ФІЗИКО-МАТЕМАТИЧНІ НАУКИ

УДК 336.72

АНАЛІТИЧНА ОБРОБКА ТЕКСТОВОЇ ІНФОРМАЦІЇ ЗА ДОПОМОГОЮ ЗАСОБІВ КЛАСТЕРИЗАЦІЇ

Деркач О.І.
Національний авіаційний університет

У статті було розглянуто засоби обробки текстової інформації, за допомогою яких інформацію створюють, редагують, досліджують та аналізують. Для аналізу інформації існують декілька методів. На сьогоднішній день основним методом аналітичної обробки текстових масивів даних залишається пошук документів за ключовими словами. Були визначені дві проблеми роботи з величезною кількістю інформації, автоматичний збір інформації, та автоматичний розбір інформації, що надійшла з даної тематики, проведений на основі аналізу тексту документа, та їх вирішення. В результаті аналізу було обрано кластеризацію як найбільш пристосований метод для дослідженя засобів обробки текстової інформації; здійснено порівняльний аналіз методів інтелектуального аналізу Data Mining; здійснено огляд програмного забезпечення, що функціонує на основі кластеризації; вивчено метод k − середніх.
Ключові слова: кластеризація, аналітична обробка, k − means, Data Mining, текстова інформація.

Постановка проблеми. В основу технології інтелектуального аналізу покладена концепція шаблонів, що представляють собою закономірності. В результаті виявлення цих закономірностей вирішуються завдання аналізу даних. Завдання аналізу іноді називають закономірностями або техніками. Єдиної думки щодо того, які завдання слід відносити до інтелектуального аналізу, немає. В основу сучасної технології *Data Mining* покладена концепція шаблонів, що відображають закономірності, властиві підвибіркам даних. Пошук шаблонів виконується методами, які не використовують ніяких вихідних припущень про ці підвибірки. Якщо при статистичному аналізі зазвичай формуються запитання: «Яке середнє число клієнтів банку, що не повернули вчасно кредит, серед неодружених чоловіків від 40 до 50 років?», то застосування *Data Mining*, як правило, має на увазі відповіді на запитання: «Чи існує типова категорія клієнтів, які не повертають вчасно кредити?». При цьому саме відповідь на друге запитання нерідко забезпечує прийняття успішного бізнес-рішення.

Важлива особливість *Data Mining* − нестандартність і неочевидність розшукуваних шаблонів. Іншими словами, методи (засоби) *Data Mining* відрізняються від інструментів статистичної обробки даних тим, що замість перевірки заздалегідь передбачуваних користувачами взаємозалежностей вони на підставі наявних даних здатні знаходити такі взаємозалежності самостійно і будувати гіпотези про їх характер.

Кластеризація в *Data Mining* набуває цінність тоді, коли вона виступає одним з етапів аналізу даних, побудови закінченого аналітичного рішення. Аналітику часто легше виділити за допомогою груп схожих об'єктів, вивчити їх особливості і побудувати для кожної групи окрему модель, ніж створювати одну загальну модель для всіх даних. Таким прийомом постійно користуються в маркетингу, виділяючи групи клієнтів, покупців, товарів і розробляючи для кожної з них окрему стратегію.

Неповний список прикладних областей, де застосовується кластеризація: сегментація зображень, маркетинг, боротьба з шахрайством, прогнозування, аналіз текстів і багато інших. Так, в медицині використовується кластеризація захворювань, лікування захворювань або їх симптомів, а також таксономія пацієнтів, препаратів. В археології встановлюються таксономії кам'яних споруд і стародавніх об'єктів. У маркетингу це може бути задача сегментації конкурентів і споживачів. У менеджменті прикладом завдання кластеризації буде розбиття персоналу на різні групи, класифікація споживачів і постачальників, виявлення схожих виробничих ситуацій, при яких виникає шлюб. У медицині − класифікація симптомів. У соціології завдання кластеризації − розбиття респондентів на однорідні групи.

Найчастіше кластеризація виступає першим кроком при аналізі даних. Після виділення схожих груп застосовуються інші методи, для кожної групи будується окрема модель.

Завдання кластеризації в тому чи іншому вигляді формулювали в таких наукових напрямках, як статистика, розпізнавання образів, оптимізація, машинне навчання. Звідси розмаїття синонімів поняття кластер − клас, таксон, згущення.

На сьогоднішній момент число методів розбиття груп об'єктів на кластери досить велике − кілька десятків алгоритмів і ще більше їх модифікацій. Однак нас цікавлять алгоритми кластеризації з точки зору їх застосування в *Data Mining*.

Аналіз останніх досліджень і публікацій. В роботі «Кластерний аналіз» Мандель досліджує використання методів кластеризації у фінансах і статистиці і доходить висновку, що перше питання, яке задається аналітиками при вирішенні багатьох завдань, полягає в тому, як

організувати дані в наочні структури, тобто розгорнути таксономії [1].

Хайдуков у своїй роботі «Применение кластерного анализа в государственном управлении» зазначив, що в результаті застосування різних методів кластерного аналізу можуть бути отримані кластери різної форми. Наприклад, можливі кластери «ланцюжка» типу, коли кластери представлені довгими «ланцюжками», кластери подовженої форми і т.д., а деякі методи можуть створювати кластери довільної форми.

Різні методи можуть прагнути створювати кластери певних розмірів (наприклад, малих або великих) або припускати в наборі даних наявність кластерів різного розміру.

Деякі методи кластерного аналізу особливо чутливі до шумів або викидів, інші – менш.

В результаті застосування різних методів кластеризації можуть бути отримані неоднакові результати, це нормально і є особливістю роботи того чи іншого алгоритму.

Дані особливості слід враховувати при виборі методу кластеризації [12].

Виділення не вирішених раніше частин загальної проблеми. Незважаючи на велику кількість досліджень в області кластерного аналізу, в цій області існує ряд актуальних проблем. Перелічимо основні проблеми.

1. Проблема обґрунтування якості результатів аналізу. Відомо, що процес угруповання в значній мірі носить суб'єктивний характер. Це виражається, зокрема, в тому, що один і той же набір об'єктів може класифікуватися по-різному в залежності від прикладної області, ступеня повноти знань про об'єкти вивчення і т.д. Тому необхідно розробляти методи, що дозволяють максимально повно враховувати існуючий довід, а також розробляти відповідні критерії якості угруповання.

2. Багато трудноформалізуемие області досліджень характеризуються недостатністю знань про досліджуваних об'єктах, що ускладнює формулювання їх математичних моделей. У їх умовах, зокрема, проблематичним стає застосування алгоритмів розщеплення суміші розподілів (наприклад, ЕМ-алгоритму [2]), що базуються на уявленні про те, що кожен клас описується деяким відомим (з точністю до параметрів) розподілом в просторі змінних.

3. Проблема аналізу великого числа різнотипних (кількісних або якісних) факторів. У разі різнотипної простору, виникає методологічна проблема визначення в ньому метрики (деякі способи введення такого роду метрик викладені в роботі [3]). З іншого боку, навіть у просторі однотипних (кількісних) змінних при збільшенні їх числа посилюється «прокляття розмірності», що може призвести до майже повної нерозрізненості точок.

Так, відстань від будь-якої точки до її «найближчого сусіда» для деяких видів відстаней може практично збігатися (з урахуванням машинної точності) з відстанню до її «далекого сусіда». Глядачеві аналогії, доречні в просторі малої розмірності, стають абсолютно неприйнятними в просторі великої розмірності. Наприклад, в 20-вимірному евклідовому просторі обсяг гіперкуба перевищує обсяг вписаною в нього гіперсфери більш ніж в 40 000 000 разів, що здається дивним з точки зору двох-або тривимірних просторів.

4. Нелінійність взаємозв'язків; наявність пропусків, похибок вимірювання змінних. Класичні методи зниження розмірності (метод головних компонент, метод незалежних компонент), які використовуються в кластерному аналізі, в основному орієнтовані на лінійні залежності між змінними. Для виявлення більш складних взаємозв'язків потрібні такі алгоритми, як нелінійні (ядерні) методи головних компонент [6] і т.п.

5. Необхідність представлення результатів аналізу в формі, зрозумілою фахівцям прикладної області. Крім гарної прогнозуючої здатності для будь-якого алгоритму аналізу даних важливо, наскільки зрозумілими і такими, що інтерпретуються є його результати. Для поліпшення інтерпретується рішень можна використовувати логічні моделі [4, 5]. Такого роду моделі використовуються для вирішення завдань розпізнавання образів і прогнозування кількісних показників, наприклад, в методах побудови вирішальних дерев або логічних вирішальних функцій.

6. Проблема пошуку глобального екстремуму у критерію якості угруповання. Критерій якості, як правило, є функцією, яка залежить від великої кількості факторів, нелінійним, що володіє безліччю локальних екстремумів. Для знаходження кластерів необхідно вирішити складну комбінаторних завдання пошуку оптимального варіанта класифікації.

Тому алгоритм повного перебору варіантів має трудомісткість, експоненціально залежить від розмірності. Якщо число груп заздалегідь невідомо, то переборного завдання стає ще складніше. Таким чином, при збільшенні розмірності таблиць даних відбувається «комбінаторний вибух». Класичні алгоритми кластерного аналізу здійснюють спрямований пошук в порівняно невеликому підмножині простору рішень, використовуючи різного роду апріорні обмеження (на число кластерів або їх форму, на порядок включення об'єктів в групи і т.д.). При цьому знаходження строго-оптимального рішення не гарантується. Для пошуку оптимального рішення застосовуються більш складні методи, такі як генетичні (еволюційні) алгоритми [7], нейронні мережі [8] і т.д. Існують експериментальні дослідження, що підтверджують переваги таких алгоритмів перед класичними алгоритмами [9]. Однак і при використанні еволюційних методів виникають проблеми [10], пов'язані зі специфікою розв'язуваної задачі кластер-аналізу: з труднощами інтерпретації використовуваних операторів рекомбінації і кросовера.

7. Проблема стійкості групіровочних рішень. У класичних алгоритмах рішення задач кластер-аналізу (наприклад, алгоритм К-середніх) результати угруповання можуть сильно змінюватися в залежності від вибору початкових умов, порядку об'єктів, параметрів роботи алгоритмів і т.п. Останнім часом різними авторами [11] пропонуються способи підвищення стійкості групіровочних рішень, засновані на застосуванні ансамблів алгоритмів. При цьому використовуються результати угруповання, отримані різними алгоритмами, або одним алгоритмом, але з різними параметрами налаштування, за різними підсистемами змінних і т.д. Після побудови ансамблю проводиться знаходження підсумкового колективного рішення.

Мета статті. Головною метою цієї роботи є дослідження методів аналітичної обробки текстової інформації, та опис модуля, що опрацьовує дані за допомогою засобів кластеризації, здійснення порівняльного аналізу методів інтелектуального аналізу *Data Mining*, здійснення огляду програмного забезпечення, що функціонує на основі кластеризації, вивчення методу k – середніх, огляд модуля аналітичної обробки «Кластеризація вхідних листів», що створено в середовищі розробки *Visual Studio* та здійснює кластеризацію вхідних листів.

Виклад основного матеріалу. Коротка характеристика підходів до кластеризації:

1) Алгоритми, засновані на поділі даних *(Partitioning algorithms)*:

– поділ об'єктів на k кластерів;

– ітеративний перерозподіл об'єктів для поліпшення кластеризації.

2) Ієрархічні алгоритми *(Hierarchy algorithms)*:

– агломерація: кожен об'єкт спочатку є кластером, кластери, з'єднуючись один з одним, формують більший кластер.

3) Методи, засновані на концентрації об'єктів *(Density-based methods)*:

– засновані на можливості з'єднання об'єктів;

– ігнорують шуми, знаходження кластерів довільної форми.

4) Грід – методи *(Grid – based methods)*:

– квантування об'єктів в грід – структури.

5) Модельні методи *(Model – based)*:

– використання моделі для знаходження кластерів, найбільш відповідних даних.

Методи кластерного аналізу можна розділити на дві групи:

– ієрархічні;

– неієрархічні.

Кожна з груп включає безліч підходів і алгоритмів. Використовуючи різні методи кластерного аналізу, аналітик може отримати різні рішення для одних і тих же даних. Це вважається нормальним явищем.

Загальноприйнятою класифікації методів кластеризації не існує, але можна виділити ряд груп підходів (деякі методи можна віднести відразу до декількох груп).

Корпорація *Accrue Software*, що займається розробкою програмного забезпечення для кластеризації даних, пропонує класифікацію, що приведена на рис. 1.

Найбільш поширений серед неієрархічних методів алгоритм k – середніх, також званий швидким кластерним аналізом. Повний опис алгоритму можна знайти в роботі Хартігана і Вонга (*Hartigan and Wong*, 1978) [13]. На відміну від ієрархічних методів, які не вимагають попередніх припущень щодо числа кластерів, для можливості використання цього методу необхідно мати гіпотезу про найбільш ймовірне кількості кластерів.

Алгоритм k – середніх будує k кластерів, розташованих на можливо великих відстанях один від одного. Основний тип задач, які вирішує алгоритм k – середніх, – наявність припущень (гіпотез) щодо числа кластерів, при цьому вони повинні бути різні настільки, наскільки це можливо. Вибір числа k може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

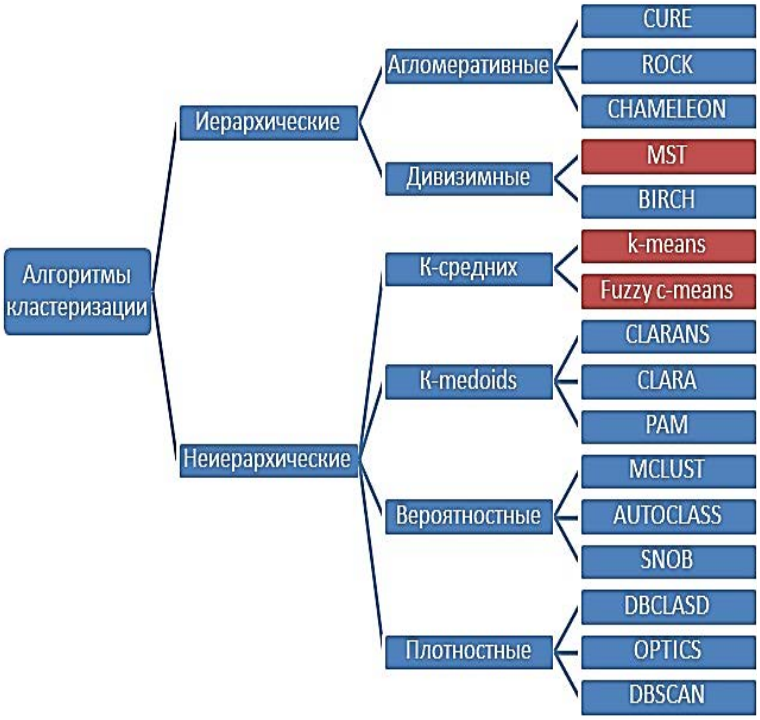

Рис. 1. Класифікація методів кластеризації

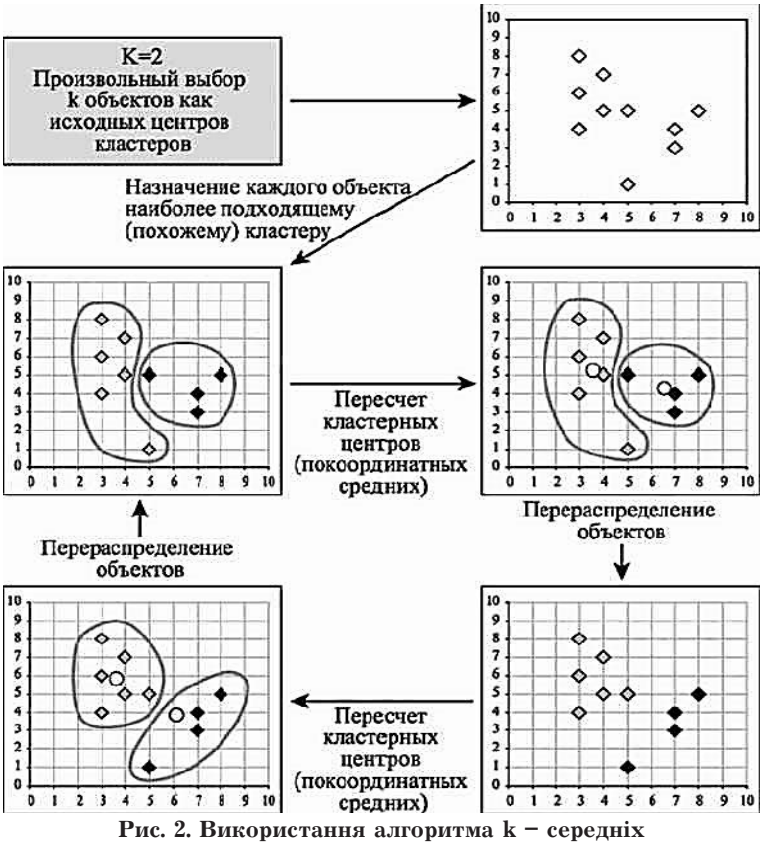

Рис. 2. Використання алгоритма k – середніх

Загальна ідея алгоритму: заданий фіксоване число k кластерів спостереження зіставляються кластерам так, що середні в кластері, центроїди, (для всіх змінних) максимально можливо відрізняються один від одного [14].

Опис алгоритму:

1) Початковий розподіл об'єктів по кластерам.

Вибирається число k, це число відповідає кількості кластерів, і на першому кроці ці точки вважаються «центрами» кластерів. Кожному кластеру відповідає один центр.

Вибір початкових центроїдів може здійснюватися в такий спосіб:

– вибір k – спостережень для максимізації початкової відстані;

– випадковий вибір k – спостережень;

– вибір перших k – спостережень.

В результаті кожен об'єкт призначений певному кластеру.

2) Ітеративний процес.

Обчислюються нові центри кластерів, за якими далі об'єкти знову перерозподіляються.

Процес обчислення центрів і перерозподілу об'єктів триває доти, поки не буде виконана одна з умов:

– кластерні центри стабілізувалися, тобто всі спостереження належать кластеру, до якого належали до поточної ітерації;

– число ітерацій дорівнює максимальному числу ітерацій.

Мета методу – розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції.

На рис. 2. наведено приклад використання алгоритму k – середніх де $k = 2$.

Вибір числа кластерів є складним питанням. Якщо немає припущень щодо цього числа, рекомендують створити 2 кластера, потім 3, 4, 5 і т.д., порівнюючи отримані результати.

Після отримання результатів кластерного аналізу методом k – середніх слід перевірити правильність кластеризації (тобто оцінити, наскільки кластери відрізняються один від одного). Для цього розраховуються середні значення для кожного кластера. При гарній кластеризації повинні бути отримані сильно відрізняються середні для всіх вимірювань або хоча б більшої їх частини.

Переваги алгоритму k – середніх:

– простота використання;

– швидкість використання;

– зрозумілість і прозорість алгоритму.

Недоліки алгоритму k – середніх:

– алгоритм занадто чутливий до викидів, які можуть спотворювати середнє. Можливим вирішенням цієї проблеми є використання модифікації алгоритму – алгоритм k – медіани;

– алгоритм може повільно працювати на великих базах даних. Можливим вирішенням цієї проблеми є використання вибірки даних.

– результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів;

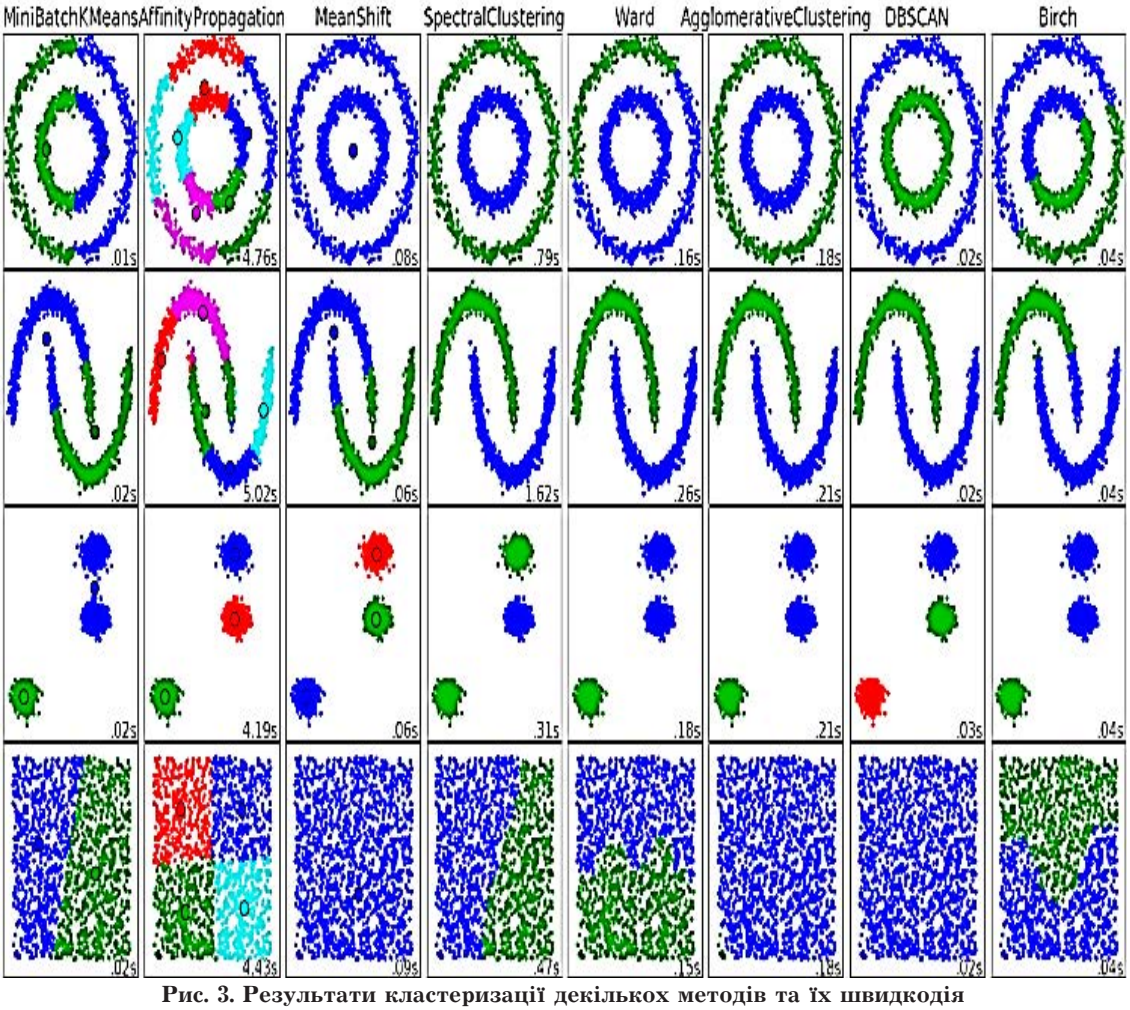

Рис. 3. Результати кластеризації декількох методів та їх швидкодія

– кількість кластерів повинна бути заздалегідь визначена дослідником;

Метод k – середніх є доволі простим і прозорим, тому успішно використовується у різноманітних сферах – маркетингових сегментаціях, геостатистиці, астрономії, сільському господарстві тощо.

В алгоритмах глибокого навчання метод k – середніх іноді застосовують не за прямим призначенням (класифікація розбивкою на кластери), а для створення так званих фільтрів (ядер згортки, словників). Наприклад, для розпізнавання зображень в алгоритм k – середніх подають невеликі випадкові шматочки зображень навчальної вибірки, припустимо, розміром 16x16 у вигляді лінійного вектора, кожен елемент якого кодує яскравість своєї точки. Кількість кластерів k задається більшим числом, наприклад 256. «Навчений» метод k – середніх за певних умов виробляє при цьому центри кластерів (центроїди), які представляють собою зручні базиси, на які можна розкласти будь вхідне зображення. Такі «навчені» центроїди надалі використовують в якості фільтрів, наприклад для нейронної мережі в якості ядер згортки або інших аналогічних систем машинного зору. Таким чином здійснюється навчання без вчителя за допомогою методу k – середніх.

На рис. 3 зображено порівняння результатів кластеризації одних і тих самих даних різними методами, з приведеною швидкодією.

Метод кластеризації k – середніх було обрано завдяки простоті та швидкості його використання, зрозумілості та прозорості його алгоритму. Тим паче існує багато методів модифікації даного алгоритму, що дозволить в майбутньому модифікувати його у відповідності до поставлених вимог. Недоліки методу – необхідність заздалегідь задавати кількість кластерів та чутливість до вибору початкових значень центрів кластерів.

Програма TextAnalyst.

TextAnalyst – комплексний змістовний аналізатор текстів, що працює з текстовою інформацією.

TextAnalyst представляє універсальний текстовий аналізатор за допомогою якого користувач може здійснювати навігацію по тексту, будувати анотацію тексту різного об'єму, будувати і редагувати семантичну мережу тексту, в тому числі редагувати семантичний вага понять і міняти структуру семантичних зв'язків за рахунок зміни їх ваг. Семантична мережа являє собою зважений граф понять з аналізованого тексту; поняття і зв'язку ранжуються за допомогою нейронної мережі. Анотація будується на основі семантичної мережі і являє собою набір найбільш значущих пропозицій. Ядром системи є бібліотека *COM* модулів *TextAnalyst SDK*.

Ключові можливості:

– Кластеризация текстів – видалення слабких посилань в семантичної мережі призводить до розбиття корпусу текстів на семантично однорідні кластери і дає можливість проводити подальший аналіз тексту.

– Виявлення змісту тексту – формування та експорт точної семантичної мережі тексту або текстової бази. Ця мережа являє зміст тексту і служить підставою для подальшого аналізу тексту.

– Точний виклад текстового матеріалу – якість короткого викладу тексту забезпечується збалансованим поєднанням методів дослідження лінгвістичної і нейронної мережі. Обсяг викладу контролюється за допомогою семантичної порогової величини.

Програма Угрупування ключових слів для *SEO/PPC*.

Щоб створити рекламну кампанію або згрупувати запити по посадковим сторінкам, підготувати комерційну пропозицію або розрахувати бюджет – потрібно перебрати 1000–5000 ключових слів. Якщо у нас всього лише 20–30 ключових слів – їх можна згрупувати вручну. Сто або двісті – за допомогою *Excel*. Однак, щодня працюючи з тисячами ключових слів – простіше і зручніше використовувати інструменти автоматизації.

Програма кластеризації визначає частоти для всього індексу документа та вираховує рейтинг. Частоти рахуються для кожного слова (після нормалізації). Якщо у нас є «відпочинок в Трускавці» то програма рахує частоти для «відпочинок» та «трускавець».

На даному етапі – рейтинг слів вибудовується від найбільш частотних до менш частотних. Навіщо це потрібно? Щоб створити основні групи (кластери). Якщо слово «Трускавець» зустрічається частіше ніж слово «готель» то пошуковий запит наприклад, «недорогі готелі Трускавця» включає слово «готель» буде віднесений до групи «Трускавець» а не навпаки. Рейтинг слів утворює назви груп. Ключові слова «прив'язуються» до відповідних груп.

Саме такий підхід працює дуже ефективно, коли основні групи створюються на основі рейтингу від найбільш частотних слів до найменш частотним, а підгрупи – від найменш частотних до найбільш частотним.

Програма дозволяє дуже швидко справлятися з великими обсягами даних. 5000-20000 ключових слів групуються за кілька секунд.

Одне цікаве застосування кластеризації – стиснення кольорового зображення. Наприклад, є зображення з мільйонами квітів. У більшості зображень, велика кількість кольорів не буде використовуватися, і, навпаки, велика кількість пікселів матиме подібні або ідентичні кольори. За допомогою *IPython Notebook* та бібліотеки *Scikit* можна зменшити кількість кольорів.

Вхідне зображення має 2563 кольорів. Вихідне зображення буде мати лише

64 кольори. Таким чином, буде знайдено *N*колір кластерів в зображені, і створено новий

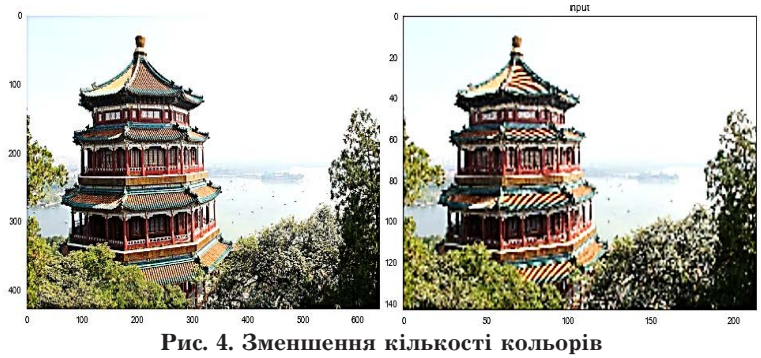

Рис. 4. Зменшення кількості кольорів

образ, де істинний введений колір замінюється кольором найближчого кластера (рис. 4).

Висновки і пропозиції. Головна ціль статті – здійснення аналітичної обробки текстової інформації за допомогою методів кластеризації, яка широко використовується в системах інтелектуального аналізу даних. Синонімами терміну інтелектуальний аналіз даних є видобуток даних *(data mining)*, виявлення знань *(knowledge discovery)*. Інтелектуальний аналіз даних пов'язаний з пошуком в даних прихованих і корисних закономірностей, що дозволяють отримати нові знання про досліджені дані. Перед фахівцями з різних областей людської діяльності постало питання про обробку даних, що збираються, та перетворення їх в знання. Відомі статистичні методи покривають лише частину потреб по обробці даних, і для їх використання необхідно мати чітке уявлення про шукані закономірності. У такій ситуації методи інтелектуального аналізу даних набувають особливу актуальність. Їх особливість полягає у встановленні наявності та характеру прихованих закономірностей в даних, тоді як традиційні методи займаються головним чином параметричної оцінкою вже встановлених закономірностей. Серед методів інтелектуального аналізу даних особливе місце займають класифікація та кластеризація. Класифікація, при відомому заздалегідь угрупованню даних на підмножини (класи), встановлює закономірність, за якою дані групуються саме таким чином. Кластеризація ж, ґрунтуючись схожості елементів, знаходить підмножини (кластери), в які групуються вхідні дані. У широкому колі завдань знайшли своє застосування методи нечіткої кластеризації, в яких елементи вхідної безлічі відносять до того чи іншого кластеру на підставі значення функції належності. Нечітка кластеризація одна з найбільш опрацьованих методик інтелектуального аналізу даних. Однак, традиційні методи нечіткої кластеризації не дають прийнятних рішень на даних зі складною внутрішньою структурою. Це пов'язано з низкою припущень, які закладаються в ці методи: кластери мають задану форму і особливу внутрішню точку – центр кластера; розбиття визначається, виходячи з взаємозв'язків між даними і центрами кластерів.

Завдання кластеризації має різні способи вирішення. Складність полягає у відсутності на момент початку аналізу будь-якої додаткової інформації про дані. У зв'язку з цим можливе безліч рішень по потужності можна порівняти з вхідним безліччю, що на практиці неприйнятно. Для якісного і швидкого вирішення завдання кластеризації необхідні методики вибору найкращих рішень.

В ході аналізу було розглянуто засоби обробки текстової інформації, за допомогою яких інформацію створюють, редагують, досліджують та аналізують. Для аналізу інформації існують декілька методів. На сьогоднішній день основним методом аналітичної обробки текстових масивів даних залишається пошук документів за ключовими словами. Були визначені дві проблеми роботи з величезною кількістю інформації, автоматичний збір інформації, та автоматичний розбір інформації, що надійшла з даної тематики, проведений на основі аналізу тексту документа, та їх вирішення.

В результаті аналізу було обрано кластеризацію як найбільш пристосований метод для обробки текстової інформації.

Для зниження впливу недоліків алгоритму *k – means* на кінцевий результат рекомендується додатковий аналіз слів. Обмеженням є те, що тексти повинні бути написані однією мовою і без помилок в словах. З додатковим використанням словника і системи перевірки на помилки, можна зробити автоматичну кластеризацію неструктурованого набору текстової інформації. Доопрацювання методу з орієнтуванням на особливості листів дозволить визначати схожість текстів по автору. Подібний метод може бути використаний при автоматичній класифікації текстів електронних бібліотек по темі і автору текстів, знайдених пошуковими системами, або в будь якій іншій сфері при використанні відповідної бази даних.

Список літератури:

1. Мандель И. Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
2. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности – М.: Финансы и статистика, 1989. – 450 с.
3. Diday, E., Simon, J. C. Clustering analysis // In: Digital Pattern Recognition, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ. P. 47–94.
4. Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений – Новосибирск: Изд-во Ин-та математики, 1999. – 212 с.
5. Michalski R., Stepp R., Diday E. Automated construction of classifications: conceptual clustering versus numerical taxonomy // IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 1983. V.5. P. 396–409.
6. Scholkopf B., Smola A., Muller K. Kernel Principal Component Analysis, Advances in Kernel Methods-Support Vector Learning, 1999.
7. Goldberg D. E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc. , Boston, MA, USA, 1989.
8. Kohonen T. Self-Organization and Associative Memory – 3rd ed. Springer information sciences series. Springer-Verlag, New York, NY. 1989.
9. K. Krishna, M. Murty. Genetic K-means algorithm // IEEE Transaction on System, Man and Cybernetics- Part B, 1999. V.29. P. 433-439.
10. Lu Y., Li S., Fotouhi F., Deng Y., Brown S. Incremental genetic k-means algorithm and its application in gene expression data analysis // BMC Bioinformatics, 2004.
11. Fern, X.Z., Brodley, C.E. Clustering ensembles for high dimensional data clustering // In Proc. International Conference on Machine Learning, 2003. P.186-193.
12. Хайдуков Д. С. Применение кластерного анализа в государственном управлении// Философия математики: актуальные проблемы. – М.: МАКС Пресс, 2009. – 287 с.
13. Hartigan, J. A. and Wong, M. A. (1978). Algorithm as 136: A k-means clustering algorithm. Applied Statistics 28, 100.

14. Олдендерфер М. С., Блэшфилд Р. К. Кластерный анализ / Факторный, дискриминантный и кластерный анализ: пер. с англ.; Под. ред. И. С. Енюкова. – М.: Финансы и статистика, 1989. – 215 с.

Деркач А.И.
Национальный авиационный университет

АНАЛИТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Аннотация
В статье были рассмотрены методы обработки текстовой информации, с помощью которых информацию создают, редактируют, исследуют и анализируют. Для анализа информации существует несколько методов. На сегодняшний день основным методом аналитической обработки текстовых массивов данных остается поиск документов по ключевым словам. Были определены две проблемы работы с огромным количеством информации, автоматический сбор информации, и автоматический разбор поступившей по данной тематике, проведенное на основе анализа текста документа, и их решение. В результате анализа были выбраны кластеризацию как наиболее приспособлен метод для исследования средств обработки текстовой информации; осуществлен сравнительный анализ методов интеллектуального анализа Data Mining; осуществлен обзор программного обеспечения, функционирующего на основе кластеризации; изучено метод k – средних.
Ключевые слова: кластеризация, аналитическая обработка, k – means, Data Mining, текстовая информация.

Derkach O.I.
National Aviation University

ANALYTICAL PROCESSING OF TEXT USING THE METHODS OF CLUSTERING

Summary
In this paper, we overview examined means of processing text information in which information create, edit, explore and analyze. To analyze the data, there are several methods. To date, the main method of analytical processing text data sets is search for documents by keyword. There were two problems by working with a lot of information, automatic data collection, and automatic analysis of information received on this subject, conducted by analyzing the document, and resolve them. The analysis was chosen as the clustering method adapted for research tools for processing text information; The comparative analysis of methods of mining Data Mining; conducted survey software that operates on the basis of clustering; studied method k – means.
Keywords: clustering, analytical processing, k – means, Data Mining, text information.