РАЗРАБОТКА ПОИСКОВОГО АЛГОРИТМА ДЛЯ СИСТЕМЫ ПОИСКА ДУБЛИКАТОВ В КОНТАКТНЫХ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Бернікович Т.Я., Голушко И.Н.

Клезович О.В. – к.п.н.,доцент

В настоящее время существует потребность в высокопроизводительных алгоритмах, которые можно использовать для поиска дубликатов в не строго форматированных данных. Их сложность заключается в отсутствии жестко заданной структуры содержания полей сравнения. Целью данной работы является разработка такого алгоритма, который сможет обойти аналоги по точности и быстродействию одновременно.

Задача поиска дубликатов может быть решена посредством алгоритмов нечеткого сравнения строк. Однако следует отметить, что большинство широко известных способов поиска имеют один существенный недостаток: они нечувствительны к контексту. В данном случае может быть применен так называемый метод n-грамм. Данный метод позволяет решать проблему смысловых ошибок поиска. Однако определение таблиц вероятностей появления грамм в соответствующих позициях фразы (при использовании данного метода) представляет собой объемный труд и является зависимым от естественного языка. В рамках реляционной БД информация уже является структурированной в соответствии с языковыми особенностями и спецификой предметной области базы данных.

Целесообразно рассмотреть применение данного способа в автоматизированной информационной системе. Так, в качестве основных путей внесения и изменения информации можно выделить: непосредственный ввод пользователя автоматизированной информационной системы; импорт данных из внешних источников. При пользовательском вводе требуется обеспечение минимального времени отклика системы. В связи с этим используемый на данном этапе алгоритм должен работать не только точно, но и предельно быстро. При этом параметр порога автоматической обработки для данной операции может быть изменен в соответствии с требованиями по скорости поиска.

Рассмотрим общий алгоритм поиска дубликатов контактных данных, который представлен в соответствии с рисунком 1:

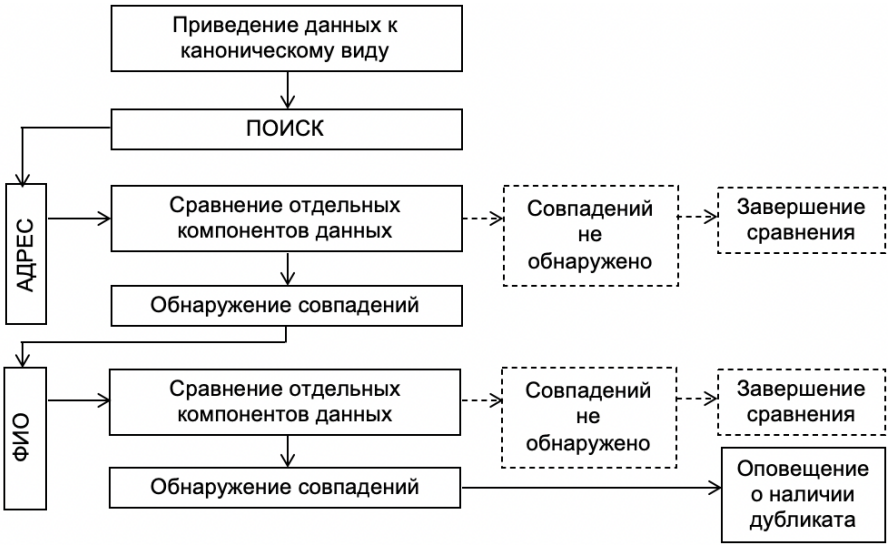

Рис. 1 – Алгоритм поиска дубликатов контактных данных в разрезе адресов

Таким образом, первоначальным элементом для сравнения является адрес, однако он не является единственным идентифицирующим полем для сравнения двух клиентов. В целях построения комплексной системы поиска целесообразно сравнение также второго поля – ФИО. Далее необходимо формирование алгоритма поиска дубликатов контактных данных в разрезе адресов.

Данное решение привело к значительному повышению быстродействия и поисковой точности на 26% выше по сравнению с алгоритмами в схожих продуктах.

Список использованных источников:

1. Тарасов, С.В. Контекстно зависимый способ поиска нечётких дубликатов в реляционных базах данных / С.В. Тарасов, В.В. Бураков // Информационно-управляющие системы. – 2015. – №2 (75). – С.76-81.

2. Сонькин, М.А. Применение алгоритмов нечеткого поиска в системах мониторинга лесопожарной обстановки / М.А. Сонькин, Ю.В. Лещик // Известия ТПУ. – 2012. – №5. – С.98-101.