A Heuristic Method for Correlating Attribute Group Pairs in Data Mining

Chua Eng Huang Cecil[1], Roger H.L. Chiang[1], and Ee-Peng Lim[2]

[1] Information Management Research Centre, School of Accountancy and Business
Nanyang Technological University, Singapore 639798
{p7408153c,ahlchiang}@ntu.edu.sg
[2] Center for Advanced Information Systems, School of Applied Science
Nanyang Technological University, Singapore 639798
aseplim@ntu.edu.sg

Abstract. Many different kinds of algorithms have been developed to discover relationships between two attribute groups (e.g., association rule discovery algorithms, functional dependency discovery algorithms, and correlation tests). Of these algorithms, only the correlation tests discover relationships using the *measurement scales* of attribute groups. Measurement scales determine whether order or distance information should be considered in the relationship discovery process. Order and distance information limits the possible forms a legitimate relationship between two attribute groups can have. Since this information is considered in correlation tests, the relationships discovered tend not to be spurious. Furthermore, the result of a correlation test can be empirically evaluated by measuring its significance. Often, the appropriate correlation test to apply on an attribute group pair must be selected manually, as information required to identify the appropriate test (e.g., the measurement scale of the attribute groups) is not available in the database. However, information required for test identification can be *inferred* from the system catalog, and analysis of the values of the attribute groups. In this paper, we propose a (semi-) automated correlation test identification method which infers information for identifying appropriate tests, and measures the correlation between attribute group pairs.

1 Introduction

Many algorithms have been developed to discover the extent that the values of a pair of attribute groups associate with each other. These *relationship discovery* algorithms include algorithms that mine for functional dependencies [13], association rules [1], and correlations [2].

Most of the recently developed algorithms discover relationships between attribute groups without taking into account information such as the measurement scale, or statistical significance of the results. Thus, the relationships discovered by these algorithms are often spurious or erroneous. For example, the support and confidence framework used by association rules may discover relationships that are contrary to the real world situation [3].

Y. Kambayashi et al. (Eds.): ER'98, LNCS 1552, pp. 29–40, 1999.

One type of information that can assist in relationship discovery is the *measurement scale* of the attribute group. Measurement scales specify the measurement properties (distinctiveness, order, distance, and zero value) in effect on an attribute group. Statisticians recognize four different measurement scales, the nominal, ordinal, interval, and ratio scales [2]. The nominal measurement scale indicates that the values of an attribute group are *distinct*. For example, `Religion` has a nominal measurement scale. All we know about the values 'Catholic', and 'Buddhist' is that they describe different religions. The ordinal measurement scale indicates that the values of an attribute group not only are distinct, but have an *intrinsic order* as well. For example, `School_Ranking` has an ordinal measurement scale. Being the '1st' in school is better than being the '2nd'. The interval measurement scale indicates that values are not only distinct, and have order, but also the distance between the values can be measured. For example, `Date_of_Birth` has an interval measurement scale. Someone born on September 4, 1976 was born 2 days after someone born on September 2, 1976. The ratio measurement scale has all of the properties of the interval measurement scale. In addition, one value of the ratio measurement scale conforms to a 'zero value'. This 'zero value' indicates the absence of the property that the attribute group measures. For example `No_of_Children` is on the ratio scale. If you have 0 children, you have no children.

An example of how measurement scale information is useful in determining the spuriousness of a relationship is given in Table 1. In Table 1, a one-to-one mapping can be established between the values of `Salary`, and `Time_Leave_Home`. This mapping seems to imply that a relationship exists between `Salary`, and `Time_Leave_Home`. However, `Salary`, and `Time_Leave_Home` have the interval measurement scale, and can be treated as continuous. We can therefore expect that a genuine relationship between these two attributes would be a continuous function. However, the Intermediate Value Theorem [16] states that for any continuous function, every value Y between some pair of values B_1 and B_2 will have a corresponding value X between the pair of values A_1, and A_2, where A_1 is associated with B_1, and A_2 is associated with B_2. In Table 1, we see no evidence that the Central Limit Theorem holds. Thus, we can suspect that the relationship between `Salary`, and `Time_Leave_Home` is spurious.

Table 1. An Artificial Relationship

Salary	Time_Leave_Home
10,425.00	7:45
15,492.66	6:30
20,485.42	8:15
15,628.23	7:54
27,154.21	7:05

Correlation tests employ the measurement scale to measure the relationship between two attribute groups. As they exploit the properties of measurement scales, relationships they discover tend to be less spurious than other relationship discovery algorithms which do not exploit this information. This is important,

as large databases may contain many spurious relationships. In addition, the results of correlation tests can also be validated by measuring the *significance* of the results. Significance measures the probability that a relationship discovered by the correlation test is a spurious one [2].

However, there are problems with using correlation tests for data mining. These problems include: 1) the identification of the measurement scale of an attribute group, and 2) the need for more information than just the measurement scales to determine the appropriate correlation test. First, to identify the appropriate correlation test, the measurement scales of the attribute groups must be known. For example, one of the correlation tests for interval measurement scales, such as the coefficient of determination should be used to measure the correlation of an attribute group pair with interval measurement scales. The correlation coefficient of an attribute group pair with nominal measurement scales should be calculated using a correlation test such as the phi coefficient. However, because of the large number of attributes in the relation, the data mining specialist cannot be expected to determine the measurement scale for every single attribute group manually.

Also, additional information needs to be derived in order to determine the appropriate correlation test from those usable for each measurement scale. For example, both the phi coefficient, and the contingency coefficient are used on attribute group pairs with nominal measurement scales. However, the phi coefficient should be employed only when both attribute groups have at most two distinct values, and the contingency coefficient should be applied when any attribute group has three or more distinct values. This information can be inferred from the values, length, and data type of the attribute.

Finally, the problem of identifying correlation tests for relationship discovery in data mining is complicated by the need to identify the appropriate correlation tests not only for the individual attribute pairs, but also the attribute group pairs. For example, while the attributes `Monthly_Salary`, `Bonus`, `Net_Yearly_Salary`, and `Taxes` may have no strong pair-wise relationships, the combination {`Monthly_Salary, Bonus`} may correlate strongly with {`Net_Yearly_Salary, Taxes`}, as the attributes are related to each other through the function `Monthly_Salary` $\times 12+$ `Bonus` = `Net_Yearly_Salary` + `Taxes`.

Therefore, a (semi-)automatic method should be developed to facilitate determining the appropriate correlation tests. In this paper, we propose and discuss a (semi-)automated correlation test identification method. The result of the appropriate correlation test (i.e. correlation coefficient) measures the strength of the relationship of the attribute group pair. While our correlation test identification method assumes that the database to be mined follows the relational database model, it can be generalized to other database models as well. In this paper, the discussion of our method focuses on the attribute groups of a single relation only.

Problem Definition- Let R be a relation with attributes $\mathtt{A}= \{\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{A}_n\}$. $G = 2^A$ is the set of attribute groups. Let $P_{ij} = (g_i, g_j)$ be a relationship between two mutually exclusive attribute groups, i.e. $g_i \in G$, $g_j \in G, g_i \cap g_j = \emptyset\}$. Given

a P_{ij}, calculate the *correlation coefficient* S_{ij} (which determines the strength of the relationship between the attribute groups of the pair). To calculate S_{ij}, we must first identify the appropriate correlation test T for (g_i, g_j).

2 Overview of the Correlation Test Identification Method

Attribute characteristics necessary for determining the appropriate test must be identified. We identify these characteristics and classify them into a set of *representations*. For each attribute, at most one representation should be identified. The representation can be identified by analyzing the instances, data type, and length of each attribute. We assume that the system catalog records the data types of attributes. If the system catalog does not contain information on the length of attributes, this information can be derived from the instances. In addition, the representations of the individual attributes are used to infer the representations of the attribute groups. The two representations of an attribute group pair will then determine the appropriate correlation test used to measure their relationship.

Data instances are used to infer the representation of an attribute. In a large relation, examining all data instances may take too much time. To speed up processing, a sample may be used for analysis as long as that sample is representative. Our method incorporates a formula which estimates the representative sample size.

There are four steps in the correlation test identification method:

Step 1: **Take a Representative Sample**. A representative sample of the relation is extracted to speed up relationship discovery. The size of the sample is determined based on the acceptable degree of error, and the acceptable probability that this degree of error will be exceeded. The formula for calculating the sample size (Step 1) is discussed in Section 3.

Step 2: **Assign Representations to All Attribute Groups**. Each attribute is assigned a representation based on the analysis of its data type, length of the attribute, and values. The representations of attributes are then used to determine the representations of attribute groups. The identification of representations for attributes, and attribute groups (Step 2) are discussed in Section 4, and 5 respectively.

Step 3: **Select Tests**. The attribute groups are paired to form all possible sets of P_{ij}. The representations of the two attribute groups in each pair are then used to identify the most appropriate correlation test. Section 6 discusses the identification of the appropriate correlation tests.

Step 4: **Execute and Evaluate the Correlation Test Result**. The correlation test is executed, and the result and its significance is analyzed. If the result of the test is strong and significant, it is validated by executing the correlation test on other samples in the same attribute group pair. The evaluation of correlation tests is discussed in Section 7.

3 Identifying a Representative Sample

Most of the time, the relation to be mined will contain a large number of instances. Any examination of all instances in the relation would take a very long time. If a random sample of the data in the relation is extracted, it often will be representative of the original data set. Thus, while mining the *representative* sample is quicker than mining the relation, the result of mining the sample will often be comparable to that of mining the relation.

Our method uses the worst case of the formula for calculating sample size from estimated proportions $n = \frac{p \times (1-p) \times Z(\alpha/2)^2}{\epsilon^2}$ to generate our sample size, where p is a value between 0 and 0.5 which indicates a degree of variability, n is the size of the sample, ϵ is a degree of error, α is a probability this degree of error will be exceeded, and Z is a function describing the area under the standard normal curve [10]. As we do not know the degree of variability, we set p to be the worst case value of 0.5. This formula can be used to estimate sample size in any relation with a large number of instances, as it assumes that the number of instances is *infinite*.

4 Assigning Representations to Individual Attributes

4.1 Representations

To identify the appropriate correlation test for an attribute group pair, we need to know more than just the measurement scale. The measurement scales are subdivided into a set of *representations*, which captures this additional information.

We are not aware of any correlation test which exploits zero value information to measure a relationship. Therefore, we do not differentiate between attribute groups with the ratio and interval measurement scales. For the interval measurement scale, two major characteristics of the data play a major part in determining the appropriate correlation test. First, the various tests for attributes groups with interval measurement scales can only handle attribute groups with certain numbers of attributes. For example, the coefficient of determination requires that of the two attribute groups being compared, one of them contains only one attribute. Second, dates differ from other kinds of attributes with the interval measurement scale, because multiplication, division, exponentiation, etc. cannot be performed on dates. Thus, two date attribute groups $\{X_1, X_2, \dots\}, \{Y_1, Y_2, \dots\}$ can relate to each other only through a linear functional form, i.e. $X_1 \pm X_2 \pm \ldots \pm C = Y_1 \pm Y_2 \pm \ldots$.

Few correlation tests exist for the ordinal measurement scale, so it is not necessary to further partition it. However, for the nominal measurement scale, special case tests exist which exploit the special characteristics of *dichotomies*. Dichotomies are attribute groups with nominal measurement scales which have only two distinct values. Since dichotomies only have two values, they have characteristics that are different from other nominal attribute groups. For example, it can be assumed that all values in the dichotomy are equidistant. In the dichotomy

`Sex` with values {M, F}, we can say that |M − F| = |F − M|. However, for a non-dichotomy like `Religion`, it cannot be assumed that |Catholic − Buddhist| = |Buddhist − Moslem|. We subdivide the nominal measurement scale into two scales, one scale for dichotomies, and one scale for non-dichotomies.

We take these characteristics of the measurement scales into account by partitioning the measurement scales into the following representations:

- Multi-attribute Numeric (MAN)- An attribute group with this representation has an interval measurement scale. This attribute group contains more than one attribute.
- Single-attribute Numeric (SAN)- An attribute group with this representation has an interval measurement scale. This attribute group contains only one attribute.
- Single-attribute Date (SAD)-This is an attribute group which represents temporal data (e.g., The attribute group {`Date_of_Birth`} is a SAD). This attribute group contains only one attribute.
- Multi-attribute Date (MAD)- An attribute group with this representation uses several attributes to describe a temporal 'fact'.
- Ordinal(Ord.)- The Ordinal representation indicates that while the values of the attribute group have a ranking order, no information is available to determine the *distance* between the values.
- Categorical(Cat.)- This representation means that the only measurement property found in the attribute group is distinctness.
- Dichotomous(Dich.)-Attribute groups with the Dichotomous representation can only contain two distinct values, e.g., {M, F}, {0, 1} etc.

The measurement scales are partitioned into the following representations: 1) The Interval measurement scale is partitioned into the MAN, SAN, MAD, and SAD representations, 2) The Ordinal measurement scale has the analogous Ordinal representation, and 3) The Nominal measurement scale is partitioned into the Categorical and Dichotomous representations. In this paper, any discussion of a measurement scale applies to all representations with that measurement scale. In addition, when we refer to a Numeric representation, we mean the MAN, and SAN representations collectively. When we refer to a Date representation, we mean the SAD and MAD representations collectively.

4.2 Data Types

The data type of an attribute is useful in determining its representation. However, different RDBMSes use different labels to describe the same data types. For the purpose of this paper, we assume that the following are the data types available in the RDBMS:

- Integers- The values of attributes with this data type differ from each other in increments of at least one unit. For example, the attribute `Years_Of_Service` is often given an Integer data type.

- Decimals - Attributes with this data type may have values which differ from each other by less than one unit. For example, the attribute `Salary` can be given the Decimal data type.
- Date- A data type where user input is restricted to specifying a day, month, and year. For example, `Date_Of_Birth` is often assigned a Date data type.
- String-Each character in this data type may have any value.

4.3 Identifying and Eliminating Representations for Single Attributes

An attribute with a particular data type can only have certain representations. The possible representations of each data type are shown in Table 2. A set of heuristic rules are then applied to the attribute to identify which of the possible representations is the correct one. Many of these heuristic rules have been adapted from rules used in SNOUT [15] to identify measurement scales from survey data.

Table 2. Initially Generated Hypotheses

Type	SAN	SAD	ORDINAL	CATEGORICAL	DICHOTOMOUS
Integer	√	√	√	√	√
Decimal	√	√			
String	√	√	√	√	√
Date		√			

The heuristic rules are listed below. Each heuristic rule uses more information to determine the correct representation than the previous rule. However, the additional information used by each rule is less reliable as compared to information added by the previous rule. Thus, representations identified by an earlier rule can be said to be more accurate than representations identified by a later rule.

1. Attributes with the Date data type have the DATE representation.
2. If the attribute has a possible DICHOTOMOUS representation, and the number of distinct values of the attribute is two or less, the attribute has the DICHOTOMOUS representation. If the number of values is more than two, the attribute can not have the DICHOTOMOUS representation.
3. If the length of the attribute varies across the values, it cannot have the SAD representation.
4. If the values of an attribute with the String data type contains non-numeric characters, then the attribute cannot have the SAN representation.
5. If an attribute does not conform to an accepted date format, it cannot have the SAD representation. For example, an attribute with the instance '231297' might have the date representation, since this instance could mean December 23, 1997. However, if the same attribute had another instance '122397', it could not have the date representation, since both instances could not indicate a date under the same format.

6. If an attribute is longer than nine characters, it cannot have the ORDINAL or CATEGORICAL representations. If an attribute has more than 25 distinct values, it cannot have these two representations either. An attribute with less than 25 distinct values cannot have the SAN representation [15].
7. Attributes with the Integer data type can not have the CATEGORICAL representation if the length of the values is greater than 2. If the difference between the minimum and maximum value of attributes with the Integer data type is greater than 25, it will not have the CATEGORICAL REPRESENTATION. This rule follows as a direct consequence of rule 6. If an attribute with the String data type has only values with characters between '0' and '9', then it must also follow this rule.
8. If the first character of any value of an attribute with the String data type differs from the first character of all other values by 3 or more, (e.g., 'D' differs from 'A' by 3), then the attribute does not have an ORDINAL representation.
9. If an attribute may have both a SAD and a SAN representation, the attribute will not have the SAN representation.
   Justification: The range of values which can indicate a SAD forms only a small proportion of the range of values which can indicate a SAN. If every single value in an attribute falls into the range of values that indicate a SAD, the attribute is more likely to represent a SAD than a SAN.
10. If an attribute can have both an ORDINAL and a CATEGORICAL representation, the attribute cannot have the ORDINAL representation.
    Justification: It is often difficult to identify whether an attribute really has an ORDINAL or a CATEGORICAL representation based on the schema and instance information alone. However, a correlation test designed for attribute groups with CATEGORICAL representations can be correctly used to compare attributes with ORDINAL representations. However, the reverse is not true. By using this rule, we err on the side of caution.
11. Attributes which may have SAD, and ORDINAL, or SAD, and CATEGORICAL representations have the SAD representation. The reasoning in this rule is similar to that of rule 9.

After an attribute has been processed using these rules, it is possible for it to have no representation. This indicates that the attribute can not be analyzed using correlation tests. After the system has discovered the representations for the individual attributes, the data mining specialist may review the results, and change any representation he or she deems incorrect.

5 Identifying Representations for Attribute Groups

The representation for an attribute group which contains only one attribute is the same as the representation of that attribute. The representation for an attribute group containing more than one attribute is determined by consulting Table 3. We discuss some of the counterintuitive derivations of attribute group representations in this section.

Table 3. Representations of Attribute Groups With More Than One Attribute

Attr. Grp 1 \ Attr. Grp 2	MAN	SAN	MAD	SAD	ORD.	CAT.	DICH.
MAN	MAN	MAN	MAD	MAD	N/A	N/A	N/A
SAN	MAN	MAN	MAD	MAD	N/A	N/A	N/A
MAD	MAD	MAD	MAD	MAD	N/A	N/A	N/A
SAD	MAD	MAD	MAD	MAD	N/A	N/A	N/A
ORD.	N/A	N/A	N/A	N/A	CAT.	CAT.	CAT.
CAT.	N/A	N/A	N/A	N/A	CAT.	CAT.	CAT.
DICH.	N/A	N/A	N/A	N/A	CAT.	CAT.	CAT.

An attribute group with the NUMERIC *representation combined with an attribute group with the* DATE *representation produces an attribute group with the* MAD *representation.* This rule reflects the situations when numbers are used to increment or decrement a date. For example, when the attribute group with SAN representation {`Project_Duration`} is combined with the attribute group with SAD representation {`Date_Started`} it produces the attribute group {`Project_Duration, Date_Started`}. The combined attribute group identifies the date the project was completed.

An attribute group with the ORDINAL *representation combined with an attribute group with the* ORDINAL *representation produces an attribute group with the* CATEGORICAL *representation:* Intuitively, the combination of two attribute groups with ORDINAL representations should result in an attribute group with the ORDINAL representation. We do not allow this for two reasons. First, the ordering priority of the two attribute groups is often not self evident. For example, the attribute group {A_1, A_2} may be ordered as (A_1, A_2) or as (A_2, A_1). Second, we have found no correlation test which allows attribute groups with ORDINAL representations to have more than one attribute. Instead, we downgrade the representation of the combined attribute group to a CATEGORICAL representation. Similarly, an attribute group with an ORDINAL representation is treated as if it had a CATEGORICAL representation when it is combined with attribute groups having INTERVAL representations.

An attribute group with the INTERVAL *representation and an attribute group with the* NOMINAL *representation cannot be automatically combined.* For the two attribute groups to be combined, the attribute group with an INTERVAL representation would have to have its values transformed to values which can be compared using a correlation test for CATEGORICAL, or DICHOTOMOUS representations. It is not possible to perform this transformation automatically.

It is not possible to convert the values of an attribute group with the INTERVAL representation to values acceptable for the DICHOTOMOUS representation, as there are too many distinct values in an attribute group with the INTERVAL representation. While a sample of the instances of an attribute group with the INTERVAL representation does not physically capture all the distinct values, missing values can still be extrapolated from the values present in the sample.

For example, while the sample may not have the value 49, the presence of 49 may still be inferred because the values 50 and 48 exist in the sample. "Dichotomization" destroys the notion of distance between values, and thus the inferred values are lost in the process. This would make the sample unrepresentative.

The conversion of an attribute with the INTERVAL representation to one with the CATEGORICAL representation cannot be automatically performed, as the appropriate categorization system, and the appropriate number of categories are often not apparent from an examination of the data. Arbitrary selection of a categorization method may lead to incorrect categorization. If categorization is performed manually, the combined representation of an attribute group with the INTERVAL representation, and one with the CATEGORICAL representation will be CATEGORICAL.

6 Measuring Relationships between Attribute Groups

Once representations for all attribute groups have been identified, all mutually exclusive attribute groups can then be paired for analysis. Depending on the representation of each attribute group in the pair, one of the tests in Table 4 is selected to measure the correlation of the attribute groups.

Table 4. Tests to Compare Relationships Among Attribute Groups

Repr.	MAN	SAN	MAD	SAD	ORDINAL	CATEGORICAL	DICHOTOMOUS
MAN	Canon. Corr.	Box-Tidwell	Canon. Corr.	Box-Tidwell	MANOVA	MANOVA	Log. Regr.
SAN	Box-Tidwell	Box-Cox	Box-Tidwell	Box-Cox	Spearman	ANOVA	Pt. Biserial
MAD	Canon. Corr.	Box-Tidwell	Canon. Corr.	Pearson Corr.	MANOVA	MANOVA	Log. Regr.
SAD	Box-Tidwell	Box-Cox	Pearson Corr.	Pearson Corr.	Spearman	ANOVA	Pt. Biserial
ORD	MANOVA	Spearman	MANOVA	Spearman	Spearman	Cntgcy. Coeff.	Ordered Log.
CAT	MANOVA	ANOVA	MANOVA	ANOVA	Cntgcy. Coeff.	Cntgcy. Coeff.	Cntgcy. Coeff.
DICH	Log. Regr.	Pt. Biserial	Log. Regr.	Pt. Biserial	Ordered Log.	Cntgcy. Coeff.	Phi Coeff.

Most of the tests described in Table 4 are common and accepted tests for comparing attribute groups with the representations described. They are discussed in most classic statistics textbooks (e.g., [2,9,14]).

7 Evaluating the Correlation Tests

Once an appropriate correlation test for a pair of attribute groups has been determined, the test is performed against the representative sample extracted from the instances of the attribute group pair. The test will yield two values, the correlation coefficient, and the statistical significance. The correlation coefficient measures the degree to which the values of one attribute group predict the values of another. The value of the ANOVA which is comparable to the correlation coefficient is the *F-ratio*, which measures the difference in the distribution of interval values associated with each categorical value.

The *statistical significance* indicates the probability that the relationship being discovered occurred as a result of chance. The smaller the significance value, the more certain we are that the correlation test found a genuine relationship.

A maximum significance threshold, and minimum correlation coefficient threshold for each test can be specified by the mining specialist prior to relationship discovery. Only attribute group pairs which have correlation coefficient values above the coefficient threshold, and significance values below the significance threshold will then be confirmed.

Like other kinds of knowledge discovered by data mining, relationships discovered by correlation tests should not be taken as gospel until they are verified and validated. While setting the significance threshold to a low value would reduce the number of discovered relationships that are spurious, some spurious relationships will still be discovered. Furthermore, setting the significance threshold to an extremely low value will cause the method to reject many genuine relationships, thus rendering the method less effective.

The results of correlation tests can be validated empirically in several ways. A quick, and reliable way would be to first re-sample values from the attribute group pair identified as having a strong correlation, and then run the same correlation test again. If repeated tests on different samples produce high correlation coefficients, and low significance values, then we are more certain that a genuine relationship exists between the two attribute groups. Of course, if time permits, the best validation would be to perform the correlation test on *all* the values of the attribute group pair.

8 Conclusion

In this paper, we propose and discuss a heuristic method for identifying correlation tests to measure the relationship between attribute group pairs. We have also discussed how correlation tests can provide information not only on the strength, but also the significance of a relationship. We are currently attempting to extend our research in several directions.

First, are in the process of standardizing the results of the various correlation tests. The possible results of the various correlation tests vary widely. For example, the Spearman's Rho, Box-Tidwell, and Canonical Correlation tests have scores ranging from -1 to 1. The minimum score of the contingency coefficient is 0, but the maximum score varies according to the sample size. The result of the F-ratio has a minimum value of 1, and a theoretically infinite maximum. We cannot expect that an untrained user will be able to interpret these varied scores.

Second, we are attempting to apply this method to the database integration problem. In developing this method, we have noted the similarity between data mining, and the attribute identification problem in database integration. Attribute identification is the sub-problem of database integration which not only deals with identifying equivalent attributes (i.e. attribute equivalence [12]), but sets of attributes as well. Finding representations for attribute identification is a problem of significantly larger scope than finding representations for data mining, since relationships between attribute groups with String and Key representations must also be accounted for. We are currently investigating the

applicability of an extended version of our correlation test identification method to the attribute identification problem.

Finally, we are looking for ways to validate our method. As it is difficult to mathematically validate heuristic methods, we are attempting to validate the heuristics employed against real world databases. Currently, we are acquiring a large variety of data sets to validate our method against. Results from preliminary tests on small, publicly available data sets (e.g., [4,8]) are encouraging. However, further tests still need to be performed.

References

1. R. Agrawal, T. Imielinski, A. Swami. *Mining Association Rules Between Sets of Items in Large Databases.* Proc. of the ACM SIGMOD Conference on Management of Data. pp. 207-216. 29
2. R.B. Burns. *Introduction to Research Methods - Third Edition.* Addison-Wesley. 1997. 29, 30, 31, 38
3. S. Brin, R. Motwani, C. Silverstein. *Beyond Market Baskets: Generalizing Association Rules to Correlations.* Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. 1997. pp. 265-276. 29
4. T.J. Biblarz, A.E. Raftery. *The Effects of Family Disruption on Social Mobility.* American Sociological Review. 1993. 40
5. B. Everitt. *Cluster Analysis.* Heinemann Educational Books. 1980.
6. J.E. Freund, R.E. Walpole. *Mathematical Statistics- Fourth Edition.* Prentice-Hall. 1987.
7. J.D. Gibbons. *Nonparametric Methods for Quantitative Analysis (Second Edition).* American Sciences Press Inc. 1985.
8. V. Greaney, and T. Kelleghan. *Equality of Opportunity in Irish Schools.* Dublin: Educational Company. 1984. 40
9. J.F. Hair Jr., R.E. Anderson, R.L. Tatham, and W.C. Black. *Multivariate Data Analysis with Readings.* Prentice Hall. 1995. 38
10. D.V. Huntsberger, and P.P. Billingsley. *Elements of Statistical Inference.* Allyn and Bacon Inc. 1987. 33
11. J.S. Long. *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications. 1997.
12. J.A. Larson, S.B. Navathe, R. Elmasri. *A Theory of Attribute Equivalence in Databases with Application to Schema Integration.* IEEE Transactions on Software Engineering. April 1989. pp. 449-463. 39
13. H. Mannila, and K.J. Raiha. *Algorithms for Inferring Functional Dependencies From Relations.* Data and Knowledge Engineering. February, 1994. pp. 83-90. 29
14. J. Neter, W. Wasserman, M.H. Kutner. *Applied Linear Regression Models. 2nd Edition.* Irwin Homewood. 1989. 38
15. P.D. Scott, A.P.M. Coxon, M.H. Hobbs, R.J. Williams. *SNOUT: An Intelligent Assistant for Exploratory Data Analysis.* Principles of Knowledge Discovery and Data Mining. 1997. pp. 189-199. 35, 36
16. G.B. Thomas, and R.L. Finney. *Calculus and Analytic Geometry.* Addison-Wesley. 1996. 30