MINGZHEN WEI, New Mexico Institute of Mining and Technology, Socorro, NM; MARTHA CATHER, New Mexico Petroleum Recovery Research Center, Socorro, NM; ANDREW H. SUNG, New Mexico Institute of Mining and Technology, Socorro, NM;

Data Integration and Data Quality Control for the New Mexico Produced Water Chemistry Database

Abstract: Access to and management of very large data sets are important issues in today's petroleum industry. Data sets are created both through modern work processes, and through the migration and integration of legacy data sets into modern database management systems (DBMS). These very large data sets can be mined to discover interesting patterns and relationships, and discovery of such information may help increase the efficiency of oil and gas production. The key to productive knowledge discovery is having quick access to reliable data, and the successful migration and integration of legacy data sets is an important part of the process that will greatly affect the success of further functional analysis or knowledge mining.

The first part of this paper presents a global schema development process for data migration and integration using a context analysis method. The paper also includes a brief discussion of schema transformations and data transformations based on the context analysis method. The second part of the paper presents a case study of methods developed and applied in data quality control for the New Mexico Produced Water Chemistry Database (PWCD). This paper presents some important considerations in integrating data from multiple sources into one database and methods for data quality enhancement after data integration is completed.

Introduction

New Mexico produces some 450 million barrels of water each year as a byproduct of oil and gas production. [1] Not only is this water costly to raise, separate, and dispose of, but produced water in lines and facilities also causes problems such as scale and corrosion that can greatly affect production efficiency. Additionally, water is increasingly seen as a potential resource to be conserved and possibly reused if it is treated to meet specific quality guidelines. Easy access to produced water chemistry data is necessary for both solving water-related problems and also for understanding where potentially useful water resources might be located. In the PWCD project, water data was collected from a variety of sources, both paper and digital. Paper reports have different report formats, different items, different units representations, and are derived from a number of reporting companies and span a large period of time. Data in the paper forms was converted into digital format via two methods: manual entry, or scanning of paper forms and converting them to text using optical character recognition (OCR) techniques. The two methods are both known to be error prone. Manual entry can produce misspellings and typographical errors, and OCR may have numerous character recognition problems. The digital data for this project was compiled from a several source files of different file types such as plain text or Microsoft Excel spreadsheets. Schematic heterogeneity also exists in the digital data sources.

Data Integration

Data integration is the process of combining data from distributed and heterogeneous data sources. In data integration, schema transformations are used to transform different data models to an equivalent global schema with respect to the same subject, in which attribute conflicts [2] in the different data sources or solved. Common attribute conflicts include:

1. Naming conflicts: two types of naming conflicts exist in data integration. Homonyms refer to two attributes that are semantically unrelated but that may have the same name, while synonyms refer to two attributes that are semantically alike, but that may have different names.

2. Representation conflicts: two semantically similar attributes may have different representations; for example the social security number of a person can be represented as "123456789" or "123-45-6789".
3. Data scaling conflicts: two semantically similar attributes may be represented using different units and measures; for example chloride concentration can be represented in either parts per million (ppm) or milligrams per liter (mg/l).
4. Data precision conflicts; two semantically similar attributes might be represented using different precisions. For example, ion concentration may be represented as a numerical value, or as a descriptive value such as " >100."
5. Default value conflicts: two semantically similar attributes might have different default values, for example "null" and "-9999.99."

Context analysis [2] is one method of comparing two data sources and analyzing attribute context and equivalency. For this project, attribute name, content, scale, representation, data type, and data constraints of attributes were selected as major comparison contexts. The context analysis method compared different attributes using these contexts, determined their similarity, and detected transformations between them if needed. Table 1 shows three attributes that were selected from three different data sources. The attributes were all used to represent chloride concentration in water samples. From the table, one can see the difficulty in integrating data from these three sources. From the context comparison of sources 1 and 2, two attributes have different names, similar contents, different measurement scales, different representations, different data types, and incomparable constraints. From the comparison, it is seen that scale and data type conflicts are the main problems in integrating this data from sources 1 and 2. Scale conflicts were solved by schema transformation. In this case the transformation was a conversion factor used to transform between chloride concentration in mg/l and in ppm to achieve the same unit of measurement. The corresponding attribute in source 3 did not have a reported measurement scale, adding new challenges to schema transformation. Some "scale missing" problems were created during the data acquisition phase by simple omission of information. For these samples, tracking data back to original sources helped to fill in the missing information. For other samples, no unit of measurement was recorded in the original document or file. Data in this category was deleted from the final database in our project because they comprised only a few records of the thousands in the project.

Solving attribute conflicts contributes to smooth and proper schema design in data integration. The context analysis method provides a means to evaluate the similarity of contexts of attributes from different sources. In this project it was anticipated that new sources of data would continually be added to the database. Therefore, a global database schema was first constructed based on examination of some of the earlier and largest data sets. In the global schema, self-explanatory attribute names and proper data types were assigned based on the nature of the attribute values, like in "chloride_mgl" for concentration of chloride in mg/l. The global schema was improved iteratively with the integration of multiple sources by solving the attribute and entity conflicts and migrating the data sets into the final database. Figure 1 shows the main steps in global schema development. In Phase 1, schema transformations are made to solve attribute conflicts and schema extensions are made to include attributes that are not in the original global schema. At the end of Phase 1, the global schema should be schematically proper for the data set. In Phase 2, the schema is fine-tuned and data are transformed by solving entity or value conflicts. Incoming data are integrated and exceptions are recorded in the global database. At the end of Phase 2, the global schema is transformed properly for all previous data sets. The entire process is repeated for each successive data set. As an example of data transformations in Phase 2 for data type conflicts, in most data sources, chloride concentration is numerical; hence the data type of chloride in the global schema is set as numerical. However in data source 2, the chloride

concentration data type was defined as a character value of length 10, which indicated the existence of some descriptive data such as "<100" even though the majority of the data in this attribute domain were actually seen to be numbers. When integrating source 2 into the global schema, the data type for the attribute chloride did not change. For the descriptive exceptions seen in source 2, null values were put into the corresponding record and a note was made in the notation column to record the actual data from the original record.

Data Quality Control

Discussions on data quality problems in the literature [4, 5, 6] show that data quality problems can be expected in most data sets unless extraordinary efforts were taken to avoid or correct them. The Produced Water Chemistry Database contained many different types of data problems such as spelling and typographical errors, character-recognition related data quality problems, serious information incompleteness, non-standardized data, different formats, and so on. Problems were caused by the data acquisition methods, the data integration process, and the original data itself. Problems are classified into two broad categories: character data problems and numerical data problems. Character data problems are mainly related to misspelling, typographical errors, entity identity problems, non-standardized attributes, and duplicate records. Numerical data problems were mainly generated from character recognition problems, typographical errors, template setup problems in the data scanning process, and some problems in the original paper report forms.

Several methods were used to fix data quality problems in each category in our project. Figure 2 lists several of the data problems in PWCD and our corresponding solutions. In this section, we briefly describe the methods used and developed in our project to control the data quality. Some results are provided to show the effectiveness of the methods.

Context-based token analysis method [7]: This method was developed to identify wrong tokens (individual words or atomic strings) in character attribute domains such as the name of well, company, etc. The motivation behind this method was the difficulty of identifying wrong tokens in a very large data set. Small data sets can be cleaned manually, but cleaning of larger data sets must be a semi-automated process. The assumption is that, for tokens appearing in a domain more than once, the correct token should have a larger frequency than its wrong variations. Put it simply, this method counts the frequencies of similar tokens and the frequencies of their occurring contexts; compares the frequencies of similar tokens and their contexts; then determines which token is wrong based on results from the above steps. The following example shows the results of a similar token pair. In Table 2, the frequencies of two similar tokens "MANZANERES" and "MANZANERAS" are 1 and 5 respectively. The context set of "MANZANERES" is a subset of that of "MANZANERAS". These two tokens do not appear in the outer database (in data cleaning, the outer database refers to the data set that contains more standardized data). Results show evidence that "MANZANERES" is an incorrect wrong variation of "MANZANERAS" in the well name attribute in the PWCD. This method successfully identified many wrong tokens in the character attribute domains and improved the effectiveness of other data cleaning methods.

Record linking [8]: Record linking (or data linking or record linkage) is to link two records in two data sets by identifying semantically equivalent entities in these two data sets. The objectives are: 1) obtaining a comprehensive profile about a real-world subject and 2) obtaining complementary data about a real-world subject. The objective of data linking in our project was the second one due to a serious lack of complete data records in the PWCD. The PWCD contains information about samples of produced water from oil and gas producers in New Mexico, and while a second database (the ONGARD database) contains well information for oil and gas producing wells in New Mexico. They overlap in containing information about well name and number, API, and

location information. ONGARD contains standardized data about names, API's, and locations, while the original PWCD contained corresponding data that were not well defined. In our project, strings concatenated from well name and number were regarded as identifying information for oil and gas wells. During the data linking process, an approximate field-matching algorithm [9] was used to compare the similarity of these identifying strings in both databases and to find the equivalent entities in both databases. Location and standardized name information in ONGARD was used update corresponding information in the PWCD once conclusive linking was determined between two records. In record linking, it is important to validate the linkage using other available information, which in this study was the sparse location information in the PWCD. Table 3 shows two examples; "Tiffany 001" and "Box Canyon 004". "Tiffany 001" is the identifying string for records in both the databases, but the location information is not at all similar. Thus, linking can not be established between these two records. Also the linking decision cannot be made between records "Box Canyon 004" in the PWCD and "Little Box Canyon 004" in the ONGARD due to the lack of validation information.

Duplicate elimination [4, 6]: Duplication of records is another very common problem in data cleaning. Duplicate elimination identifies records with the same identifying information or with duplicate values in all attributes. In our project, the purpose of duplicate elimination was the latter reason. An approximate field-matching algorithm was extended to compare record strings and calculate their similarity. Records were considered as possible duplicate records when their similarities fell in certain range, like "similarity score >0.95". Final duplicate elimination decisions were made based on user examination.

Data quality control methods for numerical data problems: For numerical data problems, multiple methods were applied and developed to identify potentially incorrect numerical data in the PWCD. General statistical methods such as histograms and box plots, and general data summarization were used to evaluate the data quality of data sets. Although these methods are widely applied and simple, they provided an excellent "big picture" view of the data set. Table 4 presents some data summary results. From these results, problems in SP values (specific gravity) are easily seen by impossibly low and high values. Another method used is the measurement correlation cross-validation, such as correlations between different scale representations of same measurements (eg., mg/l and meq/l) and correlations between different measurements. Correlation cross-validation is an efficient method to identify abnormal data, especially those caused by typographical errors. Figure 3 shows an application of this method. By examining the correlation between specific gravity and total dissolved solids (TDS) for water samples, values far off the correlation patterns are noted for further error-checking.

Spatial outlier detection [10, 11]: Spatial outlier detection methods identify abnormal data in spatial distribution through analyzing the contrast of attribute values at spatial points with those of their surrounding neighbors. Points are identified as spatial outliers when values of attributes at one location are extremely large or small compared to their surrounding neighbors. This method employs the local inconsistency to detect outliers in spatial distribution as in geological data sets. Figure 4 shows one example of the spatial outlier detection results for TDS variation in samples from the Artesia Group of the Permian Basin in New Mexico. These methods help point out data that may be questionable because of its unusual variability with respect to values in neighboring areas. Careful examination and further trend analysis may be needed to determine which data is really incorrect.

Discussion and Conclusions

With the growing demand for data analysis and knowledge mining from very large databases, successful integration of large and diverse amounts of data is becoming increasingly important. This paper presents work a case study of data integration and data quality control for the Produced Water Chemistry Database for New Mexico. In our project, a global database schema was developed by gradually integrating successive data sets through schema transformations, data transformations, and schema extension. Schema transformations based on a context analysis method were used to solve attribute conflicts, and data transformations used in solving conflicts between schema and data. Data problems were classified into character data problems and numerical data problems, and different methods were applied or developed to solve different data problems:

- A context-based token analysis method was used to identify wrong tokens in character attribute domains;
- An improved record linking process was developed to link data in two databases and to update the poorer quality database with standardized data in the outer database;
- A duplicate elimination method was applied to check the record uniqueness in the database;
- General statistical methods and data summary and correlations between measurements were applied to detect data quality problems in given data set;
- Spatial outlier detection methods were developed to identify outliers in spatial distribution using different mechanisms.

These methods greatly improve the standardization, completeness, accuracy, and reliability of the PWCD.

References

1. Robert Lee, and etc, Strategies for Produced Water Handling in New Mexico, Proceedings of 47th Annual New Mexico Water Conference, October 9-11, 2002

2. Amit Sheth, Vipul Kashyap, So Far (Schematically), Yet So Near (Semantically), IFIP Transactions, Interoperable Database Systems (DS-5), ISSN: 0926-5473

3. Gonazalo Navarro, A Guided Tour to Approximate String Matching, ACM Computing Surveys, vol. 33(1), March 2001

4. Erhard Rahm, Honghai Do, Data Cleaning: Problems and Current Approaches, IEEE Bulletin of the Technical Committee on Data Engineering, 2000, 24, 4

5. Thomas C. Redman, the impact of poor data quality on the typical enterprise, association for computing machinery, communications of the ACM, Feb 1998

6. Theodore Johnson, Data Quality and Data Cleaning: An Overview, Proceedings of the 2003 ACM SIGMOD international conference on Management of Data

7. Mingzhen Wei, Andrew H. Sung, Martha Cather, Data Quality Control for New Mexico Produced Water Chemistry Database, to be in Proceedings of 5th Canadian International Petroleum Conference, June 2004, Calgary, Alberta, Canada

8. Newcombe H. B., Kennedy J. M., Axford S. J., James A. P. Automatic Linkage of Vital Records, Science vol. 130, No. 3381, Oct. 16, 1959

9. Alvaro E. Monge, Charles P. Elkan, The Field Matching Problem: Algorithms and Applications, Knowledge Discovery and Data Mining, 1996

10. Shekhar S., Lu C. T. , Zhang P., A Unified Approach to Spatial Outliers Detection, GeoInformatica, An International Journal on Advances of Computer Science for Geographic Information Systems, Vol. 7(2), June 2003

11. Mingzhen Wei, Andrew H. Sung, Martha Cather, Detecting Spatial Outliers Using Bipartite Outlier Detection Methods, to be in Proceedings of 2004 International Conference on Information and Knowledge Engineering, June 21-24, 2004, Las Vegas, Nevada

Appendices

Tables

Table 1 Attributes "Chloride_mgl", "chlorde_ppm", and "chloride" in Different Data Sources and Their Context Analysis Results

Source	Attribute	Name	Content	Scale	Representation	Data Type	Data Constraint
1	chloride_mgl	chloride_mgl	chloride	mgl	numerical data	numerical	N/A
2	chlorde_ppm	chlorde_ppm	chlorde	ppm	numerical + descriptive	Character(10)	N/A
Context analysis Result		4*	1*	3*	0**	0**	Incomparable
1	chloride_mgl	chloride_mgl	chloride	mgl	numerical data	numerical	N/A
3	chloride	chloride	chloride	N/A	numerical data	numerical	N/A
Context analysis Result		4*	0*	incomparable	same	Same	incomparable

Note: Results with * are derived from edit distance calculation of two identifying strings. (Edit distance is the minimum operations (like insertion, deletion, and substitution) needed to convert one string to the other string. For example, edit distance between "chloride" and "chlorde" is ED ("chloride", "chlorde") =1)

Results with ** are derived from categorical data comparison

Table 2 Results of Context-Based Token Analysis Method for "MANZANERES" and "MANZANERAS"

Token	Frequency of PWCD database	Context of occurrence in PWCD	Frequency of ONGARD database	Context of occurrence in ONGARD database
MANZANERES	1	MANZANERES WATER WELL	0	N/A
MANZANERAS	3	MANZANERAS WATER WELL MANZANERAS SUSA MANZANERAS SUSA WATER WELL MANZANERAS W. W MANZANERAS WATER WELL	0	N/A

Table 3 Potential Linking Records in PWCD and ONGARD

Data Source	Oil and gas producer`s entity string	Township	range	Section	County
PWCD	Tiffany 001	32	7	1	Rio Arriba
ONGARD	Tiffany 001	19S	38E	26	Lea
PWCD	Box Canyon 004				
ONGARD	Little Box Canyon 004	20S	21E	36	Eddy

Table 4 Data Summary Results for Partial Data in PWCD

measurement	maximum	minimum	mean	median	standard deviation
pH	12.9	0	7.4093	7.37	1.151
TP	7135.52	0	3.1857	1.015	125.9095
TDS (mg/l)	397575	0	61455	55000	75464
sodium (mg/l)	1503597	-7770	21750	4218	61459
sulfate (mg/l)	282966	0	3658	1165	25635

Figures

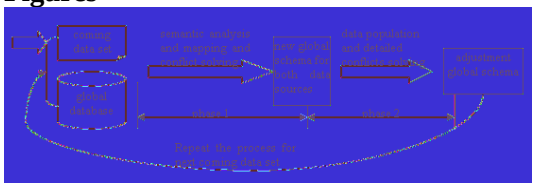

Figure 1 Incremental Schema Evolvement in Data Integration

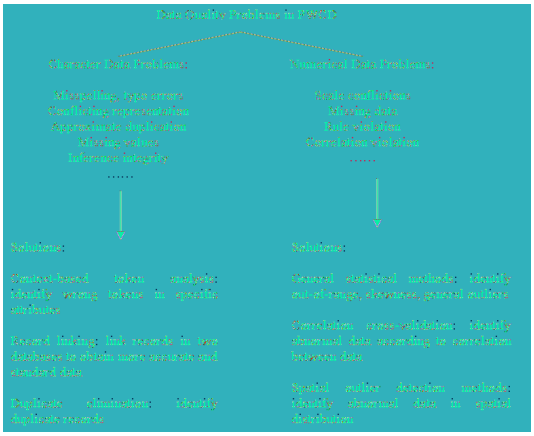

Figure 2 Data Quality Problems in PWCD and Partial Solutions

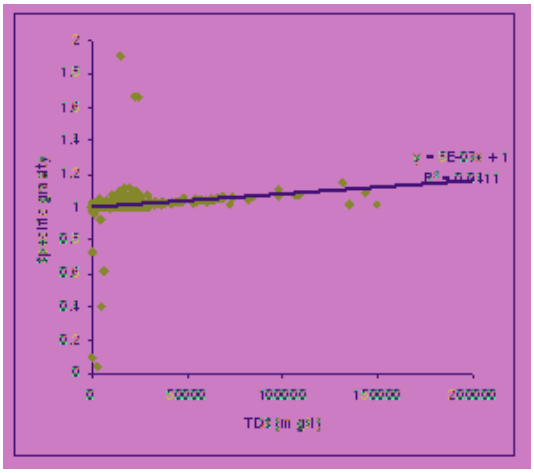

Figure 3 Correlation Cross-Validation between Specific Gravity and Total Dissolved Solids (TDS) in mg/l

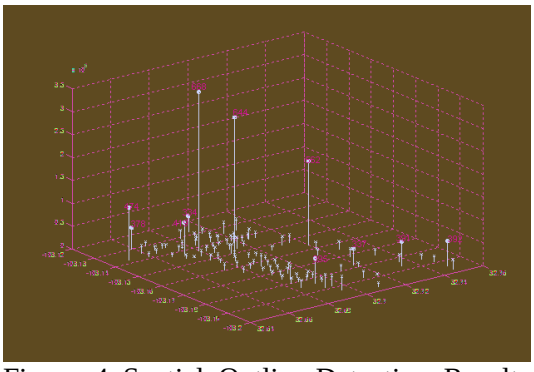

Figure 4 Spatial Outlier Detection Results (TDS Distribution in Part of Artesia Group in Permian Basin in New Mexico