УДК 519.765:519.767:004.93

Б.М. Павлишенко
Львівський національний університет імені Івана Франка,
кафедра фізики напівпровідників

КЛАСТЕРНИЙ АНАЛІЗ ПОВІДОМЛЕНЬ ГРУП НОВИН У ПРОСТОРІ СЕМАНТИЧНИХ ОЗНАК

Проаналізовано кластеризацію текстових повідомлень груп новин на основі моделі векторного простору із базисом, утвореним семантичними ознаками. Як семантичні ознаки розглянуто частотні характеристики семантичних та тематичних полів. Показано ефективність запропонованої моделі у тематичному аналізі неструктурованих текстових повідомлень.

Ключові слова: інтелектуальний аналіз текстів, кластерний аналіз, векторна модель текстів, семантичні поля, тематичні поля.

The clustering of newsgroups text messages based on the vector space model with the basis formed by semantic characteristics has been analysed in this work. As semantic attributes the frequencies characteristics of semantic and thematic fields were considered. The efficiency of the proposed model in the thematic analysis of unstructured text messages has been shown.

Key words: text mining, cluster analysis, vector space model of texts, semantic fields, thematic fields.

Вступ

Перспективним напрямом використання інтелектуального аналізу текстових документів є аналіз об'ємної бази текстових повідомлень груп новин та тематичних форумів у мережі Інтернет. Одним із методів такого аналізу є навчання інформаційної системи на основі алгоритмів кластеризації. Такі методи, на відміну від класифікаційних методів, не потребують навчальної вибірки, а формують структуру об'єднаних елементів аналізу на основі заданих критеріїв. Утворені кластери відображають структурні зв'язки подібності між аналізованими елементами [1–3].

Аналіз останніх досліджень та публікацій

У задачах аналізу текстового змісту актуальними є теорії лексичної семантики, зокрема, вчення про семантичні поля. Подібними об'єктами у комп'ютерній інформатиці є семантичні мережі, які характеризують зв'язки між різними концептами. Прикладом комп'ютерної лексикографічної системи, в якій відображено семантичну мережу зв'язків між лексемами, є система WordNet, яка розроблена у Прінстонському університеті [4]. Ця система побудована на основі експертного лексикографічного аналізу семантичних структурних зв'язків, які описують денотативні та конотативні характеристики лексемного складу словника. Глибина зв'язків у такій системі визначається експертною оцінкою лексемних комбінацій у текстових масивах і обмежується науковим досвідом експертів та обсягом проаналізованого матеріалу. Актуальною, на нашу думку, є побудова математичних моделей та алгоритмів для автоматизованого формування семантичної мережі лексем. Семантичну структурну огранізацію лексемного складу словника можна використати у відповідних алгоритмах класифікації та кластеризації текстових об'єктів з погляду зменшення розмірності задач аналізу та виявлення нових семантичних зв'язків у онтології предметної області, до якої відносять аналізований масив текстів. У роботі [5] введено поняття семантичного домена, який описує певну семантичну область розгляду тієї чи іншої теми обговорень, наприклад, економіки, політики, фізики, програмування тощо. Для розгляду алгоритмів текстової кластеризації часто використовують стандартизовані масиви текстових документів. Однією із таких колекцій є 20-Newsgroups [http://qwone.com/~jason/

/20Newsgroups/], в яку входить колекція з приблизно 20 тисяч документів 20 груп новин. Цю колекцію використовують у тестових задачах інтелектуального аналізу текстів, зокрема для класифікації та категоризації текстових масивів. Кластерний аналіз є ефективним для вивчення структури текстових масивів [6–8]. Для представлення текстових документів часто використовують модель векторного простору [9]. У цій моделі кожний документ відображено як вектор у багатовимірному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми зі словників аналізованих текстових масивів. Такий підхід породжує 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, а також за характеристиками заданих лексемних об'єднань, наприклад, семантичних полів [10]. У роботах [7, 8] текстові документи розглядають як вектори, складовими яких є частоти семантичних полів у цих документах. У роботі [10] розглянуто теоретико-множинну концепцію семантичних полів у масивах текстових даних. Показано, що семантичні класи утворюються як відношення еквівалентності. Семантичне поле визначено як сегмент, який утворюється семантичними класами, об'єднаними бінарним кластером у структурному відношенні семантичного розбиття лексемного словника текстових масивів. Розглянуто відношення, яке описує розбиття словника на семантичні класи із структурою, яка визначає семантичні поля лексемного словника. Проаналізовано утворення семантичних полів на основі лексемних відношень, зокрема, таких як сполучення у тексті лексем семантичного поля та лексем полеутворювальної множини. Показано, що використання концепції семантичних полів є ефективним у векторній моделі текстових документів внаслідок зменшення розмірності фазового простору представлення документів. У роботі [7] запропоновано модель кластеризації текстових документів у семантичному просторі, яка уможливлює новий структурний поділ документів за семантичними ознаками у просторі набагато меншої розмірності, ніж простір, утворений лексемним складом текстової вибірки. Такий структурний поділ відображає класифікацію документів за новими ознаками, зокрема за авторством текстів. В роботі [8] показано, що сингулярний розклад матриці семантичних ознак типу "частоти_семантичних_полів–документи" дає можливість аналізувати текстові документи у новому просторі семантичних концептів. Методи кластеризації у просторі семантичних ознак можуть мати ряд особливостей для окремих типів текстових масивів. Актуальним є аналіз особливостей кластеризації для текстових повідомлень груп новин та форумів. Для апробації методик інтелектуального аналізу даних існує ряд текстових колекцій, зокрема, згадана вище колекція 20Newsgroups. Візьмемо цю текстову колекцію за основу для експериментальної частини кластерного аналізу в просторі семантичних ознак. Поряд із векторним простором текстових документів, утвореним семантичними полями, ефективними для аналізу можуть бути просторові базиси на основі інших семантичних ознак, зокрема, тематичних ознак, які визначаються тематикою груп новин. За аналогією із семантичними полями можна розглянути тематичні поля на основі груп новин.

Постановка задачі

Розглянемо представлення текстових повідомлень груп новин у векторному просторі семантичних ознак. Як семантичні ознаки виберемо частотні характеристики семантичних та тематичних полів. Сформуємо матриці частот семантичних та тематичних полів для аналізованого масиву документів. Реалізуємо ієрархічну кластеризацію текстових документів у просторі семантичних ознак у випадку семантичних та тематичних полів. Побудуємо дендрограми кластерних структур для двох випадків семантичних та тематичних полів та проаналізуємо особливості утвореної кластерної структури. Проаналізуємо розподіл категорій текстових повідомлень за кластерами.

Модель текстових документів у просторі семантичних ознак

Розглянемо модель на основі теорії множин, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Нехай існує певний словник лексем, які вживаються у текстових масивах. Опишемо цей словник як впорядковану множину

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

Сукупність текстових документів опишемо такою множиною

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

Введемо множину семантичних полів

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

Під семантичним полем розуміють таку множину лексем, які об'єднані певним спільним поняттям [4, 5, 10]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття тощо. На основі лексемного складу семантичних полів сформуємо матрицю типу семантичні_ознаки-документи, у якій ознаками є частоти семантичних полів:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d}. \quad (4)$$

Частоти семантичних полів p_{kj}^{sd} визначають як суми текстових частот лексем, які входять у ці семантичні поля [7,10]. Значення цих частот пронормовані так, щоб їхня сума для кожного документа дорівнювала 1. Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}\right) \quad (5)$$

відображає документ d_j у N_s-вимірному просторі текстових документів.

Введемо поняття тематичного поля за аналогією із семантичним полем. Вважаємо, що тематичне поле утворюють лексеми словника текстових масивів, які характеризують тематику деякої групи текстів, у аналізованому випадку це тематика групи новин. Введемо деякий коефіцієнт, який відображатиме, у скільки разів деяка лексема вживається частіше у деякій групі, ніж у загальній вибірці документів усіх груп. Визначимо цей коефіцієнт як відношення частоти лексеми у документах заданої групи до частоти цієї самої лексеми у загальній текстові вибірці.

$$Kthem_{ij}^{wg} = \frac{p_{ij}^{wg}}{p_i^w}. \quad (6)$$

Назвемо $Kthem_{ij}^{wg}$ коефіцієнтом тематичної виразності. Визначимо тематичне поле як підмножину словника лексем, для яких коефіцієнт тематичної виразності більший за деяке наперед визначене значення:

$$W_k^{them} = \{w_i \mid Kthem_{ik}^{wg}(w_i) > Kthem_t\}, \quad (7)$$

де $Kthem_t$ – деяке порогове значення коефіцієнта тематичної виразності. На основі визначення множини тематичного поля сформуємо лексемний склад для кожного тематичного поля, заданого певною групою повідомлень новин. На основі сформованого складу тематичних новин визначимо частоти тематичних полів кожного документа як суми частот лексем, які належать цьому полю. Частоти тематичних полів утворюють координати текстових повідомлень у векторному семантичному просторі. Текстові документи можна представити за допомогою тематичних векторів V_j^{them}, які визначають за аналогією до семантичних векторів (5).

Введення простору семантичних та тематичних полів не тільки зменшує розмірність задачі аналізу текстів, а й також вводить 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.

Модель кластеризації текстових документів у просторі семантичних ознак

Частоти семантичних та тематичних полів утворюють векторний простір, у якому кожний документ можна представити за допомогою векторів V_j^s та V_j^{them}. Розглянемо групування документів за семантичними ознаками, використовуючи алгоритм ієрархічної кластеризації. Нехай є множина текстових документів *D*, яка описується виразом (2), та множина кластерів

$$C = \{c_m \mid m = 0,1,2...,N_c\}. \quad (8)$$

Необхідно побудувати відображення множини документів на множину кластерів:

$$U_{DC} : D \to C. \quad (9)$$

Відображення U_{DC} задає модель даних, яка є розв'язком задачі кластеризації [1, 2, 6]. Кожний елемент c_m множини кластерів C складається з підмножини текстових документів, які подібні між собою відповідно до деякої кількісної міри подібності r

$$c_m = \{d_i, d_j \mid d_i \in D, d_j \in D, r(d_i, d_j) < \varepsilon\}, \quad (10)$$

де ε – визначає деякий поріг для введення документів у кластер. Величина $r(d_i, d_j)$ є відстанню між елементами d_i та d_j, і якщо вона менша за деяке значення, то елементи вибірки вважають подібними і приналежними до спільного кластера. Оскільки на множині текстових документів введено поняття відстані, то кожен документ представлено у вигляді точки в N_s-вимірному просторі R^{N_s} семантичних ознак. У наших дослідженнях розраховуватимемо евклідову відстань

$$r_e(d_i, d_j) = \sqrt{\sum_{k=1}^{N_s} \left(p_{ki}^{sd} - p_{kj}^{sd}\right)^2}. \quad (11)$$

Розглянемо ієрархічний метод агломеративної кластеризації. На першому кроці всю множину текстових документів розглядають як множину кластерів. На наступному кроці два близьких один до одного документа об'єднують в один спільний кластер, нова множина на цьому кроці вже складається із N_d-1 кластерів. 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 та об'єднання кластерів на основі відстаней між об'єктами всередині кластера. У роботі [7] показано, що одним із ефективних методів класифікації текстових документів у просторі семантичних полів є метод Варда. У методі Варда розраховують квадрати евклідових відстаней від окремих документів до центра кожного кластера. Далі ці відстані підсумовують. У новий кластер об'єднують ті кластери, при об'єднанні яких виходить найменший приріст суми квадратів цих відстаней. Графічним зображенням результату ієрархічної кластеризації є дендрограма, на якій відображено процес агломеративного об'єднання кластерів. По осі абсцис відкладають номери кластерів, а по осі ординат – відстані між кластерами. За певних значень відстаней починають об'єднувати кластери. Зі зростанням міжкластерної відстані кластери об'єднують аж до повного злиття кластерів у один кластер. Тому для отримання інформативної кластерної структури вибирають деякий поріг міжкластерної відстані, за якого утворюється оптимальна, з погляду аналізу текстових масивів, кластерна структура.

Експериментальна частина

Для експериментального вивчення класифікації текстових документів у просторі семантичних полів ми вибрали стандартизовану текстову базу повідомлень 20NewsGroups [http://qwone.com/~jason/20Newsgroups/]. Ця база містить близько 20000 повідомлень, які рівномірно розподілені по 20 категоріях груп новин. Для формування семантичного простору вибрано лексеми, згруповані за семантичними полями іменників та дієслів у семантичній мережі WordNet [http://wordnet.princeton.edu]. Семантичні поля у мережі WordNet представлені лексикографічними файлами. У наших дослідженнях ми використали семантичні поля іменників та дієслів [4]. Семантичні поля іменників складаються із 26 лексикографічних файлів, із яких ми відібрали 54464 лексеми. Семантичні поля дієслів містять 15 лексикографічних файлів, у які ми відібрали 9097 лексем. У семантичні поля також ввійшли похідні форми лексем. За допомогою розробленого програмного забезпечення здійснено початкову обробку текстового масиву, вилучено допоміжні символи та текстові елементи, які не містять семантичної інформації. Для кожного документа та вибірки

загалом розраховано частотні словники, на основі яких обчислено матрицю M_{sd} (4) типу документ-частота_семантичного _поля. Із 20000 тисяч текстових повідомлень, розподілених по 20 категоріях, відібрано довільно 10000 документів. Вибрано агломеративний метод кластеризації із евклідовою міжкластерною відстанню. Для формування кластерів вибрано метод Варда. На рис. 1 наведено дендрограми кластерів, яка відображає процес формування кластерної структури. На цій дендрограмі зображено формування перших 20 кластерів. Процес кластеризації зупинено, коли у кластерній структурі вже 20 кластерів. Розподіл кількості документів за кластерами зображено на рис. 2.

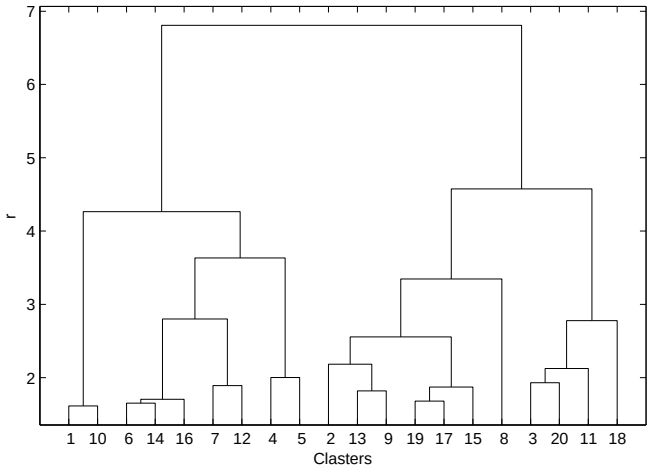

Рис. 1. Дендрограма ієрархічної кластеризації текстових документів у просторі семантичних полів

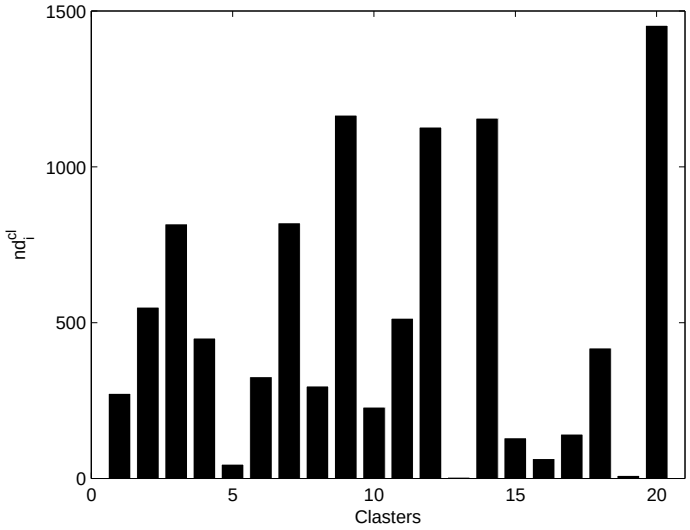

Рис. 2. Розподіл текстових повідомлень за утвореними кластерами у просторі семантичних полів

На рис. 3 наведено гістограми розподілу груп повідомлень у кластерах. Кожна гістограма відповідає окремому кластеру. Номери стовпців визначають номери груп повідомлень. Ці гістограми відображають як документи різних груп, розподілені у кожному кластері. Спостерігаються кластери, у яких домінують документи окремих категорій. Як випливає із наведених даних, деякі кластери містять документи широкого спектра категорій. Очевидно, що область цих кластерів у семантичному просторі є семантично однорідною і має низький семантично-диференціювальний потенціал. Однак також спостерігаються кластери, у яких домінують документи однієї чи декількох категорій. Семантичні просторові області цих кластерів володіють категорійно диференціювальним потенціалом і можуть бути використані в тематично-категорійному аналізі текстових документів як додатковий фактор визначення тематики документів.

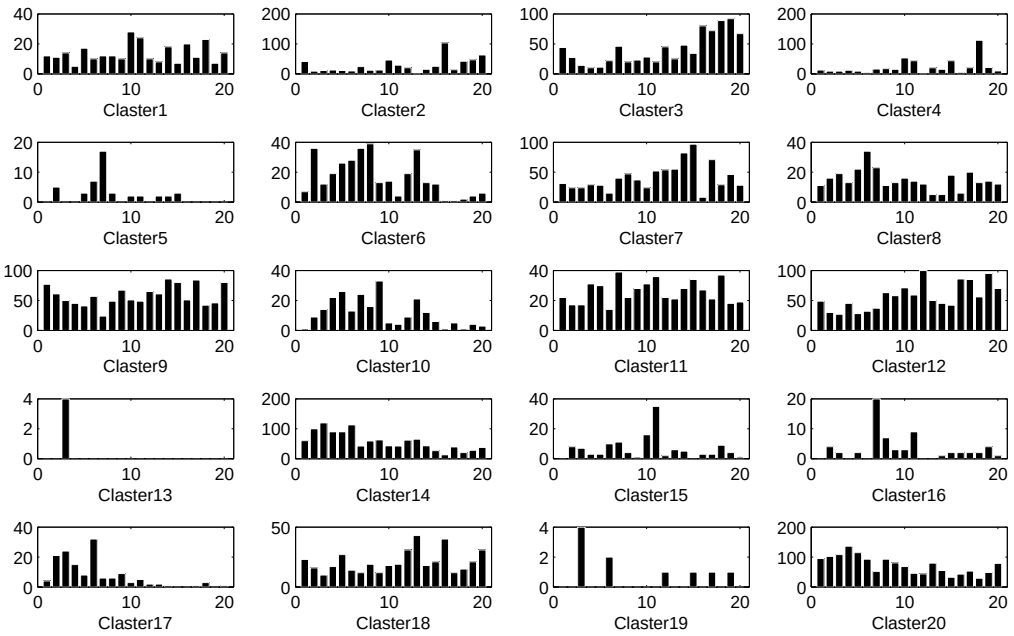

Рис. 3. Розподіл груп повідомлень у кожному кластері простору семантичних полів.

Області семантичного простору, що відповідають кластерам, у яких домінують є дві або декілька категорій, можна розглядати як області семантичного зв'язку цих категорій. На основі таких кластерів можна будувати семантичну мережу категорій, яка відображатиме ієрархічну структуру семантичних зв'язків цих категорій.

Для кластерного аналізу у просторі тематичних полів вибрано коефіцієнт $Kthem_{ij}^{wg}$, що дорівнює 2. Ми розрахували частотні словники як для окремих документів, так і для масивів повідомлень кожної окремої групи. Для кожної групи виявлено лексеми, для яких коефіцієнт тематичної виразності був більшим за 2. Ці лексеми утворюють тематичні поля, тематики яких задані кожною групою новин. На основі сформованих тематичних груп розраховано частоти тематичних полів у кожному документі. Сукупність таких частот є складовими векторного представлення кожного повідомлення у семантичному просторі.

На рис. 4 наведено дендрограму ієрархічної кластеризації текстових документів у просторі тематичних полів. На рис. 5 наведено розподіл текстових повідомлень за утвореними кластерами. На рис. 6 зображено розподіл груп повідомлень у кожному кластері. На основі аналізу розподілу груп новин у кластерах можна зробити ряд висновків. Кластери, у яких містяться повідомлення багатьох груп, характеризують у семантичному просторі області семантично нейтральних повідомлень, у яких відображені повідомлення із рівномірним семантичним розподілом лексем. Кластери, у яких є домінуючі групи новин, характеризують області семантично виразних лексем. Кластери, які містять декілька домінуючих груп, можна розглядати як області семантичних зв'язків між цими групами. Порівнюючи кластерні розподіли у просторі семантичних (рис. 3) та тематичних (рис. 4) полів, можна виявити, що простір тематичних полів є більш семантично диференціюючим для аналізованого масиву документів, порівняно із простором семантичних полів. Однак такий простір потребує додаткового формування тематичного базису векторного простору, який є ефективним для аналізованого масиву текстових документів. Крім того, формування базису тематичних полів потребує категоризованої вибірки документів, оскільки кожна категорія є основою формування заданого нею тематичного поля. Такий тип кластеризації можна вважати кластеризацією із навчальною вибіркою. Однак, на відміну від класифікаційного аналізу, в якому навчальна вибірка є необхідним елементом, сформовані тематичні поля можна використати у аналізі некатегоризованих за тематичним базисом масивах документів. Отже, тематичний базис дає можливість виявити нові групування аналізованих текстів, які не проявлялись у інших базисах.

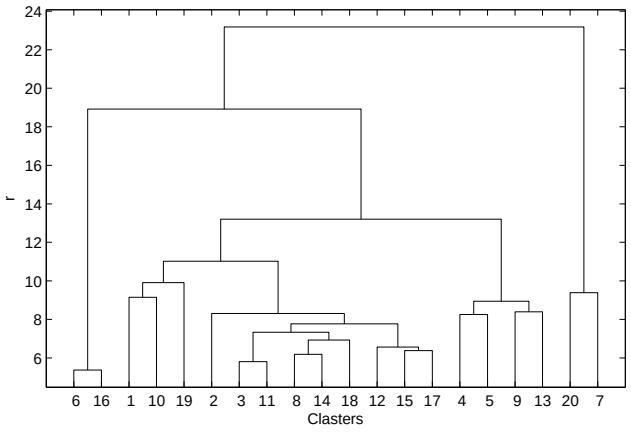

Рис. 4. Дендрограма ієрархічної кластеризації текстових документів у просторі тематичних полів

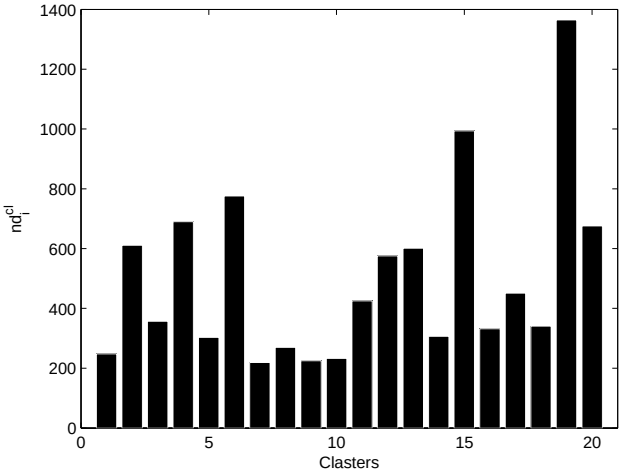

Рис. 5. Розподіл текстових повідомлень за утвореними кластерами у просторі тематичних полів

У просторі тематичних полів спостерігається більша кількість кластерів, у яких домінує лише одна група новин. Це пояснюється тим, що один із вимірів тематичного простору утворений тематично виразними лексемами певної групи. Однак спостерігаються також кластери, у яких та сама група домінує у декількох кластерах. Це свідчить про можливість розподілу цієї групи на підгрупи, у тематику яких можуть входити тематичні складові інших груп. Також проаналізована кластерна структура у просторі тематичних полів, коефіцієнт тематичної виразності якої дорівнює 3. У цьому випадку спостерігається диференційованіший розподіл тематичних груп повідомлень за різними кластерами. Також ми досліджували розбиття масиву повідомлень на більшу кількість кластерів. У цьому випадку спостерігаються кластери, у яких домінують ті самі категорії. Це може свідчити про наявність тематичних підкатегорій всередині категорій.

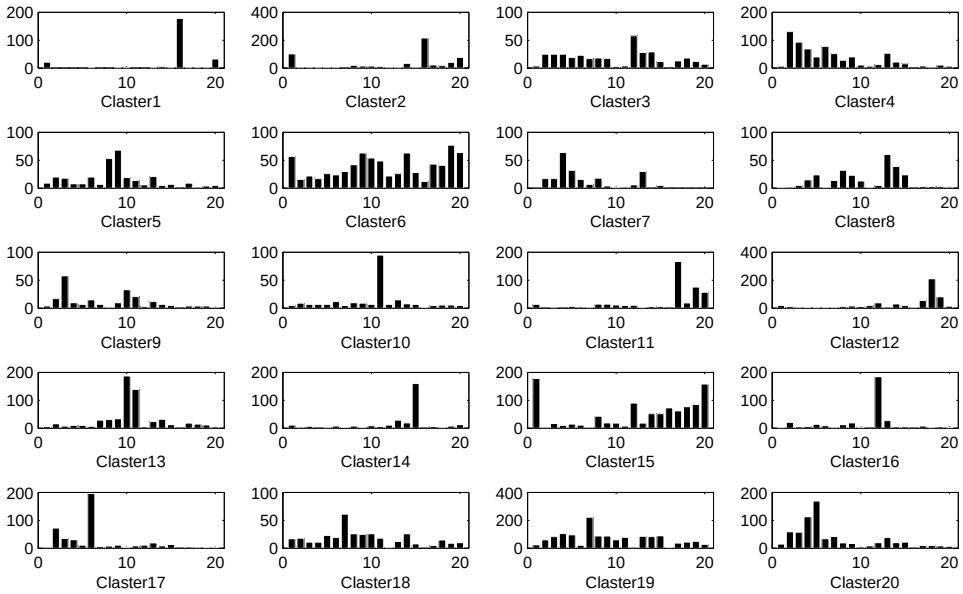

Рис. 6. Розподіл груп повідомлень у кожному кластері простору тематичних полів.

Висновки

Використання моделі векторного простору із базисом семантичних ознак є ефективним у алгоритмах кластерного аналізу текстових повідомлень груп новин. Як семантичні ознаки розглянуто частотні характеристики семантичних та тематичних полів. Тематичні групи новин утворюють тематичні поля на основі тематично виразних лексем. Аналіз розподілу груп новин у кластерній структурі показав наявність областей семантичного простору, в яких відображено

окремі групи новин, та областей, які відображають семантичні зв'язки між масивами повідомлень окремих груп. Кластерна структура повідомлень у просторі тематичних полів є семантично диференційованішою порівняно із кластерною структурою у просторі семантичних полів. Базис векторного простору на основі семантичних та тематичних полів є універсальним і не потребує експертного підбору ключових слів. Розмірність такого базису суттєво менша порівняно із методами кластеризації за ключовими словами.

1. Ким Д.О., Мьюллер Ч.У., Клекка У.Р. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.: ил. 2. Жамбю М. Иерархический кластер-анализ и соответствия: пер. с фр. – М.: Финансы и статистика, 1988. – 342 с. 3. Павлишенко Б.М. Векторизація кластерів на растрових зображеннях електронної мікроскопії // Вісник Львів. ун-ту, серія фізична. 2007 р. – Вип. 40. – С. 117–121. 4. Fellbaum C. WordNet. An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998, 432 p. 5. Gliozzo Alfio, Strapparava Carlo Semantic Domains in Computational Linguistics. Springer, 2009 – 132 p. 6. 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. – 512 с. 7. Павлишенко Б. М. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів // Електроніка та інформаційні технології. –2011. – Вип. 1.– С. 212–222. 8. Павлишенко Б. М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. – 2012. – №1. – С. 69–76. 9. Pantel Patrick, Turney Peter D. From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research.–2010. – Vol.37. – pp. 141–188. 10. Павлишенко Б.М. Використання концепції семантичного поля у векторній моделі текстових документів // Східно-Європейський журнал передових технологій. – 2011. – № 6/2(54). – С. 7–11.