Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники
Кафедра инженерной психологии и эргономики

УДК 331.101.1:004.62

Берникович
Тимур Ярославович

ЭРГОНОМИЧЕСКОЕ ОБЕСПЕЧЕНИЕ СИСТЕМЫ ПОИСКА ДУБЛИКАТОВ КОНТАКТНЫХ ДАННЫХ

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

1-23 80 08 – Психология труда, инженерная психология, эргономика

Магистрант Т.Я. Берникович

Научный руководитель
О.В. Клезович, кандидат
педагогических наук, доцент

Минск 2020

ВВЕДЕНИЕ

Достаточно часто в своей работе компании сталкиваются с дублированием данных. Дублирование данных могут подразделятся на два вида: простое (неизбыточное) и избыточное. Простое дублирование является допустимым в БД, в то время как дублирование данных избыточного характера может стать причиной ряда проблем.

Дедупликация данных представляет собой технологию, с помощью которой обнаруживаются и исключаются избыточные данные в дисковом хранилище. Это может быть осуществлено, например, посредством замены копий данных ссылками на первую копию. Это позволяет сократить объемы физических носителей для хранения тех же объемов данных.

Также дедупликация данных может быть определена как функция, которая позволяет уменьшить влияние избыточных данных на стоимость хранения. Если дедупликация данных включена, она оптимизирует свободное место в томе за счет проверки данных тома на наличие дублирующихся частей.

Дедупликация – это технология поиска повторяющихся данных на уровне файла, и замена их соответствующим указателем.

Таким образом, важной задачей является поиск избыточного дублирования информации и ее дедупликация.

Дублирование данных является показателем низкого качества БД, так как оно веден в конечном итоге к ошибочной интерпретации одного и того же объекта как двух разных.

Могут быть выделены два основных типов дублирования атрибутов в базе данных:

– дублирование атрибутов с жестко заданной структурой (форматом) содержания (коды классификаторов, идентификаторы в виде номеров телефонов, ИНН и т.п.);

– дублирование неполно структурированных атрибутов (имена собственные и названия, которые используются для идентификации: антропонимы, топонимы, названия предприятий, почтовые адреса и т. д.).

Для проверки слабоструктурированной информации на предмет дублирования данных используются алгоритмы нечетного поиска, позволяющие находить данные на основании неполного совпадения и оценки их релевантности – количественного критерия схожести. Следует учитывать, что данные алгоритмы не дают 100 %-ной гарантии от ошибок, то есть сохраняется вероятность того, что будут пропущены дублирующие данные или, наоборот, данные будут распознаны как дубликаты, не являясь таковыми. В связи с этим в рамках использования САП для достижения наилучшего результата может быть необходимо участие человека.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Актуальность данного исследования обусловлена тем, что компании, имеющие дело с контактными данными клиентов, часто сталкиваются с дубликатами и ошибками в базах с этой информацией после их ручного заполнения. Чтобы исправить проблему 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.

Цель и задачи исследования

Целью магистерской диссертации является разработка САП дубликатов данных, обладающей высокой надежностью.

Таким образом, разрабатываемая САП должна выполнять следующие функции:

– автоматический характер поиска дубликатов записей в справочнике компании и предоставление результатов, включающих информацию по заполненности и количестве использования по каждой найденной записи, для проведения анализа;

– обработка результатов поиска, путем автоматической замены ссылок с записей-дубликатов на оригинальную запись во всех записях справочников, карточках документов и вложениях задач системы;

– подробное логирование всех выполняемых операций в рамках обработки записей-дубликатов.

Для достижения поставленной цели необходимо выполнить следующие задачи:

– осуществление описания общих и эргономических принципов работы САП дубликатов данных;

– разработать структурную схему алгоритма поиска дубликатов данной в рамках создаваемой САП;

– разработать клиентскую и серверную часть приложения.

Объектом исследования являются системы автоматизированного поиска дубликатов контактных данных.

Предмет исследования – повышение производительности систем поиска дубликатов контактных данных путем повышения

Результаты исследования: практические рекомендации по разработке эргономичной и надежной САП дубликатов данных.

Основное содержание работы

Во **введении** объясняется проблематика тема, приводятся задачи и цели подобных систем. Рассматривается статьи сходного содержания.

В **первой главе** рассмотрены сущность эргономического обеспечения системы поиска дубликатов данных и их дедупликации, определены особенности проектирования современных СЧМС, проведен обзор методов поиска дубликатов данных.

Во **второй главе** приводится описание особенностей разработки САП дубликатов данных. Рассматриваются возможные СУБД и языки программирования, которые могут помочь решить поставленные задачи. Приводится обоснование методик, которые будут использованы для оценки надежности САП дубликатов данных.

В **третьей главе** описаны общие и эргономические принципы работы САП дубликатов данных. Осуществляется разработка алгоритма поиска дубликатов контактных данных, разработка серверной части, создание и подключение базы данных. Также приводятся расчеты надежности разработанной САП.

ЗАКЛЮЧЕНИЕ

В рамках изучения основ эргономического обеспечения системы поиска дубликатов контактных данных было выявлено, что целью магистерской диссертации является разработка САП дубликатов данных, обладающей высокой надежностью.

Таким образом, разрабатываемая САП должна выполнять следующие функции: автоматический поиска дубликатов записей в справочнике компании и предоставление результатов, включающих информацию по заполненности и количестве использования по каждой найденной записи, для проведения анализа; обработка результатов поиска, путем автоматической замены ссылок с записей-дубликатов на оригинальную запись во всех записях справочников, карточках документов и вложениях задач системы; подробное логирование всех выполняемых операций в рамках обработки записей-дубликатов.

В рамках теоретического обоснования разработка было проведено обоснование языка программирования и СУБД, а также приведено обоснование методик, используемых для оценки надежности программного средства – разработанной системы автоматизированного поиска.

Результатом обоснования стало то, что для разработки САП дубликатов данных будет использован язык программирования SQL. В качестве системы управления базами данных будет использована и MS Access.

Было определено, что задача поиска дубликатов может быть решена посредством алгоритмов нечеткого сравнения строк.

В рамках главы также отмечена высокая значимость надежности разрабатываемой САП и ее безотказного функционирования. Для соблюдения данного требования предполагается оценка надежности по трем моделям: модели сложности (в составе 5 типов метрик); интуитивной модели; модели Бернулли.

В рамках разработки программного модуля САП были решены все задачи, поставленные в задании, а именно: осуществлено описание общих и эргономических принципов работы САП дубликатов данных; разработана структурная схема алгоритма поиска дубликатов данной в рамках создаваемой САП; разработана серверная часть САП.

Также был разработан алгоритм поиска, в рамках которого было определено, что 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.

При обнаружении совпадений по всем компонентам такого типа контактных данных, как адрес, далее необходимо осуществление сравнение по такому атрибуту, как ФИО.

В целом, сравнение по двум полям позволят выявить дубликаты контактных данных клиентов в БД, однако для большей достоверности выявления дубликатов необходимо создание алгоритмов для нескольких других полей, которые могут быть использованы в контактных данных.

Было выявлено, что одним из способов оптимизации поиска в алгоритмы является добавление фильтраций по имени. В базу данных системы с именами к каждому имени целесообразно также добавление пола. Далее, при поиске дубликатов, записи фильтруются по полу, который должен соответствовать полу записи, дубликаты к которым ищутся.

Также была рассчитана надежность разработанного программного средства тремя способами: по модели сложности; по интуитивной модели; по модели Бернулли.

В результате расчетов выявлено, что: вероятность безотказной работы ИС по модели сложности равна 1,00; вероятность безотказной работы ИС по интуитивной модели равна 0,84; вероятность безотказной работы ИС по модели Бернулли равна 0,86.

Полученные результаты говорят о высокой надежности разработанной информационной системы, то есть достижения основной цели магистерского исследования.

Эргономическая экспертиза пользовательского интерфейса САП дубликатов контактных данных имеет высокий эргономический уровень качества оцениваемого объекта, а также его практически полное соответствие предъявляемым требованиям. При этом с целью дальнейшего повышения качества работы программы, необходимо включение в ее функционал подсказок о следующих шагах работы в системе и предупреждений о нежелательных шагах и их последствиях, а также повышение видимости важных сообщений путем добавления цветовых сигналов. Кроме того, важно изменить содержание сообщений об ошибках и добавить сведения о правилах их устранения.

Список опубликованных работ

1. Бернікович, Т. Я. Разработка поискового алгоритма для системы поиска дубликатов в контактных данных / Бернікович Т. Я., Голушко И. Н.– Репозиторий БГУИР, 2020. – [Электронный ресурс]. – Режим доступа: https://libeldoc.bsuir.by/handle/123456789/39100.

2. Бернікович, Т. Я. Автоматизированная система исправлениая и дедупликации контактных данных / Бернікович Т. Я. – Репозиторий БГУИР, 2020. – [Электронный ресурс]. – Режим доступа: https://libeldoc.bsuir.by/handle/123456789/39105.