ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

УДК 681.518.03
О. М. Роїк, д. т. н., проф.;
Ю. Я. Табачник асп.

МЕТОД ІДЕНТИФІКАЦІЇ ОБ'ЄКТІВ ДЛЯ ПРОВЕДЕННЯ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ БАГАТОВИМІРНИХ ПРОСТОРОВИХ ДАНИХ

Виконано формалізацію задачі ідентифікації просторових об'єктів у термінах теорії розпізнавання образів. Наведено базовий метод ідентифікації об'єктів та проаналізовані його недоліки. Запропоновано модифікований метод ідентифікації, який у значній мірі позбавлений недоліків базового та дозволяє збільшити відсоток об'єктів, що ідентифікуються автоматично.

Постановка задачі

Коли розв'язуються задачі класифікації операційно-територіальних одиниць (ОТО) географічні дані (векторна карта) зазвичай вже наявні, а ознаки, за якими буде проводитися багатовимірна класифікація, надаються замовником у вигляді електронних таблиць вигляду ОТО-ОТО чи ОТО-ознака в одному з розповсюджених форматів (txt, dbf, html тощо). Будемо називати таку таблицю «зовнішньою», підкреслюючи той факт, що вона не є атрибутивною таблицею векторної карти (вбудовані атрибутивні таблиці містять лише мінімальні відомості про операційно-територіальні одиниці). Зовнішня таблиця також описує ОТО у атрибутивному просторі ознак, проте інформація, що зберігається у зовнішній таблиці, є абстрагованою від географічних деталей конкретного елемента карти.

Для виконання автоматичної класифікації багатовимірних даних у геоінформаційних системах необхідно виконати зв'язок між зовнішньою таблицею (чи таблицями) та картою, щоб однозначно зіставити кожному географічному об'єкту відповідну точку у багатовимірному просторі ознак. Фізично це означає знаходження зв'язку між атрибутивною та зовнішньою таблицями.

У сучасних системах управління базами даних така задача вирішується за допомогою зовнішніх ключів [1]. У цьому випадку зв'язок між двома таблицями встановлюється за однаковими унікальними значеннями одного чи декількох полів запису, які зазвичай мають числовий тип.

Якщо кожна таблиця, що містить атрибутивні дані, має відповідний зовнішній ключ, то проблеми ідентифікації об'єктів не виникає. Проте зазвичай зовнішні таблиці отримуються з декількох джерел (статистичні управління, служби екологічного моніторингу тощо) та містять лише текстові поля з назвами операційно-територіальних одиниць.

Теоретично можна використовувати назви ОТО як зовнішні ключі, проте практика доводить, що назви одних і тих самих географічних назв можуть істотно відрізнятися. Ця різниця найбільш помітна у скороченнях типу «обл», «обл.» чи «область», «р.», «р-н» чи «район» тощо. Крім цього, можливі синонімічні назви для однієї і тієї ж операційно-територіальної одиниці, що виникли внаслідок перейменувань. Наявність російськомовних даних також утруднює ідентифікацію об'єктів.

На даний момент найбільш розповсюдженим методом ідентифікації об'єктів є застосування готових бібліотек морфологічного аналізу [2]. Такі бібліотеки використовуються, наприклад, для розширення запиту до бази даних під час пошуку заданого слова (за допомогою формування масиву різних варіантів його написання). Проте бібліотеки морфологічного аналізу не розраховані на роботу з нестандартними скороченнями та варіантами назв операційно-територіальних одиниць, а отже для значної кількості об'єктів необхідно проводити ідентифікацію вручну.

З вищенаведеного випливає необхідність розробити метод ідентифікації об'єктів, який дозволив би з мінімальним втручанням експерта проводити ідентифікацію операційно-територіальних одиниць з урахуванням нестандартних скорочень, схожих та синонімічних назв.

Формалізація задачі ідентифікації

Формалізацію задачі ідентифікації просторових об'єктів виконаємо у термінах теорії розпізнавання образів.

Нехай $X = \{X_1, \dots X_{N_X}\}$ — множина картографічних назв операційно-територіальної одиниці (стовпчик найменувань ОТО з атрибутивної таблиці, що належить векторній карті). Назвемо X множиною еталонів, N_x — кількістю еталонів.

Нехай $Y = \{Y_1, \dots, Y_{N_Y}\}$ — множина значень стовпця назв ОТО із зовнішньої таблиці. Назвемо Y множиною образів, що подаються на вхід процедури ідентифікації, N_y — кількістю образів.

Зазвичай не завжди існує можливість зібрати статистичну інформацію з проблеми, що аналізується у розрізі всіх операційно-територіальних одиниць, тому $N_x \geq N_y$ (тобто еталонів є не менше, ніж образів).

Процедура ідентифікації полягає у побудові відображення (чи функції) ідентифікації $f : Y \to X$. При цьому запис $f(y_i) = x_j,\ i \in \{1, \dots, N_y\},\ j \in \{1, \dots, N_x\}$ значить, що образ y_i відповідає еталону x_j, тобто назви y_i та x_j відносяться до однієї фактичної ОТО.

Обмеженням, що накладається на відображення f та визначає специфіку процедури ідентифікації, є обов'язкова умова його ін'єктивності, тобто

$$\forall y_i, y_j, y_i \neq y_j \Rightarrow x_a = f(y_i) \neq f(y_j) = X_b,\ x_a, x_b \in X.$$

Відображення $f : Y \to f(Y) \subset X$ є бієктивним (взаємно однозначним). Множині $X \setminus f(Y)$ відповідають ті еталони (картографічні об'єкти), для яких відсутні образи (записи у зовнішній таблиці).

Одним із стандартних підходів до розв'язання задачі розпізнавання є такий алгоритм [3]:

— формування множини ознак $V = \{v_1, \dots v_N\}$ для розпізнання образів $y_i \in Y$ та еталонів $x_j \in X$ (занурення образів та еталонів у атрибутивний простір V);

— вибір методу розрахунку відстані між образами та еталонами у побудованому просторі ознак (задання метрики у просторі V): $d : Y \times X \to R$, де R — множина дійсних чисел;

— віднесення кожного образу $y_i \in Y$ до того еталону, на якому досягається мінімальна відстань до цього образу, тобто

$$f(y_i) = x_a \Leftrightarrow d(y_i, x_a) = \min\left(d(y_i, x_j) \mid x_j \in X\right). \quad (1)$$

За наявності векторного шару карти можна вважати, що множина еталонів X вже задана.

Базовий метод ідентифікації об'єктів

Для збільшення відсотка об'єктів, які вдасться ідентифікувати автоматично, вводиться додатковий крок переходу за допомогою експерта до нової множини еталонів X'. Під новими еталонами розуміються скорочені унікальні словоформи, які обов'язково є підрядками образів. Таким чином для карти України скороченими словоформами можуть бути «Київ», «Вінниця», «Чернігівська» замість елементів множини образів «м. Київ», «місто Вінниця», «Чернігівська область» відповідно.

Подальшим етапом є формування простору ознак $V = \{v_1, \dots v_{NX'}\}$. Ознаки простору V розглядаються бінарними, тобто такими, що приймають лише булеві значення:

$$v_i \in \{0,1\} \forall i \in \{1, \dots, N_{X'}\}.$$

Значення будь-якої ознаки v_i з простору ознак V для кожного образу $y_j \in Y$ задається виразом

$$v_i\left(y_j\right) = 0, \text{ якщо рядок } y_j \text{ містить рядок } x_i^{'};\ \ v_i\left(y_j\right) = 1 \text{ — у протилежному випадку.} \quad (2)$$

Визначимо значення будь-якої ознаки v_j з простору V для кожного еталону $x_j \in X'$

$$v_i\left(x_j\right) = \begin{cases} 0, i = j; \\ 1, i \neq j. \end{cases}$$

Таким чином простір ознак сформовано та вказані методи отримання значень ознак як на еталонах, так і на образах. Як метрики простору V, тобто спосіб визначення відстаней між еталонами і образами, використаємо метрику Хеммінга [4], оскільки усі ознаки є бінарними $d\left(y_i, x_j^{'}\right) = \sum_{N=1}^{N_{X'}} \left| v_n\left(y_i\right) - v_n\left(x_j^{'}\right) \right|$, що відповідає підрахунку кількості незбіжностей значень ознак для кожного еталону та образу.

Відображення (1) задається так: $f\left(y_i\right) = x_j \Leftrightarrow d\left(y_i, x_j^{'}\right) = 0$.

Базова модель ідентифікації об'єктів має такі недоліки:

1. Необхідність залучення експерта для формування еталонів X' шару електронної карти. Ця умова не завжди є досить суттєвою, оскільки у якості експертом може виступити практично будь-який користувач (висока кваліфікація експерта не потрібна), а кількість різних електронних карт звичайно є незначною. Дуже розповсюдженою є ситуація, коли використовується одна й та сама векторна основа (або дуже обмежена їх кількість), наприклад країни світу чи регіони України. Проте для значної кількості операційно-територіальних одиниць (наприклад $N_x \geq 500$) базова модель стає практично непридатною, оскільки обробити вручну таку кількість назв для отримання унікальних словоформ досить важко.

2. Можливість ситуацій, в яких відразу декілька ОТО мають схожі назви. Наприклад, x_1 = «Київ», x_2 = «Київська область». Логічно виділити унікальні словоформи x'_1 = «Київ» та x'_2 = «Київська», але у відповідності з функцією (2) перша словоформа буде автоматично віднесена як до першого, так і до другого випадків. В результаті для y = «Київська область» мінімум функції відстані досягається відразу на двох елементах множини X'

$$d(y, x_1^{'}) = d(y, x_2^{'}) = 1 \neq 0.$$

В результаті функція f не буде визначена у точці y.

3. Можливість ситуацій, у яких одна й та сама операційно-територіальна одиниця має декілька синонімічних назв (або назв на декількох мовах). Наприклад, якщо експерт вибере x' = «Харьков» для y = «Харків», то $d(y, x_1^{'}) = 1\ \forall x_i^{'} \in X'$.

Це також означає, що функція f не буде визначена у точці y.

Вдосконалений метод ідентифікації об'єктів

Запропоновані вдосконалення до базового методу автоматичної ідентифікації об'єктів допоможуть частково позбутися першого та повністю другого та третього недоліків базового методу.

Перша відмінність від базової моделі ідентифікації полягає у формуванні розширеної множини еталонів X'':

$$X'' = \left\{x_1^{''}, x_2^{''}, \ldots, x_{Nx}^{''}\right\} = \left\{\left\{x_{11}^{'}, \ldots, x_{1N1}^{'}\right\}, \left\{x_{21}^{'}, \ldots, x_{2N2}^{'}\right\}, \ldots, \left\{x_{Nx1}^{'}, \ldots, x_{NxN_{Nx}}^{'}\right\}\right\},$$

де $x_i^{'} = \left\{x_{i1}^{'}, \ldots x_{iNi}^{'}\right\}$ — множина можливих варіантів написання i-ї ОТО; N_i — кількість варіантів написання i-ї ОТО; $N = \sum_{i=1}^{NX} N_i$ — загальна кількість варіантів написання всіх операційно-територіальних одиниць.

За аналогією з базовою моделлю сформуємо простір ознак $V = \left\{v_1, \ldots v_{Nx}\right\}$.

Для кожного $y_j \in Y$

$v_i\left(y_j\right) = 0$, якщо $\exists t = \{1, \ldots, N_i\}$ рядок y_j містить рядок x'_{it};

$v_i\left(y_j\right) = 1$ — у протилежному випадку $\forall x''_j \in X''$ $v_i(x''_j) = \begin{cases} 0, i = j; \\ 1, i \neq j. \end{cases}$

Формула для визначення відстані d залишається незмінною

$$d\left(y_i, x'_j\right) = \sum_{N=1}^{N_{X'}} \left| v_n\left(y_i\right) - v_n\left(x'_j\right) \right|.$$

Функцію ідентифікації f пропонується задавати ітераційним шляхом за допомогою такого алгоритму:

1. Покласти $N = \left|N_y\right|$;
2. $\forall y_i \in Y$ знайти $c_i = \min\left\{d\left(y_i, x''_i\right) \mid x''_i \in X''\right\}$ та сформувати множини $X''_i = \left\{x''_j \in X'' \mid d(y_i, x''_j) = c_i\right\}$.
3. $\forall y_j \in Y$: якщо $\left|X''_i\right| = 1$, то покласти $f(y_i) = x''_j \in X''_i$; $X'' = X'' \setminus X''_i$; $Y = Y \setminus \{y_i\}$.
4. Якщо $\left|N_y\right| = N$, то кінець, інакше перейти до кроку 1.

Розглянута модель частково вирішує першу і повністю третю проблеми базової моделі ідентифікації об'єктів за рахунок можливості задання декількох синонімів одного й того ж слова. Це дозволить експерту менше концентрувати власну увагу на виділенні унікальних словоформ, а також використовувати всі можливі назви операційно-територіальних одиниць.

Друга проблема базової моделі вирішується за рахунок ітераційного задання функції f. Наприклад, для розглянутого вище прикладу для образу $y_1 =$ «Київ» буде підібрано відповідний йому еталон $x'_1 =$ «Київ», а образ $y_2 =$ «Київська область» ідентифікований не буде, оскільки на першій ітерації $c_2 = d\left(y_2, x'_1\right) = d\left(y_2, x'_2\right) = 1$ та $\left|X''_2\right| = 2$. Проте на другій ітерації $x'_1 \notin X''$, тому $\left|X''_2\right| = 1$, отже для образу y_2 буде знайдено відповідний йому еталон $x'_2 =$ «Київська».

Висновки

1. Виконано аналіз задачі ідентифікації операційно-територіальних одиниць. Проведена формалізація задачі автоматичної ідентифікації об'єктів у термінах теорії розпізнавання образів.

2. Розглянуто базовий метод ідентифікації ОТО та показано його основні недоліки. На основі базової моделі побудовано вдосконалений метод ідентифікації, який дозволяє коректно ідентифікувати операційно-територіальні одиниці зі схожими та синонімічними назвами.

СПИСОК ЛІТЕРАТУРИ

1. Дейт Л. Дж. Введение в системы баз данных. Пер. с англ. — К.: Диалектика, 1998.
2. Шаши Шекхар, Санжей Чаула. Основы пространственных баз данных. — М.: Кудиц — Образ, 2004. — 309 с.
3. Горелик А. Л., Скрипник В. А. Методы распознавания. — М.: Высш. Шк., 1989. — 232 с.
4. Айвазян С. А. и др. Прикладная статистика: Исследование зависимостей. — М: Финансы и статистика, 1985. — 587 с.

Рекомендована кафедрою інтелектуальних систем

Надійшла до редакції 28.10.04.
Рекомендована до друку 16.11.04.

Роїк Олександр Митрофанович — завідувач кафедри; ***Табачник Юрій Якович*** — аспірант.
Кафедра інформаційного менеджменту, Вінницький національний технічний університет