АВТОМАТИЗИРОВАННАЯ СИСТЕМА ИСПРАВЛЕНИЯ И ДЕДУПЛИКАЦИИ КОНТАКТНЫХ ДАННЫХ

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

Бернікович Т.Я.

Клезович О.В. – к.п.н.,доцент

Компании, имеющие дело с контактными данными клиентов, часто сталкиваются с дубликатами и ошибками в базах с этой информацией после их ручного заполнения. Чтобы исправить это они нанимают специальных операторов, которые проверяют данные и вносят правки вручную. Такая проверка не отличается высокой точностью. На данный момент не существует универсального решения для этой задачи, а разработка решения специально под конкретную систему весьма затратна.

Необходимо разработать универсальную автоматизированную систему, которая будет структурировать и корректировать хранимую информацию о контактных данных. Кроме того, она должна уметь правильно объединять данные из нескольких источников. Предусмотреть возможность ручной модерации работы системы, в случае если не удается точно определить дубликат. Одними из главных требований являются – быстрота, точность и пригодность к обработке больших объемов данных. Затраты на наладку должны быть минимальными.

Система представляет собой программный блок, на вход которого из одного или нескольких (в зависимости от решаемой задачи) источников передаются контактные записи. Далее блок поиска дубликатов с помощью подблоков для сравнения определяет дубликаты записей или записи с искажениями. После определения этих записей они передаются в следующий блок, где корректируются. Опционально перед блоком коррекции может быть блок для ручного уточнения правок. Таким образом на выходе мы получаем исправленные данные. Увидеть общую структурную схему системы можно на рисунке 1.

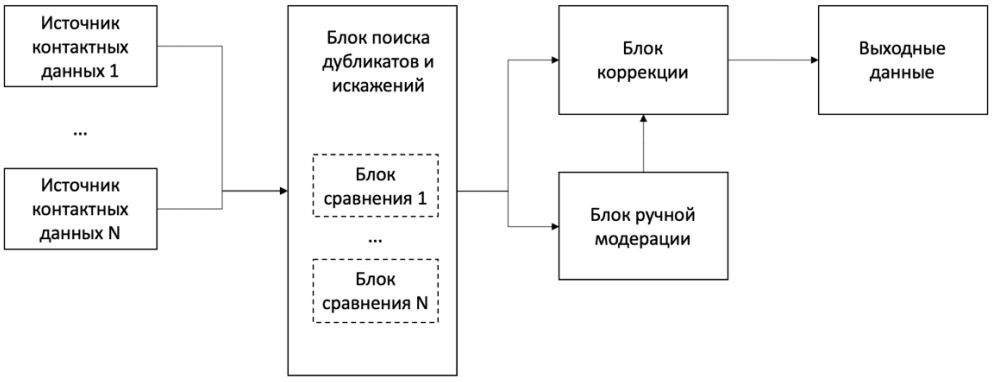

Рис. 1 – Структурная схема системы

При проектировании системы были учтены многие факторы, которые могли бы поставить ее в затруднительное положение. Следующим образом обрабатываются адреса:

1. Адреса перед сравнением приводятся к каноническому виду – порядок компонентов адреса может быть нарушен и сравнение в этом случае ничего не даст.
2. Сравниваются лишь компоненты адреса, для каждого из которых будут свои правила, – нельзя допустить ошибку при нечетком поиске в номере дома или квартиры, ведь это уже быдет совсем другой адрес.
3. Сравнение адресов происходит нечетким сравнением – так как при заполнении базы данных вручную могут быть допущены ошибки, а малейшая опечатка в названии улицы либо города не даст нам определить дубликаты.

Таким образом мы получили компонент для сравнения адресов. Но так как адрес не может быть единственным идентифицирующим полем при сравнении данных двух клиентов, сравнивать данные нужно комплексно. Поэтому подобным образом создается правила для всех существующих полей.

Созданная в результате система решает множество проблем связанных с контактными данными клиентов, т.е. может быть полезна широкому кругу компаний: банков, различным провайдерам, интернет-магазинам и другим – каждая из них найдет для себя полезное применение системы.

Список использованных источников:

1. Карвин, Б. Программирование баз данных SQL. Типичные ошибки и их устранение / Карвин Б., Райтман М. // 2011. – 336 с.

2. Гэлловей, М. Сила Objective-C 2.0. Эффективное программирование для iOS и OS X / Гэлловей М., Матвеев Е., Адуевская Л. // Питер, 2014. – 304 с.

3. Влиссидес, Д. Приемы объектно-ориентированного проектирования. Паттерны проектирования / Гамма Э., Хелм Р. // Питер, 2015. - 368 с.