УДК 004.04

АЛГОРИТМЫ И ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ПОИСКА ДАКТИЛОСКОПИЧЕСКИХ ОБРАЗОВ В ИНФОРМАЦИОННОЙ БАЗЕ ДАННЫХ

Лебедев К.Е., Привалов М.В.
Донецкий национальный технический университет

Введение

Идентификация человека на основе отпечатков пальцев – задача, далеко не новая и в настоящее время существует большое количество исследований и разработок по этой теме. Уже во многих странах системы с таким видом идентификации введены в эксплуатацию и успешно функционируют. Примерами таких систем могут служить автоматизированная дактилоскопическая идентификационная система по следам и отпечаткам пальцев и ладоней АДИС Папилон (Россия) [1], система учета рабочего времени и контроля доступа по отпечаткам пальцев BioTime (Россия) [2], Next Generation Identification (NGI) [3] разрабатываемая ФБР глобальная система идентификации.

В тоже время многие страны стремятся ввести, а некоторые уже ввели, обязательный сбор биометрических данных в таких областях, как авиация, железнодорожный транспорт, туризм, выдача виз и др. Активно вводятся биометрические паспорта. День за днем, растет объем биометрической информации, которую необходимо хранить в базах данных, размер которых также пропорционально возрастает, а в связи с этим задачи эффективного поиска биометрической информации в базах данных не теряют своей актуальности.

Постановка задачи

В любой подсистеме идентификации на основе отпечатков пальцев возникает задача хранения большого объема данных об отпечатках пальцев и их владельцах, и задача эффективного поиска в этих данных. Для решения этих задач наиболее эффективным является использование специализированной биометрической системы управления базами данных (СУБД). Но сложность и дороговизна разработки такой СУБД очень велика, поэтому, до сих пор таких разработок на всеобщий рынок выпущено не было.

Рассмотрим два подхода, для выхода из этого положения. Первый подход – использовать обычную реляционную СУБД для хранения данных, а для обработки данных разработать клиентское приложение. Этот способ является наименее эффективным, т.к. между сервером, на котором установлена СУБД, и клиентским приложением возникает большой поток данных, который приводит к лишним временным затратам. Второй подход исключает этот недостаток за счет обработки данных на сервере в самой СУБД, а клиентское приложение лишь посылает исходные данные и забирает готовый результат. Но второй подход требует от СУБД наличия определенных инструментальных возможностей (например, определение собственного типа данных и использование программной библиотеки, разработанной для эффективной работы с этим типом данных). Такие возможности есть далеко не во всех СУБД, а те, в которых они есть, в большинстве своем, являются дорогостоящими.

СУБД PostgreSQL [4] полностью подходит для второго подхода к решению задачи хранения и обработки данных об отпечатках пальцев и их владельцах, при этом является бесплатным программным продуктом с открытым исходным кодом.

Таким образом, в качестве инструментального средства для хранения и поиска отпечатков пальцев предлагается использовать СУБД PostgreSQL. Далее рассмотрим, что собой представляют данные об отпечатке пальца.

Данные об отпечатке пальца можно представить вектором характеристик особых точек, который имеет следующий вид $T = \{m_1, m_2, ..., m_N\}$ [5]. Здесь m_i - тройственный вектор вида $m_i = \{x, y, \theta\}$, который указывает координаты x и y расположения детали и угол детали θ.

Особые точки – это уникальные для каждого отпечатка признаки, определяющие пункты изменения структуры папиллярных линий, ориентацию папиллярных линий и координаты в этих пунктах. Каждый отпечаток содержит до 70 деталей.

Особые точки бывают следующих видов:

- конечная точка (ending point) – точка, где заканчивается линия гребня;
- точка ветвления (bifurcation) – точка расхождения линий гребня;
- центр (core) – точка наибольшей кривизны гребня;
- дельта (delta) – зона, где гребень разветвляется на три линии.

Гребень (ridge) – возвышающаяся линия отпечатка пальца, а бороздка (valley) – желобок между гребнями.

На рис. 1 представлены особые точки: конечная точка (круг), точка ветвления (квадрат), центр (треугольник), дельта (ромб).

Существует много методов извлечения векторов особых точек из изображения отпечатка пальца. В данной работе эти методы не рассматривались. Будем считать, что векторы особых

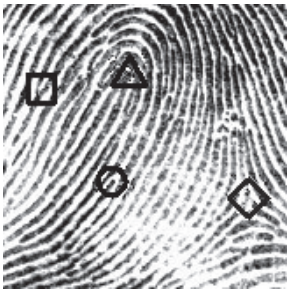

Рисунок 1 – Особые точки отпечатка пальцев

точек уже получены и занесены в таблицу базы данных в виде векторов вещественных чисел $T_i = \{x_1, y_1, \theta_1, x_2, y_2, \theta_2, ..., x_N, y_N, \theta_N\}$.

Также будем считать, что классы отпечатков пальцев тоже определены и наряду с векторами особых точек занесены в таблицу. Класс отпечатка пальца – это один из многих признаков, по которым отпечатки пальцев можно разделять на группы. Традиционной, стала система классификации Генри, согласно которой все отпечатки пальцев можно разделить на пять классов (рис. 2).

Для определения классов отпечатков пальцев также существует много методов, называемых классификаторами. Например, классификатор К-ближайший, классификатор нейронная сеть, двухэтапный классификатор, классификатор «отклонение выбора», классификатор скрытой модели Маркова, классификатор дерева решения.

Теперь, когда данные, по которым можно строить индекс, определены, рассмотрим типы индексов, встроенные в СУБД

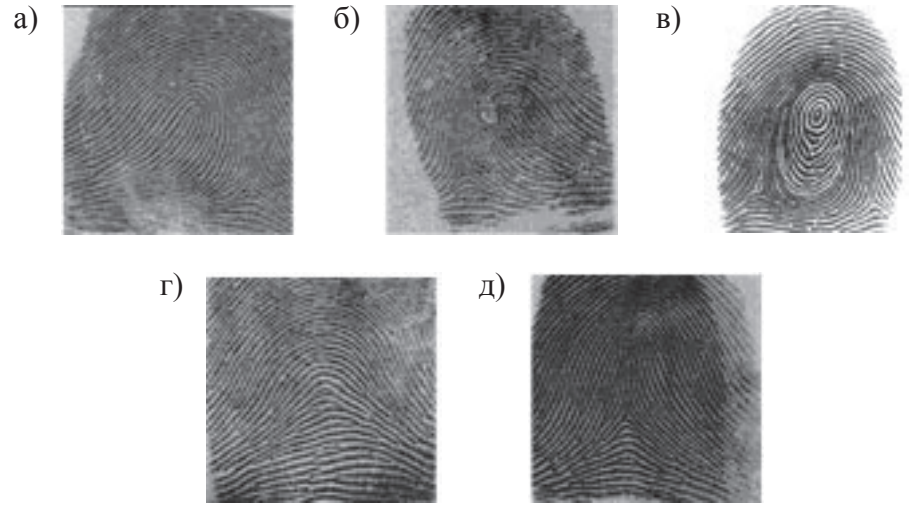

Рисунок 2 – Классы отпечатков пальцев:
а) левая петля (Left Loop); б) правая петля (Right Loop); в) завитушка (Whorl); г) дуга (Arch); д) полусфера (Tented Arch).

PostgreSQL. PostgreSQL поддерживает следующие типы индексов: B-Tree, B-Tree cluster, Hash, GIN, GiST [6].

В первом типе индекса с высокой степенью параллельности используются алгоритмы B-деревьев Лемана-Яо (Lehman-Yao). Это самый распространенный способ индексации, обладающий наибольшими возможностями. По этой причине он используется по умолчанию. Реализация хэша основана на алгоритмах линейного хэширования Литвина (Litwin), которые традиционно используются для индексов с частой проверкой равенства.

GIN (Generalized Inverted Index) – обобщенный инвертированный индекс. Это индексная структура, которая хранит множество пар (ключ, список ссылок), где «список ссылок» содержит ссылки на записи, в которых появляется этот ключ. Каждое индексируемое значение может содержать много ключей, поэтому один и тот же идентификатор записи может появляться в нескольких списках ссылок.

GiST (Generalized Search Tree) – обобщенное поисковое дерево, представляет собой сбалансированное (по высоте) дерево, листья которого содержат пары (key, rid), где key - ключ, а rid - указатель на соответствующую запись на странице данных.

Также PostgreSQL позволяет создавать функциональные индексы, т.е. осуществлять индексирование не по значению поля, а по некоторой функции этих значений.

В качестве функции для построения функционального индекса применялась дисперсия распределения особых точек вокруг их мнимого центра масс, которая вычисляется по формуле (1).

$$D[S] = \frac{1}{N-1}\sum_{i=1}^{N}\left(S_i - M[S]\right)^2 \qquad (1)$$

где S_i – расстояния от центра масс до каждой точки (2),

$M[S]$ – математическое ожидание расстояний от центра масс до каждой точки (3)

$$S_i = \sqrt{\left(X_C - X_i\right)^2 + \left(Y_C - Y_i\right)^2} \quad (2)$$

$$M[S] = \frac{1}{N}\sum_{i=1}^{N} S_i \quad (3)$$

где X_C и Y_C – координаты центра масс точек распределения, которые, при условии, что масса находится в вершинах и каждая вершина весит одинаково, вычисляются по формулам (4) и (5).

$$X_C = \frac{1}{N} \cdot \sum_{i=1}^{N} X_i, \quad (4)$$

$$Y_C = \frac{1}{N} \cdot \sum_{i=1}^{N} Y_i. \quad (5)$$

Экспериментальные исследования и анализ результатов

Исследования методов поиска и индексирования отпечатков пальцев проводились в среде СУБД PostgreSQL v8.4.0. Для ввода программных команд использовался psql – клиент базы данных, входящий в комплект поставки PostgreSQL. Для просмотра результатов выполнения команд также использовался pgAdminIII v1.10.0 – система управления серверами и базами данных PostgreSQL.

Для хранения данных об отпечатках пальцев была разработана база данных Fingerprints, а в ней создана тестовая таблица tFingerprints. Таблица содержала поля: «Идентификатор отпечатка пальца» (типа int), «Вектор особых точек» (типа double[]), «Дисперсия вектора особых точек» (типа double). Объем таблицы составлял 2000 записей, причем каждой записи соответствовал один отпечаток пальца. Для каждой записи поле «Вектор особых точек» было проинициализировано вектором вида $T_i = \{x_1, y_1, \theta_1, x_2, y_2, \theta_2, ..., x_N, y_N, \theta_N\}$, где x_i, y_i – действительные числа в интервале [0..1], θ_i – действительное число в интервале [0..360], а n – целое число в интервале [30..50]. Также для каждой

записи было вычислено поле «Дисперсия вектора особых точек» с помощью функции дисперсии от поля «Вектор особых точек».

Затем были разработаны тестовые функции для вычисления времени, затрачиваемого на поиск, в зависимости от типа используемого индекса. В этих функциях последовательно бралась каждая из 2000 записей таблицы, и затем с помощью оператора SELECT с условием WHERE равным значению поля «Вектор особых точек» или «Дисперсия вектора особых точек» осуществлялась выборка из базы данных. Время, затрачиваемое на выборку, суммировалось и результатом функции являлось общее время, затраченное на поиск 2000 записей в таблице объемом 2000 записей. Для поля «Вектор особых точек» эксперименты проводились с такими типами индексов, как B-Tree, B-Tree cluster, GIN, а для поля «Дисперсия вектора особых точек» – B-Tree, B-Tree cluster, HASH.

Не удалось провести эксперимент с индексом HASH для поля «Вектор особых точек» и GIN – для поля «Дисперсия вектора особых точек» из-за несовместимости типа данных этих полей с типом индекса.

План каждого из экспериментов был таким: 1. создавался индекс одного из типов по одному из полей; 2. вычислялось время, за-

Таблица 1

Время (с.), затрачиваемое на поиск в зависимости от поля и метода индексирования.

Имя поля \ Метод	Без индексирования	B-Tree	B-Tree cluster	HASH	GIN
Вектор особых точек	1,0096	0,15315	0,20295	– / –	3,10765
Дисперсия вектора особых точек	1,0148	0,1923	0,21905	0,1398	– / –

траченное на поиск. В табл. 1 приведено среднее время (в секундах), затрачиваемое на поиск, в зависимости от индексируемого поля и метода индексирования (для набора из 2000 записей и количества особых точек в отпечатке в интервале [30 .. 50]).

Выводы

Как видно из табл. 1, наиболее оптимальным оказалось индексирование по полю «Дисперсия вектора особых точек» с использованием метода индексирования HASH. Удовлетворительный результат показало индексирование по обоим полям с использованием метода B-Tree. Остальные методы индексирования (B-Tree cluster и GIN) оказались намного хуже и в дальнейших исследованиях рассматриваться не будут. В будущем планируется рассмотреть метод индексирования GiST, который возможно даст выигрыш по сравнению с методом HASH.

Литература

[1] Автоматизированная дактилоскопическая информационно-поисковая система АДИС ПАПИЛОН. [Электронный ресурс] – Режим доступа: http://www.papillon.ru/rus/16/?PHPSESSID=8813bf02f4ee087c82abbad19a824515

[2] Система учета рабочего времени и контроля доступа по отпечаткам пальцев BioTime. [Электронный ресурс] – Режим доступа: http://www.biotime.ru

[3] Next Generation Identification – NGI. [Электронный ресурс] – Режим доступа: http://www.fbi.gov/hq/cjisd/ngi.htm

[4] Сайт о СУБД PostgreSQL. [Электронный ресурс] – Режим доступа: http://www.postgresql.org

[5] D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar, Handbook of fingerprint recognition, New York, 2003, 348p

[6] Дж. Уорсли, Дж. Дрейк, PostgreSQL для профессионалов, Учебник, СП: «ПИТЕР», 2003, 496 с.