Assessing the Interestingness of Discovered Knowledge Using a Principled Objective Approach

Robert J. Hilderman
Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
robert.hilderman@uregina.ca

ABSTRACT

When mining a large database, the number of patterns discovered can easily exceed the capabilities of a human user to identify interesting results. To address this problem, various techniques have been suggested to reduce and/or order the patterns prior to presenting them to the user. In this paper, our focus is on ranking summaries generated from a single dataset, where attributes can be generalized in many different ways and to many levels of granularity according to taxonomic hierarchies. We theoretically and empirically evaluate twelve diversity measures used as heuristic measures of interestingness for ranking summaries generated from databases. The twelve diversity measures have previously been utilized in various disciplines, such as information theory, statistics, ecology, and economics. We describe five principles that any measure must satisfy to be considered useful for ranking summaries. Theoretical results show that the proposed principles define a partial order on the ranked summaries in most cases, and in some cases, define a total order. Theoretical results also show that seven of the twelve diversity measures satisfy all of the five principles. We empirically analyze the rank order of the summaries as determined by each of the twelve measures. These empirical results show that the measures tend to rank the less complex summaries as most interesting. Finally, we demonstrate a technique, based upon our principles, for visualizing the relative interestingness of summaries.

Keywords: data mining, diversity measures, theory of interestingness, statistics and probability, visualization

1. INTRODUCTION

An important problem in the area of data mining is the development of effective measures of interestingness for ranking discovered knowledge. In this paper, we focus on the use of diversity measures as heuristic measures of interestingness for ranking summaries generated from a single dataset, where attributes can be generalized in many different ways and to many levels of granularity according to taxonomic hierarchies. With diversity measures, the problem that we are faced with is essentially one of ranking distributions of populations of objects having some distinguishable characteristics. The problem is common to many disciplines, such as species diversity in ecology, income/consumption inequality in economics, linguistic diversity in geography, market penetration in business, genetic differences in biology, and others. The common theme is that of classifying some quantity of objects into well-defined categories according to the aforementioned distinguishable characteristics.

The question that we ultimately ask when comparing two or more populations is whether one of the categorized populations is more or less diverse than another. And the question is similar, regardless of the discipline in which it is asked. For example, in ecology, we ask whether a sample of individuals from a particular habitat is more diverse than a sample taken from a neighboring or similar habitat. In economics, we ask whether a sample of individuals in a particular region has greater equality of income distribution than a sample of individuals in another region. And in linguistics, we ask whether the possibility for communication in a sample of individuals in a geographic region is more likely than in a sample of individuals from another geographic region. The above situations are all a specific case of the general problem that can be described, as follows. Suppose $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$ are two populations of individuals, where x_i and y_j are integers representing the number of individuals classified into X_i and Y_j, respectively. Which of the distributions is more (less) diverse (or depending on the discipline, concentrated or uniform or monopolistic or specialized or dispersed)?

We introduced the use of diversity measures for ranking summaries, as described in the previous paragraph, in [26] and [27], where well-known diversity measures from information theory, statistics, ecology, and economics were proposed as heuristic measures of interestingness. Although diversity measures are frequently used in these various disciplines, their use for ranking the interestingness of summaries was a new application area. An empirical analysis found that highly ranked, concise summaries provided a reasonable starting point for further analysis of discovered knowledge. It was also shown that for selected sample datasets, the order in which some of the measures rank summaries is highly correlated, but the rank ordering can vary substantially when different measures are used. In [28], the

UBDM'06, August 20, 2006, Philadelphia, Pennsylvania, USA.

notion of a summary was extended to include other well-known forms of knowledge representation, and we showed that these other forms are also amenable to ranking using diversity measures.

We now study twelve diversity measures as heuristic measures of interestingness for ranking summaries in data mining applications, and propose five principles that any measure must satisfy to be considered useful for ranking the interestingness of summaries generated from databases. The five principles provide a foundation for an intuitive understanding of the term "interestingness" when used within this context. We perform a comparative sensitivity analysis of the twelve diversity measures to identify those that satisfy the proposed principles. Since each new measure represents an alternative definition of diversity, the choice of which measure to use may make a difference. That is, when choosing any objective candidate interestingness measure for ranking summaries, determine which of the five principles are satisfied, and then using this knowledge, judge the suitability of the candidate interestingness measure for the intended application. Essentially, this principled approach imposes a subjective bias on the objective measures by suggesting principles that objective measures should satisfy.

The remainder of the paper is organized as follows. In Section 2, we motivate the need for objective measures of interestingness in data mining systems, in general, and the need for principles of interestingness, in particular. In Section 3, we describe the twelve diversity measures empirically evaluated as measures of interestingness in this work. In Section 4, we present the foundation principles for a theory of interestingness for diversity measures used to rank summaries generated from a single dataset. In Section 5, we present experimental results from our evaluation of the twelve diversity measures. In Section 6, we demonstrate the application of the principles to the visualization of the relative interestingness of summaries. We conclude in Section 7 with a summary of our work and suggestions for future research.

2. MOTIVATION

In this section, we describe a data mining example where the task is description by summarization, the representation language is generalized relations, and the method for searching is the Multi-Attribute Generalization algorithm [31]. The problem is described, as follows. Let a *summary* S be a relation defined on the columns $\{(A_1, D_1), (A_2, D_2), \ldots, (A_n, D_n)\}$, where each (A_i, D_i) is an attribute-domain pair. Also, let $\{(A_1, v_{i1}), (A_2, v_{i2}), \ldots, (A_n, v_{in})\}$, $i = 1, 2, \ldots, m$, be a set of m unique tuples, where each (A_j, v_{ij}) is an attribute-value pair and e ach v_{ij} is a value from the domain D_j associated with attribute A_j. One attribute A_k is a derived attribute, called *Count*, whose domain D_k is the set of positive integers, and whose value v_{ik} for each attribute-value pair (A_k, v_{ik}) is equal to the number of tuples which have been aggregated from the base relation (i.e., the unconditioned data present in the original database).

A summary, such as the one shown in Table 1, can be generated from a database, such as the one shown in Table 2, using *domain generalization graphs* (DGGs) [30, 33], such as the one shown in Figure 1. For example, the DGG in Figure 1 is associated with the *Office* attribute in the database of Table 2. In Figure 1, the domain for the *Office* attribute is represented by the *Office* node. Increasingly general descriptions of the domain values are represented by the *City*, *Division*, and *ANY* nodes. A user-defined taxonomy in the form of a table is associated with every arc between the nodes in the DGG and describes a generalization relation from one domain to another in a process called *attribute-oriented generalization* (AOG) [20] (other generalization relations besides table lookups are possible, but we restrict our discussion for the sake of simplicity and clarity). The table associated with the arc between the *Office* and *City* nodes defines the mapping of the domain values of the *Office* node to the domain values of the *City* node (e.g., 1 and 2 map to Vancouver, 3 and 4 map to Los Angeles, and 5 to 7 map to New York). The table associated with the arc between the *City* and *Division* nodes can be described similarly. The table associated with the arc between the *Division* and *ANY* nodes maps all values in the *Division* domain to the special value *ANY*. The summary in Table 1 corresponds to the *Division* node of the *Office* DGG, where the corresponding values in the *Quantity* and *Amount* attributes from Table 2 are also aggregated accordingly.

When there are DGGs associated with multiple attributes, then more complex summaries can be generated (known as *multi-attribute generalization*). For example, a DGG for the *Quantity* attribute is shown in Figure 2, where the generalization space consists of three nodes. The set of all possible combinations of domains from the DGGs associated with the *Office* and *Quantity* attributes defines the generalization space for the many summaries that can be generated from Table 2. Thus, the generalization space consists of the 12 nodes shown in Figure 3 (i.e., 4 nodes in the *Office* DGG ×

Table 1: A sample summary

Office	*Quantity*	*Amount*	*Count*
West	8	$200.00	4
East	11	$275.00	3

Table 2: A sales transaction database

Office	*Quantity*	*Amount*
2	2	$50.00
5	3	$75.00
3	1	$25.00
7	4	$100.00
1	3	$75.00
6	4	$100.00
4	2	$50.00

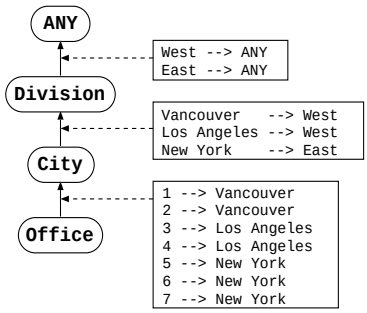

Figure 1: A DGG for the *Office* attribute

3 nodes in the *Quantity* DGG), and each node corresponds to a unique summary. For example, the *Division/Quantity* node corresponds to the summary generated by generalizing the *Office* attribute to the level of the *Division* node in the *Office* DGG, while the *Quantity* attribute remains ungeneralized (this summary is equivalent to the summary in Table 1). Similarly, the *City/Status* node corresponds to the summary shown in Table 3, and is generated by generalizing the *Office* and *Quantity* attributes to the level of the *City* and *Status* nodes, respectively. Naturally, the general technique is applicable to more than two attributes and should now be clear.

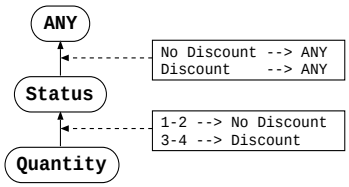

Figure 2: A DGG for the *Quantity* attribute

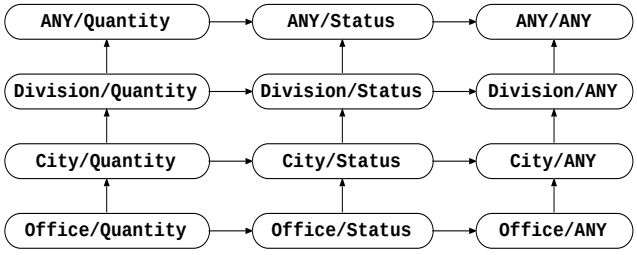

Figure 3: The generalization space defined by the *Office* and *Quantity* DGGs

Table 3: The *City/Status* summary

Office	*Quantity*	*Amount*	*Count*
Los Angeles	No Discount	$75.00	2
New York	Discount	$275.00	3
Vancouver	No Discount	$50.00	1
Vancouver	Discount	$75.00	1

Up to this point, we have only discussed summaries generated from databases using AOG and DGGs. However, alternative methods could be used to guide the generation of summaries. These include Galois lattices [17], conceptual graphs [7], or formal concept analysis [50]. Similarly, summaries could more generally include views generated from databases, characterized/generalized association rules generated from itemsets, or summary tables (i.e., data cubes) generated from data warehouses [28].

3. MEASURING INTERESTINGNESS

The tuples in a summary are unique, and therefore, can be considered to be a population with a structure that can be described by some frequency or probability distribution. Here, we review twelve diversity measures that consider the frequency or probability distribution of the values in the derived *Count* attribute (or some other similar numeric measure attribute) to assign a single real-valued index that represents its interestingness relative to other summaries.

3.1 Background

Diversity is an important concept that has seen extensive use in several different areas of research. However, although diversity is used in many disparate areas, it is widely claimed that diversity is a difficult concept to define [1, 2, 40, 44, 52]. The difficulty in defining diversity arises because it actually encompasses two separate components: the number of *classes* (also referred to in the literature as *richness*, *abundance*, or *density*) and the *proportional distribution* of the population among the classes (also referred to in the literature as *relative abundance*, *heterogeneity*, or *evenness*). Within the context of ranking the interestingness of a summary, the number of classes is simply the number of tuples in the summary; the proportional distribution is simply the actual probability distribution of the classes based upon the values contained in the derived *Count* attribute.

In a typical diversity measure, the two components are combined to characterize the variability of a population by a single value. This concept of a dual-component diversity measure was first introduced in [49]. The diversity measures considered to be most useful, and those most frequently referenced in the literature, are dual-component measures. Yet, despite the widespread acceptance and use of diversity measures, there is no single mathematical definition of diversity which has been widely accepted as the de facto standard and which has been shown to be superior to all others [1, 40, 44]. There is some general agreement, however, that a population is considered to have high diversity when it has many classes and the proportional distribution is fairly even. Similarly, a population is considered to have low diversity when it has few classes and the proportional distribution is uneven. Unfortunately, this leaves considerable room for ambiguity in measuring diversity because a population with few classes and a fairly even proportional distribution could have the same or nearly the same diversity as a population with many classes and an uneven proportional distribution.

Although there are some problems related to a precise and universally accepted definition for diversity, there are numerous research areas where the concept of diversity has been considered useful, such as ecology [1, 8, 9, 41, 42], economics [2, 4, 14, 47], genetics [36], linguistics [18, 37], business [5, 22, 23, 34], epidemiology [35], bibliometrics [45], software engineering [43], and the measurement of scientific productivity [2]. More general treatments attempt to define the concept of diversity and develop a related theory of diversity measurement [44, 52].

Here we apply twelve diversity measures to a new application area, that of ranking the interestingness of summaries generated from databases. They share three important properties. First, each measure depends only on the frequency or probability distribution of the values in the derived *Count* attribute of the summary to which it is being applied. Second, each measure allows a value to be generated with at most one pass through the summary. And third, each measure is independent of any specific units. Utilizing these heuristics for ranking the interestingness of summaries generated from databases is a natural and useful extension for these diversity measures into a new application domain.

3.2 Notation

The variables used to describe the diversity measures are now defined. Let m be the total number of tuples in a summary. Let n_i be the value contained in the derived *Count* attribute for tuple t_i. Let $N = \sum_{i=1}^{m} n_i$ be the total count. Let p be the actual probability distribution of the tuples based upon the values n_i. Let $p_i = n_i/N$ be the actual probability for tuple t_i. Let q be a uniform probability distribution of the tuples. Let $\bar{u} = N/m$ be the count for tuple t_i, $i = 1, 2, \ldots, m$ according to the uniform distribution q. Let $\bar{q} = 1/m$ be the probability for tuple t_i, for all $i = 1, 2, \ldots, m$ according to the uniform distribution q. Let r be the probability distribution obtained by combining the values n_i and $\bar{u}$. Let $r_i = (n_i + \bar{u})/2N$, be the probability for tuples t_i, for all $i = 1, 2, \ldots, m$ according to the distribution r. For example, given the sample summary shown in Table 4, we have $m = 4$, $n_1 = 3$, $n_2 = 2$, $n_3 = 1$, $n_4 = 1$, $N = 7$, $p_1 = 0.429$, $p_2 = 0.286$, $p_3 = 0.143$, $p_4 = 0.143$, $\bar{u} = 1.75$, $\bar{q} = 0.25$, $r_1 = 0.339$, $r_2 = 0.268$, $r_3 = 0.196$, and $r_4 = 0.196$.

Table 4: Another sample summary

Colour	*Shape*	*Count*
red	round	3
green	round	2
red	square	1
blue	square	1

3.3 Diversity Measures

We now describe the twelve diversity measures. Due to space limitations, examples are omitted. The interested reader is encouraged to work examples of each measure based upon the sample summary shown in Table 4.

$\mathbf{I_{Variance}}$: Based upon sample variance from classical statistics, measures the weighted average of the squared deviations of the probabilities p_i from the mean probability $\bar{q}$, where the weight assigned to each squared deviation is $1/(m-1)$. We use sample variance because we assume the summary may not contain all possible combinations of attribute values, meaning we are not observing all of the possible tuples. The *sample variance* is given by

$$I_{Variance} = \frac{\sum_{i=1}^{m}(p_i - \bar{q})^2}{m-1}.$$

$\mathbf{I_{Simpson}}$: A variance-like measure based upon the Simpson index [49], measures the extent to which the counts are distributed over the tuples in a summary, rather than being concentrated in any single one of them. The *concentration* is given by

$$I_{Simpson} = \sum_{i=1}^{m} p_i^2.$$

Let each tuple i be represented by a "commonness value" (i.e., the probability of occurrence p_i). If an individual is drawn at random from the population, the probability that it will belong to tuple i is p_i, and if it does, its commonness value is also p_i. Thus, the expected commonness values for tuple i is p_i^2, and for all tuples $i = 1, \ldots, n$ is $\sum_1^m p_i^2$. Equivalently, this can be viewed as the average commonness value that would be obtained if the experiment of drawing an individual at random were repeated many times.

$\mathbf{I_{Shannon}}$: Based upon a relative entropy measure from information theory (known as the *Shannon index*) [48], measures the average information content in the tuples of a summary. The *average information content*, in bits per tuple, is given by

$$I_{Shannon} = -\sum_{i=1}^{m} p_i \log_2 p_i.$$

Say there are n_i individuals summarized in a tuple i, out of a possible N individuals. The probability of drawing one of the individuals in tuple i is n_i/N, or p_i. The information conveyed by announcing the result of drawing a particular individual in tuple i is $-\log_2 p_i$. The total contribution of these n_i individuals to the overall average information conveyed by announcing the result is $-p_i \log_2 p_i$. Summation over all such cases for all possible individuals is given by $-\sum_{i=1}^{m} p_i \log_2 p_i$.

$\mathbf{I_{McIntosh}}$: Based upon a heterogeneity index from ecology [41], views the counts in a summary as the coordinates of a point in a multidimensional space and measures the modified Euclidean distance from this point to the origin. The *modified Euclidean distance* is given by

$$I_{McIntosh} = \frac{N - \sqrt{\sum_{i=1}^{m} n_i^2}}{N - \sqrt{N}}.$$

The value $\sqrt{\sum_{i=1}^{m} n_i^2}$ is just the Pythagorean Theorem. Since $\sqrt{\sum_{i=1}^{m} n_i^2}$ is a measure of concentration, the N-complement $N - \sqrt{\sum_{i=1}^{m} n_i^2}$ is a measure of diversity. The value $N - \sqrt{N}$ makes it a diversity measure independent of N. The greater the count in a particular class, the further that class will be from the origin. If the count is reduced, or the count is spread more evenly between class, the distance from the origin will be reduced. $I_{McIntosh}$ relates the distance between a class and the origin to the range of possible values as determined by the number of tuples in the original relation.

$\mathbf{I_{Lorenz}}$: Based upon the Lorenz curve from statistics, economics, and social science [51], measures the average value of the Lorenz curve derived from the probabilities p_i associated with the tuples in a summary. The *average value of the Lorenz curve* is given by

$$I_{Lorenz} = \bar{q} \sum_{i=1}^{m} (m - i + 1) p_i.$$

The Lorenz curve is a series of straight lines in a square of unit length, starting from the origin and going successively to points (p_1, q_1), $(p_1 + p_2, q_1 + q_2)$, ... When the p_i's are all equal, the Lorenz curve coincides with the diagonal that cuts the unit square into equal halves. When the p_i's are not all equal, the Lorenz curve is below the diagonal.

$\mathbf{I_{Gini}}$: Based upon the Gini coefficient [51], which is itself defined in terms of the Lorenz curve, measures the ratio of the area between the diagonal (i.e., the line of equality) and the Lorenz curve, and the total area below the diagonal. The *Gini coefficient* is given by

$$I_{Gini} = \frac{\bar{q} \sum_{i=1}^{m} \sum_{j=1}^{m} |p_i - p_j|}{2}.$$

$\mathbf{I_{Berger}}$: Based upon a dominance index from ecology [6], measures the proportional dominance of the tuple in a summary with the highest probability p_i. The *proportional dominance* is given by

$$I_{Berger} = \max(p_i).$$

Say a sample of individuals is taken from some population of species in a particular habitat. The number of individuals taken from each species is assumed to represent the proportional distribution of species in the actual population. I_{Berger} is called a dominance index because the index of diversity that it assigns to the sampled population is simply the proportional distribution of the most dominant species (i.e., the species with the highest proportional distribution).

$\mathbf{I_{Schutz}}$: Based upon an inequality measure from economics and social science [47], measures the relative mean deviation of the actual distribution of the counts in a summary from a uniform distribution of the counts. The *relative mean deviation* is given by

$$I_{Schutz} = \frac{\sum_{i=1}^{m} |p_i - \bar{q}|}{2}.$$

$\mathbf{I_{Bray}}$: Based upon community similarity indices from ecology [8], measures the percentage of similarity between the actual distribution of the counts in a summary and a uniform distribution of the counts. The *percentage of similarity* is given by

$$I_{Bray} = \frac{\sum_{i=1}^{m} \min(n_i, \bar{u})}{N}.$$

$\mathbf{I_{MacArthur}}$: Based upon the Shannon index from information theory [39], combines two summaries and then measures the difference between the amount of information contained in the combined distribution and the amount contained in the average of the two original distributions. The *difference*, in bits, is given by

$$I_{MacArthur} = \left(-\sum_{i=1}^{m} r_i \log_2 r_i \right) - \left(\frac{(-\sum_{i=1}^{m} p_i \log_2 p_i) + \log_2 m}{2} \right).$$

$\mathbf{I_{Theil}}$: Based upon a distance measure from information theory [51], measures the distance between the actual distribution of the counts in a summary and a uniform distribution of the counts. The *distance*, in bits, is given by

$$I_{Theil} = \frac{\sum_{i=1}^{m} |p_i \log_2 p_i - \bar{q} \log_2 \bar{q}|}{m\bar{q}}.$$

$\mathbf{I_{Atkinson}}$: Based upon a measure of inequality from economics [4], measures the percentage to which the population in a summary would have to be increased to achieve the same level of interestingness if the counts in the summary were uniformly distributed. The *percentage increase* is given by

$$I_{Atkinson} = 1 - \left(\prod_{i=1}^{m} \frac{p_i}{\bar{q}} \right)^{\bar{q}}.$$

Lower values of $I_{Atkinson}$ mean that the distribution of counts in a summary are fairly equal, or near uniform. Higher values mean the distribution is fairly uneven. As an example, say $I_{Atkinson} = 0.105$, as shown in the example below. This value means that if the counts of the tuples were uniformly distributed, then we would need only approximately 90% of the current total count to realize the same level of interestingness.

4. PRINCIPLES OF INTERESTINGNESS

We now describe a theory of interestingness against which the utility of candidate interestingness measures can be assessed. We do this through the mathematical formulation of five principles that we believe must be satisfied by any acceptable diversity measure for ranking the interestingness of summaries generated from databases using our, or a similar, technique. Through the development of these five principles, we have established some basic criteria for the measurement of interestingness within this context which provide the basis for a theoretical foundation in identifying appropriate diversity measures for ranking summaries.

Through the mathematical formulation of the five principles, we study functions f of m variables, $f(n_1, \ldots, n_m)$, where f denotes a general measure of diversity, m and each n_i are as defined in the previous section, and $(n_1, \ldots, n_m)$ is a vector corresponding to the values in the derived *Count* attribute (or numeric measure attribute) for some arbitrary summary whose values are arranged in descending order such that $n_1 \geq \ldots \geq n_m$ (except for discussions regarding I_{Lorenz}, which requires that the values be arranged in ascending order). Since the principles presented here are for ranking the interestingness of summaries generated from a single dataset, we assume that N is fixed. We begin by specifying two fundamental principles.

Minimum Value Principle (P1). Given a vector $(n_1, \ldots, n_m)$, where $n_i = n_j$, for all i, j, $f(n_1, \ldots, n_m)$ attains its minimum value.

P1 specifies that the minimum interestingness should be attained when the tuple counts are all equal (i.e., uniformly distributed). For example, given the vectors $(2, 2)$, $(50, 50, 50)$, and $(1000, 1000, 1000, 1000)$, we require that the index value generated by f be the minimum possible for the respective values of m and N.

Maximum Value Principle (P2). Given a vector $(n_1, \ldots, n_m)$, where $n_1 = N - m + 1$, $n_i = 1$, $i = 2, \ldots, m$, and $N > m$, $f(n_1, \ldots, n_m)$ attains its maximum value.

P2 specifies that the maximum interestingness should be attained when the tuple counts are distributed as unevenly as possible. For example, given the vectors $(3, 1)$, $(148, 1, 1)$, and $(3997, 1, 1, 1)$, where $m = 2, 3,$ and 4, respectively, and $N = 4, 150,$ and 4000, respectively, we require that the index value generated by f be the maximum possible for the respective values of m and N.

The behaviour of a measure relative to satisfying both *P1* and *P2* is significant because it reveals an important characteristic about its fundamental nature as a measure of diversity. A measure of diversity can generally be considered either a *measure of concentration* or a *measure of dispersion*. A measure of concentration can be viewed as the opposite of a measure of dispersion, and we can convert one to the other via simple transformations. For example, if g corre-

sponds to a measure of dispersion, then we can convert it to a measure of concentration f, where $f = \max(g) - g$. Here we only consider measures of concentration. A measure was considered to be a measure of concentration if it satisfied P1 and P2 without transformation. A measure was considered to be a measure of dispersion if it satisfied P1 and P2 following transformation. All measures of dispersion were transformed into measures of concentration prior to our analysis.

Permutation Invariance Principle (P3). Given a vector $(n_1, \ldots, n_m)$ and any permutation $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$, $f(n_1, \ldots, n_m) = f(n_{i_1}, \ldots, n_{i_m})$.

P3 specifies that every permutation of a given distribution of tuple counts should be equally interesting. That is, interestingness is not a labeled property, it is only determined by the distribution of the counts. For example, given the vector $(2, 4, 6)$, we require that $f(2, 4, 6) = f(4, 2, 6) = f(4, 6, 2) = f(2, 6, 4) = f(6, 2, 4) = f(6, 4, 2)$.

Transfer Principle (P4). Given a vector $(n_1, \ldots, n_m)$, $n_i \geq n_j$, $i < j$, and $0 < c \leq n_j$, $f(n_1, \ldots, n_i + c, \ldots, n_j - c, \ldots, n_m) > f(n_1, \ldots, n_i, \ldots, n_j, \ldots, n_m)$.

P4, adopted from [14], specifies that when a strictly positive transfer is made from the count of one tuple to another tuple whose count is greater, then interestingness increases. For example, given the vectors $(10, 7, 5, 4)$ and $(10, 9, 5, 2)$, we require that $f(10, 9, 5, 2) > f(10, 7, 5, 4)$.

Majorization Principle (P5). Given vectors $(n_1, \ldots, n_m)$ and $(n_1^{'}, \ldots, n_m^{'})$, whenever $f(n_1^{'}, \ldots, n_m^{'}) > f(n_1, \ldots, n_m)$, then $(n_1^{'}, \ldots, n_m^{'}) \succ (n_1, \ldots, n_m)$, read $(n_1^{'}, \ldots, n_m^{'})$ *majorizes* $(n_1, \ldots, n_m)$.

The majorization operator, $\succ$, is based upon the Lorenz dominance order. The *Lorenze dominance order* [21] compares vectors with different distributions and says for any two vectors $(n_1, \ldots, n_m)$ and $(n_1^{'}, \ldots, n_m^{'})$, that $(n_1^{'}, \ldots, n_m^{'}) \succ (n_1, \ldots, n_m)$ if the following fours conditions are true:

1. $n_1 \geq \ldots \geq n_m$.
2. $n_1^{'} \geq \ldots \geq n_m^{'}$.
3. $\sum_{i=1}^{j} n_i^{'} \geq \sum_{i=1}^{j} n_i$, for every $j = 1, \ldots, m$.
4. $\sum_{i=1}^{m} n_i^{'} = \sum_{i=1}^{m} n_i$.

An important property of the Lorenz dominance order is that it defines a partial order on the set of all possible vectors, a property useful and important for ranking summaries.

Those measures that satisfy the principles of interestingness are shown in Table 5. In Table 5, the *P1* to *P5* columns describe the proposed principles, and a measure that satisfies a principle is indicated by the *bullet* symbol (i.e., •).

Mathematical proofs were derived for each measure satisfying principles P1 to P5 in Table 5. However, due to space limitations, the proofs are omitted. The interested reader is referred to [29] for the complete proofs.

Table 5: Measures satisfying the proposed principles

Measure	*P1*	*P2*	*P3*	*P4*	*P5*
$I_{Variance}$	•	•	•	•	•
$I_{Simpson}$	•	•	•	•	•
$I_{Shannon}$	•	•	•	•	•
$I_{McIntosh}$	•	•	•	•	•
I_{Lorenz}	•	•		•	•
I_{Gini}	•	•	•	•	•
I_{Berger}	•	•	•		
I_{Schutz}	•	•	•		
I_{Bray}	•	•	•		
$I_{MacArthur}$	•	•	•	•	•
I_{Theil}	•		•		
$I_{Atkinson}$	•	•	•	•	•

5. EXPERIMENTAL RESULTS

A series of experiments were run using *DGG-Interest*, an extension to *DB-Discover*, a research data mining tool developed at the University of Regina [12]. *DB-Discover* generates summaries from databases according to DGGs associated with attributes and the AOG technique described in Section 2. *DGG-Interest* evaluates and ranks the summaries generated using the twelve diversity measures described in Section 3.

Input data for the experiments was supplied by the NSERC Research Awards database, freely available in the public domain, and the Customer Accounts database, a confidential database provided by a commercial research partner in the telecommunications industry. The NSERC Research Awards database contains records of Canadian government funding provided to academic and industrial researchers in the natural sciences and engineering, and has been frequently used in previous data mining research [10, 11, 19, 38]. It consists of 10,000 tuples in six tables describing a total of 22 attributes. The Customer Accounts database has also been frequently used in previous data mining research [13, 24, 32, 25]. It consists of over 8,000,000 tuples in 22 tables describing a total of 56 attributes. The largest table contains over 3,300,000 tuples representing the account activity for over 500,000 customer accounts and over 2,200 products and services. In the discovery tasks run against the NSERC database, from two to four attributes were selected for discovery, and in those run against the Customer Accounts database, from two to five attributes were selected. We refer to the NSERC discovery tasks containing two, three, and four attributes as as *N-2*, *N-3*, and *N-4*, respectively, and the Customer Accounts discovery tasks containing two, three, four, and five attributes as *C-2*, *C-3*, *C-4*, and *C-5*, respectively.

Within the context of discovery tasks that generate summaries, the discovery tasks run against the Customer Accounts database are considered large. For example, the characteristics of the DGG's associated with each attribute are shown in Table 6. In Table 6, the *No. of Paths* column describes the number of unique paths through the DGG, the *No. of Nodes* column describes the number of nodes in the DGG, and the *Avg. Path Length* describes the average path length of the unique paths. The number of summaries to be generated by a discovery task (i.e., the size of the generalization space) is determined by multiplying the values in the *No. of Nodes* column. For example, *C-5* selected attributes *C*, *D*, *E*, *F*, and *G*, generating a generalization space containing 102,816 nodes (i.e., $12 \times 17 \times 8 \times 3 \times 21$). Many of these nodes correspond to summaries that are duplicates (i.e., the count vectors are identical). Dulplicates can either occur by chance, or when the generalization of an attribute to a higher node in the associated DGG does not result in any tuples being aggregated, and this can occur quite frequently.

Since the diversity measures used to rank the vectors cannot differentiate these summaries, they are considered to be of equal interest. Consequently, the number of summaries (i.e., count vectors) actually ranked is considerably less in practice. For example, of the 102,816 summaries generated by *C-5*, there were only 493 unique vectors, but the entire generalization space still needs to be traversed to find them.

Table 6: Characteristics of the DGGs associated with the selected attributes

Attribute	*No. of Paths*	*No. of Nodes*	*Avg. Path Length*
A	5	20	4.0
B	4	17	4.3
C	3	12	4.0
D	4	17	4.3
E	2	8	4.0
F	1	3	3.0
G	5	21	4.2

We now discuss the complexity of the summaries ranked by the various measures. In analyzing a summary, whether it be a two-dimensional spreadsheet or a multi-dimensional data cube, one metric that determines how easily the information it contains may be to understand by a domain expert, is simply its physical size in terms of the number of cells (i.e., where a cell is commonly understood to be the piece of information referenced by a unique combination of labels corresponding to the data items associated with each dimension). So, for this analysis, we define the *complexity* of a summary as simply the product of the number of tuples and the number of non-ANY attributes contained in the summary. One could ask, then, why use diversity measures to rank summaries at all if complexity, as defined, is a suitable metric for comparing summaries? That is, why not simply rank the least complex summaries as most interesting? The answer to this, of course, is that complexity ignores both the number of tuples in a summary and the the proportional distribution of the tuples, while diversity measures do not. However, complexity is still a useful measure because it is easy to understand at an intuitive level, and a good indicator of the amount of information contained in a summary.

Now, in a previous study, domain experts suggested that more information is better than less, provided that the most interesting summaries are not too complex and remain relatively easy to understand [19]. This implies that useful summaries are those that are complex enough to inform, yet not so complex as to overwhelm. That is, the knowledge contained in a summary should be non-trivial, yet understandable with reasonable effort by the domain expert. Consequently, we believe a desirable property of any ranking function be that it tend to rank summaries with low complexity as more interesting. However, although we want to rank summaries with low complexity as more interesting, we do not want to lose the meaning or context of the data by presenting summaries that are either too complex or too simple.

In this experiment, we analyze the measures and evaluate whether they satisfy the complexity guidelines of our domain experts. The relative complexity of summaries ranked by each measure when grouped according to a three-tier scale of relative complexity (i.e., H=High, M=Moderate, L=Low). High, moderate, and low complexity summaries were considered to be the top, middle, and bottom 20%, respectively, of summaries as ranked by each measure. The *N-2*, *N-3*, and *N-4* discovery tasks generated sets containing 22, 70, and 214 summaries, respectively, while the *C-2*, *C-3*, *C-4*, and *C-5* discovery tasks generated sets containing 43, 91, 155, and 493 summaries, respectively. Thus, the complexity of the summaries from the *N-2*, *N-3*, and *N-4* discovery tasks is based upon the top four, 14, and 43 summaries, respectively, while the complexity of the summaries from the *C-2*, *C-3*, *C-4*, and *C-5* discovery tasks is based upon nine, 18, 31, and 97 summaries, respectively.

A graphical comparison of the complexity of the summaries ranked by the twelve measures from the *N-2*, *N-3*, and *N-4* discovery tasks and the *C-2*, *C-3*, *C-4*, and *C-5* discovery tasks is shown in the graphs of Figures 4 and 5, respectively. In Figures 4 and 5, the horizontal and vertical axes describe the measures and the complexity, respectively. Each horizontal row of bars corresponds to the complexity of the most interesting summaries from a particular discovery task. The backmost horizontal row of bars corresponds to the average complexity for a particular measure. Both figures show a maximum complexity on the vertical axes of 60.0, although the complexity of the most interesting summaries ranked by I_{Lorenz}, I_{Schutz}, I_{Bray}, $I_{MacArthur}$, and $I_{Atkinson}$ in *N-4* exceed this value (i.e., 133.6, 289.8, 289.8, 249.5, and 531.1, respectively). When the measures are ordered by complexity, from lowest to highest, they are ordered according to Figure 4, as follows (position in the ordering is shown in parentheses): I_{Max} (1), I_{Total} (2), I_{Gini} (3), $I_{Shannon}$ and $I_{Kullback}$ (4), I_{Theil} (5), $I_{Variance}$ (6), $I_{Simpson}$ and $I_{McIntosh}$ (7), I_{Berger} (8), I_{Lorenz} (9), $I_{MacArthur}$ (10), I_{Schutz} and I_{Bray} (11), and $I_{Atkinson}$ (12). They are ordered according to Figure 5, as follows: I_{Total} (1), I_{Max} (2), I_{Berger} (3), $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, $I_{McIntosh}$, and I_{Gini} (4), $I_{Kullback}$ (5), I_{Lorenz} (6), $I_{MacArthur}$ (7), $I_{Atkinson}$ (8), I_{Schutz} and I_{Bray} (9).

6. VISUALIZING INTERESTINGNESS

We now demonstrate the application of the five principles of Section 4 to the ranking of summaries. Here we use the results generated by the *N-3* discovery task described in Section 5 as the basis for the extended example as these are representative of the results obtained for all discovery tasks.

An important implication of P5 is that if $X \succ Y$, then all measures satisfying P5 will order the vectors X and Y in the same way. However, it is important to note that even when two measures order the vectors X and Y in the same way, they may not agree on the extent to which X is more concentrated than Y due to the differing range and distribution of the possible values, as described in Section 5. Consequently, the results we discuss here are valid for all of $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, $I_{McIntosh}$, I_{Gini}, $I_{MacArthur}$, and $I_{Atkinson}$.

For this example, we used an extension of *DGG-Interest* to prune the number of summaries generated by *N-3* from 70 down to 27. This extension of *DGG-Interest* utilizes the chi-squared test for independence to consider only those summaries in which the attributes are associated. These summaries and their Lorenz dominance order are shown in Table 7. In Table 7, the *ID* column describes the unique identifiers associated with each of the 27 summaries, the numbered columns describe those summaries that are majorized by the

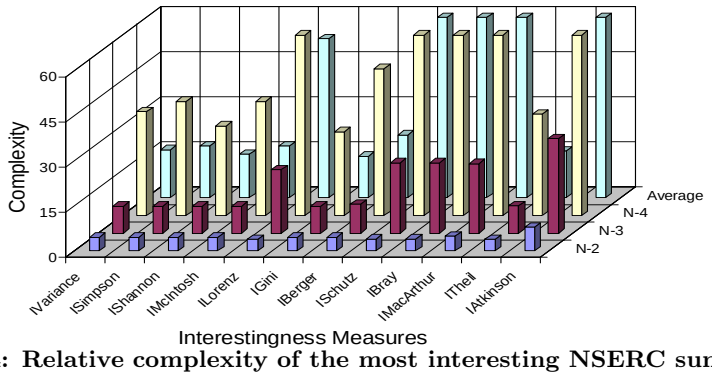

Figure 4: Relative complexity of the most interesting NSERC summaries

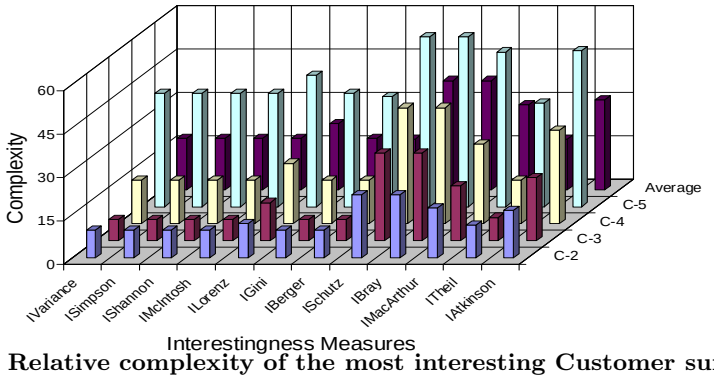

Figure 5: Relative complexity of the most interesting Customer summaries

corresponding summary in the *ID* column, and a summary that is majorized is indicated by the *bullet* symbol (i.e., •). Summaries whose count vectors were identical (i.e., identical number of tuples and identical probability distributions) are grouped together and treated as a single summary for this analysis (because if vector $X = Y$, then $X \succ Y$ and $Y \succ X$, so the vectors are indistinguishable according to the Lorenz dominance order). For example, in the second row, it is shown that summary 8 majorizes 11, 12, 33, 34, 80, 83, and 84 (equivalently $8 \succ \{11, 12, 33, 34, 80, 83, 84\}$). Since we consider majorization to be equivalent to interestingness, then essentially we consider summary 8 to be more interesting than 11, 12, 33, 34, 80, 83, and 84. Summaries 33, 34, and 84 are examples of summaries that do not majorize any other summaries.

Taking advantage of the transitive property of the Lorenz dominance order, we can discover all of the majorization relationships described in Table 7. For example, consider summary 7 in the first row. We see that $7 \succ 8$. Moving to the row beginning with summary 8, we see that $8 \succ 11$. Moving to the row beginning with summary 11, we see that $11 \succ 12$. Moving to the row beginning with summary 12, we see that $12 \succ 84$. Moving to the row beginning with summary 84, we see that 84 does not majorize any other summary. Thus, we can summarize the discovered relationship as the partial order $7 \succ 8 \succ 11 \succ 12 \succ 84$. Note that although we know from the first row that $7 \succ \{8, 11, 12, 84\}$, the first row does not tell us anything about the relationships between 8, 11, 12, and 84. We had to examine the rows corresponding to 8, 11, 12, and 84 to discover these relationships.

Table 7 actually describes 33332 possible partial orders. Using another extension to *DGG-Interest*, 96 rules were generated for consolidating these partial orders into the concise graph of Figure 6. In Figure 6, the majorization relationship of the 27 summaries can be easily determined. The shaded nodes with a bold border indicate summaries that are not majorized by any others, and are start points for traversing the graph. For example, starting at node 17/18, we can follow a path that includes nodes 7, 21/22, 11, and 33/34. Node 33/34 is a shaded node without a bold border, and indicates a stop point (i.e., 33/34 majorizes no other summaries). Similarly, starting at node 17/18, we can follow a path that includes 16, 79, 8, 80, and 33/34. Note that while summary 17/18 majorizes both summaries 7 and 16, there is no path between 16 and 7, so we cannot say anything definitive about the relative interestingness of these two summaries. However, we do know that 17/18 is more interesting than both 16 and 7.

7. CONCLUSION

The use of diversity measures for ranking the interestingness of summaries generated from databases is a new application area. Here we described twelve diversity measures used as heuristic measures of interestingness, and proposed five principles that diversity measures must satisfy to be considered useful for ranking summaries generated from a single dataset. Theoretical results show that seven measures satisfy all of the proposed principles. These include $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, $I_{McIntosh}$, I_{Gini}, $I_{MacArthur}$, and $I_{Atkinson}$. The five remaining measures did not perform as well, failing to satisfy at least one of the proposed prin-

Table 7: Summaries and their Lorenz dominance order

ID	7	8	11	12	16	17/18	21/22	27	28	29/30	33/34	52	53/54	57/58	79	80	83	84	99	100	123
7		•	•	•			•		•	•	•			•	•	•	•	•	•	•	
8			•	•							•					•	•	•			
11				•							•						•	•			
12																		•			
16		•	•	•			•		•	•	•	•		•	•	•	•	•	•	•	
17/18	•	•	•	•	•		•		•	•	•	•	•	•	•	•	•	•	•	•	
21/22			•	•							•			•			•	•			
27		•	•	•					•	•	•			•		•	•	•	•	•	
28				•							•							•		•	
29/30				•					•		•						•	•		•	
33/34																					
52		•	•	•					•	•	•			•		•	•	•	•	•	
53/54		•	•	•			•		•	•	•	•		•	•	•	•	•	•	•	
57/58				•							•						•	•			
79		•	•	•							•					•	•	•			
80				•							•						•	•			
83				•														•			
84																					
99				•					•		•						•	•		•	
100											•							•			
123		•	•	•			•		•	•	•	•		•	•	•	•	•	•	•	

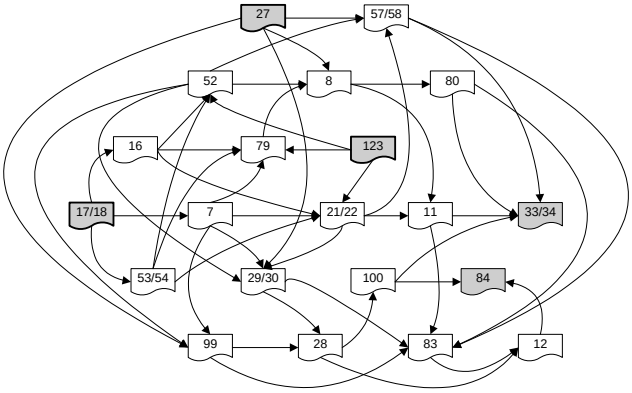

Figure 6: A graph summarizing the Lorenz dominance order

ciples. Experimental results showed that the partial order described by the Lorenze dominance order can be used to generate a graph summarizing the relative interestingness of summaries.

Considerable research remains to be done in the application of diversity measures to the problem of ranking the interestingness of summaries generated from databases. We see two major areas for future research. First, other diversity measures need to be evaluated to determine their suitability for ranking the interestingness of summaries generated from databases. There is certainly no shortage of possible candidates in the literature [45, 46, 16, 3, 15, 15]. And finally, principles of interestingness for comparing summaries generated from different databases need to be developed and evaluated.

8. REFERENCES

[1] R.V. Alatalo. Problems in the measurement of evenness in ecology. *Oikos*, 37(2):199–204, 1981.

[2] P.D. Allison. Measures of inequality. *American Sociological Review*, 43:865–880, 1978.

[3] P.D. Allison. Inequality and scientific productivity. *Social Studies of Science*, 10:163–179, 1980.

[4] A.B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.

[5] M. Attaran and M. Zwick. An information theory approach to measuring industrial diversification. *Journal of Economic Studies*, 16:19–30, 1989.

[6] W.H. Berger and F.L. Parker. Diversity of planktonic forminifera in deep-sea sediments. *Science*, 168:1345–1347, 1970.

[7] I. Bournaud and J.-G. Ganascia. Accounting for domain knowledge in the construction of a generalization space. In *Proceedings of the Third International Conference on Conceptual Structures*, pages 446–459. Springer-Verlag, August 1997.

[8] J.R. Bray and J.T. Curtis. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27:325–349, 1957.

[9] L. Bulla. An index of evenness and its associated diversity measure. *Oikos*, 70(1):167–171, 1994.

[10] C.L. Carter and H.J. Hamilton. Fast, incremental generalization and regeneralization for knowledge discovery from databases. In *Proceedings of the 8th Florida Artificial Intelligence Symposium*, pages 319–323, Melbourne, Florida, April 1995.

[11] C.L. Carter and H.J. Hamilton. Performance evaluation of attribute-oriented algorithms for knowledge discovery from databases. In *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence (ICTAI'95)*, pages 486–489, Washington, D.C., November 1995.

[12] C.L. Carter and H.J. Hamilton. Efficient attribute-oriented algorithms for knowledge discovery from large databases. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):193–208, March/April 1998.

[13] C.L. Carter, H.J. Hamilton, and N. Cercone. Share-based measures for itemsets. In J. Komorowski and J. Zytkow, editors, *Proceedings of the First European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'97)*, pages 14–24, Trondheim, Norway, June 1997.

[14] H. Dalton. The measurement of the inequality of incomes. *Economic Journal*, 30:348–361, 1920.

[15] L. Egghe and R. Rousseau. Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science*, 42(7):479–489, 1991.

[16] J. Gaston. *The reward system in British and American science*. Wiley and Sons, 1978.

[17] R. Godin, R. Missaoui, and H. Alaoui. Incremental concept formation algorithms based on galois (concept) lattices. *Computational Intelligence*, 11(2):246–267, 1995.

[18] J.H. Greenberg. The measurement of linguistic diveristy. *Language*, 32:109–115, 1956.

[19] H.J. Hamilton and D.F. Fudger. Estimating DBLearn's potential for knowledge discovery in databases.

Computational Intelligence, 11(2):280–296, 1995.

[20] J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(1):29–40, February 1993.

[21] G. Hardy, J.E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, 1952.

[22] P.E. Hart. Entropy and other measures of concentration. *Journal of the Royal Statistical Society, Series A*, 134:73–85, 1971.

[23] J.L. Hexter and J.W. Snow. An entropy measure of relative aggregate concentration. *Southern Economic Journal*, 36:239–243, 1970.

[24] R.J. Hilderman, C.L. Carter, H.J. Hamilton, and N. Cercone. Mining association rules from market basket data using share measures and characterized itemsets. *International Journal on Artificial Intelligence Tools*, 7(2):189–220, June 1998.

[25] R.J. Hilderman, C.L. Carter, H.J. Hamilton, and N. Cercone. Mining market basket data using share measures and characterized itemsets. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 159–173, Melbourne, Australia, April 1998.

[26] R.J. Hilderman and H.J. Hamilton. Heuristic measures of interestingness. In J. Zytkow and J. Rauch, editors, *Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*, pages 232–241, Prague, Czech Republic, September 1999.

[27] R.J. Hilderman and H.J. Hamilton. Heuristics for ranking the interestingness of discovered knowledge. In N. Zhong and L. Zhou, editors, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, pages 204–209, Beijing, China, April 1999.

[28] R.J. Hilderman and H.J. Hamilton. Applying objective interestingness measures in data mining systems. In D.A. Zighed and J. Komorowski, editors, *Proceedings of the 4th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pages 432–439, Lyon, France, September 2000.

[29] R.J. Hilderman and H.J. Hamilton. Principles for mining summaries: Theorems and proofs. Technical Report CS 00-01, Department of Computer Science, University of Regina, February 2000. Online at `www.cs.uregina.ca/research/Techreport/0001.ps`.

[30] R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interest*. Kluwer Academic Publishers, 2001.

[31] R.J. Hilderman, H.J. Hamilton, and N. Cercone. Data mining in large databases using domain generalization graphs. *Journal of Intelligent Information Systems*, 13(3):195–234, November 1999.

[32] R.J. Hilderman, H.J. Hamilton, R.J. Kowalchuk, and N. Cercone. Parallel knowledge discovery using domain generalization graphs. In J. Komorowski and J. Zytkow, editors, *Proceedings of the First European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'97)*, pages 25–35, Trondheim, Norway, June 1997.

[33] R.J. Hilderman, Liangchun Li, and H.J. Hamilton. Visualizing data mining results with domain generalization graphs. In U. Fayyad, G.G. Grinstein, and A. Wierse, editors, *Information Visualization in Data Mining and Knowledge Discovery*, pages 251–270. Morgan Kaufmann Publishers, 2002.

[34] A. Horowitz and I. Horowitz. Entropy Markov processes and competition in the brewing industry. *Journal of Industrial Economics*, 16:196–211, 1968.

[35] J. Iszak. Sensitivity profiles of diveristy indices. *Biometrical Journal*, 38(8):921–930, 1996.

[36] R.C. Lewontin. The apportionment of human diversity. *Evolutionary Biology*, 6:381–398, 1972.

[37] S. Lieberson. An extension of Greenberg's linguistic diversity measures. *Language*, 40:526–531, 1964.

[38] H. Liu, H. Lu, and J. Yao. Identifying relevant databases for multidatabase mining. In X. Wu, R. Kotagiri, and K. Korb, editors, *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 210–221, Melbourne, Australia, April 1998.

[39] R.H. MacArthur. Patterns of species diversity. *Biological Review*, 40:510–533, 1965.

[40] A.E. Magurran. *Ecological diversity and its measurement*. Princeton University Press, 1988.

[41] R.P. McIntosh. An index of diversity and the relation of certain concepts to diveristy. *Ecology*, 48(3):392–404, 1967.

[42] J. Molinari. A calibrated index for the measurement of evenness. *Oikos*, 56(3):319–326, 1989.

[43] D. Partridge and W. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 39:707–717, 1997.

[44] G.P. Patil and C. Taillie. Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–567, 1982.

[45] A.D. Pratt. A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, 28:285–292, 1977.

[46] J.L. Ray and J.D. Singer. Measuring the concentration of power in the international system. *Sociological Methods and Research*, 1:403–437, 1973.

[47] R.R. Schutz. On the measurement of income inequality. *American Economic Review*, 41:107–122, March 1951.

[48] C.E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949.

[49] E.H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

[50] G. Stumme, R. Wille, and U. Wille. Conceptual knowledge discovery in databases using formal concept analysis methods. In J. Zytkow and M. Quafafou, editors, *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98)*, pages 450–458, Nantes, France, September 1998.

[51] H. Theil. *Economics and information theory*. Rand McNally, 1970.

[52] M.L. Weitzman. On diversity. *The Quarterly Journal of Economics*, pages 363–405, May 1992.