УДК 519.765:519.767:004.93

Б.М. ПАВЛИШЕНКО

СИНГУЛЯРНА ДЕКОМПОЗИЦІЯ МАТРИЦІ СЕМАНТИЧНИХ ОЗНАК В АЛГОРИТМІ ІЄРАРХІЧНОЇ КЛАСТЕРИЗАЦІЇ ТЕКСТОВИХ МАСИВІВ

Анотація. *Досліджується ієрархічна кластеризація текстових документів у просторі семантичних концептів, утвореному внаслідок сингулярного розкладу матриці текстових частотних характеристик семантичних полів. Показано, що кластерна структура в такому просторі може відображати класифікації документів за різними ознаками, зокрема, за авторством текстів.*
Ключові слова: *інтелектуальний аналіз текстів, кластерний аналіз, семантичні поля, сингулярна декомпозиція матриці, латентний семантичний аналіз.*

Аннотация. *Исследуется иерархическая кластеризация текстовых документов в пространстве семантических концептов, образованном вследствие сингулярного разложения матрицы текстовых частотных характеристик семантических полей. Показано, что кластерная структура в таком пространстве может отображать классификации документов по разным признакам, в частности, по авторству текстов.*
Ключевые слова: *интеллектуальний анализ текстов, кластерный анализ, семантические поля, сингулярная декомпозиция матрицы, латентный семантический анализ.*

Abstract. *The hierarchical clusterization of the text documents in the field of semantic concepts formed as a result of singular value matrix decomposition of the text frequencies characteristics of semantic fields has been investigated. It is shown that the cluster structure can represent documents classification by different characteristics particularly text authorship.*
Keywords: *intellectual text analysis, cluster analysis, semantic fields, singular value matrix decomposition of the text, latent semantic analysis.*

1. Вступ

Алгоритми кластеризації широко використовуються в інтелектуальному аналізі даних [1–3], зокрема, при вивченні структури текстових масивів [3]. Для представлення текстових документів часто використовують модель векторного простору [3, 4]. У цій моделі кожний документ відображається як вектор у багатовимірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми із словників текстових масивів. 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. Одним із шляхів вирішення цієї проблеми є використання латентного семантичного аналізу [4–6]. 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 побудований на лінійних комбінаціях квантитативних характеристик лексем словника. Такий новий векторний простір часто називають простором концептів (в деяких статтях – простором гіпотез). 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 у порівнянні із матрицею ознак для лексем словника текстових масивів. Семантичні поля формуються на основі експертного аналізу,

© Павлишенко Б.М., 2012

ISSN 1028-9763. Математичні машини і системи, 2012, № 1

одні і ті ж лексеми можуть одночасно належати до різних семантичних полів. 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 певною ознакою текстових документів. Такою ознакою може бути, наприклад, спільний стиль або автор. Сингулярна декомпозиція матриці семантичних ознак буде ефективною у випадку відображення класифікаційної структури в кластерній структурі, утвореній у новому векторному просторі семантичних концептів.

2. Постановка задачі

Для аналізу ефективності сингулярної декомпозиції матриці семантичних ознак розглянемо утворення матриці «частоти_семантичних_полів–документи» та проаналізуємо сингулярний розклад цієї матриці. На прикладі тестової вибірки текстових документів проаналізуємо утворення ієрархічної кластерної структури у векторних просторах семантичних концептів різної розмірності. Далі співставимо класифікаційний розподіл текстових документів за авторами та утворену кластерну структуру в новому просторі семантичних концептів.

3. Утворення матриці ознак «частоти_семантичних_полів–документи»

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

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

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 [7, 8]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та ін. Документ d_j з множини текстових документів D можна представити як упорядковану множину слів, порядок елементів якої відповідає порядку слів у цьому документі:

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

Впорядкований за алфавітом словник текстового документа d_j розглянемо як мультимножину W_j^d над множиною словника W:

$$W_j^d = \{ n_{ij}^{wd}(w_i) \mid w_i \in d_j, i = 1,2..., N_w \}, \tag{5}$$

де n_{ij}^{wd} – кількість входжень лексеми w_i зі словника W у множину лексем текстового документа d_j, яку можна визначити як

$$n_{ij}^{wd} = \sum_{l=1}^{N_j^t} f_{wd}(t_{lj}, w_i), \tag{6}$$

де $$f_{wd}(t_{lj}, w_i) = \begin{cases} 1, t_{lj} = w_i \\ 0, w_{lj}^d \neq w_i \end{cases}. \quad (7)$$

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

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

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

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

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

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

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

$$n_{kj}^{sd} = \sum_{l=1}^{N_j^t} f_s(t_{lj}, s_k), \quad (11)$$

де $$f_s(t_{lj}, s_k) = \begin{cases} 1, t_{lj} \in W_k^s \\ 0, t_{lj} \notin W_k^s \end{cases}.$$

Введемо матрицю семантичних ознак типу «частоти_семантичних_полів–документи»

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

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

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

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. У семантичному базисі можуть спостерігатися якісно нові групування текстових документів. Розгляд таких групувань може бути ефективним в алгоритмах комплексного аналізу текстів.

4. Сингулярний розклад матриці частот семантичних полів

Розглянемо сингулярний розклад 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

$$(V_p^s)^T V_q^s \tag{15}$$

визначає кількісну міру близькості цих векторів у N_s-мірному семантичному просторі текстових документів. Відповідно добуток матриць

$$(M_{sd})^T M_{sd} \tag{16}$$

містить скалярні добутки векторів $(V_p^s)^T V_q^s$ всіх документів і відображає їхні кореляції у просторі семантичних векторів. Нехай існує сингулярна декомпозиція матриці M_{sd}:

$$M_{sd} = U_{sd} \Sigma_{sd} \mathrm{Y}_{sd}{}^T. \tag{17}$$

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

$$(M_{sd})^T M_{sd} = \left(U_{sd}\Sigma_{sd}\mathrm{Y}_{sd}{}^T\right)^T \left(U_{sd}\Sigma_{sd}\mathrm{Y}_{sd}{}^T\right) = \mathrm{Y}_{sd}\Sigma_{sd}{}^T\Sigma_{sd}\mathrm{Y}_{sd}{}^T. \tag{18}$$

У відповідності з теорією сингулярного розкладу матриць [5, 6] діагональна матриця Σ_{sd} містить сингулярні числа у 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}:

$$(M_{sd})_K = (U_{sd})_K (\Sigma_{sd})_K (\mathrm{Y}_{sd})_K{}^T. \tag{19}$$

Матриця $(\mathrm{Y}_{sd})_K$ відображає зв'язок між векторами документів $\hat{V}_j^s$ у новому комбінованому K-мірному семантичному просторі, який будемо називати простором семантичних концептів. Зв'язок між вектором V_j^s документа у початковому семантичному просторі та вектором $\hat{V}_j^s$ у просторі семантичних концептів можна описати так:

$$\begin{aligned} V_j^s &= (U_{sd})_K (\Sigma_{sd})_K \hat{V}_j^s, \\ \hat{V}_j^s &= (\Sigma_{sd})_K^{-1} (U_{sd})_K^T V_j^s. \end{aligned} \tag{20}$$

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.

5. Ієрархічна кластеризація текстових документів у семантичному просторі

Розглянемо групування документів за семантичними ознаками за допомогою алгоритму ієрархічної кластеризації. Нехай є множина текстових документів D, яка описується виразом (2), та множина кластерів

$$C = \{ c_m \mid m = 0,1,2..., N_c \}. \tag{21}$$

Необхідно побудувати відображення множини документів на множину кластерів:

$$U_{DC} : D \to C . \tag{22}$$

Відображення U_{DC} задає модель даних, яка є розв'язком задачі кластеризації [1–3]. Кожний елемент c_m множини кластерів C складається з підмножини текстових документів, які подібні між собою відповідно до деякої кількісної міри подібності r:

$$c_m = \{ d_i, d_j \mid d_i \in D,\ d_j \in D,\ r(d_i, d_j) < \varepsilon \}, \tag{23}$$

де ε – визначає деякий поріг для включення документів у кластер. Величина $r(d_i, d_j)$ є відстанню між елементами d_i та d_j. Якщо виконується умова

$$r(d_i, d_j) < \varepsilon , \tag{24}$$

то елементи вибірки вважають подібними і відносять до спільного кластера. В іншому випадку елементи знаходяться у різних кластерах. У наших дослідженнях будемо використовувати евклідову відстань:

$$r_e(d_i, d_j) = \sqrt{\sum_{k=1}^{N_s} \left(p_{ki}^{sd} - p_{kj}^{sd} \right)^2} . \tag{25}$$

Розглянемо послідовність агломеративної кластеризації. На першому кроці вся множина текстових документів розглядається як множина кластерів:

$$c_1 = \{ d_1 \}, c_1 = \{ d_1 \}, \ldots c_{Nd} = \{ d_{Nd} \}. \tag{26}$$

На наступному кроці два близьких один до одного документи (наприклад, d_p і d_q) об'єднуються в один спільний кластер, нова множина на цьому кроці вже складається із $N_d - 1$ кластерів і має вигляд

$$c_1 = \{ d_1 \}, c_2 = \{ d_2 \}, \ldots c_p = \{ d_p, d_q \} \ldots c_{Nd-1} = \{ d_{Nd-1} \}. \tag{27}$$

Повторюючи кроки, на яких будуть об'єднуватися кластери, отримаємо множину із N_c кластерів. Процес об'єднання кластерів завершується на тому кроці алгоритму, коли жодна пара кластерів не відповідає порогу об'єднання для міри близькості елементів. Враховуючи те, що кластери можуть складатися з декількох об'єктів, існують різні методи формування й об'єднання кластерів на основі відстаней між об'єктами в середині кластера. У наших дослідженнях ми використовували метод Варда. У цьому методі обраховують квадрати евклідових відстаней від окремих документів до центра кожного кластера. Далі ці відстані сумують. У новий кластер об'єднуються 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. Наприклад, при дослідженні можливості кластеризації текстових документів за авторами доцільно взяти таке порогове значення міжкластерної відстані, при якому утворюється кількість кластерів, рівна кількості аналізованих авторів.

6. Експериментальна частина

Для аналізу ефективності розглянутих алгоритмів кластеризації взято текстову вибірку 155 художніх творів англомовної класики 4 відомих авторів (Ч. Діккенс, Д. Лондон, В. Скотт, М. Твен). Для утворення семантичного простору сформовано 15 семантичних полів, в які входить близько 5000 неозначених форм дієслова. Деталізація літературних та лексикографічних характеристик вхідних даних не є суттєвою 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 розраховані частотні спектри семантичних полів документів. Отже, кожний документ розглядається як вектор в 15-мірному початковому семантичному просторі. Далі проведено сингулярний розклад матриці семантичних ознак. На рис. 1 наведено графічне зображення перших сингулярних чисел семантичних ознак типу «частоти_семантичних_полів–документи» у порядку спадання.

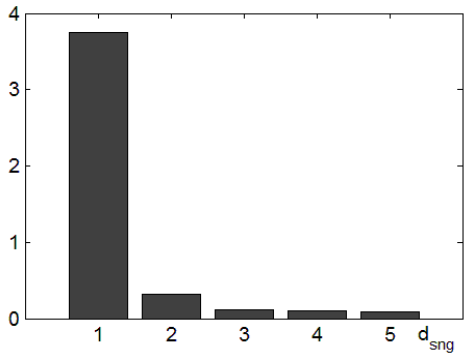

Рис. 1. Сингулярні числа матриці семантичних ознак у порядку спадання

Слід відмітити різке спадання значень сингулярних чисел, що дає можливість для апроксимації матриці семантичних ознак взяти суттєво менше значення рангу апроксимації K у порівнянні із початковою розмірністю семантичного простору. На наступному етапі була проведена агломеративна ієрархічна кластеризація документів у просторах семантичних концептів різної розмірності. Для оцінки міжкластерних відстаней використовувалась евклідова відстань (25), а кластеризацію було проведено методом Варда. На рис. 2 наведено дендрограму ієрархічної кластеризації при розмірності простору семантичних концептів $K = 10$, а на рис. 3 – при $K = 5$. По осі абсцис відкладено номери кластерів, а по осі ординат – міжкластерні відстані.

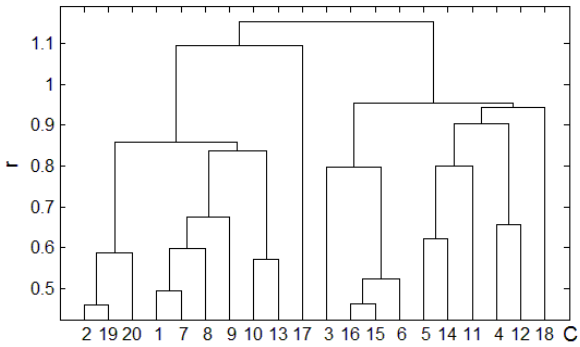

Рис. 2. Дендрограма кластеризації масиву текстових документів при $K = 10$

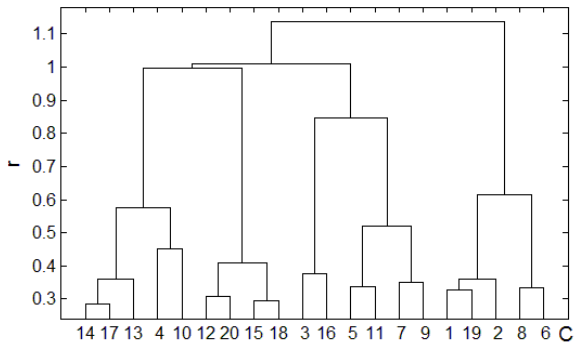

Рис. 3. Дендрограма кластеризації масиву текстових документів при $K = 5$

Наведені дендрограми обмежені рівнем із 20-ма кластерами. Як випливає з наведених рисунків, вибраний ранг апроксимації матриці семантичних ознак впливає на формування кластерної структури. Для подальших досліджень розглядається розмірність простору семантичних концептів $K = 5$ як найбільш оптимальна з точки зору утворення ієрархічної кластерної структури, яка відображає класифікаційну структуру розглянутого текстового масиву. Проаналізуємо класифікацію текстових документів за авторами. Виберемо таку порогову міжкластерну відстань, при якій утворюється кількість кластерів рівна кількості авторів текстів у досліджуваній вибірці. В аналізованому випадку це чотири кла-

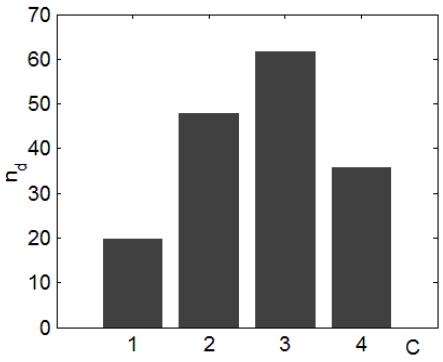

Рис. 4. Розподіл кількості текстових документів за кластерами $(K=5)$

стери. На рис. 4 наведено розподіл кількості текстових документів за чотирма кластерами, утвореними методом Варда.

На рис. 5 наведено розподіл текстів за авторами (1-Ч. Діккенс, 2-Дж. Лондон, 3-В. Скотт, 4-М. Твен) у кожному із чотирьох кластерів. Як випливає із наведених даних, тексти автора № 3 відсутні у кластерах № 1, 3, 4 і максимально сконцентровані у кластері № 2. Тексти автора №1 відсутні в кластері №1 і домінують у кластері №4. Домінуючим кластером для автора № 2 є кластер № 3 і т.д. Такий нерівномірний розподіл текстів за авторами в кластерах свідчить про те, що кластерна структура документів у просторі семантичних концептів відображає класифікаційну структуру документів за авторами.

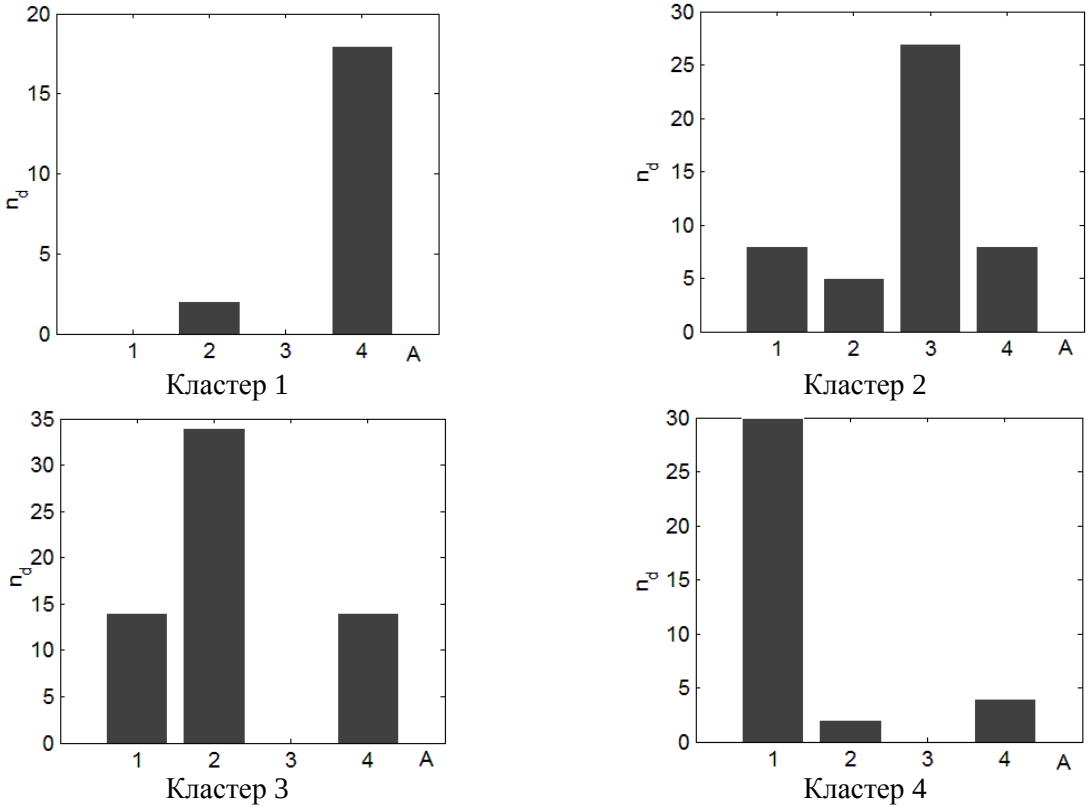

Рис. 5. Розподіл кількості текстових документів за авторами в досліджуваних кластерах $(K=5)$

У випадку кластеризації документів у просторі семантичних концептів більшої розмірності $(K>5)$ розподіл документів одного і того ж автора за кластерами може якісно відрізнятися, однак спостерігаються домінантні кластери для документів певних авторів. При низькій розмірності $K \in \{1,2,3\}$ кластери текстів з домінуючими авторами зникають і розподіл за авторами по кластерах стає більш рівномірним.

7. Висновки

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 у новому просторі семантичних концептів. Ієрархічна кластеризація документів у 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 кластерной структурі, що визначається наявністю домінуючих кластерів для документів окремих авторів.

СПИСОК ЛІТЕРАТУРИ

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

Стаття надійшла до редакції 10.06.2011