Санкт-Петербургский
государственный
университет
www.spbu.ru

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Автоматизация сверки и устранения дубликатов в персональных данных

Студент Абубакиров А. Р.
Научный руководитель ст. преподаватель Севрюков С. Ю.

Содержание

Введение в проблематику

1. Что такое дубликаты?
2. Почему они возникают?
3. Какие сферы деятельности это затрагивает?
4. Зачем нужно устранять дубликаты?
5. Почему нельзя положиться только на номер паспорта?

Постановка задачи

1. Повышение качества данных
2. Разработка алгоритм поиска дубликатов
3. Тестирование на искусственных данных
4. Применение на реальных данных
5. Оценка результатов алгоритма
6. Встраивание в ИС компании МИАЦ

Обзор данных

	Фамилия	Имя	Отчество	Дата рождения
Всего	34406	34405	31995	34406
Кол-во уникальных	16304	1315	2006	16508
Кол-во пропущенных	0	1	2411	0

Табл.1. Статистика по данным

Типы ошибок

Ошибки на уровне поля:

1. опечатки: недопустимые символы
2. различные варианты написания имён
3. различные форматы даты
4. двойные фамилии и имена

Ошибки на уровне записи:

1. пропущенные значения
2. значения, которые стоят вместо отсутствующих значений
3. несоответствия поля и значения

Трудности интерпретации: истинные значения некоторых полей тяжело определить даже человеку.

Предобработка данных

Исправление ошибок с помощью:

1. регулярных выражений;
2. эвристических правил стандартизации;
3. полуавтоматического исправления ошибок.

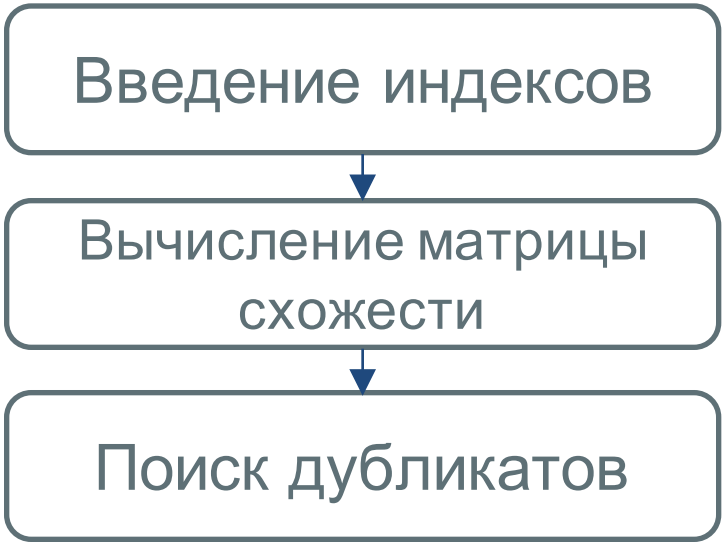
Приведение всех значений к единому формату.

Алгоритм поиска дубликатов: схема

Идея:

Вычислить “степень схожести” между записями, определить пороговое значение и выбрать те записи, которые удовлетворяют этому пороговому значению.

Схема:



Алгоритм поиска дубликатов: индексирование

Задача: уменьшить количество рассматриваемых записей.

Решение: выделить группы потенциальных дубликатов.

Условие: две записи считаются потенциальными дубликатами, если доля общих букв значений поля “Фамилия” превышает порогового коэффициента.

Алгоритм поиска дубликатов: вычисление матрицы

Редакционное расстояние:

$$dist : X \times X \to \mathbb{N}_0 \quad (1)$$

Матрица расстояний:

$$x_{ij}^{k} = dist(row_i[k], row_j[k]) \quad (2)$$

$$k \in [1, count_{fields}], i \in [1, count_{rows}], j \in index[i] \quad (3)$$

Алгоритм поиска дубликатов: нормализация

Проблема

По абсолютным значениям трудно судить о схожести двух записей.

Решение

Ввести нормализацию.

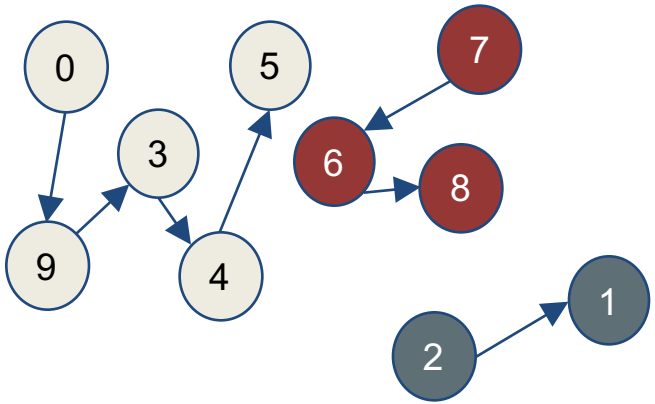
$$x_{ij}^k = \frac{length_{sum}^k - x_{ij}^k}{length_{sum}^k} \quad (4)$$

$$length_{ij}^k = length(row_i[k]) + length(row_j[k]) \quad (5)$$

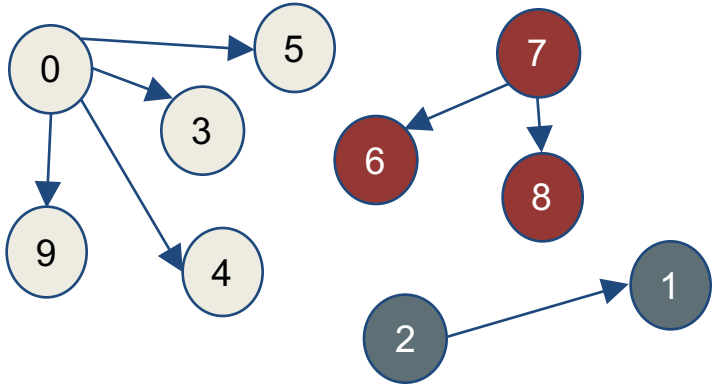
Алгоритм поиска дубликатов: поиск дубликатов

Идея: ввести пороговое значение и отобрать все записи, которые ему удовлетворяют.

Рекурсивный алгоритм



Линейный алгоритм



Устранение дубликатов

Варианты устранения:

1. Оставить наиболее полную запись
2. Оставить самую свежую запись
3. Объединить все записи в одну
4. Предоставить выбор эксперту

Применение алгоритма

	Искусственные данные	Реальные данные
Количество данных	1390	34405
Количество дубликатов	300 кластеров	422 кластера
Количество уникальных записей	1000	33976
Среднее количество ошибок	2.8	—

Табл. 2. Результаты применения алгоритма

Дальнейшие шаги

Как можно улучшить решение:

1. Добавить больше эвристических правил
2. Использовать сторонние базы данных для повышения качества данных
3. Автоматизация выбора пороговых значений
4. Автоматизация тестирования
5. Реализация функции добавления новых записей
6. Внедрение в ИС компании МИАЦ

Выводы

- Разработан набор методов для очистки данных.
- Реализован алгоритм поиска дубликатов.
- С помощью алгоритма было найдено подмножество дубликатов.
- Идеи, на которых построен алгоритм, заслуживают дальнейшего развития.

Спасибо за внимание

Исходный код: https://github.com/KirovVerst/record_linkage