IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.
www.ijiset.com
ISSN 2348 – 7968

Current Approaches to Data Cleaning in DB&DW System

1.G.Silambarasan 2.K.Mohan

[1]**Asst.professor, Dept.of.Computer Application, Annai Vailankanni Arts&Science College Thanjavur-7.**

[2]**Asst.professor, Dept.of.Computer Science, Annai Vailankanni Arts&Science College Thanjavur-7.**

Abstract

We classify data quality problems that are addressed by data cleaning and provide an overview of the main solution approaches. Data cleaning is especially required when integrating heterogeneous data sources and should be addressed together with schema-related data transformations. In data warehouses, data cleaning is a major part of the so-called ETL process. We also discuss current tool support for data cleaning.

Keyword: Data warehouse, Mining tools

I. Introduction

***Data cleaning*, also called *data cleansing* or *scrubbing*,**
deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

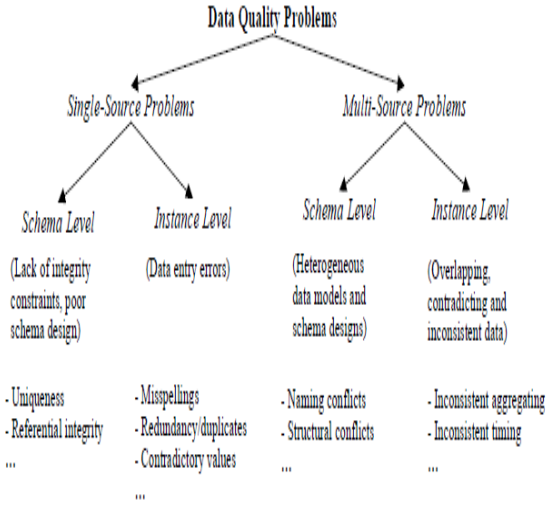
Steps of building a data warehouse: ETL process:

A data cleaning approach should satisfy several requirements. First of all, it should detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. The approach should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources. Furthermore, data cleaning should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata. Mapping functions for data cleaning and other data transformations should be specified in a declarative way and be reusable for other data sources as well as for query processing. Especially for data warehouses, a workflow infrastructure should be supported to execute all data transformation steps for multiple sources and large data sets in a reliable and efficient way.

ISSN 2348 – 7968

2 Data cleaning problems

This section classifies the major data quality problems to be solved by data cleaning and data transformation. As we will see, these problems are closely related and should thus be treated in a uniform way. Data transformations are needed to support any changes in the structure, representation or content of data. These transformations become necessary in many situations, e.g., to deal with schema evolution, migrating a legacy system to a new information system, or when multiple data sources are to be integrated. As shown in Fig. we roughly distinguish between single-source and multi-source problems and between schema- and instance-related problems. Schema-level problems of course are also reflected in the instances, they can be addressed at the schema level by an improved schema design (schema evolution), schema translation and schema integration. Instance-level problems, on the other hand, refer to errors and



inconsistencies in the actual data contents which are not visible at the schema level. They are the primary focus of data cleaning. Fig. also indicates some typical problems for the various cases. While not shown in Fig. the single-source problems occur (with increased likelihood) in the multi-source case, too, besides specific multi-source problems.

2.1 Single-source problems

The data quality of a source largely depends on the degree to which it is governed by schema and integrity constraints controlling permissible data values. For sources without schema, such as files, there are few restrictions on what data can be entered and stored, giving rise to a high probability of errors and inconsistencies. Database systems, on the other hand, enforce restrictions of a specific data model (e.g., the relational approach requires simple attribute values, referential integrity, etc.) as well as application-specific integrity constraints. Schema-related data

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date - birth date) should hold
Record type	Uniqueness violation	emp=(name="John Smith", SSN="123456") emp=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

quality problems thus occur because of the lack of appropriate model-specific or application-specific integrity constraints,

ISSN 2348 – 7968

e.g., due to data model limitations or poor schema design, or because only a few integrity constraints were defined to limit the overhead for integrity control. Instance-specific problems relate to errors and inconsistencies that cannot be prevented at the schema level (e.g., misspellings).
Examples for single-source problems at schema level (violated integrity constraints)

2.2 Multi-source problems

The problems present in single sources are aggravated when multiple sources need to be integrated. Each source may contain dirty data and the data in the sources may be represented differently, overlap or contradict. This is because the sources are typically developed, deployed and maintained independently to serve specific needs. This results in a large degree of heterogeneity w.r.t. data management systems, data models, schema designs and the actual data. A main problem for cleaning data from multiple sources is to identify overlapping data, in particular matching records referring to the same real-world entity (e.g., customer). This problem is also referred to as the object identity problem [11], duplicate elimination or the merge/purge problem [15]. Frequently, the information is only partially redundant and the sources may complement each other by providing additional information about an entity. Thus duplicate information should be purged out and complementing information should be consolidated and merged in order to achieve a consistent view of real world entities.

Customers (integrated target with cleaned data)

No	LName	FName	Gender	Street	City	State	ZIP	Phone	Fax	CID	Cno
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Examples of multi-source problems at schema and instance level

3 .Data cleaning approaches

- Data analysis
- Definition of transformation workflow and mapping rules
- Verification
- Transformation
- Backflow of cleaned data

3.1 Data analysis

Metadata reflected in schemas is typically insufficient to assess the data quality of a source, especially if only a few integrity constraints are enforced. It is thus important to analyse the actual instances to

ISSN 2348 – 7968

obtain real (reengineered) metadata on data characteristics or unusual value patterns. This metadata helps finding data quality problems. Moreover, it can effectively contribute to identify attribute correspondences between source schemas (schema matching), based on which automatic data transformations can be derived There are two related approaches for data analysis, data profiling and data mining.

3.3 Conflict resolution

A set of transformation steps has to be specified and executed to resolve the various schema- and instance level data quality problems that are reflected in the data sources at hand. Several types of transformations are to be performed on the individual data sources in order to deal with single-source problems and to prepare for integration with other sources. In addition to a possible schema translation, these preparatory steps typically include:

• ***Extracting values from free-form attributes (attribute split):***

Free-form attributes often capture multiple individual values that should be extracted to achieve a more precise representation and support further cleaning steps such as instance matching and duplicate elimination. Typical examples are name and address fields (Table 2, Fig. 3, Fig. 4). Required transformations in this step are reordering of values within a field to deal with word transpositions, and value extraction for attribute splitting.

• ***Validation and correction:***

This step examines each source instance for data entry errors and tries to correct them automatically as far as possible. Spell checking based on dictionary lookup is useful for identifying and correcting misspellings. Furthermore, dictionaries on geographic names and zip codes help to correct address data. Attribute dependencies (birth date – age, total price – unit price / quantity, city – phone area code can be utilized to detect problems and substitute missing values or correct wrong values.

• ***Standardization:***

To facilitate instance matching and integration, attribute values should be converted to a consistent and uniform format. For example, date and time entries should be brought into a specific format; names and other string data should be converted to either upper or lower case, etc. Text data may be condensed and unified by performing stemming, removing prefixes, suffixes, and stop words. Furthermore, abbreviations and encoding schemes should consistently be resolved by consulting special synonym dictionaries or applying predefined conversion rules.

4 .Tool support

A large variety of tools is available on the market to support data transformation and data cleaning tasks, in particular for data warehousing.1 Some tools concentrate on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination. Due to their restricted domain, specialized tools typically perform very well but must be complemented by other tools to address the broad spectrum of

ISSN 2348 – 7968

transformation and cleaning problems. Other tools, e.g., ETL tools, provide comprehensive transformation and workflow capabilities to cover a large part of the data transformation and cleaning process. A general problem of ETL tools is their limited interoperability due to proprietary application programming interfaces (API) and proprietary metadata formats making it difficult to combine the functionality of several tools

4.2 Specialized cleaning tools

Specialized cleaning tools typically deal with a particular domain, mostly name and address data, or concentrate on duplicate elimination. The transformations are to be provided either in advance in the form of a rule library or interactively by the user. Alternatively, data transformations can automatically be derived from schema matching tools such as described in.

• ***Special domain cleaning****:*

Names and addresses are recorded in many sources and typically have high cardinality. For example, finding customer matches is very important for customer relationship management. A number of commercial tools,

(e.g)

- ✓ **IDCENTRIC(FirstLogic),**
- ✓ **PUREINTEGRATE(Oracle),**
- ✓ **QUICKADDRESS(QASSystems), REUNION(PitneyBowes),**
- ✓ **TRILLIUM (TrilliumSoftware)**

focus on cleaning this kind of data. They provide techniques such as extracting and transforming name and address information into individual standard elements, validating street names, cities, and zip codes, in combination with a matching facility based on the cleaned data. They incorporate a huge library of prespecified rules dealing with the problems commonly found in processing this data. For example, TRILLIUM's extraction (parser) and matcher module contains over 200,000 business rules. The tools also provide facilities to customize or extend the rule library with user-defined rules for specific needs.

• ***Duplicate elimination****:*

Sample tools for duplicate identification and elimination include datacleanse (edd), merge/purgelibrary(Sagent/QMSoftware),matchit (helpITSystems), and mastermerge (PitneyBowes). Usually, they require the data sources already be cleaned for matching. Several approaches for matching attribute values are supported; tools such as

5 Conclusions

We provided a classification of data quality problems in data sources differentiating between single- and multi-source and between schema- and instance-level problems. We further outlined the major steps for data transformation and data cleaning and emphasized the need to cover schema and instance-related data transformations in an integrated way. Furthermore, we provided an overview of commercial data cleaning tools. While the state-of-the-art in these tools is quite advanced, they do typically cover only part of the problem and still require substantial manual effort or self-programming. Furthermore, their interoperability is limited (proprietary APIs and metadata representations).

ISSN 2348 – 7968

References

[1] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: *Tools for Data Translation and Integration*. In [26]:3-8, 1999.

[2] Batini, C.; Lenzerini, M.; Navathe, S.B.: *A Comparative Analysis of Methodologies for Database Schema Integration*. In Computing Surveys 18(4):323-364, 1986.

[3] Lakshmanan, L.; Sadri, F.; Subramanian, I.N.: *SchemaSQL – A Language for Interoperability in Relational Multi-Database Systems*.

[4] Chaudhuri, S., Dayal, U.: *An Overview of Data Warehousing and OLAP Technology*.

[5] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: *Cleansing Data for Mining and Warehousing*. Proc. 10th Intl. Conf.