A RULE-BASED EXPERT SYSTEM FOR HETEROGENEOUS DATA SOURCE INTEGRATION IN SMART GRID SYSTEMS

J. I. Guerrero*, PhD, Antonio García, Enrique Personal, PhD, Antonio Parejo, Francisco Pérez, PhD, and Carlos León, PhD
Department of Electronic Technology
University of Seville, Seville, Spain

ABSTRACT

The arrival of new technologies related to Smart Grids and the resulting ecosystem of applications and management systems pose challenges or problems to be solved. In this way, the compatibility with databases of the traditional and the new management systems, related to initiatives of new technologies, have given rise to different formats and architectures. Due to this, a heterogeneous data source integration system is essential to update these systems for the new Smart Grid reality. In this sense, there are several problems which need to be solved: information

* Corresponding Author Email: juaguealo@us.es.

integration, incomplete data model definition, understanding of database models, evolution of technologies and modelling information.

Additionally, it is necessary to take advantage of the information that Smart Grids provide. The Smart Grids must provide new services to the consumers and operators, integrating the information from all the partners, ensuring the information protection and security.

At first, this chapter briefly treats an analysis of the proposed problems and makes a bibliographical review. Following this review, it proposes a solution for heterogeneous data source integration in the information standard formats, based on Rule Based Expert System (RBES) to implement a metadata mining process. Later, it describes the process of automatic modelling in which the proposed RBES support in the data mining technique applications, based on the results of metadata mining process. Finally, it describes the application issues of the proposed solution in real cases.

Keywords: expert systems, metadata mining, smart grid, data integration, big data

INTRODUCTION

The Smart Grids have provided a great scope of solutions to manage and control the power grid. Additionally, the Smart Cities propose a wider vision of Smart Grids, extending this model to other utilities and services which could have a city: health, claims, transport, communication, etc. These new scenarios pose new advantages for the current society, increasing the quantity of information, the services, etc. The analysis of information and services is a very complex task. Traditionally, the analysis of information in all these cases could provide some patterns or relationships between parameters. Thus, the data mining and computational intelligence techniques are broadly applied to solve these problems. The utilization of these techniques can provide patterns or relationships between parameters, which would manually be very difficult to get.

Currently, the new technologies related to Smart Grids have increased the quantity of available information. This new information is provided by a wide variety of systems, which are mainly implemented by Intelligent

Information Systems (IISs). Thus, the intelligent analytical tools are essential to implement this type of system, coordinated with robust and secure infrastructures. However, the increasing of available information makes the manual analysis of information impossible, so the information should be analyzed using automatic and advanced techniques for data analysis. Additionally, the new technologies related to information management could improve the analysis techniques. The new architectures based on Hadoop, Spark, or other frameworks of information management are essential to implement the new scenario provided by Smart Grids. Moreover, the technologies related to High Performance Computing (HPC) increase the capability of systems analyzing, making it quicker and attaining a higher quantity of information.

However, the new ecosystem implemented in Smart Grids needs some special capabilities. These systems should be able to work with information of a traditional system, and also use this information to enrich the model with new information, so the deployment of these systems cannot be performed quickly. The update of all the facilities of the distribution grid take a lot of time, economic, and technical costs. Thus, in the starting stages of updating a power grid, traditional and modern power grids must coexist in the same grid.

PROBLEM DESCRIPTION

The traditional systems in the power distribution grids usually have databases with different data structure. The new technologies related to Smart Grids have provided the advantage of new and advanced functions. Although these new systems are based on the usage of sensor networks and information systems, the systems need the information from the older ones, integrating information from the heterogeneous data sources. In this sense, there are several problems which need to be solved: information integration, incomplete data model definition, understanding of database models, evolution of technologies, and modelling information.

Information Integration

The new systems need to take advantage of an old and new data sources. Thus, the integration of these heterogeneous data sources is very difficult, because each database has their own structure. This data source should be translated into a common format. In this way, the information standards provide a good source for a Common Information Models (CIM). Organizations, like International Electrotechnical Committee (IEC) and Distributed Management Task Force (DMTF), have provided each CIM standard. The IEC CIM is used in utility companies, mainly in power sector, and DMTF CIM is a generic model, mainly focused on the system management. Both standards are complex and demand a high level of knowledge in order to be applied on database system. Therefore, the automatic information integration could provide a supporting system in the implementation of new systems based on data from the old systems.

Additionally, the information integration could be necessary for the application of the analytic tools. In this case, the integration should be made of data warehouse structure. Thus, the system needs to generate new structures from old databases focused on one or more topics. The application of this type of integration could be also used for reporting, visualization, etc.

Incomplete Data Model Definition

The data structures and the models of relational databases are not often completely implemented. Frequently, there are several things that are lacking in the database structure: foreign keys, primary keys, constraints of specific columns, etc. The lack of any of these components make more difficult to understand stored information; although these lacks make the implementation of interfaces easier, being the joining of tables performed in queries.

There are several problems related to the incomplete data model definition. For example, in the case of lack of foreign keys, the cardinality

of the non-specified relationships could be definitive for classification of the content, and, consequently, the adaptation of this information to the new integrated model. Moreover, other type of relationships like self-relationships, could make very difficult the interpretation of the information model.

Understanding of Database Models

Each system involved in power distribution grids usually has a different structure: charging management for electrical vehicles (Richardson, Flynn, & Keane, 2012; Sousa, Morais, Vale, Faria, & Soares, 2012), energy management systems for buildings (La, Chan, & Soong, 2016; Wang, Wang, & Yang, 2012), and distribution systems (Zidan & El-Saadany, 2012). The use of information standards simplifies the understanding information stage, in any process of system, data mart or modelling development. The information standards provide a CIM to store all information about the power grid and management systems, for example: IEC with 61970, 61968 and 62325 standards, and DMTF with CIM, RedFish, etc. Although, the systems developed for new paradigm of Smart Grid are usually based on them, the old systems are usually based on third party or proprietary information models. Thus, the integration of these types of data sources usually requires one or more experts with advanced knowledge about the specific information model.

Evolution of Technologies

The evolution of power grids from a traditional model to the named Smart Grid has been provoked by several factors: the generalized liberalization of electric sector, the pressuring to adopt an environmentally sustainable system, and the technological development of Information and Communication Technology (ICT) scope.

This evolution has provided the deployment of a set of systems previously unavailable: from the massive deployment of Distributed Energy Resources (DER), including the electric vehicle; the massive deployment of Advanced Metering Infrastructure (AMI) highlighting the Smart Meters (SM), which generate an exponential growth in the volume of available data, requiring the development and application of new tools of information treatment in grid management topic (Analytics). They are the seed of new developments in new applications in this scope:

- Meter Data Management (MDM).
- Demand Response (DR).
- Distributed Energy Resources Management Systems (DERMS).
- Virtual Power Plants (VPPs) and microgrid.
- Electric Vehicles, infrastructure management.

These applications require the abilities of monitorization, modelling, simulation, analysis, and forecasting. These abilities start from heterogeneous data sources: weather, traffic, consumption measures (at level of Distributed System Operator (DSO), or client), asset information, geographical, power market data, etc.

Currently, the development of new technologies is faster than the market's ability to apply them, being more evident in the electrical distribution field. Particularly, the technologies related to the information management developed for power distribution companies needs to evolve the systems to take advantage from the new functionalities.

Modelling Information

The new technologies based on Smart Grid systems increase the volume of databases. These databases need powerful algorithms to model the information. Additionally, the information from older system provides several references in order to evaluate the impact of these new

technologies, i.e., by means of Key Performance Indicators (KPI), or to get better models.

Two popular methodologies are used in data mining: SEMMA (Sample, Explore, Modify, Model, and Assess) and CRISP-DM (Cross-Industry Standard Process for Data Mining). These two methodologies are notably different, but both methodologies have steps for data understanding and preparation. For example, in SEMMA: Sample, Explore and Modify, and in CRISP-DM: Business Understanding, Data Understanding, and Data Preparation. These steps take a long time, and to carry them out, companies usually need staff with both highly specialized background knowledge of data mining and familiarity with the domain of the problem to be solved.

Additionally, in certain cases, an integration of heterogeneous data collections should be performed to provide a sample for a data mining application or a data source for a different framework, for example, security appliances (Ruj & Nayak, 2013). In the area of power systems, Smart Grids pose new scenarios with great quantities of information that require big data infrastructures and real time processing. These infrastructures should integrate the new data generated and the old data provided by the traditional systems (e.g., SCADAs). The Smart Grid ecosystem compounds a great quantity of systems, such as charging management for electrical vehicles (Richardson et al., 2012; Sousa et al., 2012), energy management systems for buildings (La et al., 2016; Wang et al., 2012), and distribution systems (Zidan & El-Saadany, 2012).

BIBLIOGRAPHICAL REVIEW

The main research related to metadata mining applies to documents (Campos & Silva, 2000) and multimedia contents (Wong, 1999), and they are focused on knowledge discovery (Yi, Sundaresan, & Huang, 2000) or content classification (Yi & Sundaresan, 2000). Other references are focused on the usage of metadata over several types of contents: (Şah & Wade, 2012) proposed a novel automatic metadata extraction framework,

which is based on a novel fuzzy method for automatic cognitive metadata generation and uses different document parsing algorithms to extract rich metadata from multilingual enterprise content. (Asonitis, Boundas, Bokos, & Poulos, 2009) proposed an automated tool for characterizing new video files, using metadata schemas.

Some other references deal with heterogeneous data integration. (Alemu & Stevens, 2015) proposed an efficient metadata filtering in order the users apply metadata and thus enhance the findability and discoverability of the information objects. (Fermoso et al., 2009) proposed a new software tool called XDS (eXtensible Data Sources) that integrates data from relational databases, native eXtensible Markup Language (XML) databases, and XML documents. This framework integrates all information from heterogeneous databases to a XML-based format, such as MODS (Metadata Object Description Schema).

There are some references which pose the heterogeneous data integration providing models based on the analysis of available information. (Liu, Liu, Wu, & Ma, 2013) propose a Heterogeneous Data Integration Model (HDIM) based on the comparison and analysis of the current existing data integration approaches. On this HDIM, a pattern-mapping-based system called UDMP is designed and implemented. This approach tries to improve the rapid development of the Internet of Things (IoT). Moreover, (Lu & Song, 2010) proposed a heterogeneous data integration for Smart Grids. The authors described a model based on XML and ontology combined with cloud services to solve the heterogeneous problem from the syntax and semantics. They also tested with Supervisory Control and Data Acquisition (SCADA) data to validate the model. Some of these models were provided by means of an algebra. This tool is especially interesting because the mathematical description was more accurate and efficient. (Tang, Zhang, & Xiao, 2005) propose a capability object conceptual model to capture a rich variety of query-processing capabilities of sources and outline an algebra to compute the set of mediator-supported queries based on the capability limitations of the sources they integrate. This algebra is used in several works.

Additionally, there are a lot of studies and researches related to heterogeneous data integration based on, for instance, XML (Fengguang, Xie, & Liqun, 2009) (Su, Fan, & Li, 2010) (Lin, 2009), Lucene and XQuery (Tianyuan, Meina, & Xiaoqi, 2010), and OGSA-DAI (Gao & Xiao, 2013). In the same way, heterogeneous data integration has been applied on many areas, such as Livestock Products Traceability (X. d Chen & Liu, 2009), safety production (Han, Tian, & Wu, 2009), management information systems (Hailing & Yujie, 2012), medical information (Shi, Liu, Xu, & Ji, 2010), and web environments (Fan & Gui, 2007).

There are also examples of the application of data mining mixed with heterogeneous data source integration. These types of solutions increase the capability of solution to adapt it to different and heterogeneous data sources. (Cao, Chen, & Jiang, 2007) proposed a framework of a self-Adaptive Heterogeneous Data Integration System (AHDIS), based on ontology, semantic similarity, web service and XML techniques, which can be regulated dynamically. (Merrett, 2001) use On-Line Analytical Processing (OLAP) and data mining to illustrate the advantages for the relational algebra of adding the metadata type attribute and the transpose operator.

Currently, one of the main objectives of integrating the information is to analyse it. The proposed solution integrates heterogeneous data sources in specific and standard structures, and automatically applies data mining techniques. Thus, for example, some references related to this topic apply specific algorithms. (Li, Kang, & Gao, 2007) proposed a high-level knowledge modelled by Ordinary Differential Equations (ODEs) discovered automatically in dynamic data by an Asynchronous Parallel Evolutionary Modelling Algorithm (APHEMA). The data mining techniques are mainly used for forecast parameters. (J. Chen, Li, Lau, Cao, & Wang, 2010) proposed detecting automated load curve data cleansing based on the B-Spline smoothing and Kernel smoothing to automatically cleanse corrupted and missing data. (Hoiles & Krishnamurthy, 2015) proposed a nonparametric demand forecasting based on Least Squares Support Vector Machine (LS-SVM). The main lack of these references is that they did not automatically select the parameters to model, but they are

selected previously. In the proposed solution, the parameters to model are automatically selected in metadata mining stage.

GENERAL DESCRIPTION

The functional architecture of the proposed solution is shown in Figure 1. This solution is based on the use of different information analysis: metadata mining, text mining, data mining, and rule based expert system. Each of these techniques provides the ability to treat different types of data, even the metadata. The merging of these technologies allow to solve the problem of heterogeneous data source integration. Although several techniques related to the machine learning are used in this solution, it is mainly based on the knowledge, so the core of the proposed solution is the RBES. Thus, the proposed solution is limited to a knowledge domain related to Smart Grids and utilities. Although it has this limitation, it has a very big scope of the applications for present and future initiatives related to Smart Grid and utilities. In this sense, the deployment of Smart Grid ecosystem or a specific system in a Smart Grid is quicker because the proposed solution designs a specific Extract, Transform, and Load (ETL) for the new systems based on the information standards. Moreover, the integration of information can be optionally stored in data warehouses (with star or snowflake structure) to use the information in analysis processes. Additionally, the integration process can be applied in any distributed system with a high security level, due to the system only uses metadata. Finally, the proposed system includes a data and text mining engine to provide basic models for each parameter identified in the data sources, using different data and text mining techniques.

The information flow (specified by the arrows) and functional architecture is shown in Figure 1. The metadata from data sources is gathered by the Metadata Mining Engine using the query engine. The Metadata Mining Engine generates different information as parameters. Some of these parameters include the application of different techniques related to text mining, fuzzy logic, and Natural Language Processing

(NLP). The metadata are characterized and classified in different aggregation levels. The classification process is supported by the proposed Rule Based Expert System (RBES). Thus, in some stages, the RBES works as a Decision Support System (DSS). The RBES has several rules that are based on the indicators generated in the metadata mining process and the results of queries. When the system has classified all metadata from all data sources, the Dynamic ETL Engine performs the integration. There are two possibilities: according to an information standard or data warehouse (star or snowflake structure). If the user requires it, the integrated information can be modelled by the Data and Text Mining Engine. This engine performs an analysis according to the metadata mining information, to obtain the best model for each selected parameter. This process is supported by the RBES, too.

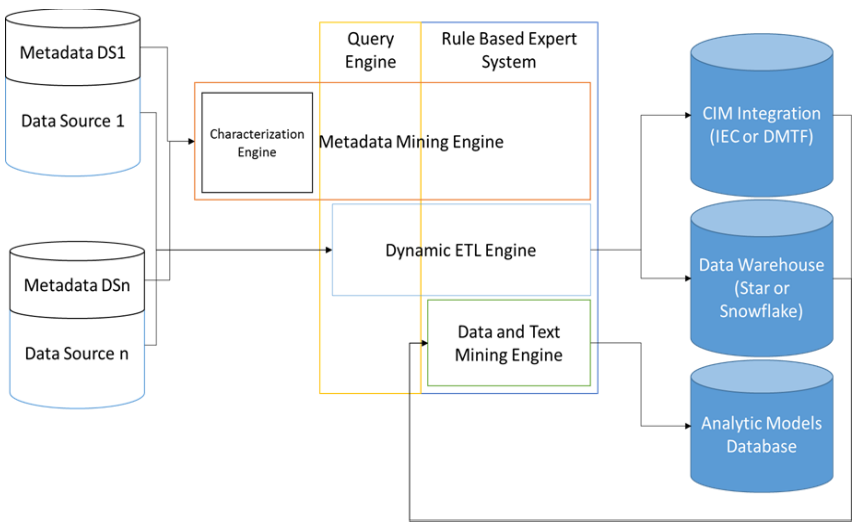

Figure 1. Flow and architecture overview.

METADATA MINING

The relational databases are one of the most widespread type of databases applied in past and present systems, even in future solution based

on Smart Grids. The metadata mining process is based on the metadata extracted from relational databases. The study of other type of databases is in research stage, but the relational databases are the beginning point. The metadata mining methodology is the same in all these cases. The flow diagram is shown in Figure 2. In the case of relational databases, this methodology has several steps:

1. Relational database identification. The proposed system has been tested with relational databases: MySQL, IBM DB2, Oracle Database, PostgreSQL, Microsoft SQL Server, and HBase. The identification of the relational database management system provides:
   a. Query language.
   b. Specific considerations about the RDBMS (Relational Data Base Management System).
   c. The name and structure of system tables.

   This process was simultaneously applied to several data sources.
2. Metadata extraction from system tables of database. The MDBS identification provides the definition of SQL (Statement Query Language) sentences to extract the metadata and, in some cases, the different aspects of SQL. This process is supported by the RBES.
3. Execution of grouping queries. After the metadata extraction, a process of the identification from the different entities is performed. For each entity, the system generates a sequence of SQL queries to extract aggregated data from each column separately. In this way, the information anonymization is warranted, because it is not possible to cross information from different columns.
4. Characterization Process. The characterization of different entities is performed over tables, columns, relationships, etc.

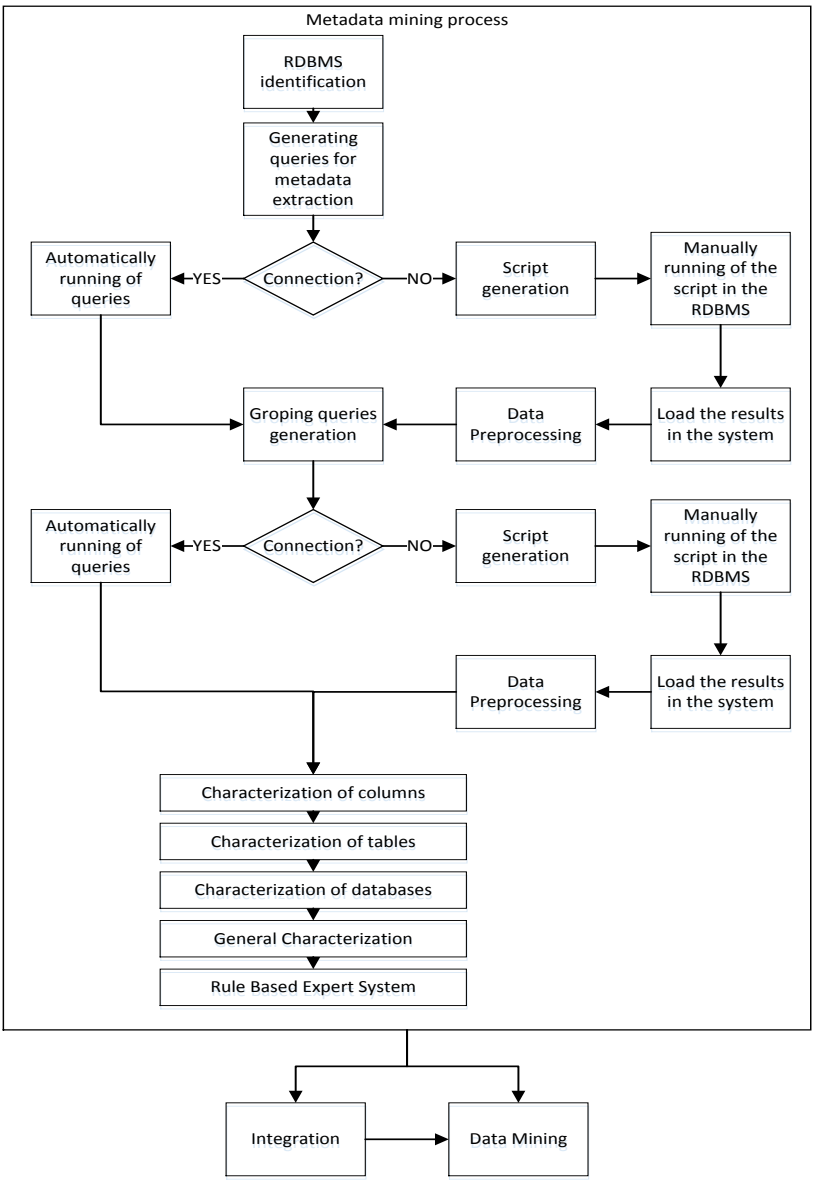

Figure 2. Metadata mining process flow chart.

5. Integration of information. The integration of information according to the results of characterization is performed, according to the information standards or data warehouse. This process is supported by RBES.

   Query design or data extraction. There are two options:

   a. Performing the integration. The queries are created and designed as an ETL, supporting by the RBES. Then, the integration is made of running the queries.
   b. Creating only the queries. The queries are created and designed, the system runs the queries when the users or other subsystems need it.

7. Data mining. After the creation of the database or scheme, the system applies a pool of data mining and text mining technique to obtain models from information. This process is supported by the RBES and the results of Metadata Mining process.
8. Results. All results are included as part of the final database scheme.

METADATA EXTRACTION FROM SYSTEM TABLES OF DATA SOURCES

Several queries to extract system tables are performed according to the database identified. These queries are automatically generated according to the identified RDBMS. The system provides two options to perform the queries:

- In a system whose RDBMS is not directly accessible, the system provides a SQL script. The user has to run this script in the command line of the RDBMS. The results of the script usually are several text files (one per system table) and these text files are

loaded in the system. This information is pre-processed in order to correct mistakes and format errors.

- The system runs the queries through a connection with RDBMS. The pre-processing step is easier than in the other case because the direct connection reduces the mistakes and errors in the interpretation of extracted information.

EXECUTION OF GROUPING QUERIES

The grouping queries are executed for each column of each table to obtain information regarding: the different values of the column, the frequency of each value, the absolute and relative frequencies of each value, and different statistical information about the distribution of values.

This information is mixed with the information from the previous step if there is a statistical information regarding the column in the system tables. Occasionally, the statistical information in the system tables is empty because the database was not analysed by RDBMS tools.

The results of grouping queries are stored in different tables. For example, the grouping query for columnA in Table 1 in standard SQL (Statement Query Language) was:

```
SELECT tableA. columnA, COUNT(*) AS counter_tableA_columnA
FROM tableA
GROUP BY tableA. columnA
ORDER BY tableA. columnA;
```

This query provides a table with two columns: columnA and counter. The first column contains all the possible values of the column. The second column contains the number of the register which has the corresponding value. This table is stored in the target database. The name of the table will be a combination of the table and column names: table1_column.

CHARACTERIZATION ENGINE

The metadata mining process is executed in several characterization stages: columns, tables, data source, and general. The general characterization is simultaneously performed with the columns, tables, and data source characterization. This means that if we have three data sources four characterizations are simultaneously performed.

Characterization of Columns

The characterization of a column depends on the data type. The first step is to classify the column in one of these categories: Numerical, Text, Timestamp, Object, Binary, or Other column. Each category is characterized according to different indexes and coefficients with statistical information about the contents of columns. They have some indexes and coefficients in common, which are related to the frequency of null, blank, and default values.

Additionally, the relationships between different columns to the other columns are calculated with two indexes:

- Fuzzy relationship coefficient with each column. This is an array of indexes, one per column in the database. Each element of this array establishes the relationship between different fields according to the name of the column. The index calculation is based in the application of a fuzzy algorithm to match the column name with other column names. The index can have a value between 0 and 1; zero indicates that there isn't any relationship, and one indicates that the columns are related.
- Relationship coefficient with each column. This is an array of indexes, one per column in the database. Each element of this array establishes the relationship between different columns according to

the registered values. First, the algorithm compares the data type, after that, the values.

These indexes and the information of the number of registers provide information about cardinality of relationships.

The profile of numerical columns contains information about data type, length, precision, column description, and constraints. Some statistical information is calculated: maximum, minimum, standard deviation, average, median, mode, and variation coefficient.

The profile of text columns contains information about data type, length, char set, column description, and constraints. Some statistical information is calculated: maximum length, minimum length, average length, standard deviation length, maximum number of words, minimum number of words, and average word length. Additionally, a dictionary is generated using text mining techniques. This dictionary is used to calculate the relationship coefficient with each column. The text mining technique attempts to elicit the text field concepts, structured or otherwise. A concept can comprise one or more words which represent an entity (e.g., action, and event). Natural Language Processing (NLP) methods are used to extract linguistic (e.g., words and phrases) and non-linguistic (e.g., dates and numbers) concepts. An interesting review of this technique and its use in information management systems is proposed in (Métais, 2002). The following set of functionalities are included:

a. Recognition of punctuation errors. These types of mistakes include the incorrect use of the accent, the period, the comma, the point and comma, the dividing bar, etc.
b. Recognition of spelling errors. A grouping fuzzy technology is applied. When concepts of the text are extracted, words with similar spelling (referring to the letters that compose it) or that are closely related are classified together. By applying this algorithm, mistakes of omission of letters, duplication of letters, or permutation of letters are corrected. This algorithm is used in the fuzzy relationship coefficient with each column calculation.

Although these mistakes are corrected before storing the concept in the dictionary, they are registered in the system in order to establish the level of wording of the column.

The profile of timestamp columns contains information about data type, format, column description, and constraints. Some statistical information is calculated: minimum, maximum, average time period between registers, minimum time period between registers, maximum time period between registers, values with the maximum number of registers, values with the minimum number of registers, average number of registers per value, standard deviation of number of registers per value, and values with the nearest number of registers to average number of registers per value. Additionally, a histogram of the number of registers per value is created. This histogram is normalized from 0 to 1, dividing the number of registers in each value by total number of registers. This information is used to calculate the relationship coefficient with each column.

The profile of object columns is used when the column contains information in a specific datatype defined in the RDBMSs. These data types are composed by different primitive types. If the system table contains information about this data type (sometimes this information is not accessible) the system associates several profiles to the column, one per primitive type, generating all the information previously described in each profile. Arrays are classified in this category.

The profile of binary column is used when the data type of a column stores binary information, for example images, documents, etc. Currently, the metadata mining only classifies the type of contents into the following categories: images, documents, video, technical, and other.

The profile of other columns is used when the column cannot be classified in the categories above. Normally, these columns are not used in the metadata mining process, and they are manually handled in order to establish a new profile. The encrypted columns are usually classified in this category.

Characterization of Tables

Each table on the selected data source is classified in one of the following categories: parametric information table, entity information table, personal information table, historical information table, complementary information table, bridge table, orphan table, and dummy table.

Additionally, some indicators are calculated for each table related to: relationships with other tables, self-relationships, coefficients and indicators of constraints, number of columns, number of each category of columns, and number of registers.

Characterization of Data Source

The characterization of the data source determines the coherence and reliability of the stored information, and it establishes the different indicators that will be used in automatic application of data mining techniques. These techniques try to establish models for prediction and classification of information. Additionally, the characterization includes information for automatic integration with other data sources.

- Database malleable indicator. This indicator establishes the potential for data analysis based on the information stored in the database. The number of columns with a high rate of useful information (columns with any possibility of application of any data mining technique) plus columns without useful information but with a high correlation coefficient with useful columns divided by the total number of columns.
- Database time analysis indicator. This indicator establishes the potential of temporal analysis. The calculation of this indicator is very similar to the “Database malleable indicator”. This indicator

considers as useful columns, those columns with any possibility of application of any time analysis technique.

- Database classification analysis indicator. This indicator establishes the potential of application of classification and clustering techniques. The calculation of this indicator is very similar to the "Database malleable indicator". This indicator considers as useful columns those columns with any possibility of application of any classification or clustering technique.
- Database forecasting analysis indicator. This indicator establishes the potential of application of forecasting techniques. The calculation of this indicator is very similar to the "Database malleable indicator". This indicator considers as useful columns those columns with any possibility of application of any forecasting technique.
- Database text analysis indicator. This indicator establishes the potential of application of text mining techniques. The calculation of this indicator is very similar to the "Database malleable indicator". This indicator considers as useful columns, those columns with any possibility of application of any text mining technique.
- Cohesion indicator. This indicator shows the information cohesion. The orphan registers and tables are used to calculate this indicator. Additionally, if statistical information about the database is available, then this indicator is modified adding columns without queries.
- Replication indicator. This indicator shows the level of redundant information.

General Characterization

This characterization is simultaneously carried on with the characterization of each data source. The general characterization establishes the relationship between all the data sources characterized

according to the method previously defined. In this characterization, all the previous steps are repeated, but considering all databases or data sources as the sole database. In this way, the characterization of tables, columns, relationships, and data sources are calculated again, but the name of the table includes the name of original data source (‘data_sourceA.tableA’ is tableA from data_sourceA). The new calculated indicators contain values according to all data sources. These new indicators have the prefix ‘general’.

RBES AND INTEGRATION OF INFORMATION

The RBES has several rules that are based on the indicators generated in the metadata mining process and the results of queries. The proposed RBES implements a DSS. The RBES has 492 rules: 30 rules in Metadata Mining Engine, 352 rules in Dynamic ETL Engine, and 110 rules in Data and Text Mining Engine. Each of these rules has been obtained from experience in collaboration in around 20 research projects with utility companies. The common problem in these projects is the existence of different relational data sources (95% were relational databases), with different: data management systems, data model, scope, and, often, without defined foreign keys. The 30 rules in Metadata Mining Engine deal with technical metadata. The 352 rules in Dynamic ETL Engine deal with technical and informational metadata to create and run the ETL. These rules could be classified into:

- Dynamic rules. The antecedent and consequence of a dynamic rule are stored on a table. This really means that each dynamic rule is applied several times, depending on the coincidences between available information and the data stored in the dynamic rule antecedent. In this sense, several sets of rules could be identified by:
  - 95 rules deal with IEC Common Information Model (CIM).
  - 83 rules deal with DMTF CIM.

  - 32 rules deal with IEC CIM extensions.
  - 36 rules deal with DMTF CIM extensions.
  - 53 rules deal with constraints.
  - 33 rules deal with foreign constraints.
- Static rules. These rules only have one antecedent and consequence. There are 20 rules which treat general topics to create and run the dynamic ETL.

The 110 rules in Data and Text Mining Engine could be classified in:

- 96 dynamic rules, which deal with the selection and application of the most adequate method for each modelling process, according to technical and informational metadata and the characterization performed.
- 14 static rules, which deal with the analysis of the results of modelling methods applied.

The integration of information from heterogeneous databases is accomplished by the application of general characterization in all classified databases. This module creates queries to integrate all information from columns and tables based on a decision support system based on the 352 rules. This DSS is part of a Dynamic ETL Engine and it is based on the information generated in the characterization of metadata mining process and on the results of several queries. The rules enable the queries to build the final query that integrates the information from different tables from different data sources. These queries are packed into ETL according to the target RDBMS. All tables with similar characterization are checked to be grouped according to the calculated cardinality. These new tables are characterized using the process previously described. The new values are compared with the original values in order to check the integration.

An example of these rules that involves several queries is shown below. This rule is used in the characterization of columns in order to calculate the cardinality of one side of the relationship. Some queries are performed to calculate it. This queries are:

- SELECT COUNT(tableA_columnA.columnA) AS count_tableA_columnA FROM tableA_columnA WHERE NOT(tableA.columnA IN (SELECT tableA.columnA FROM tableA_columnA, tableB_columnB WHERE tableA_columnA.columnA=tableB_columnB.columnB));
- SELECT MIN(counter_table.counterA) AS minA, MAX(counter_table.counterA) AS maxA, min(counter_table.counterB) AS minB, MAX(counter_table.counterB) AS maxB FROM (SELECT tableA_columnA.columnA, tableB_columnB.columnB, SUM(tableA_columnA.counter_tableA_columnA) AS counterA, SUM(tableB_columnB.counter_tableB_columnB) AS counterB FROM tableA_columnA, tableB_columnB where tableA_columnA.columnA=tableB_columnB.columnB group by tableA_columnA.columnA,tableB_columnB.columnB) AS counter_table;

The RBES uses the results of these queries and the calculated index to establish the cardinality of relation between columnA of table1 and columnB of Table2.

```
If fuzzy_relationship >=0.5 or
   relationship_coefficient >= 0.9 or
   exists defined constraint then
      If (minA==maxA and minA>1) or
         minA<maxA then
            (maximum cardinality is N)
      endif
      If (minA==maxA and minA==1) then
         (maximum cardinality is 1)
      endif
      If countA<>0 then
         (minimum cardinality is 0)
      else
         (minimum cardinality is 1)
      endif
endif
```

Currently, the process of checking the validity of the integration is performed by using several threshold parameters. These parameters are specified by the user or analyst. The automatic threshold parameter adjustment is in the research stage. Additionally, the user can filter orphan tables and bridge tables or avoid bridge tables.

GENERATION AND VALIDATION OF RULES

The rules generated for RBES were gathered from experience in several research projects related to utilities and Smart Grids. The most representative projects which take an important role in the creation of this solution have been previously published, but the main objective of these projects were not the heterogeneous data source integration, notwithstanding this capability was needed to get the results for each of these projects. Thus, the first prototype was not designed for all heterogeneous data sources, just only for the project, however, other projects showed the same problem. In this way, a general framework was created to integrate all information from different data sources in order to make easier the pre-processing stage of these projects. Some of these projects are described in:

- (Í. Monedero, Biscarri, León, Biscarri, & Millán, 2006), (León et al., 2011), (Juan I. Guerrero, León, Monedero, Biscarri, & Biscarri, 2014), and (Juan Ignacio Guerrero et al., 2016) describes The MIDAS Project. This project treats to reduce the non-technical losses applying data mining and computational intelligence to analyse the data from Endesa databases. Endesa databases have several data sources with different formats which need to be integrated in order to analyse it.
- (I. Monedero et al., 2015) describes a framework to detect water tampering in water utility, using data mining techniques. The data mining techniques are applied over data integrated from different data sources.

- (Personal, Guerrero, Garcia, Peña, & Leon, 2014) describes an application of Key Performance Indicators (KPI) Monitoring System to evaluate the integration of Smart Grid in front of traditional power grid. This system includes a Data Acquisition System. This system gathers information from different data sources in a unique data source in a format based on IEC CIM.
- (J.I. Guerrero, Personal, Parejo, García, & León, 2016) presents a framework to integrate systems in a Smart Grid ecosystem based on Web Service Mining and computational intelligence.

The first prototype of this framework only has a semi-automatic process to integrate tables, based on 239 configuration options, several of them derived from the number and nature of data sources. This tool was developed for MIDAS project and evolved and improved after several applications in MIDAS and other projects. After several applications of this prototype, several configuration rules were generated, validated, and the main structure of rules was designed. The proposed solution tries to integrate all information from all provided data sources in a specific standard format. When the framework could not integrate any part of any data source, the framework includes the necessary information to trace it. Thus, this information was manually analysed and new rules could be generated.

The proposed solution is the result of several iterations. The inference of new rules based on information from integration fails is still in research stage, but fuzzy logic and swarm intelligence methods have provided some interesting results.

INTEGRATION OF INFORMATION

The proposed system can integrate information in two different modes (the user can also configure the option of running both modes at the same time):

- According to the information of characterization. The system has been tested with several data sources. The intelligent ETL engine tries to create databases with star or extended-star architecture, in order to generate a data warehouse. This data warehouse is conditioned by the results of the characterization process, taking very high importance the indexes related to characterization of data source, and which shows where is the best information to create data mining and text mining modules. In this way, the data warehouse is optimized to support specific data mining or text mining models.
- According to the information of characterization and an information standard. Currently, the system only works with power distribution information standards. This system has been tested with information related to utilities, energy management, and information systems. The intelligent ETL engine can follow two standards: IEC CIM based on IEC 61970 and 61968 or DMTF CIM based on version 2.44.1 (but only applied to power grids). Currently, the utilization of other standards for health (HL7 and OpenEHR) are in the research stage.

The process is described in Figure 3. The integration of information includes several tables with information of characterization. This information was generated in metadata mining. The added tables are:

- GEN_CHAR. This table contains one register per data source, and contains information about the calculated indicators and data source description.
- DB_CHAR. This table contains one register per database, and contains information about the calculated indicators and database information. It is associated with data source described in GEN_CHAR.

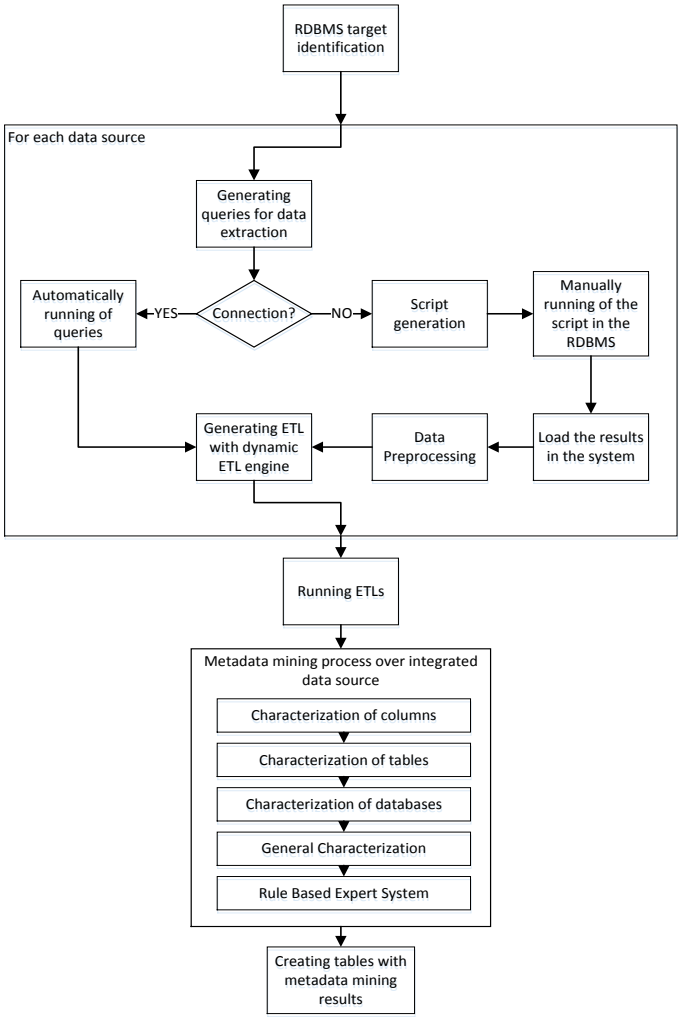

Figure 3. Integration flow chart.

- TAB_CHAR. This table contains one register per table, and contains information about the calculated indicators, relationship information, and table information. It is associated with data source (GEN_CHAR) and database (DB_CHAR).
- COL_CHAR. This table contains one register per column, and contains information about the calculated indicators, relationship information, and table information. It is associated with data

source (GEN_CHAR), database (DB_CHAR), datatype (DT_CHAR), and table (TAB_CHAR)

- CONS_CHAR. This table contains one register per constraint, and contains information about constraints and the associated table and column. It is associated with column (COL_CHAR) and table (TAB_CHAR).
- DT_CHAR. This table contains one register per component of data type, and contains information about data types.

Additionally, the information from the integrated resource is described by similar tables with 'I' prefix: I_DB_CHAR, I_TAB_CHAR, I_COL_CHAR, I_CONS_CHAR, and I_DT_CHAR.

These tables have several additional columns to store information that will be generated in the data mining stage.

DATA EXTRACTION

The data extraction stage depends on whether there is a direct connection to the data source. When the data source is protected, and it is not possible to have a direct connection or remote connection, the queries are executed by a script generated by the system, and the user runs the script in an authorized client. The script provides several text files with information from each table. The user loads these files into the system.

However, when the system has a direct or remote connection with the data source, the extraction is automatically performed with authorization from the user.

DATA MINING

The data mining module is guided by information generated in the characterization stage, supported by an intelligent system based on 110

rules. In the first place, a feature selection is performed to associate a support index to each column. This feature selection is performed for each column as a target. In this way, each column has one value associated with it.

Currently, a threshold is manually specified to use the different columns. The value assigned for each method is established based on experience, although the user could modify that value if it is necessary. Moreover, the generation of models can be personalized by: specification of time limit in model generation, specification of memory limit in model generation, manual filtering of non-desired targets, establishing a limit in the number of parameters to consider in the modelling process, and/or manual filtering of non-desired algorithms or techniques. If any model exceeds any of these restriction, it is automatically discarded. The information about the discarded models comprise the inputs, the technique or algorithm, the target, and the exceeded restrictions.

Additionally, according to the performed characterization, several methods are applied to obtain models. This module has been implemented in SPSS Modeler and Python. In this way, the applied algorithms or techniques are: Anomaly detection (Chandola, Banerjee, & Kumar, 2009), apriority (Agrawal & Srikant, 1994), Bayesian network (Pearl, 2000), C5.0, Carma (Hidber, 1999), C&R Tree (Breiman, Friedman, Stone, & Olshen, 1984), Chi-squared Automatic Interaction Detector or CHAID (Kass, 1980), Cluster evaluation (based on silhouette coefficient, sum of squares error or SSE, sum of squares between or SSB, and predictor importance), COXREG (Cox, 1972), Decision List, Discriminant, Factor Analysis (PCA) (Geiger & Kubin, 2012), Generalized Linear Models, Generalized linear mixed models (Madsen & Thyregod, 2010), K-Means (MacQueen, 1967), Kohonen (Kohonen, 1982), Logistic Regression (Freedman, 2005), KNN (Pan, McInnes, & Jack, 1996), Linear modelling (Belsley, Kuh, & Welsch, 2013), neural network (Haykin, 1994), optimal binning (Usama M. Fayyad, 1993), "Quick, Unbiased, Efficient Statistical Tree" or QUEST (Loh & Shih, 1997), linear regression, Sequence, Self-learning response model or SLRMs, support vector machine (SVM), temporal casual modelling algorithms (Arnold, Liu, & Abe, 2007), time

series (Box, Jenkins, & Reinsel, 2008), and Two Step cluster (Chiu, Fang, Chen, Wang, & Jeris, 2001).

The selection of the best technique is based on two criteria: the error rate of each generated method, and the correlation between the model and the target.

EXPERIMENTAL RESULTS

The proposed system was applied to several data sources related to power distribution. The data sources were related to (some columns were omitted because of a confidentiality agreement):

- Source A: Consumer historical information. This data source contains information about consumers: historical consumption data and contract information. This data source has four tables: contract information, historical data, and two parametric tables.
- Source B: Recharging point usage information. This data source contains information about consumption at a recharging point. This data source has seven tables: recharging point information, contractual information, vehicle information, consumption information, and three parametric tables.
- Source C: Generation data from different source types. This data source contains information about wind and photovoltaic generation data. This data source has three tables: historical information, source information, and a parametric table.

In the three cases, the foreign keys and interrelations between the tables were not established by constraints; the authors indicate the relations in order to make a better presentation of the data source.

After the metadata mining process and the characterization stage, the results for each data source is shown in Table I, whose information is only regarding databases. This information is evaluated by the RBES. The

information about columns and tables has been omitted because of a confidentiality agreement.

In Table 1, the coefficients and indicators were calculated according to the results of previous characterization processes. In this case, all the sources showed a high rate of possibilities for application of data mining techniques. They show a high rate of cohesion and low rate of replication. The best punctuation is for time analysis and forecasting. Thus, the decision support system selected the methods related to time analysis and forecasting to be applied in the data mining stage.

These data sources were in different RDBMSs: Microsoft SQL Server, MySQL, and Oracle. The integration was performed in an H Base.

Following the IEC Standards, seventeen tables were created: Power System Resources, Measurement, Terminal, Analog, Analog Value, Analog Limit Set, Accumulator, Accumulator Value, Accumulator Limit, Accumulator Limit Set, String Measurement, String Measurement Value, Discrete, Discrete Value, Value Alias Set, Value To Alias, and Measurement Value Source. There is no table about quality of measure because there was no table about quality. Additionally, the information about the different characterization process (metadata mining) was added to the database using the tables described in the Integration of Information section.

The data mining modelling was configured to forecasting methods. This configuration is selected by the system based on the nature of the parameters and the indexes calculated in the metadata mining process. However, this situation can be changed by the user adding options for outlier detection, classification, or visualizations.

Table 1. Results of characterization of data sources

Data Source	Indicators and coefficients						
	Minable	Time analysis	Classification analysis	Fore-casting	Text analysis	Cohesion	Repli-cation
A	0.82	0.60	0.53	0.70	0.05	0.87	0
B	0.93	0.72	0.70	0.50	0.10	0.70	0.30
C	0.75	0.81	0.41	0.68	0.10	0.98	0

Table 2. Results of data mining forecasting detected parameters

Data Source	Parameter	Modelling Method	Correlation	Error
A	Authorized car dealer*	Linear Regression Generalized Linear Model	0.993	0.014
A	Hotel industry*	Regression Generalized Linear Model	0.993	0.014
A	Technical advice office*	Regression Generalized Linear Model	0.992	0.017
A	General Services*	Regression Generalized Linear Model	0.996	0.007
A	Communication office*	Regression Generalized Linear Model	0.99	0.019
A	Power Generation Company office*	Regression Generalized Linear Model	0.955	0.087
A	Authorised car dealer (without garage)*	Regression Generalized Linear Model	0.971	0.058
A	Consulting office	Neural Network (multilayer perceptron)	0.961	0.046
A	Main Power Distribution office*	Regression Generalized Linear Model	0.992	0.015
A	Power distribution office	Neural Network (multilayer perceptron)	0.977	0.046
A	Temporary employment agency office*	Regression Generalized Linear Model	0.983	0.033
B	Recharging points	Not useful model		
C	20 KW Generation Plant*	Regression Generalized Linear Model	0.993	0.014
C	80 KW Generation Plant*	Linear Regression Regression Generalized Linear Model	0.99	0.019
C	100 KW Generation Plant*	Linear Regression Regression Generalized Linear Model	0.991	0.018

*: Several modelling techniques provides similar correlation and error rates. The different techniques based on regression usually provide the same model or similar.

The results of data mining modelling in each parameter are shown in Table 2. In some cases, the system selected several methods because they had the same evaluation value; nevertheless, the different methods were ordered according to the time required for the model generation process.

A regression model was created for the Recharging point, but in the test stage the generated model showed a very high error rate. The algorithm has no information about routes or drivers.

CONCLUSION

Smart Grids and the new technologies related to information management are the future of the new smart services and applications. Several services and applications of different technological levels coexist within the current utility grid. In this sense, it is necessary to establish techniques that provide the capability to integrate information from different architecture and technological levels. These technologies increase the robustness of the management systems related to the utility grid.

The metadata mining process is focused on metadata, and taking advantage of this technology it is possible to make systems that integrate the information, according to an information standard, star, or extended-star structure. Additionally, a system for automatic modelling is provided, based on previous application of a metadata mining algorithm. In this way, this technology provides an easy-to-use and adaptive platform to integrate and model information. The models could be improved by adding new information, and performing the modelling algorithm.

In this chapter, the proposed system is applied over a power distribution system, but the future research lines may include its application of this technology over other types of databases, such as document-based and key-value databases.

ACKNOWLEDGMENTS

The authors would like to thank the Smart Business Project (SBP), which provided data sources. Additionally, the authors would like to thank the IDEA Agency for providing the funds for the project.

The authors are also appreciative of the backing of the SIIAM project, which is funded by the Ministry of Economy and Competitiveness of Spain.

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487–499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645920.672836.

Alemu, G., & Stevens, B. (2015). 8 - The principle of metadata filtering. In *An Emergent Theory of Digital Library Metadata* (pp. 89–96). Chandos Publishing. Retrieved from http://www.sciencedirect.com/ science/article/ pii/B9780081003855000080.

Arnold, A., Liu, Y., & Abe, N. (2007). Temporal Causal Modeling with Graphical Granger Methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 66–75). New York, NY, USA: ACM. https://doi.org/ 10.1145/ 1281192.1281203.

Asonitis, S., Boundas, D., Bokos, G., & Poulos, M. (2009). Semi – automated tool for characterizing news video files, using metadata schemas. In M.-A. Sicilia & M. D. Lytras (Eds.), *Metadata and Semantics* (pp. 167–178). Springer US. https://doi.org/10.1007/978-0-387-77745-0_16.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2013). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, N.J: Wiley-Interscience.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control* (4 edition). Hoboken, N.J: Wiley.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis.

Campos, J. P., & Silva, M. J. (2000). ActiveXML: Compound Documents for Integration of Heterogeneous Data Sources. In J. Borbinha & T. Baker (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 380–384). Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.fama.us.es/chapter/10.1007/3-540-45268-0_45.

Cao, Y., Chen, Y., & Jiang, B. (2007). A Study on Self-adaptive Heterogeneous Data Integration Systems. In L. D. Xu, A. M. Tjoa, & S. S. Chaudhry (Eds.), *Research and Practical Issues of Enterprise Information Systems II* (pp. 65–74). Springer US. Retrieved from http://0-link.springer.com.fama.us.es/chapter/10.1007/978-0-387-75902-9_7.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Comput. Surv.*, *41*(3), 15:1–15:58. https://doi.org/10.1145/1541880.1541882.

Chen, X. d, & Liu, J. Z. (2009). Research on Heterogeneous Data Integration in the Livestock Products Traceability System. In *International Conference on New Trends in Information and Service Science, 2009. NISS '09* (pp. 969–972). https://doi.org/10.1109/NISS.2009.94.

Chen, J., Li, W., Lau, A., Cao, J., & Wang, K. (2010). Automated Load Curve Data Cleansing in Power Systems. *IEEE Transactions on Smart Grid*, *1*(2), 213–221. https://doi.org/10.1109/TSG.2010.2053052.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 263–268). New York, NY, USA: ACM. https://doi.org/10.1145/502512.502549.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Fan, H., & Gui, H. (2007). Study on Heterogeneous Data Integration Issues in Web Environments. In *International Conference on Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007* (pp. 3755–3758). https://doi.org/10.1109/WICOM.2007.929.

Fengguang, X., Xie, H., & Liqun, K. (2009). Research and implementation of heterogeneous data integration based on XML. In *9th International Conference on Electronic Measurement Instruments, 2009. ICEMI '09* (pp. 4-711-4–715). https://doi.org/10.1109/ICEMI.2009.5274686.

Fermoso, A. M., Berjón, R., Beato, E., Mateos, M., Sánchez, M. A., García, M. M., & Gil, M. J. (2009). A New Proposal for Heterogeneous Data Integration to XML format. Application to the Environment of Libraries. In M.-A. Sicilia & M. D. Lytras (Eds.), *Metadata and Semantics* (pp. 143–153). Springer US. https://doi.org/10.1007/978-0-387-77745-0_14.

Freedman, D. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Gao, J., & Xiao, J. (2013). Research on Heterogeneous Data Access and Integration Model Based on OGSA-DAI. In *2013 Fifth International Conference on Computational and Information Sciences (ICCIS)* (pp. 1690–1693). https://doi.org/10.1109/ICCIS.2013.441.

Geiger, B. C., & Kubin, G. (2012). Relative Information Loss in the PCA. *arXiv:1204.0429 [Cs, Math]*, 562–566. https://doi.org/10.1109/ITW.2012.6404738.

Guerrero, J. I., Personal, E., Parejo, A., García, A., & León, C. (2016). Forecasting the Needs of Users and Systems - A New Approach to Web Service Mining. In *The Fifth International Conference on Intelligent Systems and Applications* (pp. 96–99). Barcelona, Spain: IARIA.

Guerrero, Juan I., León, C., Monedero, I., Biscarri, F., & Biscarri, J. (2014). Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection. *Knowledge-Based Systems*, *71*, 376–388. https://doi.org/10.1016/j.knosys.2014.08.014.

Guerrero, Juan Ignacio, Parejo, A., Personal, E., Biscarri, F., Biscarri, J., & Leon, C. (2016). Intelligent Information System as a Tool to Reach Unaproachable Goals for Inspectors - High-Performance Data Analysis for Reduction of Non-Technical Losses on Smart Grids (pp. 83–87). Presented at the INTELLI 2016, The Fifth International Conference on Intelligent Systems and Applications. Retrieved from https://www.thinkmind.org/index.php?view=article&articleid=intelli_2016_4_10_60123.

Hailing, W., & Yujie, H. (2012). Research on heterogeneous data integration of management information system. In *2012 International Conference on Computational Problem-Solving (ICCP)* (pp. 477–480). https://doi.org/ 10.1109/ICCPS.2012.6384220.

Han, X. b, Tian, F., & Wu, F. b. (2009). Research on Heterogeneous Data Integration in the Safety Production and Management of Coal-Mining. In *2009 First International Workshop on Database Technology and Applications* (pp. 87–90). https://doi.org/10.1109/DBTA.2009.60.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. MacMillan Publishing Company.

Hidber, C. (1999). Online Association Rule Mining. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data* (pp. 145–156). New York, NY, USA: ACM. https://doi.org/10.1145/304182.304195.

Hoiles, W., & Krishnamurthy, V. (2015). Nonparametric Demand Forecasting and Detection of Energy Aware Consumers. *IEEE Transactions on Smart Grid*, *6*(2), 695–704. https://doi.org/10.1109/TSG.2014.2376291.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(2), 119–127. https://doi.org/10.2307/2986296.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288.

La, Q. D., Chan, Y. W. E., & Soong, B. H. (2016). Power Management of Intelligent Buildings Facilitated by Smart Grid: A Market Approach. *IEEE Transactions on Smart Grid*, *7*(3), 1389–1400. https://doi.org/10.1109/TSG.2015.2477852.

León, C., Biscarri, F., Monedero, I., Guerrero, J. I., Biscarri, J., & Millán, R. (2011). Integrated expert system applied to the analysis of non-technical losses in power utilities. *Expert Systems with Applications*, *38*(8), 10274–10285. https://doi.org/10.1016/j.eswa.2011.02.062.

Li, Y., Kang, Z., & Gao, H. (2007). Automatic Data Mining by Asynchronous Parallel Evolutionary Algorithms. In L. Kang, Y. Liu, & S. Zeng (Eds.), *Advances in Computation and Intelligence* (pp. 485–492). Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.fama.us.es/ chapter/10.1007/978-3-540-74581-5_53.

Lin, Y. (2009). Study and technological realization about heterogeneous data integration based on XML Schema. In *International Conference on Test and Measurement, 2009. ICTM '09* (Vol. 2, pp. 394–397). https:// doi.org/10.1109/ICTM.2009.5413020.

Liu, H., Liu, Y., Wu, Q., & Ma, S. (2013). A Heterogeneous Data Integration Model. In F. Bian, Y. Xie, X. Cui, & Y. Zeng (Eds.), *Geo-Informatics in Resource Management and Sustainable Ecosystem* (pp. 298–312). Springer Berlin Heidelberg. Retrieved from http://0-link.springer. com.fama.us.es/chapter/10.1007/978-3-642-45025-9_31

Loh, W. Y., & Shih, Y. S. (1997). SPLIT SELECTION METHODS FOR CLASSIFICATION TREES. *Statistica Sinica*, *7*(4), 815–840.

Lu, B., & Song, W. (2010). Research on heterogeneous data integration for Smart Grid. In *2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)* (Vol. 3, pp. 52–56). https://doi.org/10.1109/ICCSIT.2010.5564620.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, The Regents of the University of California. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512992.

Madsen, H., & Thyregod, P. (2010). *Introduction to General and Generalized Linear Models*. CRC Press.

Merrett, T. H. (2001). Attribute Metadata for Relational OLAP and Data Mining. In G. Ghelli & G. Grahne (Eds.), *Database Programming Languages* (pp. 97–118). Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.fama.us.es/chapter/10.1007/3-540-46093-4_6.

Métais, E. (2002). Enhancing information systems management with natural language processing techniques. *Data & Knowledge Engineering*, *41*(2–3), 247–272. https://doi.org/10.1016/S0169-023X(02)00043-5.

Monedero, I., Biscarri, F., Guerrero, J. I., Peña, M., Roldán, M., & León, C. (2015). Detection of Water Meter Under-Registration Using Statistical Algorithms. *Journal of Water Resources Planning and Management*, *142*(1), 04015036.

Monedero, Í., Biscarri, F., León, C., Biscarri, J., & Millán, R. (2006). MIDAS: Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques. In *Computational Science and Its Applications - ICCSA 2006* (pp. 725–734). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11751649_80.

Pan, J. S., McInnes, F. R., & Jack, M. A. (1996). Fast clustering algorithms for vector quantization. *Pattern Recognition*, *29*(3), 511–518. https://doi.org/10.1016/0031-3203(94)00091-3.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge, U.K.; New York: Cambridge University Press.

Personal, E., Guerrero, J. I., Garcia, A., Peña, M., & Leon, C. (2014). Key performance indicators: A useful tool to assess Smart Grid goals. *Energy*, *76*, 976–988. https://doi.org/10.1016/j.energy.2014.09.015.

Richardson, P., Flynn, D., & Keane, A. (2012). Local Versus Centralized Charging Strategies for Electric Vehicles in Low Voltage Distribution Systems. *IEEE Transactions on Smart Grid*, *3*(2), 1020–1028. https://doi.org/10.1109/TSG.2012.2185523.

Ruj, S., & Nayak, A. (2013). A Decentralized Security Framework for Data Aggregation and Access Control in Smart Grids. *IEEE Transactions on Smart Grid*, *4*(1), 196–205. https://doi.org/10.1109/TSG.2012.2224389.

Şah, M., & Wade, V. (2012). Automatic metadata mining from multilingual enterprise content. *Web Semantics: Science, Services and Agents on the World Wide Web*, *11*, 41–62. https://doi.org/10.1016/j.websem. 2011.11.001.

Shi, Y., Liu, X., Xu, Y., & Ji, Z. (2010). Semantic-based data integration model applied to heterogeneous medical information system. In *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)* (Vol. 2, pp. 624–628). https://doi.org/10.1109/ICCAE. 2010.5451697.

Sousa, T., Morais, H., Vale, Z., Faria, P., & Soares, J. (2012). Intelligent Energy Resource Management Considering Vehicle-to-Grid: A Simulated Annealing Approach. *IEEE Transactions on Smart Grid*, *3*(1), 535–542. https://doi.org/10.1109/TSG.2011.2165303.

Su, J., Fan, R., & Li, X. (2010). Research and design of heterogeneous data integration middleware based on XML. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)* (Vol. 2, pp. 850–854). https://doi.org/10.1109/ICICISYS. 2010. 5658689.

Tang, J., Zhang, W., & Xiao, W. (2005). An Algebra for Capability Object Interoperability of Heterogeneous Data Integration Systems. In Y. Zhang, K. Tanaka, J. X. Yu, S. Wang, & M. Li (Eds.), *Web Technologies Research and Development - APWeb 2005* (pp. 339–350). Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.fama.us.es/ chapter/10.1007/978-3-540-31849-1_34.

Tianyuan, L., Meina, S., & Xiaoqi, Z. (2010). Research of massive heterogeneous data integration based on Lucene and XQuery. In *2010 IEEE 2nd Symposium on Web Society (SWS)* (pp. 648–652). https://doi.org/10.1109/SWS.2010.5607370.

Usama M. Fayyad, K. B. I. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022–1029.

Wang, L., Wang, Z., & Yang, R. (2012). Intelligent Multiagent Control System for Energy and Comfort Management in Smart and Sustainable Buildings. *IEEE Transactions on Smart Grid*, *3*(2), 605–617. https://doi.org/10.1109/TSG.2011.2178044.

Wong, R. K. (1999). Heterogeneous data integration and presentation in multimedia database management systems. In *IEEE International Conference on Multimedia Computing and Systems, 1999* (Vol. 2, pp. 666–671 vol.2). https://doi.org/10.1109/MMCS.1999.778563.

Yi, J., & Sundaresan, N. (2000). Metadata based Web mining for relevance. In *Database Engineering and Applications Symposium, 2000 International* (pp. 113–121). https://doi.org/10.1109/IDEAS.2000.880569.

Yi, J., Sundaresan, N., & Huang, A. (2000). Metadata Based Web Mining for Topic-Specific Information Gathering. In K. Bauknecht, S. K. Madria, & G. Pernul (Eds.), *Electronic Commerce and Web Technologies* (pp. 359–368). Springer Berlin Heidelberg. Retrieved from http://0-link.springer.com.fama.us.es/chapter/10.1007/3-540-44463-7_31.

Zidan, A., & El-Saadany, E. F. (2012). A Cooperative Multiagent Framework for Self-Healing Mechanisms in Distribution Systems. *IEEE Transactions on Smart Grid*, *3*(3), 1525–1539. https://doi.org/10.1109/TSG. 2012.2198247.

BIOGRAPHICAL SKETCHES

Juan Ignacio Guerrero Alonso

Affiliation: Department of Electronic Technology. University of Seville (SPAIN)

Education: PhD, Computer Science (2011) and Master Degree in Computer Science (2006) in University of Seville (Spain). Computer Engineering Degree in University of Corduba (Spain).

Business Address: Escuela Politécnica Superior. C/Virgen de Africa nº 7. 41011 Sevilla, Spain

Research and Professional Experience: Currently, he works in University of Seville as an Assistant Professor, collaborating with department of Electronic Technology in teaching tasks and with Electronic Technology and Industrial Informatics (TIC150) research team in researching tasks. He has worked in more than 30 projects in collaboration with companies. He has several publications in reputed journals and conferences. The main research areas are related to application of big data analytics, data mining, text mining, and computational intelligence to utility sector and Smart Grids.

Professional Appointments: Assistant Professor

Publications from the Last 3 Years:

In the last three years, the author has five publications in high impact factor international scientific journals, one publication in low impact factor journal, several chapters, conferences, and seminars.

Antonio García Delgado

Affiliation: Department of Electronic Technology. University of Seville (SPAIN)

Education: B.Sc. degree in electronic physics from the University of Seville, in 1982

Business Address: Escuela Politécnica Superior. C/Virgen de Africa nº 7. 41011 Sevilla, Spain

Research and Professional Experience:

Professor of Electronic Engineering in the Electronic Technology Department since 1984. His areas of research include instrumentation, digital signal processing; fault location methods in power lines and Data analytics on Smart Grid applications.

Professional Appointments: Professor

Publications from the Last 3 Years:

4 publication in International Journals.

Enrique Personal

Affiliation: Department of Electronic Technology. University of Seville (SPAIN)

Education: industrial electronic engineering degree, automatic control and industrial electronic engineering degree, and Ph.D. degree in industrial computer science.

Business Address: Escuela Politécnica Superior. C/Virgen de Africa nº 7. 41011 Sevilla, Spain

Research and Professional Experience: Fields of interest are smart grids, fault location methods, power systems, and WSNs.

Professional Appointments: Assistant Professor, University of Seville, Spain

Publications from the Last 3 Years:
4 journal papers and 2 conference papers

Antonio Parejo Matos

Affiliation: Department of Electronic Technology. University of Seville (SPAIN)

Education: Master Degree in Electronic Engineering

Business Address: Escuela Politécnica Superior. C/Virgen de Africa nº 7. 41011 Sevilla, Spain

Research and Professional Experience: My main research areas are Smart Cities, Smart Grids and Industrial Computational Inteligence Applications. I work in University of Seville from 2015 in the research group TIC-150.

Professional Appointments: Scholarship.

Publications from the Last 3 Years: 4 research publications in aforementioned areas.

Francisco Pérez García

Affiliation: Department of Electronic Technology. University of Seville (SPAIN)

Education:
B.S. Degree in Physics (Electronic), University of Seville, 1985
Ph.D in Robotics and Digital Imaging Processing, University of Seville, 1992

Business Address: ETS Ingeniería Informática, Avenida Reina Mercedes s/n, 41012, Sevilla, Spain

Research and Professional Experience:

Dr. Francisco Pérez accepted his PhD in the field of Robotics and Digital Imaging Processing from the University of Seville, in 1992. From then, his main interest areas have been related to the field of Industrial Informatics and Industrial Communications.

In last years he has served as the Dean of the Computer Science and Engineering High Technical School (1996-2006), and as Vicerector of Teaching of the University of Seville (2007-2009).

Professional Appointments: Full Professor

Carlos León de Mora

Affiliation: Universidad de Sevilla

Education: Physic Degree. Computer Science PhD

Business Address: Escuela Politécnica Superior. C/Virgen de Africa nº 7. 41011 Sevilla, Spain

Research and Professional Experience:

Received the M.S. degree in physical electronics and the Ph.D. degree in computer science from the University of Seville, in 1991 and 1995, respectively. He is a Full Professor of Electronic Engineering with the University of Seville. His areas of research are Computational Intelligence, Knowledge-based and Cognitive systems, Data mining and Machine leraning, focusing on utility systems and complex industrial. Management. Dr. León is a senior member of the IEEE. His is co-author of more than 50 papers in JCR index Journals, has participated as author in more than 90 conferences, has co-authored 14 research book papers and has been

director of 11 PhD Thesis. His has been Director of more than 50 Research projects and have 4 patents.

Professional Appointments: Full Professor, Vicerector.