Potential Application of Contextual Information Processing To Data Mining

Gregory Vert
Texas A&M Central Texas and
Center for Secure Cyberspace, LSU University
(206) 409-1434
gvert12@csc.lsu.edu

Anitha Chennamaneni
Texas A&M University Central Texas
(254)519-5463
Chennamaneni@tarleton.edu

S.S Iyengar
Center for Secure Cyberspace Computer Science
Louisiana State University
(222) 578-1252
iyengar@csc.lsu.edu

ABSTRACT

Contextual processing is a new emerging field based on the notion that information surrounding an event lends new meaning to the interpretation of the event. Data mining is the process of looking for patterns of knowledge embedded in a data set. The process of mining data starts with the selection of a data set. This process is often imprecise in its methods as it is difficult to know if a data set for training purposes is truly a high quality representation of the thematic event it represents. Contextual dimensions by their nature have a particularly germane relation to quality attributes about sets of data used for data mining. This paper reviews the basics of the contextual knowledge domain and then proposes a method by which context and data mining quality factors could be merged and thus mapped. It then develops a method by which the relationships among mapped contextual quality dimensions can be empirically evaluated for similarity. Finally, the developed similarity model is utilized to propose the creation of contextually based taxonomic trees. Such trees can be utilized to classify data sets utilized for data mining based on contextual quality thus enhancing data mining analysis methods and accuracy.

Keywords
Contextual processing, data mining, taxonomies, data mining quality data sets

1. INTRODUCTION

Data Mining is the process of sifting through large stores of data to extract previously unknown, valid patterns and relationships that provide useful information [14]. It uses sophisticated data analysis tools and visualization techniques to segment the data and evaluate the probability of future events. By scouring large data sets for hidden patterns, these tools not only provide answers to many questions that were traditionally time consuming to resolve, but also present information in easily understandable form. Data mining can be applied to tasks such as decision support, forecasting, estimation, predictive analysis and so forth. Consequently, it is increasingly becoming an important tool allowing businesses to provide more meaningful services to customers and to make proactive, knowledge-driven decisions to increase revenue and reduce expenses. Today, it is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, trend analysis, scientific discovery and other innovative applications.

There are several data mining methods. The most common amongst them are methods that generate descriptive or predictive models to solve problems. Descriptive models provide techniques to discover the hidden patterns in the data and understand the relationships between attributes represented by the data. Predictive models on the other hand, predict future outcomes based on past data. Two approaches are commonly used to generate models: supervised and unsupervised [8]. Supervised models also known as directed learning are goal oriented. Supervised models attempt to explain the value of the target as a function of a set of independent attributes or predictors. In supervised learning, there is a pre-specified target variable and the algorithm is given many examples where the value of the target variable is provided, so that algorithm may learn which values of the target variable are associated with which values of the predictor variables. The goal of unsupervised models also known as undirected learning is pattern detection. In unsupervised modeling, there is no particular target variable as such. Instead, the data mining algorithm searches for patterns and structure among all the variables. Often, unsupervised modeling is applied to recognize relationships in the data and supervised modeling is used to explain those relationships.

Supervised Modeling Techniques

The most common supervised modeling techniques are decision trees, neural networks, naïve Bayes, K-nearest neighbor, case based reasoning, genetic algorithms, rough sets and fuzzy sets.

Decision Trees

Decision trees are tree-shaped structures that represent a series of rules that lead to the classification of a data set. The graphical output has all the basic components of a tree: the decision node, branches and leaves. The decision node specifies the test to be carried out, the branch node leads either to another decision node or a stopping point called leaf node which represent classes or class distributions. A decision tree can be employed to classify an unknown instance by beginning at the root of the tree and navigating through it until a leaf node, which provides the class for that instance. The methods used for building decision trees include classification and regression trees (CART), and chi-square automatic interaction detection (CHAID).

Neural Networks

Often referred to as black box technology, neural networks are densely interconnected networks of related input/output units where each relation has a weight associated with it. The elements of networks are called neurons. Neural networks learn during the learning phase by adjusting weights so as to able to predict the correct class label for the input data set. Neural networks involve

long training times and involve very careful data cleansing, selection, preparation and preprocessing. Advantages of neural networks include their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained. However, they are more complicated than other techniques.

Naïve Bayes

The Naïve Bayes classifier makes predictions based on Bayes' theorem with strong (naïve) independence assumptions. The classifier derives the conditional probability of each relationship, by analyzing the relationships between the independent and dependent variables. Naïve Bayes can generate a classification model by requiring only one pass through the data. As such, it is very efficient and is especially suited when the dimensionality of inputs is very high. However, it does not handle continuous data. Any independent or dependent variables that contain continuous values must be broken into categories, before naive Bayes classifier can be applied.

K-nearest neighbor

K-nearest neighbor classifiers are based on learning by analogy. The classifier classifies new objects based on the closest training samples in the feature space. The training samples are expressed by n-dimensional attributes with each sample representing a point in n-dimensional pattern space. When analyzing an unknown sample, the classifier searches the pattern space for the k training samples closest to the unknown sample. These closest k training samples then become the nearest neighbors of the unknown sample. Unlike other classifiers such as decision tree, neural networks etc, which build a generalization model before receiving new samples to classify, nearest neighbor classifiers merely store all the training samples and do not build a classifier until an unknown sample needs to be classified. This makes them lazy learners.

Case-Based Reasoning

Unlike nearest neighbors, where in training samples are stored as points in Euclidean space, case based reasoning classifiers store cases or samples as complex symbolic descriptions. It classifies a new case based on the solutions of the similar past cases. Upon receiving a new test case, a case-based reasoner first checks to see if an identical training case exists. In one exits, then the associated solution to that case is returned. If no identical case exists, the case-based reasoner searches for other training cases that may have components similar to those of the new test case. It then combines the solutions of these cases and proposes a solution for the new case.

Genetic Algorithms

Genetic algorithms include ideas of natural evolution. They employ optimization techniques that use processes such as genetic combination, mutation, crossover and natural selection. Populations of rules evolve by means of cross over and mutation operations until each rule within a population satisfies a prespecified fitness threshold.

Rough Sets

Rough set theory classifier can be used to discover structural relationships within imprecise or noisy data. A rough set is an approximation of a conventional set in terms of a pair of sets which give the lower and upper approximation of the original set. The classifier handles discrete-valued attributes. Continuous-valued attributes need to be discretized before applying the rough sets. Rough set theory is useful for rule induction from incomplete data sets, feature reduction and relevance analysis.

Fuzzy Sets

A disadvantage of the rule based systems for classification is that they involve sharp cutoffs for continuous attributes. Fuzzy sets, on the other hand, provide a very useful tool to deal with human vagueness by replacing the brittle threshold cutoffs for continuous-valued attributes with degree of membership functions.

Unsupervised Modeling Techniques

Cluster analysis is the most commonly used unsupervised technique.

Cluster Analysis

The most common method for unsupervised modeling is cluster analysis or clustering. Clustering involves segmenting a set of observations into a number of subgroups or clusters so that observations in the same cluster are similar to each other and are dissimilar to the observations in other clusters. Cluster analysis is often used as a standalone data mining tool to gain insights into the distributions of the data or as a preprocessing step for other data mining algorithms.

Outliers

Data set outliers

Outliers are data objects that are significantly different from other objects in the data set. It is a data object that does not comply with the general behavior of the data. Outliers often occur as a result of mechanical faults, changes in system behavior, fraudulent behavior, or through natural deviations in population.

There are two schools of thought on how to deal with outliers. The first school regards outliers as noise or errors and recommends removing them during preprocessing before the data analysis begins to ensure accurate data mining process. The second school of thought regards outliers as rare and interesting patterns hidden in the data that are potentially valuable for decision making. In recent decades, several methods have been proposed for outlier detection. These methods can be broadly grouped into several categories: distribution-based[9], depth-based[10], distance –based[11], density-based[12], clustering-based[13] and model-based.

Taxonomy

Taxonomy is the practice and science of classification of things or concepts according to natural relationships. The classification is often arranged in a hierarchical structure organized by generalization – specialization relationships or supertype-subtype relationships. In an inheritance relationship such as the one in a

taxonomy, the subtype relationship inherits the properties, behaviors and constraints of the supertype. In addition, the subtype has its own properties, behaviors or constraints.

2. CONTEXTUAL PROCESSING

2.1 Background

Contextual processing starts with the notion that data that is not shared often has uncorrelated inferences of meaning and criticalities of information processing in a fashion that truly serves various perspectives needs. Context driven processing is driven by the environment and semantics of meaning describing an event. Often this type of processing requires a context which may contain meta data about the events data. Meta descriptive information of leads to previously unknown insights and contextually derived knowledge. Such meta data usually has a spatial and temporal component to it but is actually much more complicated. The key is that contextual meta data describes the environment that the event occurred in such as the collection and creation of data sets for knowledge mining.

The concept of context has existed in computer science for many years especially in the area of artificial intelligence. The goal of research in this area has been to link the environment a machine exists in to how the machine may process information. An example typically given is that a cell phone will sense that its owner is in a meeting and send incoming calls to voicemail as a result. Application of this idea has been applied to robotics and to business process management [1].

Some preliminary work has been done in the mid 90's. Schilit was one of the first researchers to coin the term context-awareness [2, 3]. Dey extended the notion of a context with that of the idea that information could be used to characterize a situation and thus could be responded to [4]. In the recent past more powerful models of contextual processing have been developed in which users are more involved [5]. Most current and previous research has still largely been focused on development of models for sensing devices [6] and not contexts for information processing.

While research is evolving in the application of contextual information in security, logic, and repository management, little work has been done on the topic of how contextual processing methods might be applied to how data mining training sets might be selected and classified for quality. The model that was originally developed for context processing was that of creating a model for describing information events, storage of meta data and processing rules, thus giving them a context. This context then could then be used to control the processing and dissemination of such information in a hyper distributed global fashion. The next section will provide a general overview of the newly a part of the model defining the dimensions of contextual processing methods. The following section will propose a taxonomic model utilizing contextual dimensions to classify data sets for quality and similarity to a data mining theme.

To understand the issues connected with contexts we present some details about contextual processing.

The initial motivation for the development of a context was to examine the natural disasters of the Indian Ocean tsunami, Three Mile Island nuclear plant and 9/11 to determine what elements could be used to categorize these events. After analysis it was realized that all of them had the following categories, which refer to as the *dimensions of a context*. They are:

time – the span of time and characterization of time for an event

space – the spatial dimension

impact – the relative degree of the effect of the event on surrounding events

similarity – the amount by which events could be classified as being related or not related.

Each one of the dimensions can be attributed with meta characterizers which originally could be use to drive the processing rules. However the meta characterizers also have the potential to control the classification of data, such at that found in the selection of high quality data sets for data mining.

The time and space dimension can be described as having factors of geospatial and temporal elements applied to them to them. The geospatial domain can mean that information is collected and stored at a distance from where it may be processed and used in decision support as well as a description of the region that a context may pertain to. This means that context based data mining data set selection (CBDMDSS) processing must have a comprehensive model underlying it to useful results

Some factors that should be considered in CBDMDSS processing are referred to as information criticality factors (ICF). These factors are further developed in ongoing research but are primarily used to drive processing decision making and classification methods. They may include such attribution among other attributes as:

- time period of information collection
- criticality of importance,
- impact e.g. financial data and cost to humans
- ancillary damage of miss classification
- spatial extent data set coverage
- proximity to population centers spatially or conceptually to other related data sets.

Other factors affecting CBDMDSS classification might be based on the *quality of the data* such as:

- currency, how recently was the data collected, is the data stale and smells bad
- ambiguity, when things are not clear cut – e.g. does a degree rise in water temperature really mean global warming

- contradiction, what does it really mean when conflicting information comes in different sources
- truth, how do we know this is really the truth and not an aberration
- confidence that we have the truth

In order to analyze the above factor and their effect on CBDMDSS, it was useful to examine three different data sets describing natural and manmade disasters most people are familiar with in which selection of the best data set for analysis of the event might have lead to better response. We initially considered the 9/11 incident where information about the attackers and their operations and activities were stored everywhere from Germany, to Afghanistan to Florida. If the information could have been orchestrated into a contextual collection of data, the context and relationships of the data would have given a very different interpretation or knowledge about what was really going on. Of course the goal of our model does not examine how that information would be located and integrated, that can be the subject of future work. The model only proposes a paradigm for data set selection based on contextual factors that might affect quality impact of erroneous analysis. For the initial analysis of 9/11 we came up with the following descriptive factors which eventually lead to the derivation of the context of contextual dimension presented earlier. These were:

temporality – defined to be the time period that the event unfolded over from initiation to conclusion

damage – the relative damage of the event both in terms of casualties, and monetary loss

spatial impact – defined to be the spatial extent, regionally that the event occurs over.

policy impact – directly driving the development of IA (security) policy both within a country and among countries. This directly led to the evolution of security policy driving implementation because of the event.

2.2 Defining a Context

Contextual processing is based on the idea that information can be collected about events and objects such as data sets and that meta information about the object can then be used to control how the information is processed by a data mining methods. In its simplest form, a context is composed of a feature vector

$$F_n<a_1,..a_n>$$

where the attributes of the vector can be of any data type describing the object. Feature vectors can be aggregated via similarity analysis methods, still under investigation, into super contexts S_c. Some potential methods that might be applied for similarity reasoning can be statistical, probabilistic (e.g. Bayesian), possibilistic (e.g. fuzzy sets), support vector machines or machine learning and data mining based methods (e.g. decision trees). The goal of similarity analysis is to be able to state the following:

$$R(A|B)$$

where:

A, B - are sets of contextual vectors F_n about a data set

R() - is a relation between A and B, s.t. they can be said to be similar in concept and content

Similarity analysis should facilitate the aggregation into super sets of feature vectors describing attributes of a data set based on contextual dimensions describing the data set. This is done to mitigate collection of missing or imperfect information and to minimize computational overhead when processing contexts.

definition: A context is a collection of attributes aggregated into a feature vector describing a abstract event, object or concept.

A super context was previously described as a triple denoted by:

$$S_n = (C_n, R_n, S_n)$$

where:

C_n - is the context data of multiple feature vectors

R_n - is the meta-data processing rules derived from the event and contexts data

S_n - is controls security processing.

However, in definition of super context there is really not a need for the S_n vector unless the data set and its processing will be hyper distributed.

definition: A super context for contextual data mining is defined by the feature vectors describing the contextual dimensions of the set and the data mining methods applied to the set.

The redefinition for data mining set quality and selection might then becomes:

$$S_n = (C_n, M_n)$$

where:

C_n - is the context data of multiple feature vectors

M_n - is the meta-data processing rules and application methods for analysis quality for the data set, e.g. taxonomic methods.

The cardinality of F_n with C is still a defined in previous work:

m:1

which when substituted into S creates a (C_n, M_n) cardinality of:

m:1:1

All of the above are a *type* of feature vectors where the elements of the vector can considered as inputs to a contextual processing black box along with the sets data to produce a better selection of data mining sets for processing.

3. SUPER CONTEXTS AND TIME

3.1 Overview

Super contexts are composed of context data from many sensing event objects, Eo_i, As such contextual information collection works in a similar fashion to sensor networks and can borrow from theory in the field in application of quality metrics for data mining sets.

definition: A thematic event object (Teo) is the topic of interest for which event objects are collecting data. An example of a Teo would be the center of a tsunami.

3.2 General Operation and Concept

As previously defined contexts are composed are defined by four dimensions those of temporality, spatiality, impact and similarity. Contextual objects thus can have *meta characterizations* based on any of these areas. Some examples presented in previous work are:

- Singular – an event that happens a point in time, at a singular location
- Regional
- Multipoint Regional
- Multipoint Singular – events that occur at a single point in time but with multiple geographic locations
- Episodic – events the occurs in bursts for given fixed or unfixed lengths of time
- Regular – as suggested these events occur at regular intervals
- Irregular – the time period on these type of events is never the same as previous t
- Slow Duration - a series of event(s) that occupy a long duration, for example the eruption of a volcanoes
- Short Duration – example an earthquake
- Undetermined
- Fixed Length
- Unfixed Length
- Bounded
- Unbounded
- Repetitive - these types time events generate streams of data – graph of attributes change in value over time

These meta-characterization can be applied to data for the previously discussed original 9/11 analysis (temporality, damage, spatial impact, policy impact) and thus to the final set of dimension that were derived for characterization of contextual processing.

The above meta-characterization of context can be classified into categories that can then be utilized in mapping context to data mining quality. The classifications can be utilized with quality metrics in data mining to point the way to methods that might classify quality of data sets based on context. The previously developed categories are:

Event Class < abstract, natural>

Event Type < spatial, temporal>

Periodicity < regular, irregular>

Period < slow, short, medium, long, undeterminable, infinite, zero >

Affection<regional, point, global, poly nucleated, n point>

Activity < irregular, repetitive, episodic, continuous, cyclic, acyclic>

Immediacy < catastrophic, minimal, urgent, undetermined >

Spatiality < point, bounded, unbounded >

Dimensionality <1, 2, 3, n>

Bounding < Fixed Interval, Bounded, Unbounded, Backward Limited, Forward Limited, continuous>

Directionality < linear, point, polygonal >

Figure 2: Modeling the semantic categories of context based meta-characterizations of data in a context

Previously the above were developed into a semantic grammar that could control processing of information. However such a grammar could also be developed to classify the quality of data mining sets in a high level qualitative fashion. The syntax took the form of the following:

R1: <event class>, <event type>, <R2>

R2: (<periodicity> <period>) <R3>

R3:(<affection><activity>) <spatiality> <directionality> <bounding> <R4>

R4: <dimensionality> <immediacy>

Figure 3: Syntax for application data meta-characterizations to derive super context processing rules R in S(C, R, S).

The above grammar and syntax can be developed into sentences describe qualitatively the integrity of a data set for data mining purposes based on the theme about which the data was collected. For example, the following might be a semantic descriptor of a data set about an event in statistics, perhaps a cancer data set:

"R1 = abstract, spatial & temporal, regular-slow, episodic urgent"

In this case "abstract" defines the fact that data may be derived from naturally observable data, that the area of the cancer cluster occurs in "spatial and temporal" areas of the country, that the development of the cancers occur regularly (such as skin cancers in the southwest, and that the urgency of the data needs to be considered. If we are talking about another cancer data set, described by the following rule:

"R2 = abstract, singular event, point, undetermined"

One can see that from an HCI standpoint one might be more likely to select the data set defined by R1 as being potentially better data integrity. Any one of these contextual meta characterizations can be selected from the meta-classification categories present previously.

However, while these are mappings of contextual processing concepts onto data mining data set quality suggesting that this can be done, it does not really incorporate the issues of quality as data mining defines them with the concepts of contextual processing.

Some measures of quality in data mining sets are defined as the following:

- Relevance (Re) - degree of relationship of data to a theme
- Timeliness (Ti)- temporal proximity of the data to it T_{eo}
- Noise (No) – the degree to which data is observed versus injected by observational equipment.
- Outliers (Ou) – observed actual data that exceeds the norm
- Sparsity (Sp)– a binary representation of known data versus unknown data in a matrix of observed data
- Dimensionality (Di) – the number of observed attributed about a T_{eo}, of which some may be more relevant for analysis
- Freshness (Fr)–
- Accuracy (Ac)– the degree to which the data in the set reflects reality
- Sequentiality (Se) –
- Bias (Bi) –
- Duplication (Du) – a characterization of some data appearing to be the same and representing different objects, other times being the same and representing the same object
- Aggregation (Ag) – the degree to which data is combined where increased aggregation produces better stability in the data but loses granularity.

In considering the development of a method that could be used to taxononify data mining sets, the next step was to map the above defined measures of data mining quality onto contextual dimensions. This produced the following mapping:

$$f(Re) \rightarrow (T, Sp, Im, Si)$$
$$f(Ti) \rightarrow (T, Im)$$
$$f(No) \rightarrow (Im, Sp, Am)$$
$$f(Ou) \rightarrow (Im, Sp, Am)$$
$$f(Sp) \rightarrow (Im, Si, Am)$$
$$f(Di) \rightarrow (Sp, Im, Am)$$
$$f(Se) \rightarrow (T, Si)$$
$$f\,Bi) \rightarrow (Am)$$
$$F(Du) \rightarrow (Am, Si, Im, Sp)$$
$$F(Ag) \rightarrow (Am)$$

where:

Sp	- spatiality dimension
T	- temporality dimension
Si	- similarity dimension
Im	- impact dimension
Am	- the ambiguity dimension

Of note in this mapping, a new dimension became necessary in order to map context onto measures of data set quality in data mining. This dimension is new in context and has a bidirectional mapping back onto the area of contextual processing in general.

4. Application to Adaptive Quality Measures and Taxonification Construction

4.1 Overview

The above mapping suggests the next step in potentially developing a taxonomy for classification of data set quality. This method requires the use of a similarity matrix and an algorithm which is the subject of future research. The proposed approach is to create a similarity matrix based on the reversing the mapping previously done where for instance:

$$f(T) \rightarrow (Re, Ti, Sei)$$
$$f(Am) \rightarrow (No, Ou, Sp, Di, Du)$$
$$f(Sp) \rightarrow (Re, No, Ou, Di, Du)$$
$$f(Si) \rightarrow (Re, Sp, Du)$$
$$f(Im) \rightarrow (Re, Ti, Ou, Sp, Di, Du)$$

In this mapping the previous measures of data set quality are mapped onto the four original dimensions of context plus the new one of Ambiguity. Note, the evaluation of how one for instance measures the quality metrics is the subject of future research in this method. For instance determination of what would be appropriate measure of *freshness* or *timeliness.*

The notion of mapping data set quality measures onto the dimensions of context allow an adaptive similarity matrix to be developed where the measures of quality can be weighted based on observed errors and deviations for once a data mining method is applied to a data set. While the exact process is not currently defined and is the subject of future research, it may take the following form in an algorithm.

Algorithm in pseudo code:

// 1 Develop a similarity matrix weight at .5 where the value means the similarity of on dimensions quality measures are not known

// 2 During a training phase apply a selected data mining method iteratively to several data sets and measure the accuracy of the method

// 3 Based on step two re-weight the similarity matrix

//4 At the end of training organize low similarity dimensions into the top most levels of a classification taxonomy into the upper most levels of a taxonomic tree because they are the most general. Place the

Logic:

The tree then becomes a classification scheme against which multiple data sets could be evaluated based on quality metrics. Low quality data sets in theory would classify into root nodes higher up in the tree. Thus selection of the data sets lower in the tree could in theory indicate higher quality for the set overall.

A practical example might be the following:

Step 1: Initialization of the similarity matrix for contextual dimension sets containing data mining quality measures

	f(T)	*f(Am)*	*f(Sp)*	*f(Si)*	*f(Im)*
f(T)	1	.5	.5	.5	.5
f(Am)	.5	1	.5	.5	.5
f(Sp)	.5	.5	1	.5	.5
f(Si)	.5	.5	.5	1	.5
f(Im)	.5	.5	.5	.5	1

Note, each of the above sets *f()* represents a collection of quality metrics for data mining that has been mapped into the set previously. The determination of such values as data set freshness within a set is the subject of the next step in this research. Also of note is that the degree of similarity is 1 (true) when considering the same dimensions and .5 representing not known when considering similarity with other contextual dimensions and quality mappings.

Step 2: Apply a selected data mining methods and for each of the contextual dimension sets measure the *correlation* of each with each other sets against the observed error.

A sample might be the following for 10 data sets on a particular Teo:

f(T) = .2
f(Am) = .3
f(Sp) = .9
f(Im) = .85
f(Si) = .5

*Step 3:*The interpretation (a to be developed method) of the above might be:

f(T)≈ *f(Am)*= {.2, .3} → high similarity as low quality mapping of contextual dimension quality factor measures (.1)

f(Sp)≈ *f(Im)*= {.9, .85} → high similarity as high quality mapping of contextual dimension quality factor measures (.9)

f(Si)!≈ {*f(Sm), f(T), f(Im), f(Sp)*= {.5,} → unknown (.5) similarity to high or low quality mappings of contextual dimension quality factor measures

The similarity matrix might be reweighted in the following fashion to reflect these discovered relationships:

	f(T)	*f(Am)*	*f(Sp)*	*f(Si)*	*f(Im)*
f(T)	1	.1	.5	.5	.5
f(Am)	.5	1	.5	.5	.5
f(Sp)	.5	.5	1	.5	.9
f(Si)	.5	.5	.5	1	.5
f(Im)	.5	.5	.5	.5	1

Step 4: Construct a classification taxonomy for other data sets on the T_{eo} that organized contextual mapped classes of data mining quality from the generally poor predictors of good quality to the better predictors.

Based on Step 3's data such taxonomy might have the structural organization shown in figure 1. In this taxonomy it is important to reiterate that the data mining measures of quality defined

previously have been mapped as evaluative attributes inside each contextual dimension class.

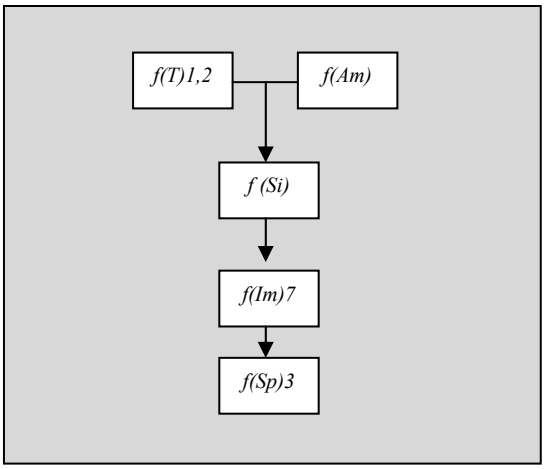

Figure n: A possible construction of a taxonomic tree for classification of data mining sets based on the similarity matrix.

In this taxonomy some data sets will not be able to be classified into lower levels of the tree and some may based on the data mining quality factor mentioned previously. Utilization of the sets that reach the bottom of the taxonomy implies that they have the highest quality and there are the best for analysis. In figure n, data set 3 classifies to the bottom of the tree based on this method and is probably the highest quality. Whereas data set 1, 2 only will classify to the top most level of the tree and is therefore not the best to conduct analysis on based on the data quality factors in the *f(T)* mapping.

This matrix maps the similarity of one dimensions of context to another based on the mapping of quality metrics onto a given dimension of context.

In this example, the value of .5 is the initial value set into the matrix meaning that the similarity among dimensions may exist or may not exist.

The matrix is adaptive in its classification method in that the following algorithm could be developed for changing the similarity weighting during a training period. Such an algorithm might have the following structure:

5. CONCLUSIONS

5.1 Future Work

The modeling of contextual processing and is a broad new area of computer science and can be the beginning of many new research threads. This paper proposes a method by which the vague attributes of data mining set quality might be mapped into the contextual model. It then provides a method to evaluate the relationships among contextual dimensions via a similarity matrix to determine which mapped contextual quality dimension might have the highest degree of relationship. Strong relationship can then be the basis for creation of a data mining set taxonomic tree can be constructed. Such a tree could be able to classify data mining sets based on contextual quality.

Much work remains to be developed on this topic. The first area to pursue would be that of determining quantitative models for the data mining set quality attributes. For instance how would *freshness* be modeled and what is its relationship with other quality measures. The development of the algorithm for interactive population of the similarity matrix would be another large area of research. The issue in this research would be how to build an algorithm that can duplicate its results consistently and reliably. Finally, the construction method for taxonomic contextual quality classification should be investigated thorough for its application and principles. It may turn out that this concept borrowed from biology might actually merge with methods in data mining to become a new approach to data mining. The concept of contextual processing is broad and new. As such, it offers potential application to data mining and a variety of other fields.

6. REFERENCES

1. Rosemann, M., & Recker, J. (2006). "Context-aware process design: Exploring the extrinsic drivers for process flexibility". T. Latour & M. Petit *18th international conference on advanced information systems* engineering. proceedings of workshops and doctoral consortium*: 149-158, Luxembourg: Namur University Press.*
2. Schilit, B.N. Adams, and R. Want. (1994). "Context-aware computing applications" (PDF). *IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'94), Santa Cruz, CA, US*: 89-101.
3. Schilit, B.N. and Theimer, M.M. (1994). "Disseminating Active Map Information to Mobile Hosts". *IEEE Network* **8** (5): 22–32. doi:10.1109/65.313011.
4. Dey,Anind K. (2001). "Understanding and Using Context".*Personal Ubiquitous Computing* **5** (1): 4–7. doi:10.1007/s007790170019.
5. Cristiana Bolchini and Carlo A. Curino and Elisa Quintarelli and Fabio A. Schreiber and Letizia Tanca (2007). "A data-oriented survey of context models" (PDF). *SIGMOD Rec.* (ACM) **36** (4): 19--26. doi:10.1145/1361348.1361353. ISSN 0163-5808. http://carlo.curino.us/documents/curino-context2007-survey.pdf.
6. Schmidt, A.; Aidoo, K.A.; Takaluoma, A.; Tuomela, U.; Van Laerhoven, K; Van de Velde W. (1999). "Advanced Interaction in Context" (PDF). *1th International Symposium on Handheld and Ubiquitous Computing (HUC99), Springer LNCS, Vol. 1707*: 89-101.

7. Tan, Pang-Ning, Steinbach, M., Kumar, V., Introduction to Data Mining, pp 27-43, Addison Wesley, 2006.
8. J. Han and M. Kamber, *Data Mining: Concept and Techniques*, Morgan Kaufmann Publishers, 2001.
9. V. Barnett. *Outliers in Statistical Data*. John Wiley, 1994.
10. J.W. Tukey. *Exploratory Data Analysis*. Addison-Wiley, 1977.
11. Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *VLDB*, pages 392-403, 1998.
12. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and JÄorg Sander.LOF: Identifying density-based local outliers. In *SIGMOD Conference*, pages 93-104, 2000.
13. Martin Ester, Hans-Peter Kriegel, JÄorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise.In *KDD*, pages 226{231, 1996
14. R. Agrawal, T. Imielinski and A. Swami, Data Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering* **5**(6) (1993), 914–925.