати як впорядковану множину слів, порядок елементів якої відповідає порядку слів у цьому документі

$$T_j^d = \{ t_{lj} \mid l = 1,2...N_j^t \} \tag{4}$$

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

$$U_{ws} : w_i \to s_k, \ i = 1,2..., N_w; k = 1,2..., N_s . \tag{8}$$

Оператор U_{ws} задамо таблицею, яка визначається експертним лексикографічним аналізом [6,7]. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\to} s_k, i = 1,2..., N_w \right\} \tag{9}$$

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

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

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

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

Введемо матрицю семантичних ознак типу "частоти_семантичних_полів–документи"

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

де p_{kj}^{sd} – частота семантичного поля s_k у лексемному складі документа d_j , яку обчислимо за формулою

$$p_{kj}^{sd} = \frac{n_{kj}^{sd}}{N_j^t} \tag{13}$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right) \tag{14}$$

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.

Кластеризація документів методом *k*–середніх у векторному семантичному просторі

Розглянемо алгоритм кластеризації документів методом *k*-середніх у просторі семантичних полів. Нехай є множина текстових документів *D*, яка описується виразом (2) та множина кластерів

$$C = \{ c_m \mid m = 0,1,2..., N_c \}. \quad (15)$$

Необхідно побудувати відображення множини документів на множину кластерів :

$$U_{DC} : D \to C. \quad (16)$$

Відображення U_{DC} задає модель даних, яка є розв'язком задачі кластеризації [1–3].

На початковому кроці вибираємо k центрів кластеризації, це можуть бути випадкові точки семантичного простору. Для кожного центру формуємо групу текстових документів, які є найближчими за евклідовою відстанню у векторному просторі до цього центру. Евклідову відстань визначимо так

$$r_e(d_i, d_j) = \sqrt{\sum_{k=1}^{N_s} \left(p_{ki}^{sd} - p_{kj}^{sd} \right)^2} \quad (17)$$

Утворені групи текстових документів формують проміжні кластери. Потім визначимо центри мас цих кластерів. Координати вектора центрів мас розрахуємо як середні значення за координатами векторів текстових документів в утворених кластерах. Отримані центри мас беремо як центри кластеризації на наступній ітерації, у якій відбувається новий перерозподіл текстових документів за найменшою відстанню до центрів кластеризації. Процес кластеризації завершується на ітерації, за якою не відбувається нового перерозподілу текстових документів. Суть кластеризації полягає в мінімізації дисперсії s на точках кластерів у векторному просторі

$$s = \sum_{m=1}^{N_c} \sum_{d_j \in c_j} r(d_j, m_m) \quad (18)$$

де m_m – центр мас для векторів V_j^s документів d_j, які належать до кластера c_j.

Апробація семантичної кластеризації на тестовій вибірці

Для аналізу ефективності розглянутих алгоритмів кластеризації взято текстову вибірку 155 художніх творів англомовної класики чотирьох відомих авторів (Ч. Дікенс, Джек Лондон, В. Скотт, М. Твен). Для утворення семантичного простору сформовано 15 семантичних полів, до яких входить близько 5000 неозначених форм дієслова. Деталізація літературних та лексикографічних характеристик вхідних даних не є істотною для аналізу можливості кластерного структурування даних, тому для подальшого аналізу будемо розглядати лише статистичні характеристики текстових документів. Для кожного документа були розраховані частотні словники, на основі яких розраховано частотні спектри семантичних полів документів. Отже, кожний документ розглянемо як вектор у 15-мірному початковому семантичному просторі. Для кластеризації документів методом *k*-середніх у векторному семантичному просторі вибрано 15 центрів кластеризації як випадкові точки у семантичному просторі. У результаті реалізації алгоритму кластеризації отримано розподіл текстових документів по 15 кластерах у семантичному просторі. На рис.1 наведено розподіл кількості текстових документів за утвореними кластерами.

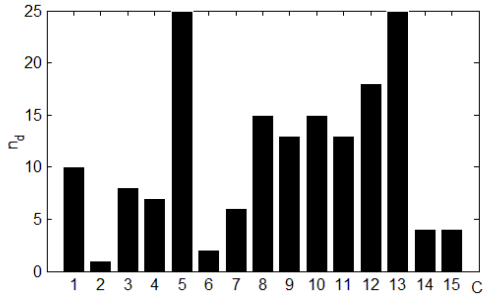

Рис. 1. Розподіл кількості текстових документів за кластерами

На рис.2 наведено розподіл текстів за авторами (1 – Ч. Дікенс, 2 – Дж. Лондон, 3 – В. Скотт, 4 – М. Твен) в кожному із 15-ти кластерів. По осі абсцис відкладено номер автора, а по осі ординат – кількість творів цього автора в кластері. Як випливає із наведених даних, тексти одних авторів домінують у деяких кластерах і відсутні в інших. Такий нерівномірний розподіл текстів у кластерах свідчить про те, що кластерна структура документів у просторі семантичних полів відображає класифікаційну структуру документів за авторами.

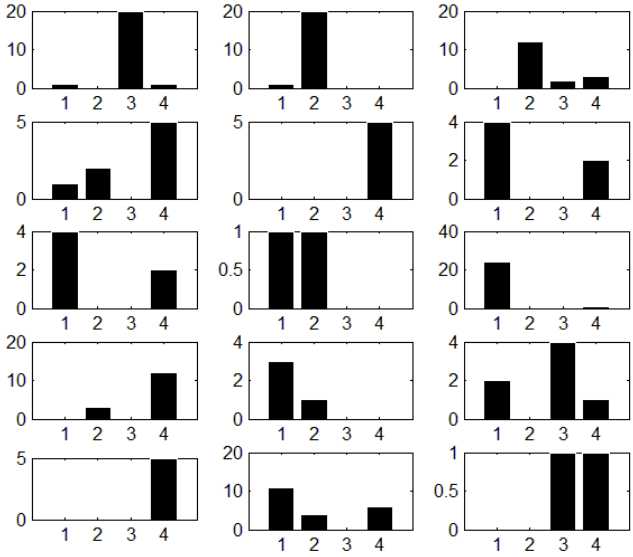

Рис.2 Розподіл кількості текстових документів за авторами в утворених кластерах

Висновки

Формування простору семантичних полів дає можливість отримувати новий структурний поділ документів за семантичними характеристиками. Кластеризація документів у такому просторі методом *k*–середніх відображає класифікаційну структуру документів за різними ознаками, зокрема, за авторством текстів. Кількість центрів кластеризації є вхідним параметром у методі *k*–середніх і може бути вибрана експериментальним шляхом, враховуючи наявність в кластерній структурі домінуючих кластерів для документів із спільними класифікаційними ознаками.

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 соответсвия: пер. с фр. – М.: Финансы и статистика, 1988. – 342 с: ил. 3. Павлишенко Б.М. Векторизація кластерів на растрових зображеннях електронної мікроскопії. Вісник Львів. ун-ту, серія фізична. 2007р., вип.40, с117-121. 4. 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. 5. 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. 6. Вердиева З.Н. Семантические поля в соверменном английском языке. – М.: Высшая школа, 1986. – 120с. 7. Левицкий В.В., Стернин И.А. Экспериментальные методы в семасиологии. – Воронеж: Изд-во ВГУ, 1989. – 192 с.