

# ARTICULATING DOMAIN PRIOR KNOWLEDGE USING THE ANALYTIC HIERARCHY PROCESS FOR MORE RELEVANT DATA MINING PATTERNS

**K. Niki Kunene**

College of Business, University of Louisville, Louisville, KY, USA  
kniki.resume@gmail.com

## ABSTRACT

The use of data mining by organizations has grown sizeably because thanks to Moore's Law, over the last decade especially, organizations have been able to gather vast amounts of data during their operations. However, the results accruing from this usage of data mining on operational data has been mixed. Data mining is intended to be a non-trivial process of identifying interesting patterns, where interesting infers valid, understandable, novel, and potentially useful patterns; and yet too frequently, the results have yielded uninteresting (irrelevant, and/or obvious) patterns. To increase the confidence of decision makers in the interestingness of discovered patterns, some researchers believe in the incorporation of domain prior-knowledge into the data mining process. In this paper, we present a new design artifact, that uses the analytic hierarchy process (AHP) to conceptualize and structure domain prior-knowledge, thus capturing a broader essence of domain knowledge on which data mining can be applied. Our method is built and evaluated using best practice design science principles and guidelines. The evaluation of the artifact occurs within the domain of brain trauma intensive care. This particular paper focuses on the design and design components of our artifact.

**Keywords:** Knowledge discovery in databases, data mining, analytic hierarchy process, decision support systems, traumatic brain injury.

## 1. INTRODUCTION

The use of data mining techniques to exploit the large amounts of data—collected by organizations for operational purposes—to support decision-making has generally been well received. Knowledge discovery and data mining (KDD), also frequently (though inaccurately) called data mining, is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data. Data mining *per se* is the application of specific search algorithms to extract patterns, or models, from data. In other words, data mining is just one step in the KDD process. The difficulty in reality is, with the exception of the mining of website semantics, pattern discovery methods frequently generate irrelevant, obvious and uninteresting patterns (Padmanabhan and Tuzhilin, 1999). This is because data mining algorithms techniques rely solely on (operational) data from a database to infer knowledge about a domain. All other domain knowledge that is not contained in the operational database is excluded. Domain knowledge refers to experiential or prior knowledge. Domain knowledge may be personalized knowledge, or already articulated and shared knowledge in a domain of practice (Tuomi, 1999; Alavi and Leidner, 2001). In fact, using the view of Tuomi, data/information<sup>i</sup> from a database constitutes domain knowledge that has been formally articulated and shared by members of the said domain of practice; however, it is a limited or partial articulation that is typically designed to serve operational activities. Clearly, for supporting decision making, there is room for formally articulating more decision-relevant knowledge into repository/database information. Such an

articulation would then allow us to perform data mining activities on a decision-relevant, knowledge enhanced repository. Such knowledge may include semantic meta-data, decision-makers' prior expectations and intuitions, as well as any formalized knowledge about the domain of practice that is employed by the data mining analyst or the domain decision-maker. Domain knowledge, therefore, also includes the knowledge used to analyze, select the required data, interpret it, and evaluate the results of the knowledge discovery process (Fayyad et al., 1996; Pohle, 2003). The base assertion followed in this research is, approaches that seek to increase the relevance and interestingness of discovered patterns from data mining activities must seek to incorporate important and decision-relevant domain knowledge into the KDD process.

In this research we proposed a new approach, a method that uses the analytic hierarchy process (AHP) to conceptualize the domain and, where required, additional decision-rules to capture relevant and important prior knowledge that is not already captured by the AHP hierarchy. The contribution of this research is in identifying and using multicriteria decision analysis (MCDA) and the AHP specifically as a candidate tool for systematically supporting the (domain) knowledge-intensive phase, i.e. problem analysis, in the area of knowledge discovery and data mining (KDD). This is a gap that was systematically identified in Kopanas, Avouri & Daskalaki (2002), and it has yet to be plugged. Simply put, this is the first combination of both data mining and an MCDA technique. This is not trivial because, both techniques have been well-tested across many domains. In this research, their strengths are combined where the AHP serves the knowledge discovery and data mining (KDD) process, and both act in concert to support decision-making.

Our method is built and evaluated using best practice design science principles in information systems (March and Smith, 1995; Hevner et al., 2004). The purpose of design science is to build and evaluate design artifacts (constructs, models, methods, and instantiations). In this research, our method is the design artifact, and its utility is evaluated by instantiating it in the domain of patient monitoring in brain trauma intensive care unit (ICU). See Hevner, et al. (2004), Walls et al. (1992), March and Smith, (1995), and Simon (1999) for more on design science in information systems.

Our research question can be framed formally as follows: Does the use of our design artifact (the method) within the KDD process result in objectively and subjectively more interesting knowledge discovery patterns than the otherwise traditional approach?

This question is of interest to two groups of people. Firstly, this work contributes to the area of knowledge discovery and data mining (KDD) in that, to the best of our knowledge, no research has been done that integrates the AHP and data mining for the purposes of capturing extra-database knowledge (i.e. prior domain knowledge) in the problem analysis phase to improve pattern interestingness, as well as processing efficiency which is associated with the resulting data reduction. In addition, our research suggests the need to think of different ways by which to conceptualize extra-database domain knowledge for machine processing; using the analytic hierarchy process (AHP) is just one way of doing so; it is not necessarily complete or without its own inherent shortcomings. There may be other ways, such as, the possible mining of ontologies. Inquiring further into which domain conceptualization works better is subject to future research.

Second, typical of design science research, this research is also of interest to the experts in the domain of practice where it is instantiated (i.e. ICU personnel and traumatic brain injury (TBI) clinicians), other domain decision-makers or personnel who use this data to support patient monitoring and management; it is also interesting for research purposes in the TBI domain. In the latter case, the discovered/predicted patterns can be used to further investigate or give additional supporting evidence to existing theories, standards, guidelines and physician opinion. In cases where the discovered patterns present totally puzzling conclusions/propositions then such patterns can be used to serve (tentatively) postulating theoretical propositions about the domain.

Finally, notwithstanding the proposed merits of our design artifact, it is important for design evaluation purposes that we demonstrate that our method affects the interestingness of discovered patterns when compared to traditional data mining approaches. The rest of this paper is organized as follows: In the next section we will present a brief literature review outlining the theoretical basis for our work. In section three, we will present our research model, hereafter referred to as our Method. In section four, we present a brief illustration of how our Method is instantiated within a domain of practice.

## 2. BRIEF LITERATURE REVIEW

The KDD process is a multi-phase design, development and implementation systems effort. Much of the published research in knowledge discovery and data mining (KDD) has however focused on just the data mining phase. Secondly, attempts to incorporate domain knowledge into the data mining have largely focused on incorporating the metadata embodied by objectively measurable relationships such as statistical relationships between the data elements. On the other hand, we have also seen some conceptual expositions that seek to articulate appropriate ways by which to ascertain the interestingness of data mining patterns to the domain user/decision-maker; these measures are inherently user-oriented and subjective, and they are important because they speak to pattern relevance.

Our research must be properly located within the entire KDD process. Studying the KDD process as a whole requires an examination of the activities that span a multidisciplinary field. Fayyad et al. (1996) proposed a unifying process-centric framework for KDD. They described a nine-step process that includes: (1) learning the application domain; (2) creating the target dataset; (3) data cleaning and pre-processing; (4) data reduction and data selection; (5) choosing the data mining technique (i.e. classification, prediction, clustering, association); (6) choosing the data mining algorithm; (7) data mining; (8) interpreting the results properly; and (9) using the discovered knowledge. Han & Kamber (2001) pointed out that although there is not as yet a formally, broadly-accepted methodology for KDD, any such methodology should contain the following steps: (1) problem analysis, (2) data preparation, (3) data exploration, (4) pattern generation (data mining), (5) pattern monitoring, and (6) pattern deployment. On the other hand, practitioners seemingly have adopted the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology as a de facto standard; and several papers (de Abajo et al., 2004) published in ACM's SIGMOD and SIGKDD conferences describe KDD implementations in different industries using CRISP-DM. The design logic of CRISP-DM methodology is not dissimilar to the framework of Fayyad et al., nor the one proposed by Han & Kamber in that CRISP-DM encompasses the following phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, and (6) deployment. What these traditional approaches to the KDD have in common is, the almost exclusive reliance on the data found within the databases for knowledge discovery purposes. The use of our method suggests that an additional step is required after the *business understanding* phase: that is, structuring and formalizing relevant domain prior knowledge. Furthermore, the data preparation phase must include this formalized prior knowledge in machine-readable format as part of the repository against which modeling (or pattern generation) is applied. See Table 1, where all three methodologies are summarized in the first three columns. The last column of Table 1 shows the interventions of our Method.

**Accounting for domain knowledge:** The predominant type of research in the literature that incorporates domain knowledge into data mining does so by including the objectively measurable relationships between the data elements, for example, the (statistical) deviations of attributes, descriptive statistics, or interfield correlations. For instance, Yoon et al. (1999) look at interfield relationships and correlations between variables, which are then used to restrict or eliminate irrelevant data mining queries. Kopanas et al. (2002) used the domain knowledge about the structure of available information and its semantic value as described by the data processing department. They use this information to eliminate irrelevant attributes, determine missing values, aggregate data, and reduce data by sampling and transaction elimination. In such cases, domain knowledge contributes to the reduction of the search space. Some web mining research also incorporates domain knowledge in data mining using readily available site semantics for more "intelligent mining" (Kosala and Blockeel, 2000; Zaki et al., 2001; Pal et al., 2002; Pohle, 2003). So, contrary to our own intuition about our method, we should expect that incorporating prior knowledge should contribute to data reduction as well. (Kopanas et al., 2002)

Table 1 : Contrasting Three Approaches to the KDD Methodology

Fayyad et al.'s (1996) KDD Process-centric framework	Han & Kamber's (2001) Proposed Methodology	The CRISP-DM Methodology	Our Method
[1] Learning the application domain	(1) Problem analysis	(1) Business understanding	(1) Learn domain of practice ,and decision problem  (2) Structure and formalize relevant domain prior knowledge
[2-4]Creating the target dataset; Cleaning and pre-processing; data reduction and data selection	(2) Data preparation (3) Data exploration	(2) Data understanding (3) Data preparation	(2) Data understanding (3) Data preparation: incorporate additional formalized prior knowledge in machine-readable format
[5-7] Choosing the data mining technique; Choosing the data mining algorithm; Data mining	(4) Pattern generation (data mining)	(4) Modeling	(4) Modeling
[8] Interpreting the results	(5) Pattern monitoring	(5) Evaluation	(5) Evaluation
[9] Using the discovered knowledge	(6) Pattern deployment.	(6) Deployment.	(6) Deployment.

**Incorporating prior knowledge:** In recent years we have seen research recognizing the importance of prior knowledge to the interestingness of discovered patterns. A sizeable chunk of these papers focuses on making the distinction between *objective* and *subjective* measures of pattern interestingness. Objective measures are based on the inherent (or mathematical) structure of the discovered patterns - such as the *support* and *confidence* statistics used in association mining. On the other hand, subjective measures represent user beliefs or biases regarding the relationships in the data (Silberschatz and Tuzhilin, 1996; Freitas, 1999; Hilderman and Hamilton, 1999, 2001). Stated differently, objective measures of interestingness do not account for the user's, or the human analyst's background knowledge about the application domain (Freitas, 1999). Subjective measures, on the other hand, seek to account for the decision-maker's individual *beliefs* about the domain. It has also been suggested that subjective measures should rely on some formalization of expectations or prior knowledge (Freitas, 1999; Mitra et al., 2002).

**Subjective Measures of Interestingness:** Several researchers in KDD have expounded on subjective measures of interestingness, proposing subjective, user-oriented measures of pattern interestingness. Freitas (1999) proposes *surprisingness*; Silberschatz & Tuzhilin (1996) propose *unexpectedness as well as actionability*. These measures are further explored in Padmanabhan & Tuzhilin (1998). The latter assume that initial beliefs about a domain have been elicited from a domain expert, or learnt from the data; they do not however explore how this is done and how this can be done relatively systematically and in a repeatable fashion — which is important to this research; see also Piatetsky-Shapiro & Matheus (1994). Silberschatz & Tuzhilin (1996) discussed how the beliefs (using constraints or rules) about the domain are elicited from the domain experts, they defined beliefs as logical statements (predicate formulae expressed in first-order logic); they then assigned a confidence factor to each belief. They suggest that the degree of confidence that the belief holds can be handled using any of several approaches, e.g. the Bayesian approach, the Dempster-Shafer approach, as well as the Cyc approach where following such an approach, a

discovered pattern is considered interesting with respect to some belief system if “it affects” this system; the more it affects this system, the more interesting it is. However, the approach proposed by Sielberschatz & Tuzhilin (1996) is not instantiated. Nevertheless, the focus on subjective interestingness measures, notwithstanding how they are measured, yields the important idea that data mining output can be reduced on the basis of whether the generated patterns surpass a subjective, user-oriented interestingness threshold. (Piatetsky-Shapiro and Matheus, 1994)

Our method is different from the above implementations of interestingness measures that incorporate objectively measurable metadata in that it seeks to formulate and add prior domain knowledge into the data mining repository thus extending the data mining repository, rather than merely using metadata to clean or reduce the data; this allows pattern discovery algorithms to derive patterns of interestingness on the expanded dataset before the fact, as it were. At the same time, our method does not however preclude the implementation of beliefs as production rules with associated rule certainty, not unlike the approaches suggested in Silberschatz & Tuzhilin (1996).<sup>ii</sup>. The strength of our method is that it transparently conceptualizes the domain in a relatively structured fashion that is both understandable (to user and analyst) and repeatable, rendering it easier to evaluate or replicate across different domains. (Yoon et al., 1999)

### Theoretical Basis for this Research

We propose using multicriteria decision analysis (MCDA) to conceptualize the domain; MCDA can inherently express how domain elements relate to each other as (subjectively) understood by our domain experts. Multicriteria decision analysis serves as a systematic framework for breaking a problem into its constituent parts in order to understand the problem, and to subsequently arrive at a decision given decision-maker preferences. Importantly, multicriteria decision analysis provides a means to investigate a number of choices (or alternatives) in the presence of conflicting priorities. By structuring a problem within the MCDA framework, alternatives can be ranked according to preestablished preferences to achieve user-defined objectives. Specifically, we employ the analytic hierarchy process (AHP), a multicriteria decision analysis approach, technique and tool.

The analytic hierarchy process (AHP) uses hierarchical structures of the form, *goal-criteria-subcriteria-alternatives*, to represent a decision problem, it then assigns priorities for the decision alternatives based on user-judgments. The latter are captured progressively as pairwise comparisons of the criteria (with respect to the goal), the subcriteria (with respect to each criterion), and of the alternatives (with respect to the subcriteria). In this example the hierarchy is assumed to have three levels. However, a problem may be represented through more levels in the structure if necessary. The elements of one level of a hierarchy, say level 3, can be compared in a pairwise fashion with respect to one element, *e*, of the next higher level (level 2). The numbers reflecting the comparisons are entered into a symmetrical reciprocal matrix; the eigenvector corresponding to its largest eigenvalue gives the local priority ordering of the compared elements with respect to the root of the cluster of the elements (Saaty, 1990a). When all of the local priorities in the hierarchy have been computed, the relative weights of the criteria and subcriteria are synthesized, yielding the composite priorities of all criteria and ultimately the relative, global weights of the alternatives. For the purposes of our method, we could feasibly have more than a single hierarchy —and therefore multiple global priorities per alternative, —provided the alternatives remained the same across all hierarchies. For a detailed treatment of the AHP, see (Saaty, 1990b, 1992; Saaty, 1994).

The AHP is a well-tested multicriteria decision making technique. There are hundreds of documented applications of the AHP. To the best of our knowledge, the AHP has not been used to articulate domain prior knowledge in KDD processes.

### 3. THE DESIGN OF OUR PROPOSED METHOD

Figure 1 represents a high-level view of our method. In Figure 2, we show one of the initial conceptualizations of the domain hierarchy. We show only one representation to contain the presentation within the limitations of this paper (the hierarchy in reality had six levels, and even so does not purport to be complete, partly because expert understanding of their domain is itself incomplete, and also because we have had to scope this to fit into this written research paper<sup>iii</sup>).

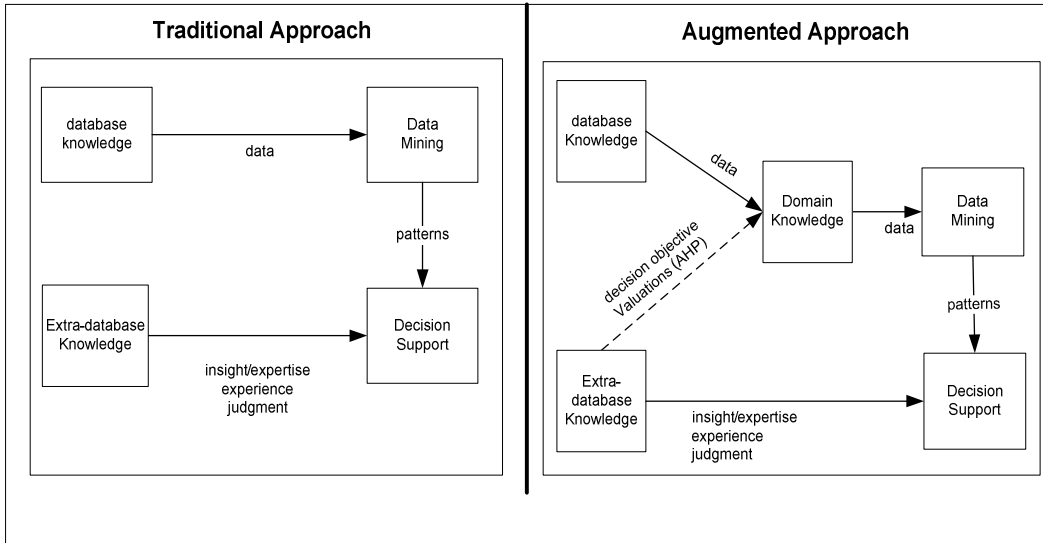


Figure 1. The Traditional Approach versus the Proposed Method

The left-hand-side (LHS) of Figure 1 represents the traditional approach to data mining, and the right-hand-side (RHS) represents our proposed method. In the traditional approach, the LHS, (domain) knowledge which is passed to the data mining process consists of knowledge derived solely from the database; consequently the pattern discovery process is bounded by the design and functional limitations of the database.

On the other hand, with our method, the RHS, we propose creating an augmented repository of domain knowledge which incorporates both database knowledge and prior domain knowledge (labeled, extra-database knowledge). Data mining takes its input from this knowledge-enhanced repository. The extra-database knowledge is informed through the structuring of the AHP problem hierarchy which incorporates the expertise, experience and judgment of the decision-maker (Saaty, 1994), and any additional decision rules or beliefs, their attributes and rule-certainty that are significant to the domain problem but are not captured by the hierarchy can be incorporated into the domain knowledge repository.

In general, when conceptualizing the domain beyond just the database, we are interested in the rules or heuristics used by the decision-makers, their decision goals, and their preference assessments. The structuring of the problem hierarchy must be done in consultation with domain experts<sup>IV</sup>, and ratified by them before value judgments can be accounted for. Once the structure of the hierarchy is developed, the domain experts generate the pairwise comparisons which can be processed using the Expert Choice<sup>TM</sup> [v].

### The Data Mining Phase

A data mining algorithm infers a model from dataset (in a repository). For the purposes of evaluating our artifact we instantiated three different repositories on which data mining would be performed. The first repository assumed the traditional approach; the other two instantiations used our approach as in the RHS of Figure 1 where one included the hierarchy structure with AHP global priorities arising from pairwise comparisons; for the last instantiation we used the hierarchy to construct the rules and claims about the domain; we did not however generate AHP global priorities.

Data mining algorithms/techniques employ either supervised or unsupervised learning. When supervised learning is employed – which refers to *classification* and *prediction* techniques – the user is required to define two or more classes in the database. In our case, we can use patient outcome, measured using the Glasgow Outcome Scale at six months (GOS6M). Thus, the data mining input consists of all of the data tuples/rows (with their specified attributes) where the data rows are classified on the basis of a selected *class label attribute*, i.e. GOS6M. The output is, therefore, a set of class labels where, each class label corresponds to a unique pattern, the

associated *class description*; and where a class is defined by a combination of values for the predicted attributes; or more generally, a class is defined by a condition on the attributes. Prediction techniques are similar to classification in that they also use a “seen” classification model called a *training set*, and unseen data (a *test set*) is then used to predict the *class label* of each row. The resulting prediction is then compared to the known class label to measure model accuracy. The attributes denoting the tuple’s *class label* are called the *predicted attributes*. The remaining attributes are called the *predicting attributes*. Using our Method, the nature and number of predicting attributes is different to what it ordinarily would be using the traditional approach to data mining.

With *unsupervised learning* techniques such as clustering, the class label of each training set is unknown. In addition, the number of classes to be learned may not be known in advance. Clustering allows the data mining tool to learn from observation, or learn by discovery. The system has to find its own classes in a set of states without the help of a “teacher” (Holsheimer and Siebes, 1991). The data mining tool is thus supplied with an object, but no classes are defined. The system has to observe the examples, and recognize patterns (i.e. class descriptions) by itself. The result of an unsupervised learning process is a set of class descriptions, one for each discovered class (pattern), that together cover all objects in the domain environment. The descriptions form a high-level summary of the object in the domain environment.

Below, we briefly describe the domain application of our KDD method.

#### 4. A BRIEF ILLUSTRATION USING OUR APPLICATION DOMAIN CASE STUDY

We instantiated our method as a KDD application in a NeuroScience ICU of a large academic hospital. The overall goal of the Neuroscience ICU unit is to maximize patient outcome as measured by the Glasgow outcome scale (GOS) at six months. Patient outcome is affected by several factors including pre-hospital management, direct trauma centre transport, triage (measured using the GCS), and the participation in trauma education programs. *Cerebral ischemia* (occurring within the ICU) is considered the single most important secondary event affecting patient outcome by the clinicians. Note, while the primary insult to the brain occurs at the site of the accident and is beyond the control of the ICU, secondary insults occur during intensive care, and the ICU’s treatment objective is to avoid/prevent these by all means.

The ICU maintains a database which stores recorded monitoring information (oxygenation, microdialysis, and hemodynamic data), as well as epidemiological, drug treatment (e.g. mannitol, barbitol etc.), patient outcome data, and other systemic variables. Physicians and staff rely on their own expertise, continuous assessments and scientific evidence to continuously improve treatment and drug intervention for better patient outcome. Our data analysis is retrospective and aims to produce results that could lead to improvements in patient care.

From Figure 2, the *goal* of the system or domain is the overall management of patient neurological outcome which is measured using the Glasgow outcome scale at six months (GOS6M). The factors affecting patient outcome are: *initial management*, the *Glasgow coma scale (GCS) triage*, the patient’s *age*, results of *computed tomography (CT) scan*, and *cerebral ischemia*<sup>vi</sup>. The prevention of cerebral ischemia in the ICU is considered the single most important factor once a patient has been admitted into the brain trauma ICU in patient monitoring and maintenance. Two key monitoring strategies are identified as subfactors: microdialysis monitoring and cerebral blood flow (CBF), where *intracranial pressure (ICP) monitoring* and *cerebral perfusion pressure (CPP) monitoring* are sub-subfactors of the latter. On the other hand, the level 2 factor, *initial management*, while identified as important to patient outcome includes events that occur outside of the brain trauma ICU. This affects both data gathering and preference assessments. The factors, *Age* and *GCS*, are fairly well-researched and there is some understanding of the degree of clinical certainty about the claims made about age and GCS (separately and jointly) with respect to patient outcome. For such claims, we can construct the necessary decision-rules with their concomitant level of rule-certainty, and/or structure a second hierarchy that represents the rule set, its elements and rule-certainty. This works particularly well in this domain, because knowledge is categorized by the community according to varying degrees of clinical certainty: standard, guideline, or option (an opinion) in descending order. This inherently represents rule-certainty.

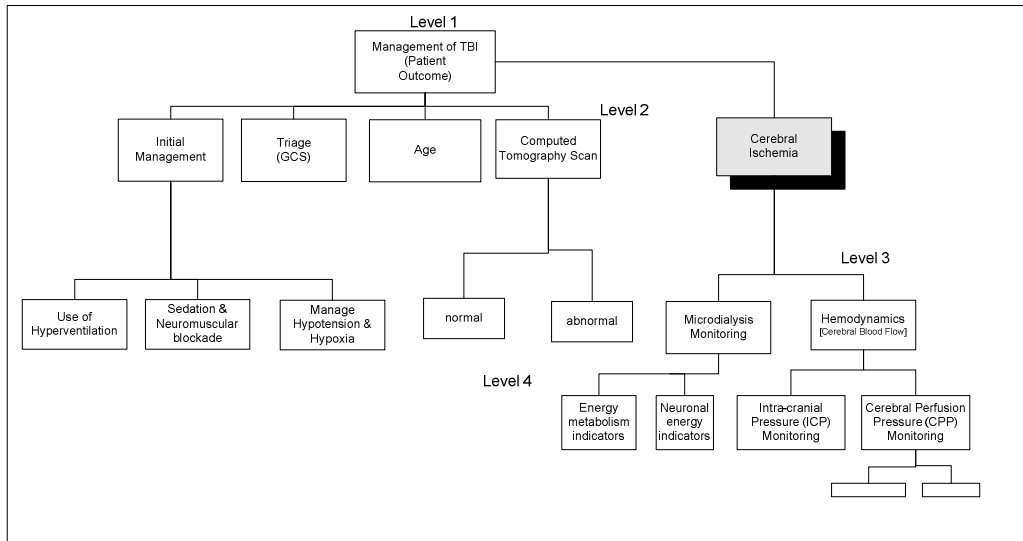


Figure 2. A Partial Hierarchy of our Case Study Domain

The goal of the system is to predict/describe patient outcome using the AHP to formulate the relationship between the decision elements. In our formulation of the problem, the alternatives correspond to individual patients. Through a process of pairwise comparison of the factors (criteria) in a cluster of the hierarchy and subsequent eigenvector calculations, the relative weights of the factors (with respect to the root of the cluster) are obtained. This is performed for all clusters and levels in the hierarchy. Through the process of hierarchical synthesis (Saaty, 1990a) global priorities are generated for every element of the hierarchy. As a result of this process, each patient is assigned a global weight. We assigned the weight as an additional attribute of the patient entity which we incorporate into the augmented database of our domain knowledge. In other words, each patient will be assigned a numeric score, corresponding to the weight calculated through the multi-criteria problem formulation.

Technically, pattern generation/ data mining can be deployed against this added data alone. In fact, it would be interesting to know whether the AHP designated global weight of a patient is a better predictor of patient outcome than the ischemic score alone, the latter which is currently the best single predictor of patient outcome. Data mining, classification in this case, is subsequently applied to a constructed domain repository that contains the assigned global weights to individual patients. Where production rules about beliefs exist, further processing is performed that results in the addition of further columns of data to the repository, but that discussion is beyond the scope of this paper.

### Mining Patient Data

The application of data mining can thus be used principally to predict/classify patient outcome and/or cerebral ischemia, or for classification of cases with respect to specific patient outcome events according to GOS6M, (values: *dead*, *vegetative*, *severe disability*, *moderate disability*, or *good recovery*). In medical terms, the prediction of the probable course of outcome of a disease is what we understand to be the prognosis. Thus, patterns identified may be considered for prognostic indicators. Using the traditional approach, the KDD process uses as its input only the data from the database. With our method, the extra-database domain knowledge we have articulated and incorporated into the repository is also used to influence the data mining phase. The details and results of this process can be found in Kunene and Weistroffer (forthcoming)

## 5. CONCLUSION

This paper described a new method for incorporating domain prior-knowledge into data mining activities using the analytic hierarchy process (AHP), a multicriteria decision analysis approach, technique, as well as tool. We also briefly described the environment in which the method is being instantiated. Our method is designed and built using the best practice principles for conducting design science research in information systems where systematic evaluation is a part (see Hevner et al., 2004). The contribution of our method to knowledge discovery and data mining is it addresses the research lacuna where data mining applications do not incorporate domain prior knowledge, and as a result generate uninteresting and irrelevant patterns. Our method uses multicriteria decision making as a mechanism with which to explore and express domain prior knowledge. MCDA inherently relies on decision-makers expressing both objective and subjective decision factors. Importantly, our method is instantiated that is, evaluated empirically. Initial findings showed that both implementations using our method generated more interesting patterns than the traditional approach.

The contribution of method to the domain relates to the following: it highlighted some structural data collection issues which need to be changed (actionability); it also generated new patterns that pose questions about current understandings; this is in contrast to the traditional method which did not produce any surprising results.

Our hierarchy conceptualization does not claim to be complete, notwithstanding that we can use production rules to articulate important domain-specific prior knowledge that is not captured in the hierarchy. We do not in fact at this stage strive for completeness; because our domain is complex and its understanding by domain experts (a different problem on its own that is beyond our case-study site) is necessarily incomplete as the human brain is arguably the most complex system known to man.

The primary limitation of our method is that it relies heavily on the AHP for both problem structuring and the measurement of decision-maker preferences, and thus is delineated by its strengths and weaknesses. There has been some spirited debate between the proponents (Harker and Vargas, 1987, 1990; Saaty, 1990a; Saaty and Vargas, 1994; Vargas, 1994) and opponents (Dyer, 1990b, 1990a) of the AHP. Some of the issues include, whether or not the AHP can handle uncertainty, the alleged ambiguity of the questions asked of the decision maker must answer when conducting pairwise comparisons; whether AHP generates inconsistent judgments; and lastly whether or not rank-reversal and the transitivity of preferences are legitimate concerns. The proponents of the AHP have presented strong arguments to argue against AHP criticisms. While, it is not our intention to resolve these arguments ourselves, our system provides for the ability to capture decision-relevant factors via decision rules not expressed by the AHP structure such as rule-certainty, we also checked judgments for inconsistency, and we instantiated a model that uses the AHP for structuring only where no pairwise assessments are made. Notwithstanding, the AHP has been widely used internationally, and its use is well-documented. For this research, the most compelling feature of the AHP is its ease of use for decision makers unfamiliar with decision aids; it aids them to effectively structure the decision problem and assign their preferences. It is much easier describing and explaining the AHP to decision-makers than describing competing decision aids. Nevertheless, if the AHP were inappropriate for the manner in which we are employing it, this would be evident when we compare pattern interestingness results with the traditional approach, as well as the model that does not use the AHP preference assignments. Nonetheless, the KDD field would benefit from research that used alternative conceptualizations such as the multi-attribute utility theory (MAUT).

Future work can focus on exploring our intuitive belief that in less complex business domains our KDD process method may prove to be a more structured and comprehensive articulation of domain prior knowledge that is readily usable for data mining purposes. Since our method intervenes in the first phase of the KDD process, it can employ existing interestingness measures without the need for creating new measures.

### **Acknowledgement:**

This research was completed while the author studied towards her PhD at the Virginia Commonwealth University in Richmond, Virginia.

## 6. REFERENCES

- Alavi, M. and Leidner, D. E. (2001), Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues, *MIS Quarterly* 25(1), 107-136.
- de Abajo, N., Diez, A. B., Lobato, V. and Cuesta, S. R. (2004), ANN Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study, *The 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, ACM Press.
- Dyer, J. S. (1990a), A Clarification of Remarks on the Analytic Hierarchy Process, *Management Science* 36(3), 274-275.
- Dyer, J. S. (1990b), Remarks on the Analytic Hierarchy Process, *Management Science* 36(3), 249-259.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM* 39(11), 27-34.
- Freitas, A. A. (1999), On Rule Interestingness Measures, *Knowledge-Based Systems* 12(5-6), 309-315.
- Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufman, San Francisco.
- Harker, P. T. and Vargas, L. J. (1987), The Theory of Ratio Scale Estimation: Saaty's Analytic Hierarchy Process, *Management Science* 33(11), 1383-1403.
- Harker, P. T. and Vargas, L. J. (1990), Reply To: Remarks on the Analytic Hierarchy Process, *Management Science* 36(3), 269-323.
- Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004), Design Science in Information Systems Research, *MIS Quarterly* 28(1), 75-105.
- Hilderman, R. J. and Hamilton, H. J. (1999), Knowledge Discovery and Interestingness Measures: A Survey, Accessed: 2004
- Hilderman, R. J. and Hamilton, H. J. (2001), *Evaluation of Interestingness Measures for Ranking Discovered Knowledge*, The 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Hong Kong, Springer-Verlag.
- Holsheimer, M. and Siebes, A. (1991), Data Mining: The Search for Knowledge in Databases, *Technical Report CS-R9406, Amsterdam - Netherlands*, Pages, Accessed: 2004
- Kopanas, I., Avouris, N. M. and Daskalaki, S. (2002). The Role of Domain Knowledge in a Large Scale Data Mining Project, in Vlahavas, I. P. and Spyropoulos, C. D. (eds), *Methods and Applications of Artificial Intelligence: Lecture Notes in AI -LNAI*. Berlin, Springer-Verlag, 288-299.
- Kosala, R. and Blockeel, H. (2000), *Web Mining Research: A Survey*, *SIGKDD Explorations* 2000, ACM Press.
- Kunene, K. N. and Weistroffer, H. R. (forthcoming), An Approach for Predicting and Describing Patient Outcome Using Multicriteria Decision Analysis and Decision Rules, *EJOR* XX(XX).
- March, S. T. and Smith, G. F. (1995), Design and Natural Science Research on Information Technology, *Decision Support Systems* 15, 251-266.
- Mitra, S., Pal, S. K. and Mitra, P. (2002), Data Mining in Soft Computing Framework: A Survey, *IEEE Transactions on Neural Networks* 13(1), 3-14.
- Padmanabhan, B. and Tuzhilin, A. (1999), Unexpectedness as a Measure of Interestingness in Knowledge Discovery, *Decision Support Systems* 27(3), 303-318.
- Pal, S. K., Talwar, V. and Mitra, P. (2002), Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on Neural Networks* 13(5), 1163-1177.
- Piatetsky-Shapiro, G. and Matheus, C. J. (1994), *The Interestingness of Deviations*, The AAAI-94 Workshop on Knowledge Discovery in Databases.
- Pohle, C. (2003), Integrating and Updating Domain Knowledge with Data Mining, [citeseer.ist.psu.edu/668556.html](http://citeseer.ist.psu.edu/668556.html), Accessed: December 2004.

- Saaty, T. L. (1990a), An Exposition of the Analytic Hierarchy Process in Reply to the Paper 'Remarks on the Analytic Hierarchy Process', *Management Science* 36(3), 259-268.
- Saaty, T. L. (1990b), *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, RWS Publications, Pittsburgh.
- Saaty, T. L. (1992), *Multicriteria Decision Making - the Analytic Hierarchy Process*, RWS Publications, Pittsburgh.
- Saaty, T. L. (1994), *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, RWS Publications, Pittsburgh.
- Saaty, T. L. and Vargas, L. G. (1994), The Legitimacy of Rank Reversal, *OMEGA* 12(5), 513-516.
- Silberschatz, A. and Tuzhilin, A. (1996), What Makes Patterns Interesting in Knowledge Discovery Systems, *IEEE Transactions on Knowledge and Data Engineering* 8(6), 970-974.
- Simon, H. A. (1999), *The Sciences of the Artificial*, Cambridge, MA, MIT Press.
- Tuomi, I. (1999), Data is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory, *The 32nd Annual Hawaii International Conference on System Sciences (HICSS-32)*, Maui, Hawaii, USA.
- Vargas, L. G. (1994), Comparison of Three Multi-Criteria Decision Making Theories: The Analytic Hierarchy Process, Multi-Attribute Utility Theory and Outranking Methods, *The International Symposium on the Analytic Hierarchy Process*, Washington, DC, USA.
- Yoon, S.-C., Henschen, L. J., Park, E. K. and Makki, S. (1999), Using Domain Knowledge in Knowledge Discovery, *The 8th International Conference on Information and Knowledge Management*, ACM Press.
- Walls, J. G., Widmeyer, G. R. and El Sawy, O. A. (1992), Building an Information Systems Research Design Theory for Vigilant EIS, *Information Systems Research* 3(1), 36-59.
- Zaki, M. J., Punin, J. and Krishnamoorthy, M. (2001), LOGML for Web Usage Mining, *INFORMS Annual Meeting*, Miami Beach, FL.

---

## ENDNOTES

<sup>i</sup> For clarity, we must point out throughout this paper we use the terms data and information interchangeably, which is not unusual in the field of information systems, even though strictly speaking we are referring to information.

<sup>ii</sup> This treatment is not addressed in this paper for space reasons

<sup>iii</sup> In subsequent refinements and discussions with domain clinical experts the hierarchy was decomposed into two hierarchies.

<sup>iv</sup> We consulted with two clinical experts continually, even though we periodically learnt different aspects about the domain from five clinicians.

<sup>v</sup> AHP and Expert Choice™ allow for the synthesis of group judgments if the problem is handled by a team instead of a single expert.

<sup>vi</sup> The term "cerebral ischemia" is somewhat of a proxy representation of multiple conditions and metrics that signify ischemia.

**Received:** June 12<sup>th</sup> 2005

**Accepted in final format:** October 30<sup>th</sup> 2006 after one revision.

---

**About the author:**

*Niki Kunene* is an Assistant Professor at the University of Louisville in the Department of Computer Information Systems. She received her PhD in information systems from the Virginia Commonwealth University in 2006. Her current research interests are knowledge management and knowledge discovery, and collaborative technologies and the successful deployment of these technologies.