УДК 621.3

АЛГОРИТМ ИДЕНТИФИКАЦИИ РЕКВИЗИТОВ ФИЗИЧЕСКИХ ЛИЦ В БАЗАХ ДАННЫХ НА ОСНОВЕ МЕТРИКИ ЛЕВЕНШТЕЙНА

Н.И. Лиманова, М.Н. Седов

При передаче данных от одного учреждения к другому возникает проблема персональной идентификации физических лиц, у которых частично или полностью не совпадают реквизиты. В работе предложен оптимальный алгоритм идентификации на основе метрики Левенштейна, позволяющий выполнять поиск физических лиц в базе данных на основе нечеткого сравнения. Алгоритм реализован на языке PL-SQL в СУБД Oracle 11g.

Ключевые слова: межведомственный информационный обмен, идентификация, нечеткое сравнение, поиск реквизитов физических лиц, функция интеллектуального сравнения, персональный идентификационный номер.

Введение

В процессе межведомственного информационного обмена возникает проблема согласования основных реквизитов (ФИО, даты рождения, адреса, паспортных данных и т.п.) физических лиц в базах данных различных ведомств, обменивающихся информацией. Проблема персональной идентификации приобретает наибольшую актуальность для физических лиц, у которых частично или полностью не совпадают реквизиты.

Для удобства обработки данных каждому набору реквизитов в базах данных присваивается так называемый персональный идентификационный номер (ПИН). В случае обработки или передачи данных о физическом лице вся привязка осуществляется именно к этому ПИНу. В России, к сожалению, пока нет единой базы с реквизитами всех жителей, поэтому в разных ведомствах ведется свой отдельный реестр физических лиц и заводятся свои ПИНы. Проблема возникает при осуществлении обмена информацией о жителях между организациями, так как необходимо выполнить привязку входящих реквизитов к уже имеющимся. Для однозначной привязки необходимо выполнить интеллектуальный поиск физического лица в базе-приемнике, который должен учитывать множество факторов: и потенциальные ошибки при ручном вводе, и отсутствующие или устаревшие реквизиты, и т.п. Естественно предположить, что подобный поиск целесообразно реализовать в виде специализированного программного обеспечения [1].

Традиционно данная проблема решается путем анализа тождественности основных реквизитов физического лица. Таких реквизитов несколько: фамилия, имя, отчество, дата рождения, серия, номер паспорта и адрес. Однозначно определив совпадение существующих и новых реквизитов, можно выполнить идентификацию физического лица в базе данных. Данный метод поиска выполняется вручную только в том случае, когда объем передаваемой информации невелик (количество физических лиц не более 30). При больших объемах передаваемых данных используется автоматизированное сравнение тождественности реквизитов. Такой подход позволяет определить в среднем 50–60% от общего числа идентифицируемых физических лиц. Оставшиеся 40–50% представляют собой персональные данные, в которых частично или полностью не совпадают реквизиты. Такую информацию вручную обрабатывать еще сложнее.

Неправильная идентификация может привести также к большому количеству данных в отчете для ручной отработки, к присвоению ПИНа не тому человеку и к добавлению излишних данных. Последствия таких ошибок в худшем случае могут полностью парализовать работу учреждения на неопределенное время, в лучшем – отнять более 10% рабочего времени специалистов на исправление ошибок. Анализ существующего программного обеспечения показал, что единого идентификатора нет, универсальный алгоритм идентификации также отсутствует. Так как большинство реквизитов физических лиц имеет строковый тип, то естественно предположить, что необходимый алгоритм должен анализировать именно строковые значения.

Математическая модель

Известно несколько видов метрик, отражающих интуитивное понятие схожести строк. Наиболее распространены расстояния Хемминга, метрика Левенштейна и расстояние редактирования [2–4].

Расстояние Хемминга определяется для строк одинаковой длины и задается как число позиций, в которой символы не совпадают. Фактически расстояние Хемминга рассчитывается как минимальная цена преобразования одной строки в другую, когда возможна только одна операция редактирования строк – замена.

В случае, когда требуется произвести сравнение строк разной длины, используются метрика Левенштейна или расстояние редактирования. Эти две метрики очень похожи по построению и фактически являются одной и той же метрикой, несколько модифицированной для каждого случая. Так, например, метрика Левенштейна определяется как минимальная цена преобразования одной строки в другую с использованием трех операций – вставки, замены и удаления символа, причем все три операции имеют одинаковый вес.

Расстояние редактирования является модификацией метрики Левенштейна на случай, когда разрешены всего две операции – вставки и удаления.

В связи с изложенным была выбрана именно общая метрика Левенштейна, которая поддерживает все три операции со строкой. Для дальнейшей работы была построена лингвистическая переменная «схожесть строк». Решено выделить следующие термы: «строки совпадают», «строки почти совпадают», «строки похожи», «строки похожи и не похожи одновременно», «строки не похожи».

В результате анализа функций принадлежности лингвистических термов возникла необходимость модификации метода вычисления метрики Левенштейна. Потребовалось модифицировать метрику таким образом, чтобы расстояние между строками зависело, в том числе, и от длины сравниваемых строк.

Теорема. Обозначим величиной $p(s_1,s_2)$ метрику Левенштейна, а величиной $||s_i||$ – длину строки s_i. Тогда функция

$$r(s_1,s_2)=\frac{p(s_1,s_2)}{\max\{||s_1||,||s_2||\}}, \qquad (1)$$

является метрикой.

Доказательство. Поскольку $p(s_1,s_2)$ – метрика, то имеем

$p(s_1,s_2)\geq 0$,

$p(s_1,s_2)=p(s_1,s_2)$,

$p(s_1,s_2)+p(s_2,s_3)\geq p(s_1,s_3)$

для любых строк s_1, s_2 и s_3.

Учитывая эти соотношения и равенство (1), приходим к выводу, что $r(s_1,s_2)$ удовлетворяет первым двум аксиомам, определяющим метрику. Остается доказать, что для любых строк s_1, s_2 и s_3 функция $r(s_1,s_2)$ удовлетворяет неравенству треугольника, т.е.

$r(s_1,s_2)+r(s_2,s_3)\geq r(s_1,s_3)$.

Запишем это неравенство в виде:

$$\frac{p(s_1,s_2)}{\max\{||s_1||,||s_2||\}}+\frac{p(s_2,s_3)}{\max\{||s_2||,||s_3||\}}-\frac{p(s_1,s_3)}{\max\{||s_1||,||s_3||\}}\geq 0.$$

Возможны следующие случаи:

1. $||s_1||\leq||s_2||\leq||s_3||$;
2. $||s_2||\leq||s_3||\leq||s_1||$;
3. $||s_3||\leq||s_1||\leq||s_2||$;
4. $||s_2||\leq||s_1||\leq||s_3||$;
5. $||s_1||\leq||s_3||\leq||s_2||$;
6. $||s_3||\leq||s_2||\leq||s_3||$.

Рассмотрим первый случай. Имеем:

$$\frac{p(s_1,s_2)}{\max\{||s_1||,||s_2||\}}+\frac{p(s_2,s_3)}{\max\{||s_2||,||s_3||\}}-\frac{p(s_1,s_3)}{\max\{||s_1||,||s_3||\}}=\frac{p(s_1,s_2)}{||s_2||}+\frac{p(s_2,s_3)}{||s_3||}-\frac{p(s_1,s_3)}{||s_3||}\geq$$

$$\geq\frac{1}{||s_3||}(p(s_1,s_2)+p(s_2,s_3)-p(s_1,s_3))\geq 0.$$

Таким образом, для первого случая неравенство треугольника выполняется. Поскольку второй случай аналогичен первому, на основании подобных выкладок делаем вывод, что для второго случая неравенство треугольника также выполняется.

Перейдем к рассмотрению третьего случая, где

$$r(s_1,s_2)+r(s_2,s_3)-r(s_1,s_3)=\frac{1}{||s_2||}(r(s_1,s_2)+r(s_2,s_3))-\frac{1}{||s_1||}r(s_1,s_3)). \qquad (2)$$

Рассмотрим вопрос о том, когда достигается минимум функции, находящейся в правой части этого равенства. Понятно, что если выражение $r(s_1,s_2)+r(s_2,s_3)$ достигает минимума, а $r(s_1,s_3)$ максимума, то значение всего выражения будет минимальным. Указанные два условия могут выполняться одновременно, если одновременно выполняются два следующих утверждения:

1. строки s_1 и s_3 не имеют общих символов;
2. строки s_1 и s_3 входят в качестве подстрок в s_2.

Тогда

$r(s_1,s_3)=\max\{||s_1||,||s_3||\}=||s_1||$,

$r(s_1,s_2)=||s_3||+||C||$, $r(s_2,s_3)=||s_1||+||C||$,

и, таким образом, минимальное значение выражения (2) запишется в виде

$$\frac{||s_3||+||C||+||s_1||+||C||}{||s_3||+||s_1||+||C||}-\frac{||s_1||}{||s_1||}=\frac{||C||}{||s_3||+||s_1||+||C||}\geq 0.$$

Следовательно, в третьем случае для функции $r(s_1,s_3)$ также выполняется неравенство треугольника. Остальные случаи аналогичны уже рассмотренным. Таким образом, функция $r(s_1,s_2)$ является метрикой, заданной на множестве строк. Теорема доказана.

Замечание. Функция $r(s_1,s_2)$ принадлежит отрезку [0,1] для любых строк s_1 и s_2.

В предложенном алгоритме данная метрика применяется для работы со строковыми реквизитами физических лиц, к которым относятся ФИО, адрес, документ и т.д. В связи с этим построенная с использованием данной метрики лингвистическая переменная позволяет обрабатывать запросы поиска для человека, похожего на другого человека по реквизитам. Приняв от пользователя такой запрос, мы фактически получаем два значения – значение искомого реквизита и радиус поиска.

Алгоритм идентификации реквизитов физических лиц

Укрупненная блок-схема разработанного алгоритма идентификации реквизитов физических лиц в базах данных представлена на рисунке.

В реализации алгоритма на языке PL-SQL СУБД Oracle 11g за предварительную выборку всех записей, отдаленно похожих на искомую, отвечает блок «Запрос количества идентичных людей в базе данных». Этот блок работает по алгоритму прямого частичного сравнения разных наборов реквизитов, например, имени, отчества и даты рождения, формируя тем самым рабочйй набор данных для рассматриваемого алгоритма идентификации. Затем в работу вступает «Блок сравнения реквизитов», ключевые функции которого отводятся логически выделенным процедурам COMPARISON_STRING и COMPARISON_NUMBER, созданным на основе модифицированного метода вычисления метрики Левенштейна, которые позволяют проводить интеллектуальное сравнение двух похожих строк или чисел с учетом возможных неточностей или ошибок ввода. С помощью указанных процедур программа формирует набор совпадений и по результатам обработки предлагаемой и искомой записи выносит решение об идентичности. Например, у человека совпадает имя, отчество, дата рождения и номер паспорта, а в фамилии допущена ошибка в одну букву. В данном случае программа однозначно идентифицирует реквизиты. Данные процедуры могут применяться не только для идентификации реквизитов, но также везде, где требуется полнотекстовый поиск с нечетко заданными входными данными.

Алгоритм идентификации аккумулирует так называемый «опыт прошлых идентификаций» и записывает его в специально отведенное место в базе данных для использования в последующих идентификациях. Это позволяет сохранить не только результаты автоматической работы программы, но и решения операторов после отработки ими оставшихся не найденных реквизитов.

Для сравнительного анализа разработанного алгоритма рассмотрим технологию идентификации на основе прямого сравнения. При использовании данной технологии предпочтение отдается скорости обработки записей, а не качеству принятия решения системой. В итоге, после окончания работы процедуры на основе прямого сравнения, остается много данных (около 20–30% от общего количества строк), не связанных с исходными, которые необходимо отрабатывать вручную, что крайне затруднительно при больших объемах обрабатываемых данных.

В качестве тестовой среды были выбраны сводные базы данных населения города с количеством записей примерно 800 000, СУБД Oracle 11g, сервер HP ProLiant DL160 G6.

При сравнении рабочих показателей двух алгоритмов выявлены особенности алгоритмов:

- алгоритм прямого сравнения
  - скорость обработки данных – примерно 100 000 строк в час;
  - точность идентификации (вероятность точного поиска реквизитов) – примерно 80%;
- алгоритм идентификации на основе нечеткого сравнения
  - скорость обработки данных – примерно 80 000 строк в час;
  - точность идентификации (вероятность точного поиска реквизитов) – примерно 99%.

Таким образом, разработанный алгоритм, хотя и демонстрирует несколько меньшую скорость обработки, но за счет интеллектуальной системы принятия решений минимизирует ручную работу оператора по отработке результатов, чего не может предложить алгоритм прямого сравнения.

Заключение

Рассмотренный метод идентификации персональных данных на основе нечеткого сравнения позволяет быстро определять людей, используя данные ранее проведенного поиска. Встроенная система приоритета реквизитов позволяет идентифицировать человека в таких случаях, как смена фамилии, имени, переезд, ошибки при ручном вводе данных, а также при частично отсутствующих реквизитах.

Рассмотренную в работе процедуру идентификации можно рассматривать как часть системы поддержки принятия решений. Процедура не требует вмешательство оператора, накапливает опыт в процессе работы, позволяя тем самым полностью освободить специалистов от низкопрофильной, неэффективной, ручной работы напрямую с наборами реквизитов физических лиц, хранящимися в базах данных.

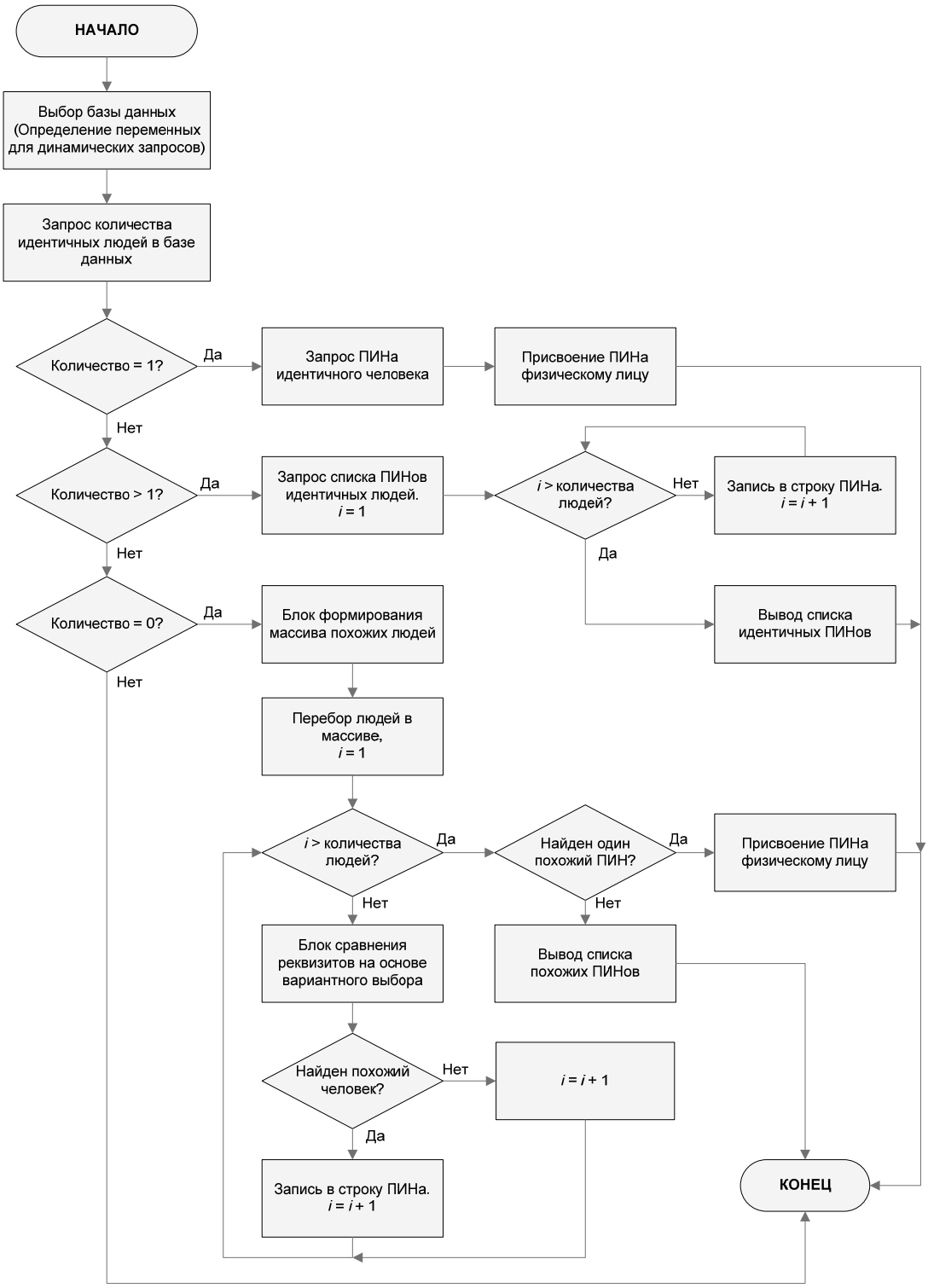

Рисунок. Укрупненная блок-схема алгоритма идентификации реквизитов физических лиц в базах данных

В перспективе данный алгоритм обладает возможностью успешного внедрения в системы глобального объединения хранилищ государственных или коммерческих организаций для ведения единой базы данных населения любой страны мира. Логическая структура разработанного алгоритма позволяет реализовать его на любом популярном языке программирования. Масштабируемость алгоритма позволяет применять программные процедуры на его основе как в малых организациях, так и в крупных корпорациях – везде, где ведется и актуализируется реестр данных физических лиц. Возможные примеры использования – портал госуслуг, медицинские электронные системы, кадровые и бухгалтерские системы учета служащих, банковские системы хранения данных о клиентах и т.п.

Алгоритм реализован на языке PL-SQL системы управления базами данных Oracle 11g. Разработанное программное обеспечение, реализующее метод автоматизированного поиска персональных дан-

ных на основе нечеткого сравнения, внедрено и успешно функционирует с 2007 г. в муниципальном учреждении «Городской информационный центр» г. Тольятти Самарской области.

Литература

1. Международный фонд автоматической идентификации. Технологии автоматической идентификации [Электронный ресурс]. – Режим доступа: http://www.fond-ai.ru/art1/art223.html, свободный. Яз. рус. (дата обращения 16.06.2012).
2. Хемминг Р.В. Теория кодирования и теория информации: Пер. с англ. / Под ред. Б.С. Цыбакова. – М.: Радио и связь, 1985. – 176 с.
3. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // ДАН СССР. – 1965. – Т. 163. – № 4. – С. 845–848.
4. Бойцов Л.М. Анализ строк [Электронный ресурс]. – Режим доступа: http://itman.narod.ru/articles/infoscope/string_search.1-3.html, свободный. Яз. рус. (дата обращения 16.06.2012).

Лиманова Наталья Игоревна – Тольяттинский государственный университет, доктор технических наук, профессор. Natalya-Igorevna@yandex.ru

Седов Максим Николаевич – Мэрия городского округа Тольятти, инженер-программист, SedovMN@inbox.ru