Database Systems for the Smart Grid

Zeyar Aung

Abstract In this chapter, two aspects of database systems, namely database management and data mining, for the smart grid are covered. The uses of database management and data mining for the electrical power grid comprising of the interrelated subsystems of power generation, transmission, distribution, and utilization are discussed.

1 Introduction

Since the smart grid reply on modern information and communication technology (ICT) infrastructure, database systems, which are one of the vital components of ICT, are indispensable in the smart grid. Database systems allows the data in the smart grid to be stored in a systematic manner and enable them to be retrieved, processed and analyzed either immediately (i.e., online data processing/analysis) or later (i.e., historical data processing/analysis).

Because of the involvement of database systems, the smart grid is no longer a business dominated by utility companies and electricity hardware companies alone. Several big software companies in data-centric business such as Teradata [10], Oracle [6], SAS [8], SAP [7], IBM [4], Microsoft [5], and Google [3] are active players in the smart grid arena now.

There are two main aspects of a database system, namely database management (data storage, transaction processing, and querying), and data mining (analysis of data to gain certain knowledge or facilitate certain decision making). These two aspects are naturally interrelated and are like the two sides of a coin. Both are essential for the business process of the smart grid's operations.

Zeyar Aung
Computing and Information Science Program, Masdar Institute of Science and Technology, Block 3 Masdar City, Abu Dhabi 54224, United Arab Emirates. e-mail: zaung@masdar.ac.ae

In this chapter, we will cover the applications of database management and data mining in the smart grid for power generation, transmission, distribution, and utilization (consumption). Again, these four application areas are interrelated and somewhat overlapping especially because of the interconnected nature of the smart grid.

The development of smart grid is an evolutionary process. During the smart grid's introduction phase, the two generations of technologies will coexist [24]. For ICT components (both software and hardware), a majority of legacy systems are first to be integrated into the smart grid and later phased out and replaced by the newer technologies. However, for power system components, the introduction of smart gird will not even drastically change the basic mechanisms of the power system's mechanical and electrical equipment (except that they will now be more intelligent and responsive because of incorporation of ICT). For example, a gas turbine will still operate just in the same way to convert natural gas into electrical power as it did in the old non-smart grid — albeit it may now use less amount of gas because of a more intelligent control system. So, a database recording the operations of such a gas turbine will be more or less the same in both the traditional grid and the smart grid.

For the aforementioned reasons, we believe that both the earlier systems for systematic power grid data management/mining even before the word smart grid was coined as well as the newer systems which were explicitly proposed for the smart grid are worth covering. As such, in this chapter, we will include the literature on power grid database systems both before and after the concept of the smart grid was conceived.

In the following two sections, database management and data mining for power grids will be respectively covered.

2 Power Grid Database Management

In this section, we will cover the database management technologies in general and then the applications of database management for a power grid in its four subsystems: generation, transmission, distribution, and utilization.

2.1 Database Management Technologies

In modern days, management of data in an ICT system is centered around a proper database management system (DBMS) or sometimes simply a file system (FS). In both cases, the basic operations of data management are:

- Schema creation: defining format of data and relationships among data.
- Data insertion: populating the database/files with data.
- Data maintenance: updating or deleting existing data.

- Querying and reporting: retrieval of stored data as per users' business requirements.
- Performance optimization: making the retrieval process faster by using indexes, etc.
- User account management: defining which user has a right to do which operations on which data.
- Backup and recovery: preventing accidental loss of data.

For DBMS, relational database (composing of tables which are mathematically termed "relations") is the most common standard. Some commonly used relational DBMS are Oracle (proprietary), Microsoft SQL Server (proprietary), IBM DB2 and Informix (proprietary), SAP Sybase (proprietary), MySQL (open source), and PostgreSQL (open source). Structured query language (SQL) is a common interface to retrieve data from relational DBMS.

Recently, post-relational database systems called NoSQL (Not only SQL) [76] become more and more common. NoSQL database systems include document-oriented databases (e.g., MongoDB), XML databases (e.g., BaseX) graph databases (e.g., InfiniteGraph), key-value stores (e.g., Apache Cassandra), multi-value databases (e.g., OpenQM), object-oriented databases (e.g., db4o), RDF (resource description framework) databases (e.g., Meronymy SPARQL), tabular databases (e.g., BigTable), tuple databases (e.g., Jini), and column-oriented databases (e.g., c-store). NoSQL database systems use conventional programming languages like C++, C#, Java, and Erlang, or XQuery in the case of XML databases in order to interface and retrieve data from the databases.

In addition to NoSQL databases, parallel and distributed file systems such as Apache Hadoop [1] and Google MapReduce [75] become increasingly popular. Since the smart grid by its own nature is distributed and the resources (like smart meters, meter data concentrators, substation transformers, etc.) in it are geographically scattered, distributed file systems can potentially be very useful for the smart grid.

Generally, databases are stored on centralized or distributed magmatic hard disk drives. However, new paradigms of databases stored on main memory (such as voltDB) and solid state drives (such as [59]) are emerging because of the increased availability of high-capacity main memory and solid state equipment at low costs.

Another increasing popular approach nowadays is to store databases in the cloud. Cloud computing and cloud database [71] are also the emerging trends that are much relevant to the smart grid. A cloud database can be in the form of either a virtual machine instance which can be purchased for a limited time or a database as a service in which the service provider installs and maintains the database, and application owners pay according to their usage. Amazon's DynamoDB and SimpleDB are examples of database as a service.

"Big data" (meaning several tera- to peta-bytes of data) is one of the current hot topics. Big data is a crucial issue for the smart grid it since an enormous volume of data is expected to be generated from its large number of connected devices and sensors at every short time interval. IBM Netezza is one of the examples of DBMS that can handle big data. The parallel/distributed data management techniques of Hadoop and MapReduce are also highly relevant to deal with big data because usually the big data is not centralized but distributed among several computing resources.

Finally, data integration is an important issue for complex systems with multiple components like the smart grid. Data from different sources, probably by different vendors, having different formats and semantics are to be systematically integrated to form a single uniform data source, which can be either virtual or physical. Such an integrated data source can facilitate an integrated information system that streamlines various business processes in a utility company. Most common data integration techniques are data warehousing, XML, and ontology-based techniques.

A high-level diagram illustrating the interrelationships among the various modern database management technologies and their applications in the different areas of power grid data management is shown as Figure 1.

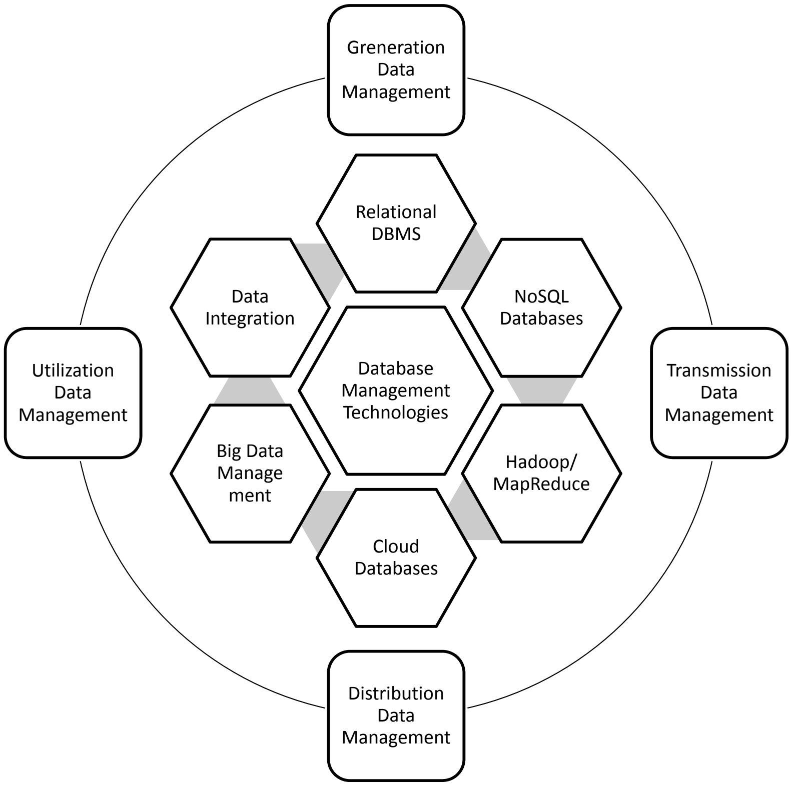
Greneration Data Management
Relational DBMS
Data Integration
NoSQL Databases
Utilization Data Management
Database Management Technologies
Transmission Data Management
Big Data Manage ment
Hadoop/ MapReduce
Cloud Databases
Distribution Data Management

Fig. 1 Modern database management technologies and their applications in power grid database management.

2.2 Generation Data Management

Power plants generate electrical power from traditional sources such as natural gas, petroleum, coal, nuclear, or hydro power as well as modern renewable sources such as wind or solar power. Database systems for power plants have different structures and contents depending on the type of the energy source.

Li *et al.* [45] describes a database system for a coal-based power plant which records and processes the data specific to coal-operated stream turbines (such as main steam pressure, feed water temperature, reheater spray, flue gas temperature, excess air coefficient, and condenser vacuum).

Huang *et al.* [36] discusses data management systems for a hydro power plant, particularly the automatic generator tripping and load shedding system installed at the Churchill Falls hydro power plant in Labrador, Canada by Hydro-Quebec.

Swartz *et al.* [66] proposes a wireless sensors network-based data collection and management system for wind farms to provide information about the dynamic behavior of the wind turbines and their response to loading.

2.3 Transmission and Distribution Data Management

After the power has been generated, it is transformed into high-voltage electricity using step-up transformers, and is transmitted along the transmission lines to multiple substations. At a substation, the electricity's voltage is transformed again to a level suitable for consumption by using a step-down transformer. Then, the electricity is distributed to the consumers for utilization.

Early examples of database systems for power transmission/distribution systems in the literature are [15] and [56].

Generally, distributed control system (DCS) and supervisory control and data acquisition (SCADA) are employed to operate various equipment used in power transmission and distribution. DCS and SCADA are usually proprietary systems from big industrial players in the power industry such as GE [2] and Siemens [9]. Being proprietary systems, they are closed and sometimes can be legacy systems. In some cases, the data format they provide can be non-standard (especially for old legacy systems). Thus, acquiring data from all these systems to build a common database system can be sometimes difficult. In the worst cases, manual data entry can be required [64].

It is not uncommon to have systems from multiple vendors in a single power facility. In order to provide a standardized interface and allow easy exchange of data among different prices of software by different vendors, common information model (CIM) [72, 63], generic substation events [74], and substation configuration language (SCL) [78] have been proposed.

Depending on the nature of application, the data generated by various pieces of power system equipment have to be stored in different formats [51]. They are:

- Raw waveforms (voltage and currents) sampled at relatively high sampling frequencies.
- Pre-processed waveforms (e.g., RMS) typically sampled at low sampling frequencies.
- Status variables (e.g., if a relay is opened or closed) typically sampled at low sampling frequencies.

A number of white papers and research articles on the database systems for power transmission/distribution systems exist in the literature. Some examples, which are by no means complete, are as follows.

Simpson [64] describes a power system database recording transformer name plate data, single line diagrams, measured data, protective device coordination, harmonic analysis, transistent calculation, load flow calculation, and short circuit calculation. Martinez *et al.* [49] gives detailed descriptions about comprehensive archiving and management of power system data for real-time performance monitoring using CERTS (Consortium for Electric Reliability Technology Solutions) architecture. Qiu *et al.* [57] proposes a system of real-time and historical (archived) databases to allow operations, controls, and analysis of power transmission and distribution. An example of a practical database schema to be used for in transmission utility enterprise-wide framework using ArcGIS, ArcSDE, Microsoft SQL Server and .NET is given in [55]. In [47] and [65], the issues of data integration in power systems are discussed. In [81], Zheng *et al.* proposes a cloud computing and cloud database framework for substations of the smart gird. Rusitschka *et al.* [60] discusses the use of cloud data management for outage management [77] and virtual power plant [14].

A comprehensive list of monitoring subsystems whose measurement data are to be collected and stored in the database for a modern power transmission/distribution system of the smart grid is provided by Kaplan *et al.* [38]. These collected data allows advanced tools to analyze system conditions, perform real-time contingency analysis, and initiate a necessary course of action as needed. These monitoring subsystems as described in [38] are:

- **Wide-area monitoring system:** GPS (global positioning system)-based phasor monitoring unit (PMU) that measures the instantaneous magnitude of voltage or current at a selected grid location. This provides a global and dynamic view of the power system.
- **Dynamic line rating technology:** measures the ampacity of a line in real time.
- **Conductor/ compression connector sensor:** measures conductor temperature to allow accurate dynamic rating of overhead lines and line sag, thus determining line rating.
- **Insulation confirmation leakage sensor:** continuously monitors leakage current and extracts key parameters. This is critical to determining when an insulator flashover is imminent due to contamination.
- **Backscatter radio:** provides improved data and warning of transmission and distribution component failure.

- **Electronic instrument transformer**: replaces precise electromagnetic devices (such as current transformers and potential transformers) that convert high voltages and currents to manageable, measurable levels.
- **Other monitoring systems:**
  - Fiber-optic, temperature monitoring system.
  - Circuit breaker real-time monitoring system.
  - Cable monitor.
  - Battery monitor.
  - Sophisticated monitoring tool: combines several different temperature and current measurements.

2.4 Utilization Data Management

The distributed electricity is utilized (consumed) at the consumers' end. Consumers may be of several types: residential (e.g., individual houses and apartment buildings), commercial (e.g., banks), industrial (e.g., factories) , transportation (e.g., subways), emergency services (e.g., hospitals), and governmental services (e.g., police), etc. Obviously, power utilization is most visible aspect of a power grid for the public.

In the old traditional grid, a traditional meter on customer's premises is read by a meter reading staff at a regular interval (e.g., once a month), and the meter readings (utilization data) are manually entered into the database system in the utility company. This utilization data is quite passive and is mainly used for the purpose of billing. It has no or little use in real-time monitoring and control of the power system in operation because of a very long time lag (e.g., up to one month) between actual power utilization and data gathering.

However, in the era of the smart grid, smart meters are installed in consumers' premises. Among its many functionality, the main function of a smart meter is to record and transmit the utilization data to the utility company at relatively short time intervals (e.g., every 5, 10, or 15 minutes). The utilization data can be either fine-grained (separate data for individual appliances or groups of appliances in the same electrical circuits) or coarse-grained (aggregated data for the whole premises). A smart meter may he equipped with a small local storage (e.g., SD card) to store some intermediate utilization data.

The data collection is hierarchical in nature. The power utilization data from a number of smart meters are first transmitted to a data concentrator, and a number of data concentrators relay the data to the central server at the utility company where the data is stored in the utilization database covering a large number of consumers.

The above process of data collection is called automatic meter reading (AMR) [70]. It is a one-way communication process in which the data is transmitted from the smart meter end to the server end through the data concentrator. Later AMR is improved into a more sophisticated system named advanced metering infrastructure (AMI) [69, 32]. AMI allows two-way communication between the smart meter and the server end. The server can send messages regarding real-time pricing, control commands to switch on/off certain appliances, etc. to the smart meter.

In a smart home environment, where modern technologies such as smart appliances, intelligent heating, ventilation, and air conditioning (HVAC), roof-top solar generation, and electric/hybrid vehicles coexist, a smart meter alone will not be able to handle all the data regarding the operations and interactions among those equipment. In addition to the smart meter, there requires a local PC/server to host an integrated information management platform. Its purpose is to store, process, and manage the data from all those smart installations and to communicate with the utility to exchange the relevant information regarding them. Lui *et al.* [48] describes in detail such a platform namely Whirlpool Integrated Services Environment (WISE), which is a proprietary system.

Since every customer connected to the smart grid is expected to generate a large volume of data from his/her smart meter as well as from the other multiple smart equipment, there is a pressing need for the smart grid to handle the big data (as also discussed above in Section). In [37], the application of IBM's big data technologies for smart meters is discussed.

Kaplan *et al.* [38] provides the following detailed list of customer-focused applications (for each of which the relevant utilization data are needed to be recorded and processed).

- **Consumer gateway:**
  - Bi-directional communications between service organizations and equipment on customer premises.
  - Advanced meter reading.
  - Time-of-use and real-time pricing (RTP).
  - Load control.
  - Metering information and energy analysis via website.
  - Outage detection and notification.
  - Metering aggregation for multiple sites and facilities.
  - Integration of customer-owned generation.
  - Remote power-quality monitoring and services.
  - Remote equipment performance diagnosis.
  - Theft control.
  - Building energy management systems.
  - Automatic load controls integrated with RTP.
  - Monitoring of electrical consumption of total load and, in some cases, various load components.
  - Functions embodied in meters, cable modems, set-top boxes, thermostats, etc.
- **Residential consumer network:** subset of consumer gateway concept.
  - Reads the meter, connects controllable loads, and communicates with service providers.

  - End-users and suppliers monitor and control the use and cost of various resources (e.g., electricity, gas, water, temperature, air quality, secure access, and remote diagnostics).
  - Consumers monitor energy use and determine control strategies in response to price signals.

- **Advanced meter:**
  - Employs digital technology to measure and record electrical parameters (e.g., watts, volts, and kilowatt hours).
  - Communication ports link to central control and distributed loads.
  - Provides consumption data to both consumer and supplier.
  - Switches loads on and off in some cases.

At the utility side, billing is the most important application for the utilization data. Arenas-Martinez *et al.* [12] developed a smart grid simulation platform to study the pros and cons of different database architectures for massive customer billing. These architectures are single relational database, distributed relational database, key-value distributed database storage, and hybrid storage (DBMS and FS).

Another utility-side application replying on the utilization data is real-time pricing to facilitate demand response by having the consumers reduce their demand at critical times or in response to market prices [23].

3 Power Grid Data Mining

In this section, we will cover the data mining technologies in general and then applications of data mining for a power grid in its four subsystems: generation, transmission, distribution, and utilization.

3.1 Data Mining Technologies

The purpose of data mining is to uncover the knowledge or interesting patterns of data that lie within a large database and use them for decision support at various levels (strategic, tactical, or operational). Data mining is also known by other names such as data analytics, knowledge discovery, and statistical data analysis. Data mining is closely related to database management, machine learning, artificial intelligence and statistics.

The most common data mining tasks are:

- **Frequent pattern mining:** to discover some sub-patterns or motifs that occur frequently in a dataset. (Note: a dataset means a collection of data organized in rows and columns. It can be a table in relational DBMS or just a comma-separated values (CSV) file in FS. A row represents an instance and a column represents an attribute.) Some well-known frequent pattern mining algorithms include *a priori*, *FP-tree*, and *Eclat*.

- **Association rule mining:** to uncover which causes usually lead to which effects in a dataset. The association rules can generally be derived from the frequent patterns described above.
- **Classification:** to classify instances in a dataset into pre-defined groups (called class labels). Classification is a supervised learning process in which we first have to train the classifier with instances whose class labels are know. Then, we use this training classifier to predict the class labels of the new instances whose labels are not know yet. Some popular classification algorithms are *decision tree*, *naive Bayes*, *artificial neural networks*, *hidden Markov model*, *support vector machine*, and *k nearest neighbors*.
- **Clustering:** to organize similar instances in a dataset into groups which are not predefined. Clustering is an unsupervised learning process in which we do not know the class labels of all the instances in the data set in advance. The number of groups (clusters) may or may not be pre-defined depending on the clustering algorithm. Some widely-used clustering algorithms are *k-means*, *fuzzy c-means*, *expectation maximization*, *DBSCAN*, *BIRCH*, and *hierarchical clustering*.
- **Regression:** to predict the value of the target attribute (called dependent variable) of an instance based on the values of other attributes (independent variables). Regression is also a type of supervised learning which works in the similar way as classification. Their main difference is that while the outputs of classification are class labels (discrete values), those of regression are real numbers (continuous values). Some common regression algorithms are *Gauss-Newton algorithm*, *logistic regression*, *neural network regression*, and *support vector regression*, and *autoregressive integrated moving average (ARIMA)*.
- **Outlier detection:** to identify anomalous instances, which might be interesting or indicate errors and require further investigation. It can be supervised, unsupervised, or semi-supervised learning. Some popular methods are *local outlier factor*, *single-class support vector machine*, *replicator neural networks*, and *cluster analysis*.

Data can rarely be mined in their raw forms as originally stored in the DBMS or FS. We usually need to perform one or more of the following data processing tasks [31] before performing a data mining task.

- **Data cleaning:** to fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- **Data integration:** to integrate multiple databases, data cubes, or files.
- **Data reduction:** to obtain reduced representation in volume but produces the same or similar analytical results. It may be in the form of dimensionality reduction, numerosity reduction, or data compression. Data reduction is usually done for the sake of efficiency and/or better generalization.
- **Data transformation and discretization:** to normalize data, aggregate data, and generate concept hierarchy.

After the data mining task has been performed, the result can be optionally presented in a visual format in order to better facilitate decision making by the user.

Some popular data mining software are SAS Enterprise Miner (proprietary), IBM SPSS Modeler (proprietary), Oracle Data Mining (proprietary), Microsoft Analysis Services (proprietary), Weka (open source), RapidMiner (open source), and ELKI (open source).

In addition to the traditional data mining paradigm on static and centralized data, the new paradigms of distributed data mining [67], data stream mining [29], and time-series data mining [41] are much relevant to the smart grid because of its very nature of distributiveness and having to deal with numerous data streams and time series data from various data sources: smart meters, sensors, and power system machinery.

Privacy is one of the top concerns in the smart grid's deployment especially from consumer's perspective [43]. Thus, privacy preserving data mining techniques [46] are much relevant for mining the data in the smart grid. An example of a proposed framework for privacy-preserving data integration and subsequent analysis for the smart grid is [40].

A high-level diagram depicting the interrelationships among the various data mining technologies and their applications in the different subsystems of power grids is shown as Figure 2.

3.2 Data Mining for Generation

In a similar manner as discussed above in Section 2.2, the data mining applications for power generation can be quite diverse because of the different natures of power sources. Li *et al.* [45] proposes a fault diagnosis system for a coal-based power plant using association rule mining. In [44], the operational performance and the efficiency characteristics for photovoltaic power generation are analyzed against various environmental conditions using statistical analysis.

For fossil fuel-based power plants where the amount of power produced can be fully controlled, the amount of generation (supply) is much dependent on the amount of electricity load (demand). So, forecasting the future load enables them to plan for the required fuel accordingly, and consequently, accurate forecasting can save utility companies millions of dollars a year [22]. Also, for renewable energy generations, load forecasting can help the utilities to plan ahead to shave the peak load by means of demand response mechanisms [23] so that the demand will not exceed the available power output from the renewable source.

Load forecasting can be for very-short term (24 hours ahead of the present time), short term ($\sim$2 weeks), medium term ($\sim$3 years), and long term ($\sim$30 years) [34]. Some examples of load forecasting methods in the literature are: Deng and Jirutitijaroen [19] using the time series models of multiplicative decomposition and seasonal ARIMA, Hong [34] using multiple linear regression, Zhang *et al.* [80] using

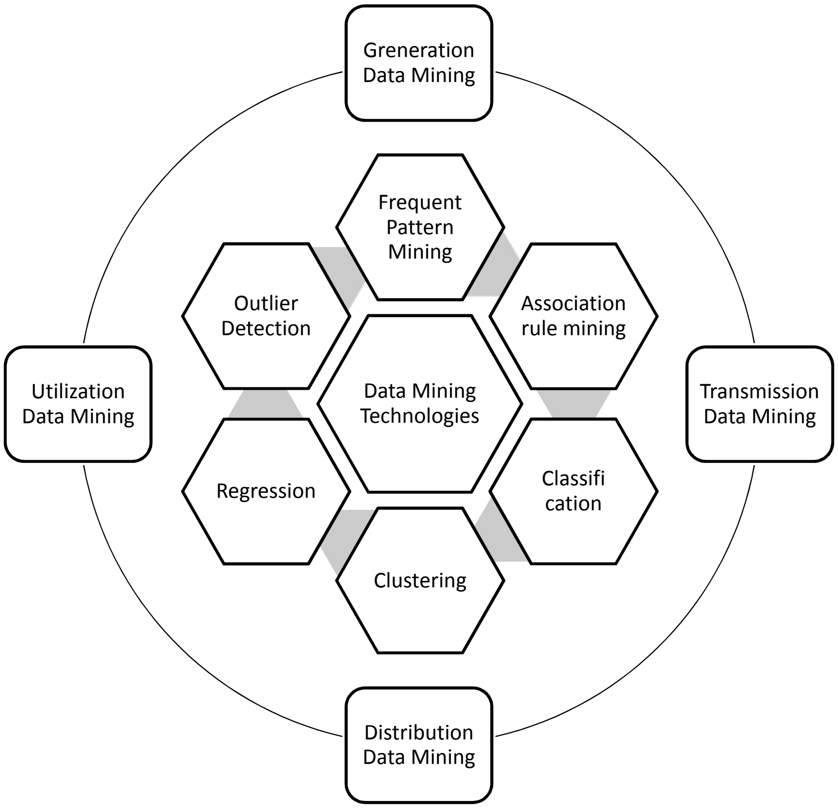

Fig. 2 Data mining technologies and their applications in power grids.

artificial neural network, and Aung *et al.* [13] using least-square support vector regression. Taylor [68] provides a good survey and evaluation of several existing load forecasting methods.

3.3 Data Mining for Transmission and Distribution

The prospects and challenges of data mining for the smart grid, particularly in the areas of transmission, distribution, and utilization are highlighted in [39]. Similarly, Ramchurn *et al.* [58] discusses the uses of artificial intelligence and data mining solutions to provide “smartness” to the smart grid.

There exists a number of papers in the literature regarding the application of data mining for power transmission and distribution systems. Some examples, which are by no means exhaustive, are as follows.

Dissolved gas analysis (DGA) [73] is the study of dissolved gases in transformer oil (insulating oil which is stable at high temperatures and possesses excellent electrical insulating properties). The information about the gases being generated by a particular transformer unit can be very useful in fault detection and maintenance. Sharma *et al.* [62] provides a survey on artificial intelligence and data mining techniques for DGA.

Power system state estimation provides an estimate for all metered and unmetered quantities throughout the whole power system. It is useful in ensuring the stability of the grid and preventing blackouts. Chen *et al.* [17] describes computation of power system state estimation using weighted least-square method on a high-performance computing platform. Zhong *et al.* [82] tries to solve a more specific problem of state assessment for transformer equipment using association rule mining and fuzzy logic.

Islanding detection is also important for the stability of a grid in which multiple small distributed renewable energy generation sources are integrated into the main grid. Islanding occurs when part of the network becomes disconnected from the grid, and is powered by one or more distributed generations only. Such an event can potentially lead to problems in the grid. Samantaray *et al.* [61] proposed an islanding detection system using a rule-based approach that employs fuzzy membership functions. In [53], naive-Bayes classifier is used to solve the problem of islanding detection.

Again, fault identification and fault cause identification are obviously important problems for power systems. Calderaro *et al.* [16] uses Petri Nets to solve the fault identification problem. Xu *et al.* [79] tries to identify fault causes in a power distribution system using a fuzzy classification algorithm.

Contingency analytics is to understand the impact of potential component failures and assess the power system's capability to tolerate them. Adolf *et al.* [11] develops a filtering technique based on multi-criteria optimization to address it.

Power quality is another important issue in the power system especially in the smart grid era. Common problems that can disturb the quality of power are sags (undervoltages), harmonics, spikes, and imbalances [38]. He *et al.* [33] proposes a self-organizing learning array system for power quality classification based on wavelet transform. Hongke and Linhai [35] describes a practical data analysis platform for power quality using Microsoft SQL Server and OLAP (Online Analytical Processing).

The reliability of the power distribution network is an important issue especially for the old networks that were first setup nearly a century ago. Gross *et al.* [30] develops a support vector machine-based model to rate the feeder lines in New York City for their reliability and identify the ones that needs maintenance or replacement.

Morais *et al.* [51] presents a good survey of 13 research articles on data mining for power systems for various purposes such as fault classification and location, detection and diagnosis of transient faults, power quality detection for power system disturbances, etc. Similarly, Mori [52] provides a list of 42 research papers on various applications of data mining for power systems.

Apart from the physical power system, the logical energy market draws much attention recently especially after its deregulation. Price forecasting is an indispensable tool for both the energy wholesaler and the retailer in such a market. Arenas-Martinez *et al.* [50] presents a price forecasting model using local sequence patterns, while Neupane *et al.* [54] tackles price forecasting by means of artificial neural networks.

3.4 Data Mining for Utilization

At the power utilization (demand) side, load forecasting for large commercial and residential buildings plays a crucial role. Building load forecasting is an integral part of a building management system. It enables the building operator to plan ahead, shave loads if required, and carry out fault identification and diagnosis in the building's electrical system if necessary. Fernandez *et al.* [27] presents a study on building load forecasting using autoregressive model, polynomial model, neural network, and support vector machine. Edwards *et al.* [20] compares the performance of seven machine learning/data mining methods for load forecasting in buildings.

Customer profiling is also related to the demand-side load forecasting task mentioned above. It is useful both for customer behavior prediction for appliance scheduling automation and dynamic pricing of electricity to suit individual customers' usage patterns. Proposed research works for customer profiling using data mining techniques include [18], [26], and [28].

Finally, security is one of the major concerns for the smart grid's deployment at the customer side [42]. To partially address this problem, Faisal *et al.* [21] presents an intrusion detection system for advance metering infrastructure (AMI) using data stream mining methods. Fatemieh *et al.* [25] applies classification techniques to improve the attack resilience of TV spectrum data fusion for AMI communications.

4 Conclusion

Database systems are one of the keystones of the ICT infrastructure that provides smartness to the smart gird. In this chapter, we have discussed both the conventional and the state-of-the-art database systems technologies regarding database management and data mining and their applications to the smart grid. We hope our chapter to be useful as a reference material for both the researchers and the practitioners of the smart grid.

Acknowledgements The author thank the Government and Abu Dhabi, United Arab Emirates for sponsoring this research trough its funding of Masdar Institute–Massachusetts Institute of Technology (MIT) collaborative research project titled "Data Mining for Smart Grids".

References

1. Apache Hadoop. http://hadoop.apache.org
2. GE Power Controls. http://www.gepowercontrols.com
3. Google Inc. http://www.google.com
4. IBM corporation. http://www.ibm.com
5. Microsoft Corporation. http://www.microsoft.com
6. Oracle Corporation. http://www.oracle.com
7. SAP AG. http://www.sap.com
8. SAS Institute, Inc. http://www.sas.com
9. Siemens Energy. http://www.energy.siemens.com
10. Teradata Corporation. http://www.teradata.com
11. Adolf, R., Haglin, D., Halappanavar, M., Chen, Y., Huang, Z.: Techniques for improving filters in power grid contingency analysis. In: Proceedings of the 7th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), pp. 599–611 (2011)
12. Arenas-Martinez, M., Herrero-Lopez, S., Sanchez, A., Williams, J.R., Roth, P., Hofmann, P., Zeier, A.: A comparative study of data storage and processing architectures for the smart grid. In: Proceedings of the 1st IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 285–290 (2010)
13. Aung, Z., Toukhy, M., Williams, J., Sanchez, A., Herrero, S.: Towards accurate electricity load forecasting in smart grids. In: Proceedings of the 4th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA), pp. 51–57 (2012)
14. Awerbuch, S., Preston, A.M.: The Virtual Utility: Accounting, Technology and Competitive Aspects of the Emerging Industry. Kluwer Academic Publisher (1997)
15. Ben-Yaacov, G.Z.: Interactive computation and data management for power system studies. The Computer Journal **22**, 76–79 (1979)
16. Calderaro, V., Hadjicostis, C., Piccolo, A., Siano, P.: Failure identification in smart grids based on Petri Net modeling. IEEE Transactions on Industrial Electronics **58**, 4613–4623 (2011)
17. Chen, Y., Huang, Z., Liu, Y., Rice, M.J., Jin, S.: Computational challenges for power system operation. Proceedings of the 2012 Hawaii International Conference on System Sciences (HICSS) pp. 2141–2150 (2012)
18. Chicco, G., Napoli, R., Postolache, P., Scutariu, M., Toader, C.: Customer characterization options for improving the tariff offer. IEEE Transactions on Power Systems **18**, 381–387 (2003)
19. Deng, J., Jirutitijaroen, P.: Short-term load forecasting using time series analysis: A case study for Singapore. In: Proceedings of the 2010 IEEE Conference on Cybernetics and Intelligent Systems (CIS), pp. 231–236 (2010)
20. Edwards, R.E., New, J., Parker, L.E.: Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings **49**, 591–603 (2012)
21. Faisal, M.A., Aung, Z., Williams, J.R., Sanchez, A.: Securing advanced metering infrastructure using intrusion detection system with data stream mining. In: Proceedings of the 2012 Pacific Asia Workshop on Intelligence and Security Informatics (PAISI), pp. 96–111 (2012)
22. Fan, S., Chen, L., Lee, W.: Short-term load forecasting using comprehensive combination based on multi-meteorological information. In: Proceedings of the 2008 IEEE/IAS Industrial and Commercial Power Systems Technical Conference (ICPS), pp. 1–7 (2008)
23. Fan, Z.: Distributed demand response and user adaptation in smart grids. In: Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 726–729 (2011)
24. Farhangi, H.: The path of the smart grid. IEEE Power and Energy Magazine **8**, 18–28 (2010)
25. Fatemieh, O., Chandra, R., Gunter, C.A.: Low cost and secure smart meter communications using the TV white spaces. In: Proceedings of the 2010 IEEE International Symposium on Resilient Control Systems (ISRCS), pp. 1–6 (2010)
26. Fernandes, R.A.S., Silva, I.N., Oleskovicz, M.: Identification of residential load profile in the Smart Grid context. In: Proceedings of the 2010 IEEE Power and Energy Society General Meeting, pp. 1–6 (2010)

27. Fernandez, I., Borges, C.E., Penya, Y.K.: Efficient building load forecasting. In: Proceedings of the 16th IEEE Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–8 (2011)
28. Figueiredo, V., Rodrigues, F., Vale, Z., Gouveia, J.B.: An electric energy consumer characterization framework based on data mining techniques. IEEE Transactions on Power Systems **20**, 596–602 (2005)
29. Gama, J.: Knowledge Discovery from Data Streams. Chapman and Hall/CRC (2010)
30. Gross, P., Boulanger, A., Arias, M., Waltz, D., Long, P.M., Lawson, C., Anderson, R., Koenig, M., Mastrocinque, M., Fairechio, W., Johnson, J.A., Lee, S., Doherty, F., Kressner, A.: Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In: Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence (IAAI), vol. 2, pp. 1705–1711 (2006)
31. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2011)
32. Hart, D.G.: Using AMI to realize the Smart Grid. In: Proceedings of the Conference on Power and Energy Society General Meeting — Conversion and Delivery of Electrical Energy in the 21st Century, pp. 20–24 (2008)
33. He, H., Starzyk, J.: A self-organizing learning array system for power quality classification based on wavelet transform. IEEE Transactions on Power Delivery **21**, 286–295 (2006)
34. Hong, T.: Short term electric load forecasting. Ph.D. thesis, North Carolina State University, USA (2010)
35. Hongke, H., Linhai, Q.: Application and research of multidimensional data analysis in power quality. In: Proceedings of the 2010 International Conference on Computer Design and Applications (ICCDA), vol. 1, pp. 390–393 (2010)
36. Huang, J.A., Vanier, G., Valette, A., Harrison, S., Wehenkel, L.: Application of data mining techniques for automat settings in emergency control at Hydro-Quebec. In: Proceedings of the 2003 IEEE Power Engineering Society General Meeting, vol. 4, pp. 2037–2044 (2003)
37. IBM Software Group: Managing big data for smart grids and smart meters. Tech. rep., IBM Corporation (2012)
38. Kaplan, S.M., Sissine, F., Abel, A., Wellinghoff, J., Kelly, S.G., Hoecker, J.J.: Smart Grid: Modernizing Electric Power Transmission and Distribution; Energy Independence, Storage and Security; Energy Independence and Security Act of 2007 (EISA); Improving Electrical Grid Efficiency, Communication, Reliability, and Resiliency; Integrating New and Renewable Energy Sources. TheCapitol.Net, Inc. (2009)
39. Krishnaswamy, S.: Energy analytics: When data mining meets the smart grid (2012). http://smartgrid.i2r.a-star.edu.sg/2012/slides/i2r.pdf
40. Kursawe, K., Danezis, G., Kohlweiss, M.: Privacy-friendly aggregation for the smart-grid. In: Proceedings of the 11th International Symposium on Privacy Enhancing Technologies (PETS), pp. 175–191 (2011)
41. Last, M., Kandel, A., Bunke, H.: Data Mining in Time Series Databases. Word Scientific Press (2004)
42. Li, D., Aung, Z., Williams, J., Sanchez, A.: Efficient authentication scheme for data aggregation in smart grid with fault tolerance and fault diagnosis. In: Proceedings of the 2012 IEEE PES Conference on Innovative Smart Grid Technologies (ISGT), pp. 1–8 (2012)
43. Li, D., Aung, Z., Williams, J., Sanchez, A.: P3: Privacy preservation protocol for appliance control application. In: Proceedings of the 3rd IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 294–299 (2012)
44. Li, D.H.W., Cheung, G.H.W., Lam, J.C.: Analysis of the operational performance and efficiency characteristic for photovoltaic system in Hong Kong. Energy Conversion and Management **46**, 1107–1118 (2005)
45. Li, J.q., Wang, S.l., Niu, C.l., Liu, J.z.: Research and application of data mining technique in power plant. In: Proceedings of the 2008 International Symposium on Computational Intelligence and Design (ISCID), vol. 2, pp. 250–253 (2008)
46. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology (CRYPTO), pp. 36–54 (2000)

47. Lu, B., Song, W.: Research on heterogeneous data integration for smart grid. In: Proceedings of the 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 3, pp. 52–56 (2010)
48. Lui, T.J., Stirling, W., Marcy, H.O.: Get smart: Using demand response with appliances to cut peak energy use, drive energy conservation, enable renewable energy sources and reduce greenhouse-gas emissions. IEEE Power and Energy Magazine **8**, 66–78 (2010)
49. Martinez, C., Huang, H., Guttromson, R.: Archiving and management of power systems data for real-time performance monitoring platform. Tech. rep., Consortium of Electric Reliability Technology Solutions (2005)
50. Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C., Aguilar-Ruiz, J.S.: Energy time series forecasting based on pattern sequence similarity. IEEE Transactions on Knowledge and Data Engineering **23**, 1230–1243 (2011)
51. Morais, J., Pires, Y., Cardoso, C., Klautau, A.: An overview of data mining techniques applied to power systems. In: J. Ponce, A. Karahoca (eds.) Data Mining and Knowledge Discovery in Real Life Applications. I-Tech Education and Publishing (2009)
52. Mori, H.: State-of-the-art overview on data mining in power systems. In: Proceedings of the 2006 IEEE PES Power Systems Conference and Exposition (PSCE), pp. 33–34 (2006)
53. Najy, W., Zeineldin, H., Alaboudy, A.K., Woon, W.L.: A Bayesian passive islanding detection method for inverter-based distributed generation using ESPRIT. IEEE Transactions on Power Delivery **26**, 2687–2696 (2011)
54. Neupane, B., Perera, K.S., Aung, Z., Woon, W.L.: Artificial neural network-based electricity price forecasting for smart grid deployment. In: Proceedings of the 2012 IEEE International Conference on Computer Systems and Industrial Informatics (ICCSII), pp. 1–6 (2012)
55. Owoola, M.A.: A generic spatial database schema for a typical electric transmission utility. In: Proceedings of the Geospatial Information and Technology Association's 27th Annual Conference (GITA), pp. 1–12 (2004)
56. Papadakis, M., Hatzjargyriou, N., Gazidellis, D.: Interactive data management system for power system planning studies. IEEE Transactions on Power Systems **4**, 329–335 (1989)
57. Qiu, J., Liu, J., Hou, Y., Zhang, J.: Use of real-time/historical database in smart grid. In: Proceedings of the 2011 International Conference on Electric Information and Control Engineering (ICEICE), pp. 1883–1886 (2011)
58. Ramchurn, S.D., Vytelingum, P., Rogers, A., Jennings, N.R.: Putting the "smarts" into the smart grid: A grand challenge for artificial intelligence. Communications of the ACM **55**, 86–97 (2012)
59. Rizvi, S.S., Chung, T.S.: Flash memory SSD based DBMS for high performance computing embedded and multimedia systems. In: Proceedings of the 2010 International Conference on Computer Engineering and Systems (ICCES), pp. 183–188 (2010)
60. Rusitschka, S., Eger, K., Gerdes, C.: Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain. In: Proceedings of the 1st IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 483–488 (2010)
61. Samantaray, S.R., El-Arroudi, K., Joós, G., Kamwa, I.: A fuzzy rule-based approach for islanding detection in distributed generation. IEEE Transactions on Power Delivery **25**, 1427 –1433 (2010)
62. Sharma, N.K., Tiwari, P.K., Sood, Y.R.: Review of artificial intelligence techniques application to dissolved gas analysis on power transformer. International Journal of Computer and Electrical Engineering **3**, 577–582 (2011)
63. Simmins, J.J.: The impact of PAP 8 on the Common Information Model (CIM). In: Proceedings of the 2011 IEEE/PES Power Systems Conference and Exposition (PSCE), pp. 1–2 (2011)
64. Simpson, R.H.: Power system database management. In: Conference Record of 2000 Annual Pulp and Paper Industry Technical Conference (PPIC), pp. 79–83 (2000)
65. SISCO, Inc: Integration of substation data.
http://cimug.ucaiug.org/KB/Knowledge Base/Integration of Substation Data ver 06.pdf
66. Swartz, R.A., Lynch, J.P., Zerbst, S., Sweetman, B., Rolfes, R.: Structural monitoring of wind turbines using wireless sensor networks. Smart Structures and Systems **6**, 1–14 (2010)

67. Talia, D., Trunfio, P.: How distributed data mining tasks can thrive as knowledge services? Communications of the ACM **53**, 132–137 (2010)
68. Taylor, J.W.: An evaluation of methods for very short term electricity demand forecasting using minute-by-minute British data. International Journal of Forecasting **24**, 645–658 (2008)
69. Wikipedia: Advanced metering infrastructure (2013). http://en.wikipedia.org/wiki/Advanced_Metering_Infrastructure
70. Wikipedia: Automatic meter reading (2013). http://en.wikipedia.org/wiki/Automatic_meter_reading
71. Wikipedia: Cloud database (2013). http://en.wikipedia.org/wiki/Cloud_database
72. Wikipedia: Common information model (electricity) (2013). http://en.wikipedia.org/wiki/Common_Information_Model_(electricity)
73. Wikipedia: Dissolved gas analyis (2013). http://en.wikipedia.org/wiki/Dissolved_gas_analysis
74. Wikipedia: Generic substation events (2013). http://en.wikipedia.org/wiki/Generic_Substation_Events
75. Wikipedia: MapReduce (2013). http://en.wikipedia.org/wiki/MapReduce
76. Wikipedia: NoSQL (2013). http://en.wikipedia.org/wiki/NoSQL
77. Wikipedia: Outage management system (2013). http://en.wikipedia.org/wiki/Outage_management_system
78. Wikipedia: Substation configuration language (2013). http://en.wikipedia.org/wiki/Substation_Configuration_Language
79. Xu, L., Chow, M.Y., Taylor, L.S.: Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm. IEEE Transactions on Power Systems **22**, 164–171 (2007)
80. Zhang, H.T., Xu, F.Y., Zhou, L.: Artificial neural network for load forecasting in smart grid. In: Proceedings of the 2010 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 6, pp. 3200–3205 (2010)
81. Zheng, L., Chen, S., Hu, Y., He, J.: Applications of cloud computing in the smart grid. In: Proceedings of the 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp. 203–206 (2011)
82. Zhong, W., Sun, Y., Xu, M., Liu, J.: State assessment system of power transformer equipments based on data mining and fuzzy theory. In: Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation (ICICTA), vol. 3, pp. 372–375 (2010)