УДК 004

ОБЗОР ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ ОЧИСТКИ ДАННЫХ

Осипова Ю.Г., Фонотов А.М.
Донецкий национальный технический университет
кафедра автоматизированных систем управления
E-mail: rubynchik@yandex.ru, anastasf@gmail.com

Аннотация

Осипова Ю.Г., Фонотов А.М. *Обзор интеллектуальных методов очистки данных. Рассмотрены причины возникновения ошибок и их типы. Рассмотрены этапы очистки данных. Рассмотрены алгоритмы и средства очистки данных, которые применяются при интеграции в современных корпоративных информационных системах. Проанализированы инструменты Data Mining (Data Cleansing) и их применения для решения поставленной задачи.*

Общая постановка проблемы

Проблема интеграции данных возникает при:

объединении подсистем предприятия;

интеграции баз данных нескольких фирм, отделов, филиалов;

слиянии баз данных с целью их обработки и анализа.

При интеграции информации в единый источник данных (например, в Хранилища данных (ХД)), возникает проблема извлечения достоверных данных. Актуальность задачи повышения достоверности данных вытекает из целого ряда причин:

разносторонность объединяемой информации;

плохого качества хранимых данных из-за орфографических ошибок, неточностей ввода, сокращений;

отличия в структуре данных;

отличия в доменах, которые созданы для хранения одной и той же информации;

дублирование информации. Базы данных часто содержат поля значений и записи, которые относятся к одной и той же сущности, но не являются синтаксически идентичными, потому, что источники часто содержат избыточные данные в различных представлениях.

Проблема поиска дублирования данных, является важной помехой на пути к интеграции и очистки данных. Еще в большей степени это касается слабо структурированных данных, таких как печатные документы, статьи, веб-страницы. [1, 2].

Поэтому, для аналитической обработки существующих данных, применяют различные методы по их коррекции, исключении дубликатов и очистки. Таким образом, задача очистки данных в корпоративных информационных системах является актуальной.

Причины возникновения ошибок и их типы

Существуют три причины возникновения ошибок в данных:

В основном личные сведения вводятся людьми вручную. Из-за невнимательности допускаются опечатки в словах, не заполняются обязательные поля при анкетировании, сокращаются названия районов, улиц или других объектов, заносятся сведения не в те поля и т.д.

Не во всех существующих программах, в которые вносятся сведения, для вводимых данных настроены ограничения для их значения.

В огромных корпорациях сбор информации о клиентах ведется несколькими подразделениями, следовательно, при слиянии всех этих сведений в одну большую базу данных возникают проблемы с форматами однотипных данных.

Существует множество видов ошибок, которые не зависят от предметной области. Таких ошибок выделяют шесть типов:

противоречивость информации
аномальные значения
пропуски данных
шум
несоответствие форматов данных
ошибки ввода данных или опечатки
дублирование

Противоречивостью информации называется такая информация, которая не соответствует законам, правилам или действительности. Сначала решается, что именно необходимо считать противоречивым. Например, по законам Украины пенсионную карту меняют в случае изменения Ф.И.О., но если человек родился мужчиной, а вышел на пенсию женщиной, противоречивость отсутствует [1].

Аномальными значениями называются такие значения, которые сильно выбиваются в целом из общей картины. Чаще всего такие значения корректируют вручную. Это связано с тем, что такие средства прогнозирования ничего не знают о природе процессов. Поэтому любая аномалия будет восприниматься как совершенно нормальное значение. Из-за этого будет сильно искажаться картина будущего. Какой-то случайный провал или успех будет считаться закономерностью [1].

Пропусками данных называется такой тип ошибок, если в полях для заполнения отсутствуют данные или заполнены не до конца. Эта проблема считается очень серьезной для большинства ХД. Большинство методов прогнозирования исходят из предположения, что данные поступают равномерным постоянным потоком. На практике такое можно встретить редко. Поэтому одна из самых востребованных областей применения ХД - прогнозирование - оказывается реализованной некачественно или со значительными ограничениями [3].

Шум – это данные, в которых показания значительно выше или ниже оптимальных значений. Часто при анализе данных сталкиваются с шумами. Он не несет никакой ценной информации, только мешает четко разглядеть картину.

Несоответствие форматов данных. Несоответствие форматов данных называются однотипные данные, имеющие разные форматы представления.

Ошибки ввода данных или опечатки преобладают в любых данных, т.к. вводятся человеком. Опечатки – это такой тип ошибок, когда данные содержат пропущенные, лишние символы или искаженные данные.

Дублирование – это повторяющиеся данные. Повторение различных данных - самая распространенная ошибка при работе с данными, которые заносятся в ХД.

Этапы очистки данных.

Очистку данных делят на пять этапов [8, 9]:

анализ данных
определение порядка и правил преобразования данных
подтверждение
преобразования
противоток очищенных данных

На первом этапе подробно анализируют данные, чтобы выявить подлежащие удалению виды ошибок и неточностей. Используется два вида проверок данных: вручную или специальными программами. На этом этапе получают метаданные о свойствах и качества данных.

На втором этапе определяется порядок и правила преобразования данных. В зависимости от числа источников данных, степени их неоднородности и загрязненности, данные могут требовать достаточно обширного преобразования и очистки. Иногда для отображения источников общей модели данных используется трансляция схемы; для хранилищ данных обычно используется реляционное представление. Первые шаги по очистке могут уточнить или изменить описание проблем отдельных источников данных, а также подготовить данные для интеграции. Дальнейшие шаги должны быть направлены на интеграцию схемы или данных и устранение проблем множественных элементов, например, дубликатов. Для хранилищ в процессе работы по определению ETL (*Extract, Transform, Load* — дословно «извлечение, преобразование, загрузка» [5]) должны быть определены методы контроля и поток данных, подлежащий преобразованию и очистке.

На третьем этапе определяются два вещи: правильность и эффективность процесса и определение преобразования. Это осуществляется путем тестирования и оценивания. При анализе, проектировании и подтверждении может потребоваться множество итераций, например, в связи с тем, что некоторые ошибки становятся заметны только после проведения определенных преобразований.

На четвертом – осуществляется выполнение преобразований либо в процессе ETL для загрузки и обновления ХД, или же при ответе на запросы по множеству источников.

На пятом этапе происходит замена загрязненных данных в исходных источниках на очищенные. Это необходимо осуществить для того чтобы улучшенные данные попали также в унаследованные приложения и в дальнейшем при извлечении не требовали дополнительной очистки. Для хранилищ очищенные данные находятся в области хранения данных [1].

Методы и средства очистки данных в современных корпоративных информационных системах.

На сегодняшний день существуют огромное количество методов по очистке данных от ошибок и неточностей. Никто из специалистов не скажет, какой из них является самым эффективным, потому что каждый метод совершенно по-разному подходят к этой проблеме.

Данную проблему решают тремя разными способами[5, 6]:

простыми методами

методами, которые основываются на понятиях математической статистики

средства ETL

Простые методы (регулярные выражения, строгие формальные правила и т.д.) очень примитивны и могут решить данную задачу только частично, поэтому ученые решили задействовать математическую статистику.

Рассчитываются необходимые показатели по всем данным, которые есть в наличии, т.е. охватывает весь диапазон значений и принимаемых признаками. На основе полученных результатов одни методы могут выделить подозрительную информацию, которая сильно отличается от остальных, а другие – вычислить величины, которые предположительно более всего похожи на истинные. Таким образом, анализируя сведения с помощью статистических характеристик, оценивают общую картину данных и уже на ее фоне определяют возможные ошибки с последующим их исправлением на подобранные похожие значения.

Выделяют такие методы очистки данных: Устраняет такие типы ошибок вычисление частот появления значений. как аномалии, пропуски, неправдоподобие данных и опечатки. В этом методе подсчитываются частоты появления определенного значения в имеющихся данных. Сначала подсчитывают какое количество раз различные значения были введены. Затем сортируются их частоты по убыванию. Следовательно, в конце списка будут значения, которые реже всего пользователь вводил. Возможно, что в данных допускались опечатки, наведены значения или введены аномальные значения. Поэтому такие поля

подвергают дополнительной обработке и последующей замене. После обнаружения данных с низким качеством используют простой метод - анализ строк (данный метод позволяет подобрать неправильно введенному слову такое правильное значение, которое будет максимально похоже на него), с помощью него восстанавливают вероятные значения.

Вычисление средних значений. Устраняет пропуски. Вычисляют 3 типа значений: мода, медиана и среднее арифметическое значение. Если данные содержат большой разброс значений, то метод средних применяется не к отдельному объекту, а к целой группе. Все данные в этом случае разбиваются на группы, содержащие приблизительно однородные элементы с похожими признаками. Внутри каждой из них рассчитывается средняя величина, которая будет типична именно для тех объектов, которые входят в эту группу.

Интервальный метод. Используется, если данные являются не разнородными. Этим методом вычисляют сначала доверительный интервал, между границами которого с заданной вероятностью находятся истинные значения оцениваемых параметров. Доверительный интервал с вероятностью 95% для большого объема данных, подчиняющихся нормальному закону распределения, определяют по формуле:

$$\bar{x} - \frac{1.96xc}{\sqrt{n}} < x_i < \bar{x} + \frac{1.96xc}{\sqrt{n}},$$

где x_i – исследуемый ряд данных,

$\bar{x}$ – среднее арифметическое значение совокупности данных,

c – среднеквадратическое отклонение,

n – количество исследуемых данных.

Значения, не попавшие в этот интервал, отмечаются как потенциальные ошибки их заменяют уже подобранными значениями (например, средней арифметической величиной). Метод применяют для однородных данных [2].

Третий способ решения задачи является использование ETL средств для ХД.

ETL средства включают в себя три основных процесса:

извлечение данных из внешних источников

преобразование данных и их очистка

загрузка в ХД. [3]

Такие средства обеспечивают возможность сложных преобразований и большей части технологического процесса преобразования и очистки данных. Общей проблемой средств ETL являются ограниченные за счет собственных API и форматов метаданных возможности взаимодействия, усложняющие совместное использование различных средств. Процесс преобразования выполняется либо системой, интерпретирующей специфические преобразования в процессе работы, либо откомпилированным кодом [5].

Зачастую инструменты ETL применяют для очищения персональных данных (Ф.И.О., адреса, с исключением дубликатов). Преобразования осуществляется двумя способами: либо в форме библиотеки правил заранее, либо пользователем в интерактивном режиме. Также данные могут быть автоматически получены и с помощью средств согласования схемы. Такие средства, за счет ограниченной области своего применения, обычно очень эффективны, но и имеют свой недостаток: они нуждаются в дополнении другими инструментами для работы с широким спектром проблем преобразования и очистки.

Очистка данных может выполнять одну или несколько функций. Такие как:

парсинг. Т.к. в зачастую инструменты ETL используют для очищения персональных данных поэтому при парсинге имя и адрес клиента будут хранится свободным форматом в в текстовых полях. Парсинг – это грамматический или лексический анализ текста. При выполнении парсинга ведется деление полей на атомарные значения.

проверка допустимости. В России и США существуют стандартные электронные каталоги, с помощью которых можно проверить правильность адресов как внутри страны так и международных. Некоторые приложения объединяются с такими программами, с помощью которых можно сверить данные.

стандартизация. Во многих странах существуют общепринятые сокращения, например для Почтовой службы, поэтому различные сокращения преобразуются в более понятные значения для этих служб (Улица или ул., или ул).

согласование и консолидация. После очищения персональных данных, таких как имя и адрес, для устранения дублирования информации о клиентах из разных источников применяется программа согласования. Почти все средства содержат алгоритмы расстановки приоритетов между полями (в процессе согласования) и контроля очередности сравнения полей.

улучшение. Существуют программы позволяющие по именам определять пол или по адресу определять долготу и широту указанной местности. Но больше всего популярностью пользуется программы, которые предоставляют по клиентским профайлам психографическую и демографическую информацию.[4]

Заключение

Не смотря на то, что существуют множество платформ, систем, инструментов для преобразования и очистки данных, их все равно не хватает. Эти средства идеально не уберут дублирование, потери данных, не соответствия. Поэтому и сейчас специалисты пытаются найти оптимальные вариации для решения очистки данных

Список литературы

1 Беликова Александра, Очистка данных, Интернет-ресурс, - Режим доступа URL: http://www.basegroup.ru/library/cleaning/person_data_part1/

2 Алексей Арустамов, Предобработка и очистка данных перед загрузкой в хранилище, Интернет-ресурс, - Режим доступа URL: http://www.basegroup.ru/library/dw_olap/dataclearing/

3 Роналд Фоурино, Электронное качество данных: скрытая перспектива очистки данных, Интернет-ресурс, - Режим доступа URL: http://www.iso.ru/print/rus/document5820.phtml

4 Mikhail Bilenko and Raymond J. Mooney. Adaptive Duplicate Detection Using Learnable String Similarity Measures // Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), - Washington DC, -2003, -C. 39-48.

5 U. Y. Nahm and R. J. Mooney. Using information extraction to aid the discovery of prediction rules from texts. // In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining, - Boston, - MA, -2000.

6 W. W. Cohen, H. Kautz, and D. McAllester. Hardening soft information sources. In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000), -Boston, -MA, -2000.

7 J. Joanne Zhu and Lyle H. Ungar. String Edit Analysis for Merging Databases. Интернет-ресурс, - Режим доступа URL: citeseerx.ist.psu.edu

8 Erhard Rahm, Hong Hai Do, Data Cleaning: Problems and Current Approaches, Интернет-ресурс, - Режим доступа URL: http://dbs.uni-leipzig.de

9 Heiko Müller, Johann-Christoph Freytag. Problems, Methods, and Challenges in Comprehensive Data Cleansing. Интернет-ресурс, - Режим доступа URL: www.dbis.informatik.hu-berlin.de