АЛГОРИТМ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА ДАННЫХ НА ОСНОВЕ НЕЧЕТКОГО СРАВНЕНИЯ

Лиманова Н.И., Седов М.Н.

Поволжский государственный университет телекоммуникаций и информатики,
каф. программного обеспечения и управления в технических системах,
Россия, 443090, г. Самара, Московское шоссе, 77, тел. (846) 228-00-13,
E-mail: Nataliya.I.Limanova@gmail.com

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 сопоставления реквизитов физических лиц в одной БД аналогичным реквизитам в другой. В работе предложен алгоритм поиска данных, позволяющий выполнять интеллектуальное сравнение двух наборов данных и выявлять закономерности в схожих данных на основе нечеткого сравнения для повышения достоверности поиска. Алгоритм включает следующие стадии: свободный поиск (в том числе валидацию), прогностическое моделирование и анализ исключений. На первой стадии осуществляется исследование исходного набора данных с целью поиска скрытых закономерностей. Достоверность предварительных гипотез относительно вида закономерностей здесь не определяется. Применительно к рассматриваемому алгоритму 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. Вторая стадия базируется на результатах работы первой стадии. Здесь обнаруженные закономерности используются непосредственно для прогнозирования. На данном этапе разработанный алгоритм поиска аккумулирует так называемый «опыт прошлых идентификаций» и записывает его в специально отведенное место для использования в следующий раз. На третьей стадии анализируются исключения или аномалии, выявленные в найденных закономерностях. Действие, выполняемое на этой стадии — выявление отклонений. 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, использование которой даст неверное заключение об идентификации подобных наборов данных. Поэтому в предложенном алгоритме заключительным этапом проводится именно анализ исключений.

Алгоритм интеллектуального поиска данных реализован на языке PL-SQL СУБД Oracle 11g и успешно работает в настоящее время в муниципальном учреждении «Тольяттинский городской информационный центр».