



англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.: ил.

3. Борисов В.В., Круглов В.В., Федулов А.С. Нечёткие модели и сети [Текст]. – М.: Горячая линия – Телеком, 2007. – 284 с.: ил.

Н.И. Лиманова, М.Н. Седов

ИДЕНТИФИКАЦИЯ ПАЦИЕНТОВ В БАЗАХ ДАННЫХ МЕДУЧРЕЖДЕНИЙ НА ОСНОВЕ МЕТРИКИ ЛЕВЕНШТЕЙНА

(Поволжский государственный университет
телекоммуникаций и информатики)

Введение

В процессе обработки информации о пациентах (физических лицах) в медицинских учреждениях для удобства обработки данных каждому набору реквизитов физических лиц (таких как ФИО, адрес, номера паспорта и полиса, СНИЛС и т.п.) 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.

При осуществлении обмена информацией о пациентах между различными медицинскими учреждениями возникает проблема сопоставления реквизитов физических лиц из одной базы данных реквизитам в другой. Для однозначной привязки необходимо выполнять интеллектуальный поиск физического лица в базе-приёмнике, который должен учитывать множество факторов: и потенциальные ошибки при ручном вводе, и отсутствующие или устаревшие реквизиты и т.п. Подобный поиск целесообразно реализовать в виде специализированного программного обеспечения.

Для решения данной проблемы был разработан интеллектуальный алгоритм поиска физических лиц в базах данных на основе нечеткого сравнения, другими словами, алгоритм идентификация реквизитов физических лиц.

Алгоритм идентификации

Предлагаемый алгоритм 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 укрупнённая выборка, включающая около 300-500 наборов, отдалённо похожих на искомым.

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

4. Оцениваются и выбираются подходящие модели, наборы данных которых в наибольшей степени совпадают с искомыми реквизитами.



5. В соответствии с выбранными моделями, определяются наборы данных. Не найденные исходные наборы реквизитов выводятся в отчёт для ручной отработки оператором.

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



Рис. 1. Укрупненная блок-схема алгоритма автоматизированного поиска реквизитов физических лиц в базах данных

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, рабочей 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 строк. 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.

Математическая модель

Для использования метрики Левенштейна для задач идентификации, потребовалось модифицировать метрику таким образом, чтобы расстояние между строками зависело, в том числе, и от длины сравниваемых строк [1-3].

Теорема 1: Обозначим при помощи величины $p(s_1, s_2)$ метрику Левенштейна, а величиной $\|s_i\|$ – длину строки s_i . Тогда функция

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

является метрикой.

Доказательство: поскольку $p(s_1, s_2)$ – метрика, то имеем:

$$p(s_1, s_2) \geq 0, \quad p(s_1, s_2) = p(s_2, s_1), \quad p(s_1, s_2) + p(s_2, s_3) \geq p(s_1, s_3)$$

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. Остаётся доказать, что для любых строк s_1, s_2 и s_3 функция $r(s_1, s_2)$ удовлетворяет неравенству треугольника: $r(s_1, s_2) + r(s_2, s_3) \geq r(s_1, s_3)$.

Запишем это неравенство в виде:

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

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. $\|s_3\| \leq \|s_2\| \leq \|s_1\|$



Рассмотрим первый случай. Имеем:



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.

Перейдём к рассмотрению третьего случая. Итак, в третьем случае имеем:

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

Следовательно, в третьем случае для функции $r(s_1, s_3)$ также выполняется неравенство треугольника. Остальные случаи аналогичны 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 алгоритме идентификации данная метрика применяется для работы со строковыми реквизитами физических лиц, к которым относятся ФИО, адрес, документ и т.д. Поэтому 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.

Заключение

Рассмотренный в статье алгоритм идентификации позволяет решить проблемы сопоставления реквизитов пациентов медицинских учреждений при информационном обмене между учреждениями. Реализованный в виде программного обеспечения алгоритм 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 базах данных. Данный алгоритм реализован на языке PL-SQL СУБД Oracle 11g и успешно функционирует с ноября 2007 г. в ряде муниципальных и государственных учреждений г. Тольятти Самарской области.

Литература

1. Хемминг Р.В. Теория кодирования и теория информации, пер. с англ. под ред. Б.С. Цыбакова, – М.: Радио и связь, 1985. – 176 с.
2. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов, – М.: Доклады АН СССР, 1965. т.163.4, 845-848 с.



3. Бойцов Леонид. Анализ строк [Электронный ресурс] /Л. Бойцов. – Электрон. текстовые дан. – Москва: [б.и.], 2004. – Режим доступа: http://itman.narod.ru/articles/infoscope/string_search.1-3.html, свободный.

Д.Р. Максимова, А.В. Калукова, М.М. Тюрина, Д.В. Хазов

МЕТОДЫ ДИАГНОСТИКИ НЕЕСТВЕННОГО ПОЛОЖЕНИЯ СТОПЫ

(Казанский национальный исследовательский технический университет им А.Н. Туполева – КАИ)

Плоскостопие – деформация стопы, следствием которой является неестественное для человека перераспределение нагрузки на весь опорно-двигательный аппарат.

В настоящее время российская медицина предлагает лечение деформации стопы двумя методами: систематическими массажными воздействиями и тренировками, включающими ношение специальной обуви или хирургическими процедурами.

Для поиска альтернативных методов лечения плоскостопия необходимо провести поиск и анализ способов выявления неестественного положения стопы и неправильной мышечной активности при положении стоя и при ходьбе. Предложенный анализ позволит выделить наиболее оптимальные автоматические и автономные методы, позволяющие систематически определять положение стопы. Автономная и автоматическая регистрация положения стопы предопределяет возможность своевременного напоминания пациенту о нарушениях походки или неестественном для стопы положении стоя, что способно обеспечить принятие мер. Это поспособствует созданию новой концепции устройства для лечения плоскостопия.

Патентный поиск позволил вывить следующие результаты:

Устройство определения положения точки нулевого момента при ходьбе без сгибания стопы [1] позволяет составить математическую модель походки человека. Оно создано для дальнейшего обучения антропоморфной части движения, но, как утверждают авторы изобретения, данную разработку можно использовать для выявления патологий нижних конечностей человека, путем сравнения полученных экспериментальных данных с патологиями, полученных путем многократных измерений математической модели передвижения здорового человека.

Устройство состоит из двух подошв, по углам которых размещены 4 резистивных тензодатчика силы. Оно стационарное, что не позволяет производить измерения автономно, во время повседневной жизни пациента, однако его показания можно использовать для составления модели естественных перемещений человека, которую можно будет использовать при дальнейшей разработке устройства для лечения плоскостопия.