УДК 004.832

О.І. СУПРИГАН, І.С. МОРГУН

ВИБІР ВІДЕО-ФАЙЛІВ ЗА ІНДИВІДУАЛЬНИМИ ДАНИМИ КОРИСТУВАЧА

Вінницький національний технічний університет,
21021, м. Вінниця, вул.Хмельницьке шоссе, 95, Україна,
тел. (0432) 560-848

Анотація. Розглянуто метод класифікації та кластеризації, а також проведено порівняльний аналіз метрик відстані між об'єктами для предметної області підбору відео-файлів за індивідуальними даними користувача. Запропоновано новий підхід до розв'язання поставленої задачі.

Ключові слова: файл, відео-файл, кластеризація, класифікація, користувач, ознаки, агломеративний метод, дивізимний метод, процес паралельної обробки, метрика Евкліда.

Аннотация. Рассмотрен метод классификации и кластеризации, а также проведен сравнительный анализ метрик расстояния между объектами для предметной области подбора видео-файлов за индивидуальными данными пользователя. Предложен новый подход к решению поставленной задачи.

Ключевые слова: файл, видео-файл, кластеризация, классификация, пользователь, признаки, агломеративний метод, дивизимний метод, процесс параллельной обработки, метрика Евклида.

Abstract. The method of classification and clustering, and comparative analysis of metrics of distance between objects for the subject field of selection of video files on individual user data. A new approach to solving this problem.

Keywords: file, video file, clustering, classification, user features, ahlomeratyvnyy method dyvizymnyy method, the process of parallel processing, the Euclidean metric.

ВСТУП

Одним із головних процесів інтелектуальної діяльності людини є процес прийняття рішень. У вузькому розумінні прийняття рішення - це процес вибору кращого рішення з чисельних альтернатив. Однак процес прийняття рішень складається не тільки з вибору кращого варіанту, а й вміщає такі етапи:

1. Діагноз проблеми: виявлення та опис проблемної ситуації, встановлення мети вирішення проблемної ситуації, ідентифікація критеріїв прийняття рішення;
2. Накопичення інформації про проблему;
3. Розробка альтернативних варіантів - розробку, опис та складання переліку усіх можливих варіантів дій, що забезпечують вирішення проблемної ситуації;
4. Оцінка альтернативних варіантів;
5. Прийняття рішення.

АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Розглянемо, актуальну на даний час, задачу вибору відповідних відео-файлів.

Для вибору відповідного відео-файла користувач керується певними критеріями. Кількість критеріїв пошуку є необмеженою, і якщо вводити їх кожного разу – це займе певну кількість часу. Отже, для швидкого вибору та економії часу потрібно мати оптимізовану систему каталогізованих відео-файлів. Розглянемо існуючі способи каталогізації та пошуку відео-файлів.

Домашні колекції відео-файлів характеризуються низкою недоліків:

1. Відшукати потрібний фільм досить складно;
2. Займає надто багато місця.

Більшість кіноманів користуються онлайн-базами відео-файлів, розташованими в мережі Internet. До таких відносять: www.videoguide.ru, www.film.ru, www.kinopoisk.ru і www.rmdb.ru. Тут можна подивитися відео-фільми в онлайн-режимі, а також завантажити їх. Це займає досить велику кількість часу, тому що потрібно самостійно вводити параметри пошуку відео-фільмів, вибирати підходящі з запропонованих. Отже, можна покращити можливості існуючих систем, для чого потрібно ввести функцію самостійного підбору системою відео-файлів для кожного користувача індивідуально. Система,

що забезпечує підбір відео-файлів кожному користувачу індивідуально, повинна виконувати такі фунуції: опрцювання індивідуальних даних про користувача та формування критеріїв підбору відповідних файлів, по закінченню аналізу виводити запропоновані відеофайли відповідно критеріям пошуку, можливість зберігання інформації, наданої користувачу. Також система не вимагає присутності користувача під час підбору файлів.

ОБГРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ ЗАДАЧІ

Для вирішення задачі підбору відповідних файлів користувачу існує багато методів, з яких потрібно вибрати оптимальний. Насамперед це має бути метод, який працює шляхом розбиття масива даних на класи. Але основним недоліком процесу класифікації є те, що об'єкти при цьому розподіляються по певним класам за вже відомими, наперед визначеними ознаками, що є не відповідним для даної задачі, оскільки введені параметри можуть не відповідати заздалегідь сформованим класам. Більш відповідним являється кластерний аналіз, який являє собою сукупність методів класифікації багатовимірних спостережень чи об'єктів, які базуються на визначенні поняття віддалі між досліджуваними об'єктами з наступним виділенням в них подібних груп. При цьому не вимагається апріорної інформації про розподіл генеральної сукупності [1].

Кластер – це група, клас однорідних одиниць сукупності. Основне завдання кластерного аналізу – формування таких груп у багатовимірному просторі. Однорідність сукупності задається правилом обчислення певної метрики, що характеризує ступінь подібності (схожості) j-ї та k-ї одиниць сукупності. Такою метрикою може бути відстань між ними С або коефіцієнт подібності Rjk. Близькі, схожі за вибраними метриками одиниці вважаються належними до одного типу, однорідними. Вибір метрики є вузловим моментом кластерного аналізу, від якого залежить кінцевий варіант поділу сукупності на класи [2]. Основною метою кластерного аналізу є розділення багатовимірної сукупності вхідних даних на однорідні групи так, щоб об`єкти всередині групи були подібними між собою згідно з деяким критерієм, а об'єкти із різних груп відрізнялися один від одного. Причому класифікація об'єктів проводиться одночасно за декількома ознаками на основі введення певної міри сумарної близькості за всіма ознаками класифікації [3-4].

Відстань між двома об'єктами позначається як $d(xi, yi)$ – це не негативна функція близькості, задається при наступних умовах [5]:

1) Вона завжди більше або дорівнює нулю.
2) Відстань від точки X до точки Y така сама, як і від Y до X.
3) Якщо числові значення факторів двох об'єктів однакові, відстань між ними дорівнює нулю.
4) Нехай існує третя точка U. Тоді сума відстаней між точками XU та YU завжди більша ніж відстань поміж точками XY.

У вигляді формули це записується так:

$$\left.\begin{cases} d(x_i, y_i) \geq 0 \\ d(x_i, y_i) = d(y_i, x_i) \\ d(x_i, y_i) = 0 \Leftrightarrow x_i = y_i \\ d(x_i, y_i) \leq d(x_i, u_i) + d(u_i, y_i) \end{cases}\right\} \forall \{i\} \in N$$

Характеристику метрик відстаней наведено у таблиці 1.

Таблиця 1.

Характеристика метрик відстаней

Назва	Недоліки	Переваги
Метрика Евкліда $\mathrm{d}_{\mathrm{e}}(x_i, y_i) = \sqrt{\sum_{i=1}^{Nf} (x_i - y_i)^2}$	Не враховуэ знакові розходження	1.Пропорційно збільшує відстань між об'єктами у випадку різних абсолютних значень показників. 2. Збільшується розмірність кластерного поля, об'єкти штучно віддаляються друг від друга. 3. Границі між кластерами стають більш чіткими і точними.
Метрика Хемінга $\mathrm{d}_{\text{хем}}(x_i, y_i) = \sum_{i=1}^{Nf} (x_i - y_i)$	Утрачаються важливі знакові характеристики розходжень.	1. Використовується, коли знакові розходження характеристик об'єктів мають принципове значення. 2. За рахунок нівелювання знакових розходжень показників об'єкти сконцентруються навколо області ядра кластера.

(продовження таблиці 1)

Метрика L-норма $d_{хем}(x_i, y_i) = \sum_{i=1}^{Nf} \lvert x_i - y_i \rvert$	Не враховуються знакові розходження.	1. Збільшується розмірність кластерного поля. 2. Об'єкти штучно віддаляються друг від друга 3. Границі між кластерами стають більш чіткими і точними.
Метрика Чєбишева $d_{sup}(x_i, y_i) = SUP \lvert x_i - y_i \rvert$	Неправомірно змінюэ картину класифікації через зневагу усіма факторами крім одного.	З усіх різниць значень факторів, взятих по модулю, обирається одна – найбільша, саме вона буде характеристикою відстані між об'єктами. Отже, чітко формується однакова відстань між об'єктами.

Загалом, кількість мір відстаней буде становити: $M = \frac{N_0(N_0 - 1)}{2}$, де *N0* – кількість об'єктів, відстань між якими ми розраховуємо.

Методи кластерного аналізу можна розділити на дві групи: ієрархічні та неієрархічні.

Неієрархічні методи виявляють більш високу стійкість по відношенню до шумів і викидів, некоректного вибору метрики, включенню незначущих змінних в набір, який бере участь в кластеризації. З їх допомогою можливо опрацьовувати великі бази даних. Основним недоліком неієрархічних методів є те, що на вхід потрібно задавати кількість кластерів або кількість ітерацій.

Для предметної області підбору відео-файлів користувачу немає припущень щодо числа кластерів або кількості ітерацій, тому доцільним буде використання ієрархічного метода, так як ієрархічні методи, на відміну від неієрархічних, будують повне дерево вкладених кластерів.

В свою чергу ієрархічні методи кластерного аналізу поділяються на агломеративні та дивізивні (Рис. 1).

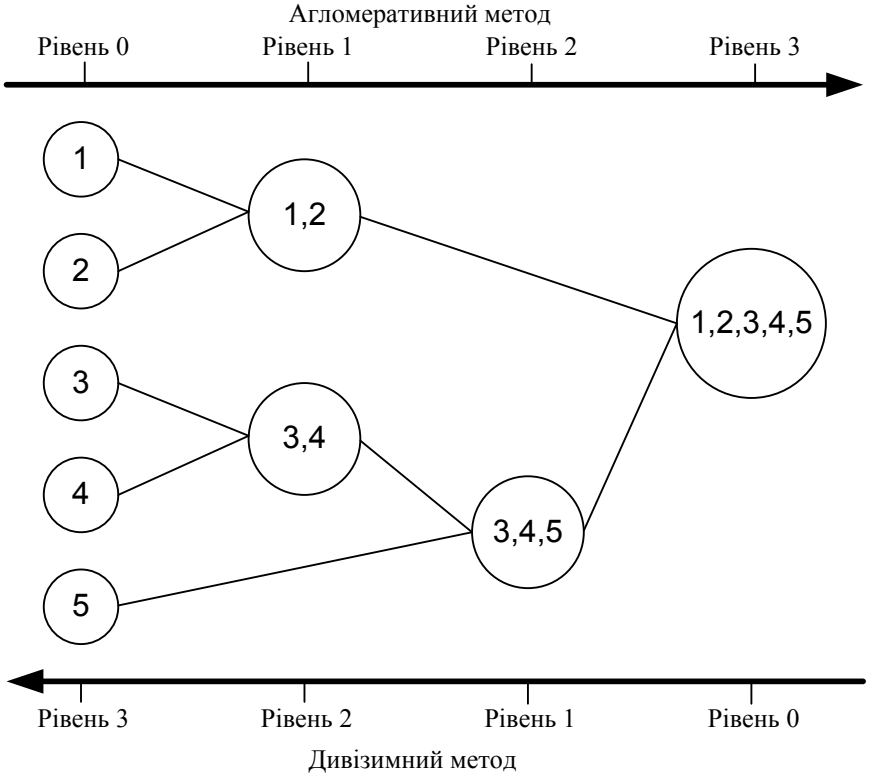

Рис. 1. Ієрархічні методи кластерного аналізу.

Розглянемо їх для правильного вибору метода розв'язання задачі підбору відео-файлів користувачу.

Ієрархічні агломеративні методи (Agglomerative Nesting, AGNES) Ця група методів характеризується послідовним об'єднанням початкових елементів і відповідним зменшенням числа кластерів. На початку роботи алгоритму всі об'єкти є окремими кластерами. На першому кроці найбільш схожі об'єкти об'єднуються в кластер. На наступних кроках об'єднання

продовжується до тих пір, поки всі об'єкти не будуть складати один кластер.

Ієрархічні дивізимні (ділені) методи (DIvisive ANAlysis, DIANA) Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об'єкти належать одному кластеру, який на подальші кроки ділиться на менші кластери, в результаті утворюється послідовність розщіплених груп.

Отже, оптимальним вирішенням поставленої задачі є агломеративний метод, так як відповідні класи заздалегідь не сформовані, а будуть створені за допомогою сформованих критеріїв шляхом об’єднання початкових кластерів. На кожному кроці об’єднання кластерів буде здійснюватись суто за сформованими ознаками підбору.

Для створення відповідного класу файлів потрібно визначити певний набір ознак. При виборі відео-файлів головними ознаками, якими керується користувач є: жанр, режисер, головні актори, країна випуску, рік випуску, наявність спец-ефектів та ін. Класифікацію об'єктів потрібно проводити не поступово за кожним критерієм, що може привести до втрати важливої частини файлів, а одночасно за декількома ознаками, що фактично є процесом паралельної обробки і дає можливість більш точного формування класу файлів для вирішення поставленої задачі [6]. Метод виділення великої кількості ознак схематично зображено на рисунку 2.

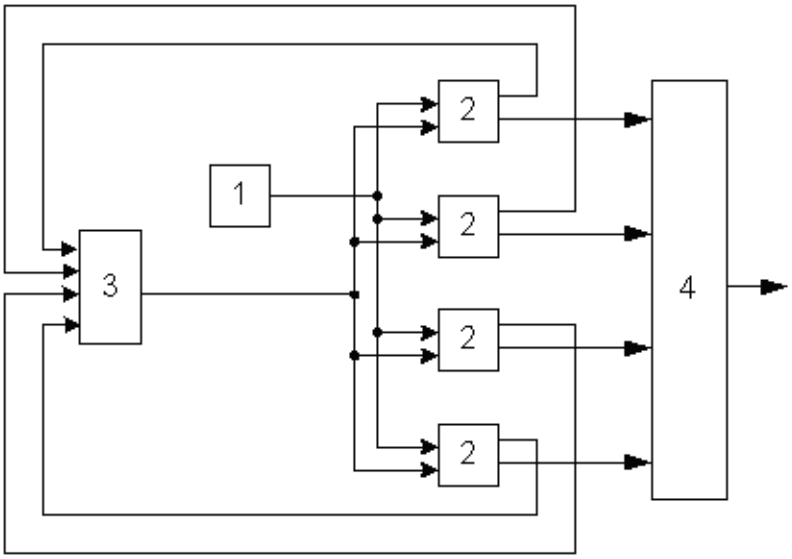

Рис. 2. Виділення ознак шляхом паралельної обробки: 1- вхід інформації, 2- блок віднімання, 3- блок порівняння, 4- блок додавання.

Під час обробки вхідної інформації (для даної предметної області – індивідуальні дані користувача) відбувається інтегрування ознак. Отримавши набір ознак файлів, потрібно порівняти їх із ознаками файлів існуючої бази даних. Для цього використовується розрахунок матриці відстаней між об’єктами. Найбільш доцільним буде використання метрики Евкліда, що дасть можливість точніше та ефективніше класифікувати відео-файли шляхом збільшення відстані між об’єктами, в результаті чого збільшується розмірність кластерного поля. При розрахунку метрики Евкліда відбувається порівняння сусідніх відео-файлів x та y за сформованими ознаками, в результаті чого відбувається зсув першого із них (x) вліво та порівняння наступної пари, де об’єкт y займає попереднє місце об’єкта x:

$$\mathrm{d_e}(x_i, y_i) = \sqrt{\sum_{i=1}^{Nf} (x_i - y_i)^2}$$

Об’єднання кластерів відбувається до того моменту, поки не буде сформовано клас файлів, який буде відповідати сформованим ознакам. Саме цей клас і буде оптимальною альтернативою на підставі ознак, ідентифікованих на етапі діагнозу проблеми.

Слід відмітити, що використання паралельної обробки початкової інформації дозволить значно пришвидшити процес прийняття рішення. Це випливає з факту отримання загальних ознак за всіма початковими даними одночасно, а також дозволяє виконувати кластерізацію не чекаючи закінчення процесу формування загальних ознак. Тобто визначення результуючого кластерного поля буде відбуватись майже одночасно з формуванням системи ознак.

ВИСНОВОК

Використання об’єднання таких методів дасть оптимальний розв’язок поставленої задачі тому, що метод паралельної обробки дає змогу чітко виділити ознаки для підбору файлів, а використання метрики Евкліда спрощує розбиття на кластери шляхом збільшення розмірності кластерного поля. Таким чином, обробка введених індивідуальних даних користувача надасть, відповідний сформованим ознакам, клас файлів.

СПИСОК ЛІТЕРАТУРИ

1. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков. – М.:Финансы и статистика, 1989. – 607 с.
2. Дюран Б. Кластерный анализ: Пер. с англ / Б. Дюран. – М.: Статистика, 1977. – 128 с.
3. Ивахненко А.Г. Алгоритмы метода группового учета аргументов при непрерывных и бинарных признаках: [Препринт] / А.Г. Ивахненко. – К.: Ин-т кибернетики им. В.М.Глушкова, 1992. – 49 с.
4. Гитис П.Х. Статистическая классификация и кластерний анализ / П.Х. Гитис. – М.: Московский государственный горный университет, 2003. – 157 с.
5. Пістунов І.М., Кластерний аналіз в економіці / І.М. Пістунов, О.П. Антонюк, І.Ю. Турчанінова.– Д: НГУ, 2008. – 87с.
6. Суприган В.А. Схемотехнічні засоби побудови оптоелектронних інтегральних схем обробки зображень: автореф. дис.к-та техн. наук 05.13.05 // Вінницький держ. техн. ун-т. – Вінниця, 2000. – 19с.

Надійшла до редакції 15.06.2010р.

СУПРИГАН ОЛЕНА ІВАНІВНА – к.т.н., доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, Вінниця, Україна

МОРГУН ІВАН СЕРГІЙОВИЧ – студент 4 курсу кафедри комп'ютерних наук, Вінницький національний технічний університет, Вінниця, Україна