УДК 343.982.34 : 004.932.2 : 681.3.06

С.С. Танянский[1], О.В. Вахета[2]

[1]*Харьковский национальный университет радиоэлектроники*
[2]*Харьковский национальный университет внутренних дел*

ПОИСК ЧАСТИЧНО СОВПАДАЮЩИХ ДАННЫХ В ЗАДАЧАХ ИДЕНТИФИКАЦИИ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ

Предложен алгоритм формирования вектора признаков информационного объекта. Рассмотрен метод множественного сравнения последовательностей характеризующих признаков в задаче идентификации информационных объектов для повышения эффективности поиска в базах данных. Обоснованы направления дальнейших исследований в области поиска наиболее «близких» векторов признаков, сохраненных в базе данных.

задачи идентификации информационных объектов

Введение

Постановка проблемы. В последние время все больше организаций используют для хранения и обработки данных автоматизированные системы управления. Неотъемлемой частью любой такой системы является система управления базой данных (СУБД). Соответственно, обрабатываются большие массивы жестко структурированных данных. Однако часто возникает проблема поиска и выборки необходимой информации из неструктурированного массива, содержащего, например, графические, мультимедийные и другие данные.

В основе проблемы неточности описания одинаковых объектов лежат вопросы приведения неструктурированных объектов к единому структурному виду. Так как часто на практике один и тот же объект в силу объективных причин может быть поразному закодирован (оцифрован), то необходимо иметь возможность определять информационную близость кодированных представлений как можно точнее.

В настоящее время приобретает все большее прикладное значение мягкие вычисления (нечеткая логика) в связи с возможностью решения целого ряда практических задач, не имевших разрешения традиционными методами.

Таким образом, методы и средства поиска, частично совпадающих данных, рассмотренные в статье являются *актуальными*.

Анализ литературы. Нечеткие запросы к базам данных (БД) - перспективное направление в современных системах обработки информации. Большая часть исследований в области нечеткой реляционной алгебры принадлежит западноевропейским ученым Д. Дюбуа и Г. Праде [1]. Значительный вклад в развитие теории и практики нечетких систем с эволюционной адаптацией внесла группа испанских исследователей во главе с Ф. Херрера [2]. Нечеткие методы кластеризации, в отличие от четких методов (например, нейронные сети Кохонена), позволяют одному и тому же объекту принадлежать одновременно нескольким кластерам, но с различной степенью. Наиболее распространены: алгоритм нечеткой самоорганизации c-means и его обобщение в виде алгоритма Густафсона-Кесселя.

В работе анализируются графические объекты, такие как изображения отпечатков пальцев. Однако предложенный ниже метод может применяться с использованием различных информационных объектов, характеризующихся определенным набором (вектором) признаков.

В исследовании дактилоскопии большой вклад сделали: Г.Л. Грановский [3], который в своих научных работах отразил много важных вопросов дактилоскопии, например, оценку частных признаков папиллярных узоров, локализации следов папиллярных линий и т.д.; Л.Г. Эджубов и С.О. Литинський [4] разработали способ автоматизированного сравнения папиллярных узоров и устройство для его практического использования; Н.Е. Гнидин, В.Е. Корноухов [5] и др. улучшили некоторые методы обнаружения, фиксации и обработки следов рук.

Цель статьи. При решении задачи идентификации отпечатков пальцев приходится оперировать с геометрическими объектами определенной природы

с определенными свойствами. Такими объектами являются, в частности, элементы, размещенные на изображении отпечатка пальца.

Множественное сравнение последовательностей имеет актуальное и важное значение в задачах идентификации. В связи с этим, *целью статьи* является исследование метода поиска наиболее «похожего» вектора признаков информационного объекта в базе данных.

Так как при изъятии отпечатков пальцев некоторые признаки могут быть утеряны (в следствие избыточного или недостаточного количества типографской краски, объективных и субъективных причин оперативной обстановки), при идентификации изображений отпечатков пальцев эксперт может иметь неполный вектор признаков, характеризующий (описывающий) изображение отпечатка пальца, который будет сохранен в базе данных. Классические методы точного сопоставления информационных характеристик в данном случае не могут дать ожидаемой точности поисковых систем. Таким образом, необходимо перейти в область неточного (приближенного) сопоставления и выравнивания последовательностей при поиске информации в базе данных векторов признаков изображений отпечатков пальцев.

Формирование информационных признаков

При анализе изображения отпечатка пальца (рис. 1) получим определенный набор признаков:

- тип узора: завитковый;
- окончаний: 36;
- вилок: 4;
- крючков: 0;
- обрывков: 2;
- перерывов: 4;
- глазков: 0;
- мостиков: 1.

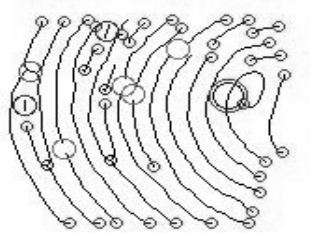

Рис. 1. Выделенные дактилоскопические признаки отпечатка пальца (в результате проведения процедур предобработки)

На данном этапе формируется вектор признаков, что позволит хранить в БД лишь необходимую для поиска и идентификации информацию, а не весь графический объект.

Преимущества такого подхода очевидны: сокращается время поиска объектов (так как нет необходимости повторной обработке графических объектов), уменьшаются требования к объему выделяемой памяти под один объект хранимой информации и т.д.

Множество всех признаков обозначим

$$P = \{p_1, \dots, p_{16}\},$$

множество количественных характеристик признаков обозначим

$$M = \{m_1, \dots, m_{15}\}.$$

Множество качественных характеристик определим следующим образом:

$$z_i = \{o_1,\dots,o_n\},$$

где i = 1, ... , 16; n – количество качественных характеристик каждого признака, определенных ниже.

Сформируем полный вектор признаков изображения отпечатка пальца, который будем использовать в дальнейшем для идентификации. Изображение T_k (при k = 1, … , *l*; где *l* – количество изображений) будет описывать вектор признаков:

$$S_k = \{P_k, M_k, z_1,\dots,z_{16}\}.$$

Полученный вектор будем хранить в БД для идентификации поступающих на вход автоматизированной дактилоскопической системе (АДИС) изображений отпечатков пальцев.

Метод сравнение частично совпадающих данных

Для формирования вывода о полном или частичном совпадении последовательности признаков, сохраненной в векторе признаков, вычислим редакционное расстояние между последовательностями (так как в ЭВМ любой символ может быть представлен в числовом значении, то можем оперировать термином «редакционное расстояние между строками»).

На практическом примере рассмотрим метод неточного сопоставления двух векторов признаков, содержащих определенный набор характеристик изображений отпечатков пальцев.

В задаче поиска наиболее «близкого» вектора признаков будем использовать понятия операций редактирования:

- вставка (insertion) значения в вектор признаков (обозначение: **i**, по аналогии со вставкой символа в строку);
- удаление (deletion) значения из вектора признаков (обозначение: **d**);
- замена (replace) значения в векторе признаков (обозначение: **r**);
- совпадение (match) значений анализируемых векторов признаков (обозначение: **m**).

Для нахождения редакционного расстояния будет использованы методы динамического программирования в терминах взвешенного редакционного графа [6].

Определение: Пусть заданы векторы S_1 и S_2 длиной n и m признаков соответственно. Взвешенный редакционный граф имеет $(n + 1) \times (m + 1)$ вершин, все они помечены непересекающимися парами (i,j) $(0 \le i \le n, 0 \le j \le m, i \ne j)$.

В задаче идентификации редакционный граф содержит дугу из каждой вершины (i, j) в каждую из вершин (i, j + 1), (i + 1, j) и (i + 1, j + 1), если эти вершины существуют. Веса первых двух дуг равны 1, вес третьей (диагональной) дуги равен t(i + 1, j + 1). Вес каждой дуги равен 1, кроме двух дуг нулевого веса, отмеченных на рис. 2 на пересечении P_1 и P^1_2, P_2 и P^1_3.

Основное свойство редакционного графа заключается в том, что любой кратчайший путь (полный вес которого минимален) от начальной вершины (0, 0) до вершины назначения (n, m) определяет редакционное предписание с минимальным числом операций. Эквивалентно, любой кратчайший путь определяет глобальное выравнивание минимального полного веса [7].

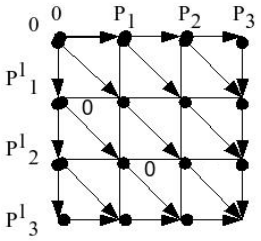

Рис. 2. Редакционный граф для последовательностей P_1, P_2, P_3 и P^1_1, P^1_2, P^1_3

Необходимо также отметить, что при решении задачи сопоставления будут использоваться произвольные веса, стоимости или балловые оценки (score), приписываемые каждой редакционной операции, в том числе и совпадению. Таким образом, любое включение или удаление имеет вес, обозначенный через **d**, замена имеет вес **r**, а совпадение – вес **e** (который обычно мал в сравнении с другими весами и часто равен нулю).

Рассмотрим общий случай вычисления операционно-взвешенного редакционного расстояния:

Пусть D(i,j) обозначает минимум полного веса редакционных операций, переводящих $S_1[1..i]$ в $S_2[1..j]$. Используем обозначение t(i,j), чтобы охватить им и замену и равенство, так что теперь t(i,j) равно **e** при $S_1(i) = S_2(j)$ и **r** – в противном случае. Теперь базовыми условиями будут:

$$D(i,0) = i \times d; \qquad D(j,0) = j \times d.$$

Общее рекуррентное соотношение выглядит так:

$$D(i, j) = \min\{D(i, j - 1) + d, D(i - 1, j) + d, D(i - 1, j - 1) + t(i, j)\}.$$

Задача об операционно-взвешенном редакционном расстоянии также может быть представлена и решена как задача о кратчайшем пути на взвешенном редакционном графе, в котором веса дуг естественным образом соответствуют весам редакционных операций.

Выводы

Описанный метод поиска и сопоставления векторов признаков информационных объектов, сохраненных в БД, позволит решить ряд задач, что определяет *практическую значимость* полученных результатов. Среди таких результатов можно выделить увеличение скорости процесс поиска информации в БД и сократить требования, выдвигаемые АДИС к ресурсам ЭВМ.

Научными результатами являются как методы формирования векторов признаков неструктурированных информационных объектов, так и метод нахождения “близких” векторов, что определяет точность поиска данных в БД. Это позволяет оперировать не только с «полными» изображениями графических объектов, но и с частичным отображением.

Для расширения применения предложенных методов, в дальнейшем необходимо провести исследования, которые позволят найти решение для задачи поиска наиболее «близких» векторов признаков, в которых могут отличаться не только количество имеющихся признаков, но и значения самих признаков. Например, при сопоставлении изображений отпечатков пальцев, полученных с определенного рода повреждениями, что послужило причиной изменения некоторых геометрических или топологических характеристик исследуемых дактилоскопических признаков.

Список литературы

1. Дюбуа Д. Теория возможностей: Приложения к представлению знаний в информатике. Пер. с фр. / Д. Дюбуа; А.Прад. – М.: Радио и связь, 1990. – 287 с.

2. Cordon O., Herrera F., Peregrin A. Applicability of the fuzzy operators in the design of fuzzy logic controllers. – Fuzzy Sets and Systems, 86, 1997, P. 15-41.

3. Грановский Г.Л. Классификация и оценка частных признаков папиллярных узоров. – М., 1956. – 340 с.

4. СТАТИСТИЧЕСКАЯ дактилоскопия: (Методологические проблемы) / Л.Г. Эджубов, В.З. Поляков, В.Н. Елисеев и др. – М.: Городец, 1999. – 183 с.

5. Корноухов В.Е., Орлов Ю.К., Журавлева И.А. Судебная экспертиза: Спец. курс. – Красноярск: Краснояр. гос. ун-т, 1998. – Ч. 1. – 332 с. – Ч. 2. – 334 с.

6. Гасфилд Дэн. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология. – С.-Пб.: БХВ-Петербург, 2003. – 654 с..

7. Оре О. Теория графов. – М.: Наука, 1980. – 336 с.

8. Baron J.L., Fleet D.J.,Beauchemin S.S. Performances of optical flow techniques. Int. Journal of Computer Vision, 12:1, 1994. – P. 43-77.

Поступила в редколлегию 30.03.2007

Рецензент: д-р техн. наук, проф. И.В. Гребенник, Харьковский национальный университет радиоэлектроники, Харьков.