Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2005 Proceedings

European Conference on Information Systems (ECIS)

2005

Expanding the Knowledge Base for More Effective Data Mining

Niki Kunene
Virginia Commonwealth University, kukunene@vcu.edu

H. Roland Weistroffer
Virginia Commonwealth University, hrweistr@vcu.edu

Follow this and additional works at: http://aisel.aisnet.org/ecis2005

Recommended Citation
Kunene, Niki and Weistroffer, H. Roland, "Expanding the Knowledge Base for More Effective Data Mining" (2005). *ECIS 2005 Proceedings*. 58.
http://aisel.aisnet.org/ecis2005/58

This material is brought to you by the European Conference on Information Systems (ECIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2005 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

EXPANDING THE KNOWLEDGE BASE FOR MORE EFFECTIVE DATA MINING

Kunene, Niki, Virginia Commonwealth University, 1015 Floyd Avenue, School of Business, P.O. Box 844000-4000, Richmond VA, USA, knkunene@vcu.edu

Weistroffer, H Roland, Virginia Commonwealth University, 1015 Floyd Avenue, School of Business, P.O. Box 844000-4000, Richmond VA, USA, hrweistr@vcu.edu

Abstract

Traditionally, data mining, as part of the knowledge discovery process, relies solely on the information contained in the database to generate patterns. Recently, there has been some recognition in the field that expanding the knowledge passed to the pattern generation phase by including other domain knowledge, may have beneficial effects of the interestingness and actionability of the resulting patterns. In this paper, we present a new knowledge discovery method that uses additional decision rules and the analytic hierarchy process (AHP) to conceptualize and structure the domain, thus capturing a broader notion of domain knowledge upon which data mining can be applied. Based on design science guidelines, we design, develop and implement our method within the domain of a brain trauma intensive care unit.

Keywords: Knowledge discovery, data mining, decision support, brain trauma.

1 A NEW METHOD FOR KNOWLEDGE DISCOVERY

1.1 Introduction

Knowledge discovery in databases (*KDD*) is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data; and *data mining*, the application of specific algorithms for the extraction of patterns or models from data, is only one step in this process. Notwithstanding, much of the published research in KDD has focused on the data mining step. Studying the KDD process in its entirety requires an examination of activities that span a multidisciplinary field. Fayyad et al. (1996), in what they loosely call a unifying, process-centric, framework for KDD, articulate a nine-step process that includes, learning the application domain, creating the target dataset, data cleaning and pre-processing, data reduction and data selection, choosing the data mining technique (i.e. classification, prediction, clustering, association), choosing the data mining algorithm, data mining, proper interpretation of the results, and use of the discovered knowledge. Han and Kamber (2001) point out that there is not, as yet, a broadly accepted methodology for KDD and data mining, but that 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. From a practitioner's viewpoint, the *CRoss Industry Standard Process for Data Mining* (*CRISP-DM*) methodology has become a de facto standard, and several papers (e.g. de Abajo et al. 2004) published in ACM's SIGMOD and SIGKDD conferences describe KDD technology implemented in different industries using CRISP-DM. The design logic of CRISP-DM methodology is not dissimilar to the framework of Fayyad et al. or that of Han & Kamber, in that it encompasses the following phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment. What these traditional approaches to the KDD process have in common, is the sole reliance on the data found within the databases upon which the pattern discovery is performed. In our research, based on design science, we develop a KDD method that articulates a way to incorporate the knowledge about the domain that lies outside of the database, knowledge that decision makers use in their day-to-day application of the data, user experience and user judgement. We demonstrate our method within the domain of a brain trauma intensive care unit of a major academic hospital.

1.2 The Proposed Method

Figure 1 below gives a high-level depiction of our Method. The left-hand side represents the traditional practice and the right-hand side our proposed method. In the traditional approach the domain knowledge passed to the knowledge discovery process consists of knowledge derived solely from the database; thus the pattern discovery process is bounded by the functional and design limitations of the database. Even in those studies (Yoon et al. 1999, Hotz et al. 2001, de Abajo et al. 2004) that have attempted some incorporation of domain knowledge into the KDD process, this is done only by explicating knowledge that can be inferred about the data in the database, without recourse to what the users of the data know about the data or their domain, e.g. inter-field relationships and attribute correlations. In our method we propose passing, what we call extra-database knowledge, to an augmented database or repository, which we have labelled *domain knowledge* in Figure 1, for further pattern generation processing. This extra-database knowledge is derived from the expertise, experience and judgment of the decision makers, using the *analytic hierarchy process* (*AHP*) (Saaty 1994) and additional decision rules.

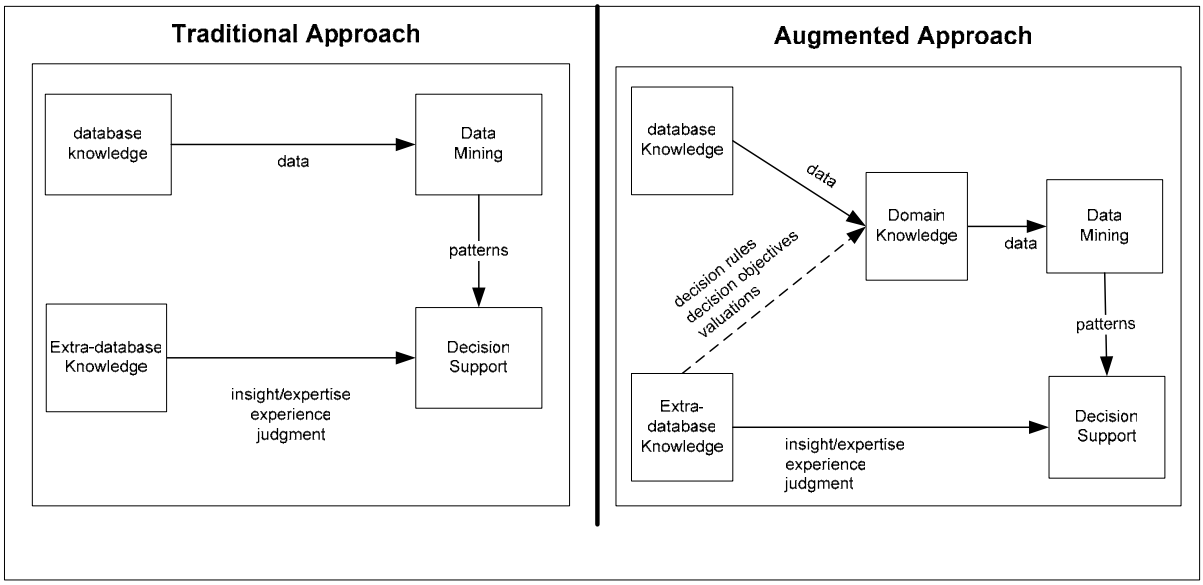

Figure 1. The Traditional Approach versus the Proposed Method.

1.3 Using AHP to Conceptualize the Domain Problem

Our Method relies on the conceptualization of the domain in a structured fashion. For this purpose we employ the AHP to formulate the problem structure, of which the database data form only a part. The resultant domain knowledge, as shown on the right hand side of Figure 1, is an AHP hierarchy. The AHP is a multi-criteria decision method that uses hierarchical structures of the *goal – criteria – alternatives* form to represent a decision problem, and then assigns priorities for the decision alternatives based on user judgments.

In conceptualizing the domain beyond only the database data, we are interested in the rules or heuristics used by the decision makers, their decision objectives, and their preference assessments. Once we have constructed the AHP decision hierarchy, our expert decision maker is asked to complete the hierarchy by making pairwise comparisons of the items (criteria) with respect to the level above, the goal as it were. Similarly, subcriteria at the next level are assigned weights with respect to the level above them. The acquired relative weights of criteria (and subcriteria) are synthesized, ultimately yielding the composite priorities of all criteria and eventually the relative (global weights) of the alternatives.

1.4 The Incorporation of Production Rules

There exists additional knowledge about each element of the hierarchy. This knowledge can be incorporated by constructing decision rules deemed important to the outcome. These rule-based inferences are of the form: IF <*condition(s)*> true THEN <*consequence(s)*> true; or alternatively, IF <condition(s)> true THEN do <*action(s)*>. These rules can be equivalently expressed as *object-attribute-value triplets* (*OAV-triplets*). In our study, no inferences will be made about items not captured in the AHP derived decision hierarchy.

2 ILLUSTRATION USING A CASE STUDY

2.1 The Application Domain

In this section, we describe the domain wherein our KDD method is developed and tested according to the design science guidelines set out by (March & Smith 1995; Hevner et al. 2004). Our case study is conducted at the brain trauma intensive care unit (ICU) of a large academic hospital. The overall goal of the ICU 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 center transport, triage (measured using the *Glasgow coma scale* (*GCS*)), and the participation in trauma education programs. *Cerebral ischemia* is considered the single most important secondary event affecting patient outcome. Primary insult to the brain occurs at the site of the accident and is beyond the control of the ICU, however secondary insults occur during patient care, and the ICU's treatment objective is to avoid these by all means.

The ICU maintains a database which stores monitoring information (oxygenation, microdialysis, and hemodynamic monitoring), as well as epidemiological, drug treatment (mannitol, barbital, etc.), outcome data and other systemic variables. Physicians and staff rely on their own expertise, continuous assessments, and scientific evidence to continuously improve surgical as well as drug intervention for better patient outcome. Our data analysis, however, is retrospective, that is nothing we do is currently used to intervene on patient care.

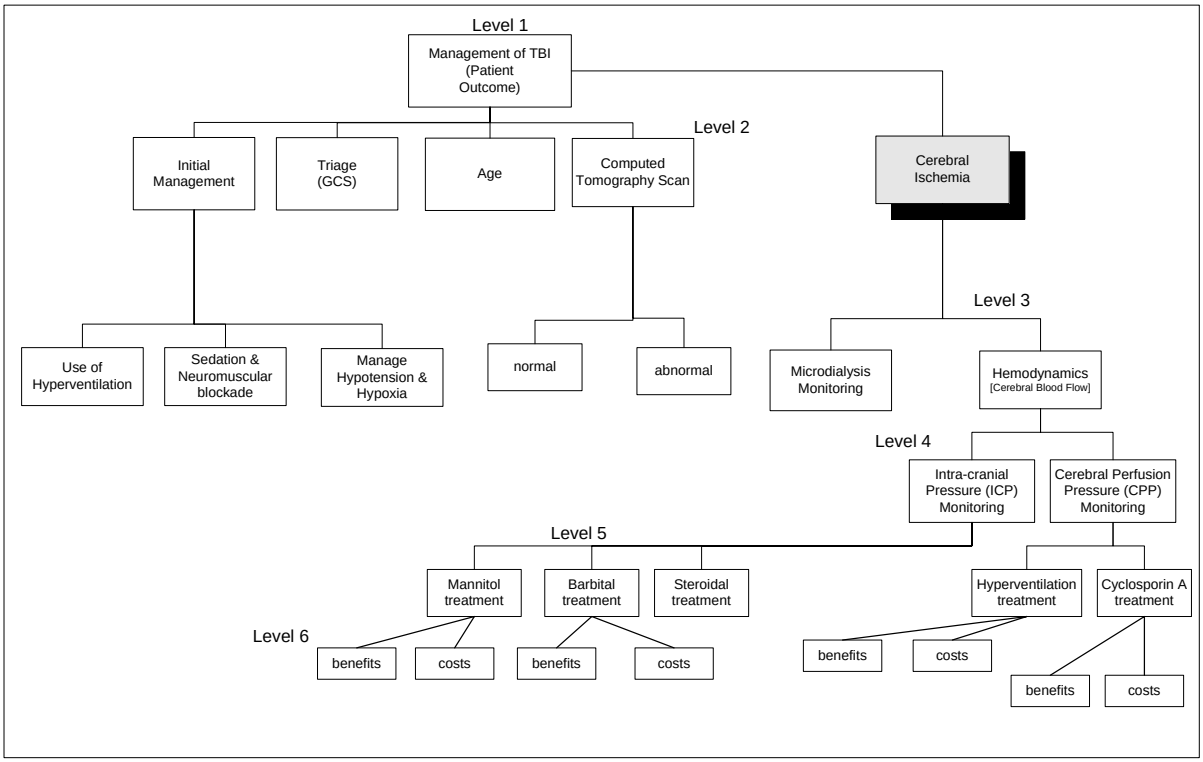

Figure 2. Conceptualization of the Domain – a Decision Hierarchy.

On Figure 2, Level 1 of the hierarchy is the root which represents the overall goal of the domain decision problem. Level 2 of the hierarchy represents the main factors or criteria that affect patient outcome: i*nitial management*, *the GCS triage*, *the patient's age*, and *cerebral ischemia*. Levels 3 and 4 represent subcriteria, and Level 5 reflects (drug) treatment interventions. Each drug treatment has both effects and side-effects which can be evaluated as benefits or costs. At the lowest level of the

hierarchy (not shown) are the alternatives (patients as it were). With the AHP the acquired relative weights of factors (and sub-factors) are synthesized, ultimately yielding the composite priorities of all criteria and eventually the relative (global weights) of the alternatives. Thus each patient is assigned a global weight, as a result of AHP structuring and assessment. This weight we transform and construct as an additional attribute of patient, which we incorporate into the augmented database of our domain knowledge. To arrive at the weights of the factors and the global weights of the alternatives, using the AHP, pairwise comparisons are performed with the pairwise assessments provided by the chief neurosurgeon.

To illustrate our Method, we will restrict ourselves to elements up to the second level of the hierarchy, which coincide broadly with what are known as epidemiological factors. In the case of initial management and cerebral ischemia, we use well-established proxy measures.

2.2 An Illustration Using GCS and Age

The GCS is regarded as an objective measure of consciousness, used as a clinical measure of the severity of injury in patients with severe traumatic brain injury. In general, there is an increasing probability of poor outcome (GOS) with a decreasing GCS in a continuous, stepwise manner. GCS is measured after pulmonary and hemodynamic resuscitation, and after pharmacologic and paralytic agents have been metabolized. Mortality means the GOS = 1. Age is an important factor in the neurological outcome of traumatic brain injury patients; the brain is believed to have a decreased capacity to repair itself with age. In general, there is an increasing probability of poor outcome with increasing age in a stepwise manner. Its predictive value is at about 70%. Younger patients tend to fair better, with age 40 being a significant marker. There is also a significant increase in probability of poor outcome above 60 years of age; in other words this is another critical age threshold for worsening prognosis. While there are certain issues that compound the observation of GCS, no observer reliability problems exist for age. Thus we treat the GCS domain largely as guideline knowledge and age as standard knowledge.

In general, we define a class of rule-objects, where a rule-object has the following attributes: RuleName, RuleCertainty (weight), RuleCondition and RuleConsequence. A rule either applies to or is not applicable to a patient, in the case of the former RuleConsequence, and RuleCertainty values must be recorded as impacting on the relevant decision factor by the specified level of certainty. For example, if a patient is 55 years old, then the expected mortality (i.e. GOS = 1) is 0.46 * 0.50. By adding an attribute called expected_mortality and assigning it the value 0.23 for the patient in the augmented database, we capture incidence of the rule. Pattern discovery is then deployed on the augmented database.

2.3 The Data Mining Phase

After having expanded the knowledge base, as described above, we now use data mining. A data mining tool infers a model from the repository. When supervised learning is employed, which generally includes *classification* and *prediction* techniques, the user is required to define one or more classes in the database. In our case, we could use GOS, for example. Thus, input consists of the rows of data with all of the defined attributes, and the data rows are classified on the basis of the *class label attribute*, GOS. The output is a set of class labels, where each class label corresponds to a unique pattern, the class description. 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, that is, using the classification model, unseen data is used to predict the *class label* of each row. The prediction is then compared to the known class label to measure model accuracy. The attributes that denote the *class label* of a tuple are called the *predicted attributes (target*

variables). 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 a traditional application of data mining.

With the decision problem and the data as described above, the application of data mining can thus be used principally to predict patient outcome and/or cerebral ischemia, or for classification of cases with respect to specific patient outcome events according to GOS, i.e. 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 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.

2.4 Measures of Interestingness

There are two types of measures of interestingness used to evaluate the data mining output. These are objective (Hilderman & Hamilton 2000) and subjective measures of interestingness (Hilderman & Hamilton 2001). Objective measures are standard quantitative assessments used in data mining; their selection naturally depends on the data mining tool used (that is, classification/prediction, clustering or association). For classification or prediction, we are interested in *accuracy*, *stability*, *lift*, and *complexity*. *Accuracy* measures how well objects were correctly classified by the tool for both seen and unseen data. *Stability* speaks the tools consistency in classifying or predicting both seen and unseen data; we measure stability as an interval number from 0 to 1, where 0 means completely stable and 1 means completely unstable. Lift charts plot the relative predictive/classification improvement from the baseline (chance); decision tree *complexity* is measured by a combination of the depth of the tree, and the number of leaves. We limit the depth so as not to exceed six for all models, thus we focus on the number leaf nodes. With respect to subjective measures of interestingness, we use *surprisingness* and *actionability* (Silberschatz & Tuzhilin 1995, 1996). Surprisingness speaks to the unexpectedness of the patterns or class descriptions generated, whereas actionability ascertains whether the resulting class description is something decision makers can do anything about, it is measured on a scale of 1 to 5, where 1 is not actionable and 5 readily actionable.

2.5 Some Preliminary Results

Our example uses only the epidemiological portion of the problem structure, that is, AGE, GCS, Hypotension (a proxy measure for initial management), and the Ischemic Score (a clinically used proxy measure for cerebral ischemia). In a preliminary testing of our model, we developed some data mining decision tree classification models using our approach and compared the results against an otherwise traditional application of data mining. We will only discuss accuracy, stability, and complexity, and provide some general comments on subjective interestingness.

Table 1 shows a marked increase in objective interestingness measures between the data mining model using the traditional approach (Model 1) and the data mining model employing our Method which includes AHP structuring (Model 2). Correct classification rates improve by 33% and 41% for training and validation data, and stability improves by 37%. Model 2 also generates a more complex model. An overly simple model has little descriptive or explanatory utility. In our case, Model 1 uses only the ischemic score as the predicting attribute, and throws the other attributes out. Model 2, on the other hand, uses four-fifths of the predicting variables to split the data. The significance of this is that by utilizing a greater proportion of the available predicting attributes, Model 2 is a better descriptive model. For most applications the worth or descriptive value of a single attribute is usually already

investigated and understood, using statistical approaches. Thus, Model 2 outperforms Model 1 with respect to objective interestingness measures.

	Model 1 (Traditional Approach)	Model 2 (our Method)
OBJECTIVE INTERESTINGNESS MEASURES		
Accuracy–Classification rate:		
Training Data set	43%	64%
Validation Data set	24%	41%
Stability [0-1]	.30	0.22
Complexity	LOW [2 leaf nodes]	MODERATE [9 leaf nodes]
SUBJECTIVE INTERESTINGNESS MEASURES		
Surprisingness	No	Yes
Actionability	N/A	3-4

Table 1. Summary of Interestingness Measures for a Preliminary Model.

The patterns generated by Model 1 were not surprising, and thus actionability was a moot point. Some leaf nodes in Model 2 were judged surprising, others not. Overall we focus on the surprisingness of patterns and weigh their actionability. Intuitively, one expects that the more complex models are subjectively more interesting than very simple models. More descriptive models may be more expensive to translate into action; however their potential actionability is more readily ascertainable.

IF AHP_100 < 16.45 AND Ischemia IS ONE OF: 0 1 2 AND GCS IS ONE OF: 6 7 8 9 THEN NODE : 32 5 : 66.7% 4 : 16.7% 3 : 16.7% 2 : 0.0% 1 : 0.0%	IF 16.45 <= AHP_100 < 24.85 AND Ischemia IS ONE OF: 0 1 2 AND GCS IS ONE OF: 6 7 8 9 THEN NODE : 33 5 : 0.0% 4 : 14.3% 3 : 14.3% 2 : 42.9% 1 : 28.6%
LHS of node split	RHS of node split

Table 2. Two Leaf Node Examples for Model 2, Using our Method.

The LHS of Table 2 shows that if a patient's AHP global weight is *less* than the lower boundary threshold of 16.45, GCS is greater or equal to 6, and ischemic less than or equal to 2 then there is a 67% chance of good recovery (GOS=5), a total of 83.4% chance of independent survival (GOS = 4, 5) and odds are the patient will not die (GOS =1) or be vegetative (GOS=2) when all three cohere. The RHS of Table 2 shows that if a patient's AHP global weight is *greater* than 16.45 but less than 24.85, with ischemia and GCS unchanged from the rule on the LHS, then the odds of good recovery are nil; there is also a 29% mortality rate, and a 85.8% chance of poor outcome (GOS 1, 2, 3). Here, the lower the AHP score the greater the risk of mortality.

In contrast, results of the first Model 1 show if the ischemic score is equal to or greater than 2, there is a 64.7% chance of mortality. This is correct, but not surprising.

3 CONCLUSION

In this paper, we illustrate a new method for knowledge discovery and illustrate it using a brain trauma ICU case study. Our method relies on the conceptualization of a domain using AHP and rule-based systems to articulate additional domain knowledge. In general, this conceptualization can be applied

to a broad class of problems and domains. The use of AHP in a multiplicity of environments is well-documented, ranging from resource allocation, planning and site location, in health-care, economics, governance, and natural resource problems (Zahedi 1986). Data mining and knowledge discovery in databases (KDD) has also been applied to a variety of application domains, such as sales/marketing, quality control, cost/utilization, and fraud detection to support decision making in these domains. The incorporation of extra-database information into the data mining process in this way uses two well-known techniques (AHP, a decision making technique, and data mining) that dovetail with each other to support a broad class of decision making problems, especially when dealing with large datasets.

References

de Abajo, N., Diez, A. B., Lobato, V. & 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.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). "The KDD Process for extracting useful knowledge from volumes of data." Communications of the ACM 39(11): 27-34.

Han, J. & Kamber, M. (2001). Data Mining: Concepts and Techniques. San Francisco (CA), Morgan Kaufmann.

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly 28(1): 75-105.

Hilderman, R. J. & Hamilton, H. J. (2000). Applying Objective Interestingness Measures in Data Mining Systems, Citeseer Citation Index. 2004.

Hilderman, R. J. & Hamilton, H. J. (2001). Evaluation of Interestingness Measures for Ranking Discovered Knowledge, Citeseer Citation Index. 2004.

Holsheimer, M. & Siebes, A. (1991). Data Mining: The Search for Knowledge in Databases. Technical Report CS-R9406, Amsterdam - Netherlands, CWI. 2004.

Hotz, E., Grimmer, U., Heuser, W., Nakhaeizadeh, G. & Wieczorek, M. (2001). REVI-MINER, a KDD-environment for deviation detection and analysis of warranty and goodwill cost statements in automotive industry. The seventh ACM SIGKDD international conference on Knowledge discovery and data mining (2001), San Francisco (CA), ACM Press.

March, S. T. & Smith, G. F. (1995). Design and Natural Science Research on Information Technology. Decision Support Systems 15: 251-266.

Saaty, T. L. (1994). Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process. Pittsburgh, PA, RWS Publications.

Silberschatz, A. & Tuzhilin, A. (1995). On Subjective Measures of Interestingness in Knowledge Discovery, Citeseer Citation Index. 2004.

Silberschatz, A. & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering 8(6): 970-974.

Yoon, S.-C., Henschen, L. J., Park, E. K. & Makki, S. (1999). Using Domain Knowledge in Knowledge Discovery. The Eighth International Conference on Information and Knowledge Management, ACM Press.

Zahedi, F. (1986). The analytic hierarchy process: a survey of the method and its application. Interfaces 16: 96-108.