УДК 621.3

ПОИСК РЕКВИЗИТОВ ФИЗИЧЕСКИХ ЛИЦ В БАЗАХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ DATA MINING

Н. И. Лиманова,
доктор техн. наук, профессор
Тольяттинский государственный университет
М. Н. Седов,
инженер-программист I категории
Мэрия городского округа Тольятти

При передаче данных от одного учреждения к другому возникает проблема персональной идентификации физических лиц, у которых частично или полностью не совпадают реквизиты. В работе оптимальный алгоритм идентификации, позволяющий выполнять поиск физических лиц в базе данных на основе нечеткого сравнения, представлен в виде процесса Data Mining. 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

Для оптимального управления большими массивами данных, связанных с реквизитами физических лиц, необходимо обеспечивать централизованные регламенты хранения таких характеристик, как ФИО, дата рождения, адрес, паспортные данные и т. д. В последнее время различные ведомства — держатели локальных баз данных (БД) стремятся объединить массивы для упрощения и повышения качества работы. Но возникает проблема: как сопоставить реквизиты физических лиц из одной БД реквизитам в другой? На помощь приходит интеллектуальный алгоритм поиска физических лиц в БД, или идентификация реквизитов физических лиц.

Для удобства обработки данных каждому набору реквизитов в БД присваивается так называемый персональный идентификационный номер (ПИН) [1]. 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 [2]. Для однозначной привязки необходимо провести 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 [3].

В результате поиска решения описанной проблемы была найдена технология для построения и реализации процедуры идентификации физических лиц в БД — Data Mining [4].

Data Mining — мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория БД и др.

Данная технология хорошо подходит для реализации решения описанной проблемы межведомственного информационного обмена, так как позволяет сформировать алгоритм с встроенной системой принятия решений для повышения качества идентификации реквизитов физических лиц. Также инструменты этой технологии позволяют производить интеллектуальное сравнение двух наборов данных и выявление закономерностей в схожих данных для повышения качества поиска на основе нечеткого сравнения.

Представление алгоритма идентификации физических лиц в виде процесса Data Mining

Идентификация физических лиц в БД призвана решить одну из важнейших проблем при межведомственном информационном обмене, возникающую в результате неимения единого ПИНа, — отсутствие однозначного соответствия физического лица с набором его основных реквизитов (ФИО, дата рождения, адрес и т. д.).

Проектирование алгоритма идентификации с учетом технологии Data Mining позволило создать самообучающуюся процедуру, работающую в автоматическом режиме.

В работе предложен алгоритм поиска, включающий следующие стадии: свободный поиск (в том числе валидацию) → прогностическое моделирование → анализ исключений.

Свободный поиск (Discovery). На этой стадии осуществляется исследование исходного набора данных в целях поиска скрытых закономерностей. Предварительные гипотезы относительно вида закономерностей здесь не определяются.

Применительно к рассматриваемой процедуре идентификации данная стадия реализована в виде расширенного поиска по запросу, возвращающему данные, отдаленно схожие с набором реквизитов искомого физического лица. Именно на этом этапе ищутся закономерности, позволяющие потом, при следующих идентификациях, применить найденное правило, что ускоряет весь процесс в десятки раз.

Прогностическое моделирование (Predictive Modeling). Вторая стадия использует результаты работы первой стадии. Здесь обнаруженные закономерности используются непосредственно для прогнозирования.

На данном этапе разработанный алгоритм идентификации аккумулирует так называемый «опыт прошлых идентификаций» и записывает его в специально отведенное место для использования в следующий раз.

Анализ исключений (forensic analysis). На третьей стадии анализируются исключения или аномалии, выявленные в найденных закономерностях. Действие, выполняемое на этой стадии, — выявление отклонений (deviation detection). Для выявления отклонений необходимо определить норму, которая рассчитывается на стадии свободного поиска.

На третьей стадии процедура идентификации удаляет из набора выявленных закономерностей все данные, полученные ошибочным путем. Например, некорректность и отсутствие некоторых реквизитов у физического лица может привести к выявлению ошибочной закономерности, которая в свою очередь при условии ее использования даст неверные выводы об идентификации подобных наборов данных. Поэтому в предложенном алгоритме заключительным этапом проводится именно третья стадия — анализ исключений.

Классификация технологических методов алгоритма

Все методы рассматриваемого алгоритма с учетом классификаций Data Mining [4] подразделяются на две большие группы по принципу работы с исходными обучающими данными:

1) непосредственное использование данных — проявляется на этапе прямой работы с выборками из БД и внесения необходимых изменений в наборы реквизитов физических лиц;

2) выявление и использование формализованных закономерностей, или дистилляция шаблонов — используется при нечетком поиске конкретного физического лица из множеств наборов данных, присутствующих в базе.

В разработанном алгоритме идентификации применяются оба вида технологических методов.

Предлагаемый алгоритм, блок-схема которого представлена на рис. 1, включает следующие этапы.

1. Определение списка полностью идентичных наборов реквизитов физических лиц.

Производится поиск физических лиц по прямому сравнению реквизитов.

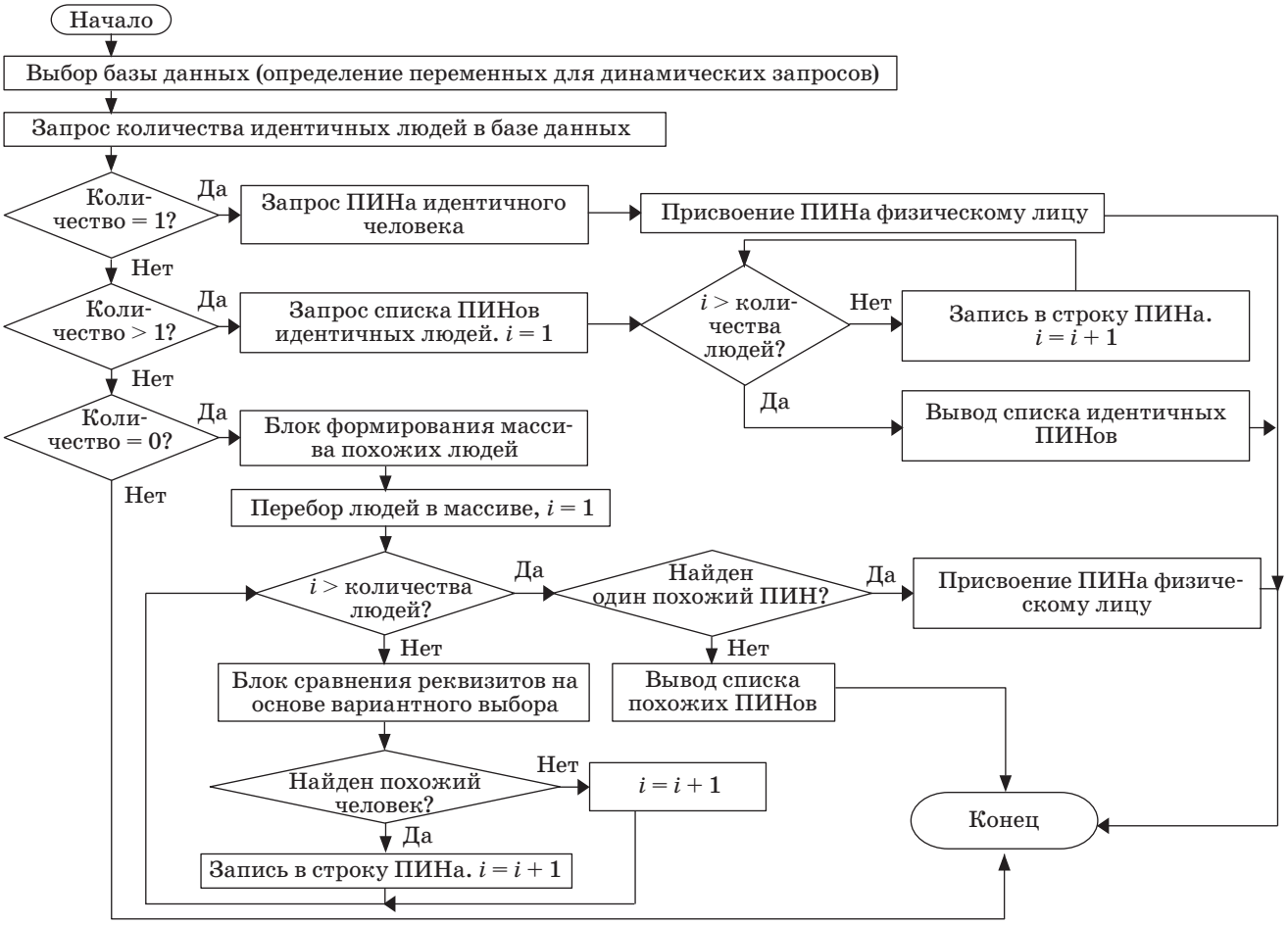
2. Подготовка данных для анализа.

В случае отсутствия реквизитов, полностью идентичных искомым, производится укрупненная выборка, включающая около 300–500 наборов, отдаленно похожих на искомый. Пример подобного запроса:

```
CURSOR persons
  IS
   SELECT p.person_id, p.lastname, p.firstname, p.patronymic, p.birthdate
     FROM work.person p
    WHERE (((SOUNDEX(TO_TRANSLIT(p.lastname)) =
              SOUNDEX(TO_TRANSLIT(fo_Lastname)))
      AND (SOUNDEX(TO_TRANSLIT(p.firstname)) =
             SOUNDEX(TO_TRANSLIT(fo_Firstname))))
       OR ((SOUNDEX(TO_TRANSLIT(p.lastname)) =
             SOUNDEX(TO_TRANSLIT(fo_Lastname)))
      AND (SOUNDEX(TO_TRANSLIT(p.patronymic)) =
             SOUNDEX(TO_TRANSLIT(fo_Patronymic))))
       OR ((SOUNDEX(TO_TRANSLIT(p.firstname)) =
             SOUNDEX(TO_TRANSLIT(fo_Firstname)))
      AND (SOUNDEX(TO_TRANSLIT(p.patronymic)) =
             SOUNDEX(TO_TRANSLIT(fo_Patronymic)))));
```

3. Вариантное сравнение похожих реквизитов.

Последовательный перебор массива похожих наборов и присвоение им моделей закономерностей. На этом этапе также производится выявление новых закономерностей.



■ ***Рис. 1.*** Укрупненный алгоритм поиска физических лиц в БД на основе нечеткого сравнения

4. Оцениваются и выбираются подходящие модели, наборы данных которых наиболее сравнимы с искомыми реквизитами.

5. В соответствии с выбранными моделями определяются наборы данных.

Ненайденные исходные наборы реквизитов выводятся в отчет для ручной отработки оператором.

6. На основе результатов ручной отработки корректируются хранимые модели для улучшения качества поиска в следующих сеансах идентификации.

В разработанной реализации алгоритма на языке PL-SQL СУБД Oracle 11g [5] ключевые функции отводятся логически выделенным процедурам COMPARISON_STRING и COMPARISON_NUMBER, которые позволяют проводить интеллектуальное сравнение двух похожих строк или чисел с учетом возможных неточностей или ошибок ввода. Данные процедуры могут применяться не только для идентификации реквизитов, но и везде, где требуется полнотекстовый поиск с нечетко заданными входными данными.

Технические и экономические показатели алгоритма

Для сравнительного анализа разработанного алгоритма рассмотрим технологию идентификации на основе прямого сравнения. При использовании данной технологии упор делается на скорость обработки записей, а не на качество 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 БД населения города с количеством записей ~ 800 0000, СУБД Oracle 11g, сервер HP ProLiant DL160 G6.

Рабочие показатели двух алгоритмов приведены в таблице.

Отсюда можно сделать вывод, что у разработанного алгоритма минимизирована работа опе-

Показатель	Алгоритм	
	прямого сравнения	нечеткого сравнения
Скорость обработки данных, строк/ч	~100 000	~80 000
Точность идентификации (вероятность точного поиска реквизитов), %	~80	~99,9

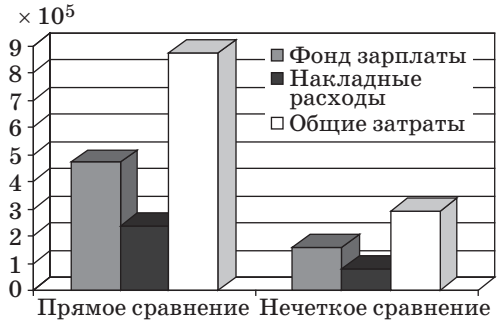


Рис. 2. Диаграмма для сравнительного анализа стоимостных показателей при использовании методов прямого и нечеткого сравнения

ратора по ручной отработке результатов, т. е. хотя скорость обработки несколько меньше, но алгоритм позволяет существенно разгрузить операторов за счет интеллектуальной системы принятия решений, чего не может предложить алгоритм прямого сравнения.

При сравнении экономических характеристик разработанного программного обеспечения на основе описываемого алгоритма с процедурой прямого сравнения для годового объема идентификации 1 200 000 физических лиц были получены следующие данные (рис. 2): трудовые затраты на обработку информации по методу нечеткого сравнения по сравнению с методом прямого сравнения уменьшены в 6,7 раза, абсолютное снижение трудовых затрат составило 1446 ч, годовые затраты при использовании метода нечеткого сравнения уменьшились в 3 раза по сравнению с аналогичным периодом применения метода прямого сравнения, а годовой экономический эффект превысил 580 тыс. р.

Заключение

Самообучающиеся системы позволяют освободить человеческие ресурсы для выполнения творческих задач. В этой области технология Data Mining предоставляет полный набор теоретических и практических средств для выбора, разработки или использования интеллектуальных компьютерных систем.

Рассмотренную в статье процедуру идентификации можно расценивать как часть системы поддержки принятия решений. Процедура не требует вмешательства оператора, накапливает опыт и самообучается в процессе работы, позволяя тем самым полностью освободить специалистов от неэффективной ручной работы напрямую с наборами реквизитов физических лиц, хранящимися в БД. Данная процедура реализована на языке PL-SQL СУБД Oracle 11g и успешно функционирует с ноября 2007 г. в муниципальном учреждении «Тольяттинский городской информационный центр». Логическая структура разработанного алгоритма позволяет реализовать его на любом популярном языке программирования.

В перспективе данный алгоритм может быть внедрен в системы глобального объединения хранилищ государственных или коммерческих организаций для ведения единой БД населения любой страны мира. Масштабируемость алгоритма позволяет применять программные процедуры на его основе как в малых организациях, так и в крупных корпорациях, т. е. 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.

Литература

1. **Положение** о персональном идентификационном номере граждан Российской Федерации, проживающих или пребывающих на территории Санкт-Петербурга. http://iac.spb.ru/shablon.asp?subpage=171&id=40&dir=0 (дата обращения: 11.06.2012).
2. **Отчет** о выполнении научно-исследовательской, опытно-конструкторской работы «Разработка механизмов однозначной идентификации данных о физических лицах и объектах недвижимости, хранящихся в различных информационных системах органов государственной власти и местного самоуправления». http://www.nisse.ru/business/article/article_464.html (дата обращения: 11.06.2012).
3. **Международный** фонд автоматической идентификации. Технологии автоматической идентификации. http://www.fond-ai.ru/art1/art223.html (дата обращения: 11.06.2012).
4. **Чубукова И. А.** Data Mining: учеб. курс. Изд-во Интернет-университета информационных технологий, 2006. http://www.intuit.ru/department/database/datamining/ (дата обращения: 11.06.2012).
5. **Скотт У.** ORACLE 9i Программирование на языке PL/SQL: учеб. пособие. — М.: Лори, 2004. — 528 с.