Energy Resource Management Based on Data Mining and Artificial Intelligence

Raj Bhatnagar, University of Cincinnati
Chandan Rao, Graphet, Inc.

ABSTRACT

Market conditions, global competition and environmental stewardship have created a need for improving energy efficiencies. Therefore, a significant role has emerged for tools and technologies that enable efficient management of energy resources. Technologies and mathematical models for such analyses have existed and are being developed in the fields of data mining and pattern recognition. Implementation of tool sets based on current state of the art techniques developed in the academic world will result in a powerful suite of methods, applications and tools to provide insights into patterns and modes of efficient and inefficient usage and thereby facilitate significant savings of energy. These tools are based on ideas derived from the recent research in the fields of data mining and artificial intelligence. We consider the problem of mining databases containing time series type of energy consumption data for discovering typical temporal characteristics and thus identifying usage patterns. Temporal profiles of a utility (electricity or gas) consumption during a day are recorded and such records for a number of years may be stored in a database. Data mining algorithms described here can determine typical temporal behaviors of utility consumption and the frequency of their occurrence. An interesting pattern is a sequence of observed values that either has a high frequency of occurrence among all the daily profiles, and thus signifies a stable operating mode, or is infrequent but signifies special operating requirements. The capability of these methods to identify operating modes and quantify the duration of these modes provides the ability to objectively specify design requirements and evaluate the energy impact of proposed solutions.

Introduction

Efficient management of energy resource by large industrial users is emerging as a critical aspect. In the last few decades, energy resource management did not receive significant attention due to the relatively low share of energy costs in the operations of industrial plants. Energy management includes corporate commitment, appropriate energy management practices/processes promoted through energy awareness and training and finally meaningful metrics for tracking results, maintaining accountabilities and responding in a timely manner to variations observed. It is becoming important for industries to not only economize on the energy usage but to have a sustainable dimension to it to retain the savings long term. Market conditions, global competition and environmental stewardship have created an urgent need for improving energy efficiencies. A significant market and technological opportunity has emerged for tools and technologies that enable efficient management of energy resource. This paper addresses energy data mining tools and techniques and the potential contribution they can make in setting energy metrics, benchmarking and in assessing potential opportunities.

Results presented in this paper have been developed specifically for energy data by adapting mathematical foundations of some recent advancement in the field of sequential data mining. The research described here has provided encouraging prototypes and will result in a refined and efficiently usable suite of methods, applications and tools to address this emerging market opportunity. Annual energy cost expenditure for the industrial sector in the US is estimated to be in excess of $150 billion per year[DOE EIA 2001]. DOE studies have shown that energy cost reduction in the range of 5 to 15% per annum can be achieved with attractive returns on investment [DOE EERE ITP]. At a 35% return on investment, this represents an ability to justify energy efficiency improvement expenditures in excess of $1.0 billion dollars.

In this paper we outline the potential of some of the recent research in the area of data mining algorithms for discovering frequent patterns in temporal (time series) data. This research is very relevant for the situations encountered in energy consumption domains. The problem of discovering typical sequential patterns in a database, which contains a time series profiles of energy consumption, is of significant interest from the perspective of optimizing the energy consumption patterns. From the daily utility consumption profiles of an industrial plant, it would be of interest to know typical usage patterns and the days on which these patterns are observed. A frequently occurring pattern of inefficient usage during some fixed hours of a day, on some of the days, may point to a problem that may be fixed.

Current Research

Current research and developments in the data mining area are making available some algorithms that can efficiently discover frequent and interesting patterns in databases of large sequential datasets. A sequential dataset in the database may be a profile of electric power consumption measured at 15-minute intervals. The objective is to discover relevant usage patterns and the identities of profiles that correspond to each pattern. That is, the algorithms should discover patterns in the form of subsequences of the daily complete sequences and also identify the days and times at which these subsequences occur. For example, let us consider the following five temporal profiles: abacbde, adabbde, aaaabde, baacbde, and acadbde. Each profile represents observations at seven successive time points. The subsequence, "a-a-bde" is shared by profile # 1, 2, 3, and 5. The subsequence "-bde" is shared by all the five profiles. A "-" at a location signifies a time-point at which the value is not shared by the profiles sharing the rest of the pattern subsequence.

A number of clustering algorithms can be employed to cluster the patterns according to their similarity throughout the observed time interval. In many applications and domains the pattern of interest is described by any subsequence of the observed time points. The task of determining interesting subsequences and their corresponding sets of profiles becomes significantly more complex from the perspective of required computational resources.

Related Data Mining Research

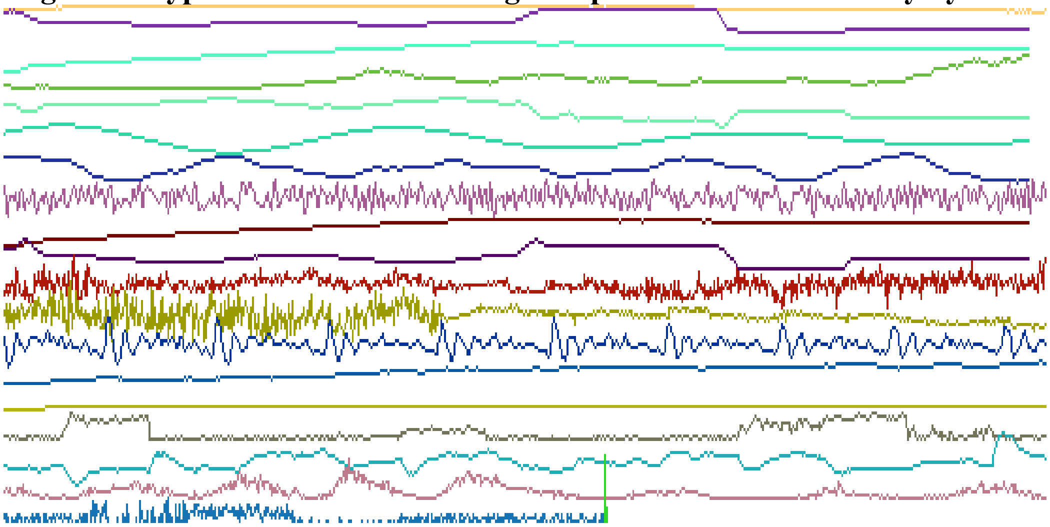
Highly rigorous mathematical models for time-series analysis are not very applicable to many datasets from the energy consumption domain because of relatively high noise levels and frequent dynamic shifts in energy consumption patterns. Also, simultaneous considerations of a number of energy related time-series data profiles is needed to get insightful patterns. These data sources may be electric input measured as voltage and current, gas inflow, temperatures, compressor power, or final product flow from a plant. Only some recent work in data mining can provide a means to extract potentially useful information about a facility's energy consumption.

A number of methodologies have been employed to cluster time-series data. Results of these approaches are presented in [Eise 98, Holt 00, Spel 98, Wen 98]. All these methods employ traditional clustering algorithms and use Euclidean distances or various correlation coefficients as metrics of distance between temporal profiles. When a pattern is defined by a subsequence, it is not possible to compute a distance metric unless the subsequence is identified. These distances are used to drive the hierarchical, agglomerative, or other graph based clustering algorithms [Eise 98] but are not available when the distance between profiles depends on yet to be identified subsequences.

Methodology

A sequential dataset is a table in which each row is a profile of a day's energy consumption. An example of a typical dataset is shown in Figure 1 below:

Figure 1: Typical Dataset Containing Temporal Profiles for Utility Systems



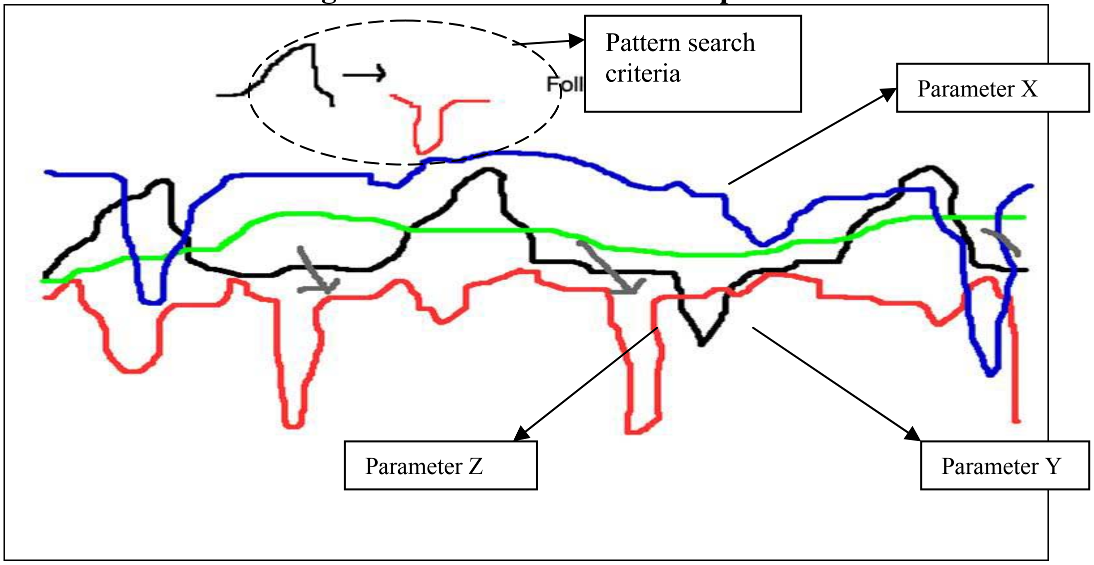
It shows recordings of eighteen different measurements taken at fixed time intervals for one day. Our database may contain such recordings for a number of days going up to a few months or even a couple of years, thus generating a significant amount of data. The interesting patterns to look for in such a database can be of any of the following types:

1. A fixed pattern in a single measured quantity that occurs for some period of time and repeats very often, possibly multiple occurrences during a day, and also during a number of days. Patterns of very large durations signify stable modes of operation and shorter patterns point towards interesting transients that may sometime trigger changes in stable operating modes.
2. A fixed pattern in a single measured quantity that occurs at approximately the same time during a significantly large number of days.
3. A pattern described over a number of different measured quantities as follows: A fixed pattern is observed in some measured quantity "X" and then immediately preceding, following, or overlapping with it another fixed pattern is observed in some other measured quantity "Y". Such synchronized behavior of patterns over a large number of measured quantities can be of significant value.

4. The above three behaviors are described in terms of fixed patterns in time being observed across multiple measurements. Given the noise in such systems and in the measurements, we would like algorithms that can discover the above types of interesting patterns even when the patterns match only approximately.

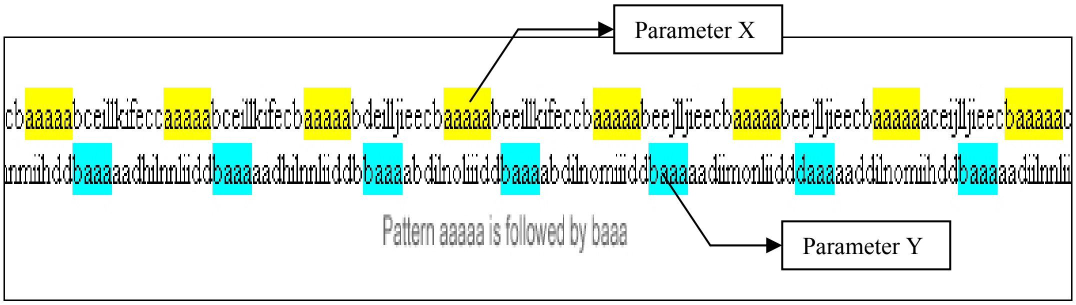
Examples of some such patterns are shown in Figure 2 below.

Figure 2: Search Pattern Example



In the above figures there are patterns in the measured parameters X, Y and Z that recur with certain frequency. Our objective is to identify these patterns even when the match between two occurrences is approximate. In the traditional literature on time series analysis and Fourier analysis one can discover such recurrences but only if the frequency with which they repeat is maintained for large lengths of time. Occurrences of such patterns at random and only at few times in the whole database is of significant interest from our perspective but the traditional analysis techniques are not suitable for discovering these occurrences. The parameter X in the above figure also has a repeating pattern with a DC-shift. We would like our algorithms to discover these patterns also, despite the DC-shift. Figure 3 shows an example of a different type of pattern of interest:

Figure 3: Search Routine Recognition of a Pattern of Interest


Parameter X
cbaaaaabceillkifeccaaaaabceillkifecbaaaaabdeilljieecbaaaaabeeillkifeccbaaaaabeejlljieecbaaaaabeejlljieecbaaaaaaceijlljieecbaaaaaac
nmiihddbaaaaadhilnnliiddbaaaaadhilnnliiddbbaaaabdilnoliiddbaaaabdilnomiiiddbaaaaadiimonliidddaaaaaddilnomiihddbaaaaadilnnli
Pattern aaaaa is followed by baaa
Parameter Y

A measured quantity shows a pattern “aaaaa” occurring with some frequency and another measured quantity shows the pattern “baaa” immediately following in time. Such patterns can shed significant light on the behavior of the system being analyzed.

Technical Overview of Mining Algorithms

A brief outline of the algorithms that help us discover such patterns is given in the following paragraphs.

Phase 0

Conversion to Symbolic Profiles: Most of the sequence mining algorithms in the current literature and also our own development work with symbolic profiles. That is, the quantitative measures are quantized and converted into symbols. That is, the numeric values in the database are converted into a small set of value-ranges and each value-range is assigned a symbol (from the alphabet). In the examples considered here we have used symbols d, c, b, a, *, A, B, C, D to represent the value ranges of monotonically increasing numeric values such that an ‘*’ represents a zero or a near zero value, lower case alphabets represent negative values, and upper case alphabets represent positive values. We use the letter x to represent a missing value in a profile. The quanta typically do not evenly divide the range between 0 and the maximum magnitude. Quantization has been done to match the distribution of all the values contained in the database. In our example datasets the distribution of values is very close to Gaussian and therefore, frequency equalizing boundaries have been chosen for the quantas resulting in smaller quantization steps near the zero magnitude and increasingly bigger steps for quanta of larger magnitudes. A very small example to illustrate our methodology is given below and it consists of the following four strings:

ABacxBC ABCD*bd ABaaxbd x*a**bd

These four strings represent four daily consumption profiles containing observations at seven time points each. The task of identifying interesting temporal hypotheses is completed in the following three more phases:

Phase-1

Determine the set S of all shared contiguous substrings of the original set of profiles such that all occurrences of each string are documented. Data structures called Generalized Suffix Tree for a set of strings can be generated by various recently developed algorithms in time directly proportional to the total size of all the strings taken together. This algorithm will generate all possible substrings of the four example strings and arrange them in a compact information-packed data structure for exploitation by further phases. This is very powerful and encouraging means for string processing algorithms. Typically, for many interesting problems in Computer Science, the time required to process a dataset can increase as a polynomial of the size of the dataset, or in the worst case as an exponential function of the size of the dataset. Having an algorithm that needs time that is a linear function of the size of the dataset is a great advantage and facilitates solving problems with very large datasets. A dataset with five million recorded values can be quantized and then converted into the trees of all possible substrings and their locations in less than five minutes on a standard laptop.

Phase-2

In this phase we use various clustering algorithms from the domain of pattern recognition and artificial intelligence to cluster the sets of substrings derived in Phase-1 such that a cluster contains a number of those substrings that have many approximately matching substrings in them. This phase seeks to maximize the number of days from the point of view of their sharing approximately matching energy consumption patterns. Various tuning parameters of these clustering algorithms can make significant difference in the extent of the approximate match, the length of the pattern substring considered, the nature of approximations to be allowed, and whether or not to incorporate the DC-shift while comparing the patterns. *These algorithms typically require time which is a very small polynomial function of the number of strings to be processed. Therefore, one must make some choices at this stage about the kinds of patterns one is looking for and maintain a narrow scope of investigation. In successive runs, the scope can be altered based on the results obtained in the previous runs and therefore is an interactive procedure where the expertise of an energy domain expert guiding the process is very much required.* In the context of the example containing four strings, this phase will discover facts such as the existence of substring “AB” in first two locations in 3 of the four strings; and also of the existence of the substring “bd” at the last two positions in three of the four strings. More complex substrings with gaps representing unmatched locations can also be easily discovered. These frequent patterns can provide significant insight into the nature of some repetitive phenomena embedded in the observed data.

Phase 3

Having clustered the substrings according to their approximate similarity, we now have obtained patterns for only one measured quantity. The same process can be repeated for each of the several measured quantities. The task in this phase is to discover dependences among the clusters of subsequences found for different measured quantities. We have developed algorithms for such data mining tasks and the research work is continuing. Some of the results with these algorithms are shown in Figure-3. This phase can discover patterns in different measurements that have a time-relationship among each other. The nature of time relationship can be as precise as “follows after 30 seconds” or as open-ended as “follows at any time in future.” These differences can also be made by appropriately adjusting the tuning parameters of the clustering and temporal relationship algorithms of this phase.

Results with Utility Dataset

This methodology has been tested with energy consumption data for an industrial plant. In this dataset electric power (kWh) being consumed is recorded every 15 minutes for a year. The daily usage profiles for the industrial plant are shown in Figure 4 and Figure 5 below. Each profile consists of 96 time points, all 15 minutes apart from each other within a day. The two dips around time point numbers 25 and 75 in each plot are around the start time of a shift in the morning and the end of shift time in the evening. The second dip around time point number 75 in

Figure 4 (profile 1) takes a sharp ‘V’ kind of shape and in Figure 5 (profile 2) it takes a ’U’ kind of shape.

Figure 4: Electric Energy Usage Profile 1

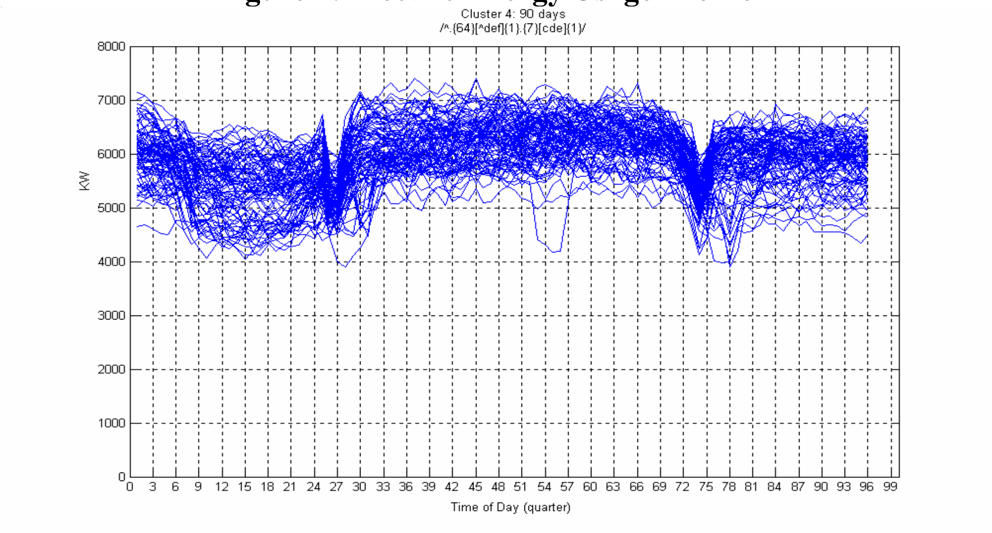
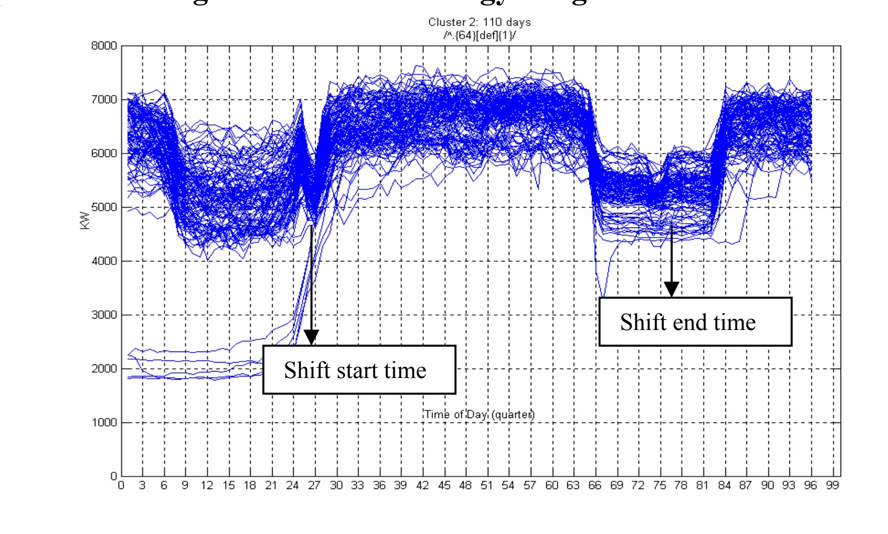


Figure 5: Electric Energy Usage Profile 2



The pattern in Figure 4 ‘V’ shape is identifiable as the result of an inefficient shutdown procedure and the pattern in Figure 5 ‘U’ shape is the result of a proper shutdown procedure. A further examination of the profiles included in each case reveals that the days examined in Figure 4 are weekend days (Friday through Sunday) and all the days examined in Figure 5 are the weekdays. The discrepancy is due to the difference in shutdown procedures during the weekend operating mode.

Energy usage profiles in time series data converted to ‘strings’ (noncontiguous subsequences) can be identified by our methodology.

Figure 6: Electric Usage Profiles

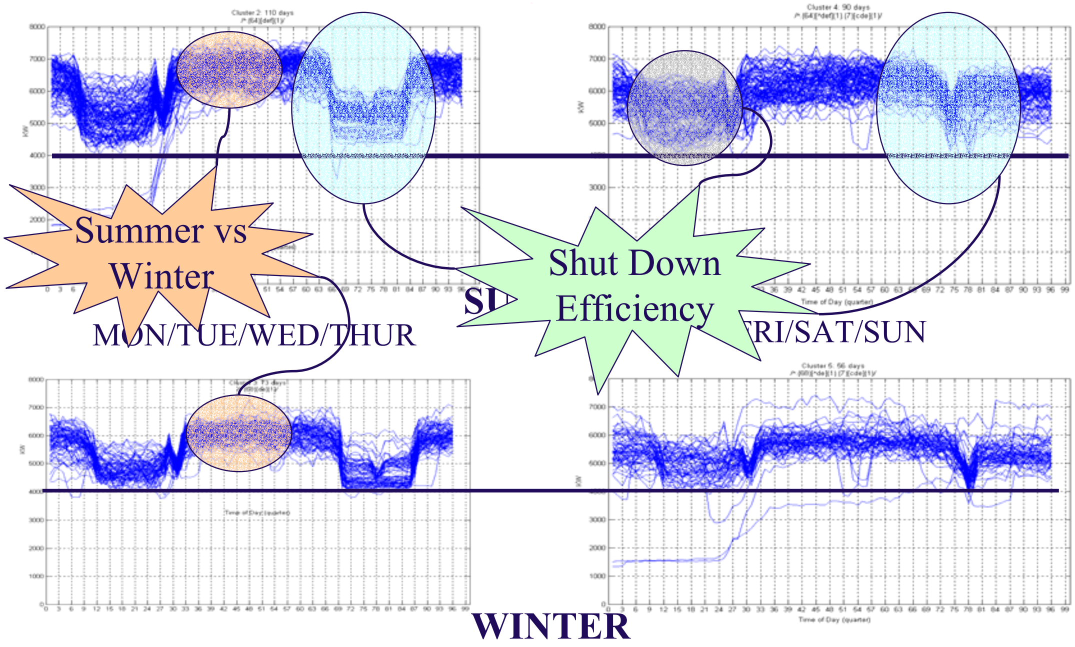


Figure 6 shows some other distinct patterns discovered from a year's worth of data records. Figure 6 shows clusters for winter and summer days demonstrating distinctly identical patterns but lower peaks for winter for electric usage as expected.

Impact on Energy Management

We have demonstrated the viability and mathematical foundations of a tool that can help us identify energy usage patterns and quantify the frequency of occurrences in a very reliable and effective way. The tools show promise in processing large metered datasets and in identifying:

- Energy usage patterns and stable operating modes
- Frequency and duration of operating modes
- Ability to identify transitions/system upsets and critical system requirements for design.

Cost effective management of energy requires qualified resources and enabling technologies that facilitate effective and efficient energy management. Industries find themselves in a position where they require new technology based approaches to first understand the information hidden in the data being collected, and then gain a competitive advantage based on this information. Vast amounts of metered or sequential data are being collected today and stored within existing systems. Energy data mining tools, such as the ones described here, provide an effective means to convert data into information and useful knowledge. Pilot studies performed with the techniques developed by us and outlined here have shown promising results. The insights gained by these energy data mining tools is enabling industries to develop, justify, implement and manage energy conservation projects with confidence.

Conclusions

We have presented a methodology for mining energy usage patterns / transient operating modes from energy data sources. Our approach considers firstly the characteristics we are looking for such as stable mode /transients / correlation between parameters that is identifiable as patterns embedded in a set of profiles observed during some process. The pattern search criteria can be refined by generalization and specialization operations. This capability provides significant power to an investigator looking for energy usage patterns. The retail and consumer industries have reaped significant benefits by being able to identify frequent buying patterns. Similarly, benefits of identifying energy usage patterns can enhance energy conservation measures. Results shown with the energy consumption dataset demonstrate that current data mining algorithms are applicable to the energy domains and can provide very promising and effective results. Meaningful energy usage patterns can be found by our methodology. Such tools can help point towards enhanced energy savings and discovery of inefficient operations in many situations where data is observed and recorded. Further effort is required to transform these tools for identifying energy usage patterns, critical system requirements and provide meaningful energy tracking metrics for sustainable energy conservation. For instance, complex utility systems such as ammonia refrigeration and compressed air are used in most industrial applications. Compressed air requires a huge amount of energy, however the overall efficiency of a typical compressed air system can be as low as 10-15% (DOE 2001 Compressed Air Market Evaluation). A piecemeal approach to building a compressed air system serves to worsen the problem through leaks, mismatched supply/demand, and inappropriate uses. The result is unreliability, energy waste, reduced productivity, and higher operating costs. The ability to monitor such systems and track events associated with their operations would be of significant importance to control energy costs. Only data mining techniques can provide a means of extracting previously unknown and useful knowledge about energy consumption through the discovery of usage patterns occurring in the data. Data mining is an interactive and iterative process. The scenario that is most applicable for energy data mining is where a query retrieves an energy dataset, chooses the right data mining algorithm, and returns results in a form of frequent patterns to the user. This will guide more efficient operating standards both within the facility and the industry as a whole.

References

[Agra 93] Agrawal, R., Faloutsos, C., Swami, A. Efficient Similarity Search in Sequence Databases, in Lecture Notes in Computer Science 730, Springer Verlag, 1993, 69-84.

[Das 98] Das, G., Smyth, P. et. al. Rule Discovery from Time Series. In Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining, pp. 16-22, AAAI Press 1998.

[DOE EIA 2001] Emission of Greenhouse Gases in the United States – 2001, U.S. Department of Energy, Energy Information Administration DOE/EIA-0573(2001), December 20, 2002

[DOE EERE ITP] U.S. Department of Energy, Energy Efficiency and Renewable Energy – Industrial Technologies Program – Home Page

[Doro 00] A Practical Suffix-Tree Implementation for String Searches. Dorohonceanu, B and Nevill- Manning, C. 133-140, Dr. Dobb's Journal, July 2000.

[Eise 98] Cluster Analysis and Display of Genome- Wide Expression Patterns. Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. PNAS, USA, Vol. 95, pp. 14863-14868, December 1998.

[Gusf 97] Algorithms on Strings, Trees, and Sequence - Computer Science and Computational Biology. Gusfield, D. Cambridge University Press, 1997.

[Holt 00] Holter, N.S., Fedoroff, N.V. et. al. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. PNAS, USA July 18, 2000 Vol.97 no.15 8409-8414.