МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ И КОМПЬЮТЕРНЫХ СЕТЕЙ

УДК 004. 652.5

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ОЧИСТКИ ДАННЫХ ПРИ РЕИНЖИНИРИНГЕ БАЗ ДАННЫХ

А. И. Баранчиков, д.т.н., доцент, профессор кафедры ЭВМ РГРТУ, Рязань, Россия;
orcid.org/0000-0003-4133-7489, e-mail: alexib@inbox.ru
И. И. Яковлев, аспирант РГРТУ, Рязань, Россия;
orcid.org/0000-0002-3813-0455, e-mail: ivan.format@gmail.ru
И. А. Клюева, программист РГРТУ, Рязань, Россия;
orcid.org/ 0000-0002-0392-3228, e-mail: i.klyueva-job@yandex.ru

Целью работы *является применение алгоритмов очистки данных, ориентированных на сбор данных для хранилищ данных, к решению задачи реинжиниринга баз данных. Основными* ***задачами*** *являются выбор метода очистки, его модификация для новых целей и применение при реинжиниринге.* ***Актуальность*** *исследования заключается в применении известных методов работы с данными к решению новых задач. Приводится классификация проблем реинжиниринга в зависимости от количества баз данных, анализируются различные уровни работы алгоритма, и даётся описание основных его этапов. Производятся выбор и модификация алгоритма очистки для реинжиниринга баз данных, оценка перспектив его использования. Приводится пример работы алгоритма.* ***Результатом*** *работы является алгоритм, позволяющий применить методы очистки данных, использующиеся при организации и работе хранилищ данных, к задачам реинжиниринга баз данных.*

Ключевые слова: *данные, база данных, структура данных, очистка данных, реинжиниринг, предметная область, атрибут, профайлинг, data mining.*

DOI: 10.21667/1995-4565-2021-76-35-41

Введение

В настоящее время объёмы накопленной информации постоянно увеличиваются, а вместе с ними нарастают сложность и размер хранилищ и баз данных. Со временем работа баз данных может перестать удовлетворять потребностям заказчика по целому ряду причин: рост совокупной стоимости владения информационной системой, потеря актуальности логической и физической структуры, устаревание механизмов и технологий, изменение требований к информационной инфраструктуре, ошибки при проектировании, накопление некорректных данных, медленная работа и т.д. Решением этих проблем является перепроектирование или реинжиниринг баз данных – совокупность действий по миграции и модернизации данных и структур данных. При этом реинжиниринг может затронуть не только наследственные системы, но и новые разработки. В данной статье рассматривается применение алгоритмов очистки данных, ориентированных на сбор данных для хранилищ данных, и их адаптация для задачи реинжиниринга данных

Теоретическая часть

В настоящее время реляционные системы, теоретические основы которых были заложены Эдгаром Коддом, являются одним из наиболее популярных принципов организации данных. В трудах многих исследователей (Мейер [1], Дейт [2, 3], Дарвен [2], Крёнке [4], Ахо,

Хопкрофт, Ульман [5], Конноли, Бэгг, Страчан [6]) были заложены теоретические и практические основы реляционных баз данных, их проектирования и реализации, структуры и алгоритмы данных. Исследования проблем семантического анализа предметной области и реинжиниринга ключей и атрибутов производились в работах [7-10].

При реинжиниринге информационной системы с продолжительным временем жизни и развития данные в ней могут быть представлены в нескольких базах данных, разработанных в разное время, разными разработчиками, с использованием различных платформ и СУБД, на разных операционных системах, но оперирующих терминами одной предметной области. Возникает целый ряд проблем, связанных с неоднородностью данных и структур данных.

– В базах данных может отличаться структура (наименования полей, их типы, размеры и т.д.).

– В базах данных может различаться принцип классификации данных.

– В базах данных может различаться временная градация.

– В базах данных в качестве ключевого значения, однозначно идентифицирующего один и тот же объект предметной области, могут быть выбраны различные атрибуты.

– Одинаковые атрибуты в разных базах данных могут быть названы по-разному, и наоборот: одноимённые атрибуты могут оказаться различными по своей сути.

Одной из основных проблем как при очистке, так и при реинжиниринге данных является идентификация атрибутов из различных баз данных как эквивалентных. Для каждого имени атрибута может быть представлено всё множество различных значений, которое он принимает. Это множество и является доменом атрибута. Эквивалентные атрибуты − относящиеся к одной предметной области и имеющие одинаковые домены [1].

Целью данной работы является применение существующих алгоритмов очистки данных, использующихся при организации и работе хранилищ данных, к задачам реинжиниринга баз данных.

Применение алгоритмов очистки

В качестве предварительного этапа реинжиниринга, полностью сохраняющего свою актуальность, предлагается проведение очистки данных. Проблемы очистки и преобразования данных тесно связаны и должны решаться совместно и параллельно. Реструктуризация базы данных необходима при изменении схемы данных при изменении предметной области, при изменении технических требований, при переходе на новую СУБД или при интеграции множественных источников данных. При этом можно выделить классы проблем в зависимости от количества баз данных, участвующих в реинжиниринге (рисунок 1).

Рисунок 1 – Классификация проблем реинжиниринга в зависимости от количества БД
Figure 1 – Classification of reengineering problems depending on the number of databases

Решение проблем уровня схемы (актуализация, трансляция, интеграция) неизбежно отражается и на элементах данных, а вот проблемы уровня элемента могут быть незаметны на уровне схемы. При этом проблемы отдельной базы данных с высокой вероятностью встречаются при обработке множества баз данных одной предметной области.

Из-за большого количества потенциальных различий между базами данных, очистка должна быть проведена на следующих уровнях.

– на уровне ячейки таблицы реляционной базы данных (орфографические ошибки, опечатки, отсутствующие, закодированные или фиктивные значения, логически неверные значения).

– на уровне записи (противоречие данных в разных полях одного объекта предметной области).

– на уровне таблицы базы данных (кардинальные различия информации, хранящейся в одной таблице и описывающей разные объекты: нарушение уникальности, дублирующиеся и противоречивые записи).

– на уровне одиночной базы данных (нарушения целостности данных).

– на уровне множества баз данных (неоднородность структур данных) [11].

В качестве основных этапов работы алгоритма могут быть выделены следующие.

– Этап выявления проблем в данных (профайлинг данных – грубый анализ значений отдельных атрибутов для получения их характеристик: типа, длины, спектра, частот дискретных значений, наличия неопределённых значений и т.д.; data mining – процессы в больших объёмах однородной информации: группировка, обобщение, поиск последовательностей и ассоциативных связей, извлечение функциональных зависимостей между атрибутами).

– Этап составления правил очистки данных (правила должны учитывать количество источников данных, степень их различия и загрязнённости данных; поэтапно должны быть проведены очистки на всех уровнях, а данные подготовлены к интеграции с помощью обработки программными средствами).

– Этап проверки составленных правил (невозможно сразу сгенерировать идеальные правила очистки, этот процесс является итерационным или циклическим с возвратом к начальным значениям и исследованием результата преобразований для достижения результата, который может быть признан наилучшим).

– Этап очистки данных (преобразование данных по выше определённым правилам – сначала внутри каждой базы данных, потом непосредственная интеграция множества баз данных, которая становится этапом реинжиниринга).

Над отдельными базами данных производятся следующие действия.

– Расщепление атрибутов (для повышения точности и устранения дублирования информации атрибуты строкового типа преобразуются к отдельным значениям, после чего работа идёт уже с новыми атомарными атрибутами).

– Проверка допустимости и исправления (элементы проверяются на наличие пустых значений и ошибок, которые при возможности исправляются на основе заранее подготовленных справочников или найденных атрибутивных зависимостей).

– Стандартизация (приведение всех однотипных записей к установленному формату, наиболее удобному для дальнейшей работы и интеграции).

После завершения работ внутри каждой базы данных производится замена исходных данных очищенными. Это позволяет сэкономить ресурсы при проведении повторных очисток, повышает качество данных и позволяет избежать избыточных объёмов данных.

Выбор и модификация алгоритма для реинжиниринга

Для решения задач реинжиниринга адаптируется следующий этап работы алгоритма очистки данных – интеграция источников данных. Его работу можно разделить на следующие основные процедуры.

– Консолидация данных, относящихся к одной сущности предметной области (для того, чтобы определить, являются ли записи эквивалентными, используются идентификационные атрибуты или комбинации атрибутов – простые или составные первичные ключи и другие атрибуты и их сочетания, которые позволяют однозначно определить объект).

– Объединение объектов (в отдельную новую запись собирается вся информация по данной сущности; в случае необходимости информация корректируется и дополняется из различных баз данных; таким образом формируется новая структура данных, являющаяся результатом работы алгоритмов очистки и реинжиниринга баз данных).

– Удаление дубликатов (производится как на этапе работы с отдельными базами данных, так и после объединения; главная цель – не допустить появления в итоговой базе данных нескольких записей, соответствующих одной сущности предметной области).

Схема модифицированного алгоритма выглядит следующим образом (рисунок 2).

Рисунок 2 – Схема модифицированного алгоритма
Figure 2 – Modified Algorithm Scheme

Таким образом, преобразование структур разнородных баз данных, построенных на разной основе и с применением различных технических средств, разными разработчиками и для разных систем управления базами данных с применением алгоритмов очистки данных позволяет достичь крайне важных и полезных при реинжиниринге баз данных результатов.

Пример работы алгоритма

Рассмотрим пример работы модифицированного алгоритма очистки данных, ориентированного на объединение схем нескольких баз данных с целью объединения предметных областей и решения одной из основных проблем очистки и консолидации данных из множества источников – выявления эквивалентных сущностей, также называемой проблемой удаления или проблемой идентичности объекта. При этом часть информации об одном и том же объ-

екте будет дублирующейся (избыточной, которая должна удаляться), а часть дополняющей (которая должна объединяться, чтобы объект получил наиболее полное и точное представление). В таблицах 1 и 2 представлена информация об одном и том же объекте предметной области (клиентах) из разных реляционных баз данных.

Таблица 1 – Клиент
Table 1 – Client

Client ID	*Name*	*Street*	*City*	*Sex*
22	Emily Johnson	Surf street, 24	Sacramento	0
33	John Johnson	Catalina St 3	Los Angeeles California	1

Таблица 2 – Клиент
Table 2 – Customer

Customer's No.	*Surname*	*Name*	*Address*	*Genus*	*Phone*
22	johnson	olivia		F	+0 (683) 55 96 87 45
555	Johnson	J.	3 Catalina Street LA 65432-9876,	M	69876485245

Разберём данный пример поэтапно. Первым шагом работы алгоритма является выявление проблем. В данных таблицах мы видим ряд аномалий данных. Рассмотрим их на различных уровнях. На уровне ячейки таблицы: различные обозначения полов (0/1, M/F), опечатки (Angeeles), пустые значения (адрес в первой строке второй таблицы), сокращения (имя во второй строке второй таблицы). На уровне таблиц: различные представления адреса в первой странице, разный формат записи телефонных номеров во второй таблице. На уровне баз данных: конфликт имён таблиц (Clients и Customers), конфликты имён атрибутов (одинаковые по сути, но различные по названию Sex и Genus, одинаковые по названию, но разные по сути Name и Name), различие схем данных (различное представление имён и адресов). Также очевидно, что в разных базах данных используются разные уникальные идентификаторы (Client ID и Customer's No.), при этом предполагаемо дублирующимся записям соответствуют разные номера (33 и 555), а очевидно различным – одинаковые (22).

На этапе реинжиниринга и очистки можно прийти к решению проблем данных. При этом сначала разрешаются противоречия схем данных; после расщепления атрибутов (поля, содержащие имя должны быть разделены на атомарные: отдельные имя и фамилия; содержащие адрес – на индекс, штат, город, улицу и номер дома), проверки допустимости, исправления ошибок и стандартизации (приведение сведений о поле к общему виду, создание нового уникального идентификатора удаление не несущих теперь информации старых) производится анализ на наличие возможности дублирования данных. При сопоставлении данных об адресе, имени и значеняй Sex/Genus алгоритмом будет выявлено наличие эквивалентных сущностей (вторая запись из первой и второй таблицы соответственно: у них совпадают имена, пол и адреса). После консолидации данных, объединения объектов и удаления дублирования итоговая таблица будет выглядеть следующим образом (таблица 3).

Таблица 3 – Клиенты
Table 3 – Customers

No.	*Surname*	*Name*	*Gender*	*Building*	*Street*	*City*	*State*	*Index*	*Phone*	*Client ID*	*Customer's No.*
1	Johnson	Emily	F	24	Surf Street	Sacramento	California			22	
2	Johnson	Olivia	F						835-596-8745		22
3	Johnson	John	M	3	Catalina Street	Los Angeles	California	65432-9876	987-648-5245	33	555

Приведём описание одного из шагов преобразования в формате SQL:99, производимых с первой базой данных:

```
CREATE VIEW Customers
(Surname, Name, Gender, Building, Street, City, State,
Index, Phone, CID, CNo)
AS
SELECT LastNameExtract(Name), FirstNameExtract (Name),
Sex, Street, CityExtract (City),
StateExtract (City),
ZIPExtract (City), CID
FROM Customer
```

Заключение

Таким образом, полученный алгоритм позволяет применить методы очистки данных, использующиеся при организации и работе хранилищ данных, к задачам реинжиниринга баз данных. Перспективой использования данного алгоритма является его применение для качественного выявления эквивалентных атрибутов, что поможет значительно повысить качество извлечения семантических зависимостей [12].

Библиографический список

1. **Мейер Д.** Теория реляционных баз данных: пер. с англ. М.: Мир. 1987. 608 с.

2. **Дейт К. Дж., Дарвен Хью.** Основы будущих систем баз данных. Третий манифест / Пер. С. Д. Кузнецов, Т. А. Кузнецова. Издательство Янус-К. 2004. 656 с.

3. **Дейт К.** Руководство по реляционной системе DB2. М.: Финансы и статистика. 1988. 320 с.

4. **Крёнке Д.** Теория и практика построения баз данных. 8-е изд. СПБ: Питер. 2003. 800 с.

5. **Ахо Альфред В., Хопкрофт Джон Э., Ульман Джеффри Д.** Структуры данных и алгоритмы: пер. с англ. М.: Вильямс. 2000. 384 с.

6. **Конноли Т., Бэгг К., Страчан А.** Базы данных: проектирование, реализация и сопровождение. Теория и практика. 2-е изд.: пер. с англ. М.: Изд. Дом «Вильямс». 2000. 1120 с.

7. **Баранчиков А. И., Алпатова А. В.** Реинжиниринг ключей в отношениях реляционных баз данных // Вестник Рязанского государственного радиотехнического университета. 2008. № 26. С. 58-62.

8. **Баранчиков А. И., Кухарев С. Э.** Алгоритмы реинжиниринга атрибутов конфиденциальности в реляционных базах данных // Вестник Рязанского государственного радиотехнического университета. 2011. № 35. С. 72-75.

9. **Баранчиков А. И., Нгуен Н. З.** Реинжиниринг существующих баз данных информационных систем // Методы и средства обработки и хранения информации: межвуз. сб. науч. тр. 2017. С. 62-64.

10. **Баранчиков А. И., Нгуен Н. З.** Разработка алгоритма определения семантических зависимостей в информационных структурах // Наука, образование, инновации: актуальные вопросы и современные аспекты: сб. статей междунар. науч.-практ. конф. МЦНС «Наука и просвещение». 2020. С. 33-36.

11. **Барсегян А. А., Куприянов М. С., Холод И. И., Тесс М. Д., Елизаров С. И.** Анализ данных и процессов: учеб. пособие. 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.

12. **Баранчиков А. И., Нгуен Н. З.** Алгоритм обнаружения закономерностей в данных, позволяющих проверить отсутствие функциональной зависимости // Cloud of Science. 2019. Т. 6. №. 4. С. 693-700.

UDC 004. 652.5

EMPLOYMENT DATA CLEANING METHODS IN DATABASE REENGINEERING

A. I. Baranchikov, Dr. in technical sciences, full professor, RSREU, Ryazan, Russia;
orcid.org/0000-0003-4133-7489, e-mail: alexib@inbox.ru
I. I. Yakovlev, post-graduate student, RSREU, Ryazan, Russia;
orcid.org/0000-0002-3813-0455, e-mail: ivan.format@gmail.ru
I. A. Klyueva, programmer, RSREU, Ryazan, Russia;
orcid.org/0000-0002-0392-3228, e-mail: i.klyueva-job@yandex.ru

The aim of the work *is to apply data cleansing algorithms focused on collecting data for data warehouses to solve the problem of database reengineering.* ***The main tasks*** *are the choice of a cleaning method, its modification for new purposes and its application in reengineering.* ***The relevance*** *of the research lies in the application of well-known methods to work with data in order to solve new problems. The article provides a classification of reengineering problems depending on the number of databases, analyzes various levels of the algorithm, and describes its main stages. The selection and modification of the cleaning algorithm for database reengineering is carried out, and the prospects for its use are assessed. An example of the algorithm operation is given.* ***The result*** *of the work is an algorithm that allows you to apply the data cleaning methods used in organization and operation of data warehouses to the tasks of database reengineering.*

Key words: *data, database, data structure, data cleansing, reengineering, subject area, attribute, profiling, data mining.*

DOI: 10.21667/1995-4565-2021-76-35-41

References

1. **Mejer D.** *Teorija reljacionnyh baz dannyh*, Moscow, Mir, 1987, 608 p. (in Russian).
2. **Dejt K. Dzh., Darven H'ju.** *Osnovy budushhih sistem baz dannyh*, Moscow, Janus-K, 2004, 656 p. (in Russian).
3. **Dejt K.** *Rukovodstvo po reljacionnoj sisteme DB2*. Moscow: Finansy i statistika, 1988, 320 p. (in Russian).
4. **Krjonke D.** *Teorija i praktika postroenija baz dannyh*. Issue 8. Saint-Petersburg: Piter, 2003, 800 p. (in Russian).
5. **Aho Al'fred V., Hopkroft Dzhon Je., Ul'man Dzheffri D.** *Struktury dannyh i algoritmy*. Moscow: Vil'jams, 2000, 384 p. (in Russian).
6. **Konnoli T., Bjegg K., Strachan A.** *Bazy dannyh: proektirovanie, realizacija i soprovozhdenie. Teorija i praktika*, Issue 2. Moscow, Vil'jams, 2000, 1120 p. (in Russian).
7. **Baranchikov A. I., Alpatova A. V.** Reinzhiniring kljuchej v otnoshenijah reljacionnyh baz dannyh. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 2008, no. 26, pp. 58-62. (in Russian).
8. **Baranchikov A. I., Kuharev S. Je.** Algoritmy reinzhiniringa atributov konfidencial'nosti v reljacionnyh bazah dannyh. *Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta*. 2011, no. 35, pp. 72-75. (in Russian).
9. **Baranchikov A. I., Nguen N. Z.** Reinzhiniring sushhestvujushhih baz dannyh informacionnyh sistem. *Metody i sredstva obrabotki i hranenija informacii: mezhvuz. sb. nauch. tr.* 2017, pp. 62-64. (in Russian).
10. **Baranchikov A. I., Nguen N. Z.** Razrabotka algoritma opredelenija semanticheskih zavisimostej v informacionnyh strukturah. *Nauka, obrazovanie, innovacii: aktual'nye voprosy i sovremennye aspekty: sb. statej mezhdunar. nauch.-prakt. konf. MCNS «Nauka i prosveshhenie».* 2020, pp. 33-36. (in Russian).
11. **Barsegjan A. A., Kuprijanov M. S., Holod I. I., Tess M. D., Elizarov S. I.** *Analiz dannyh i processov*: ucheb. posobie, issue 3, Saint-Petersburg: BHV-Peterburg. 2009. 512 p. (in Russian).
12. **Baranchikov A. I., Nguen N. Z.** Algoritm obnaruzhenija zakonomernostej v dannyh, pozvoljajushhih proverit' otsutstvie funkcional'noj zavisimosti. *Cloud of Science*. 2019, vol. 6, no. 4, pp. 693-700. (in Russian).