Heuristic Measures of Interestingness

Robert J. Hilderman and Howard J. Hamilton

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{hilder,hamilton}@cs.uregina.ca

Abstract. The tuples in a generalized relation (i.e., a summary generated from a database) are unique, and therefore, can be considered to be a population with a structure that can be described by some probability distribution. In this paper, we present and empirically compare sixteen heuristic measures that evaluate the structure of a summary to assign a single real-valued index that represents its interestingness relative to other summaries generated from the same database. The heuristics are based upon well-known measures of diversity, dispersion, dominance, and inequality used in several areas of the physical, social, ecological, management, information, and computer sciences. Their use for ranking summaries generated from databases is a new application area. All sixteen heuristics rank less complex summaries (i.e., those with few tuples and/or few non-ANY attributes) as most interesting. We demonstrate that for sample data sets, the order in which some of the measures rank summaries is highly correlated.

1 Introduction

Techniques for determining the interestingness of discovered knowledge have previously received some attention in the literature. For example, in [5], a measure is proposed that determines the interestingness (called surprise there) of discovered knowledge via the explicit detection of Simpson's paradox. Also, in [22], information-theoretic measures for evaluating the importance of attributes are described. And in previous work, we proposed and evaluated four heuristics, based upon measures from information theory and statistics, for ranking the interestingness of summaries generated from databases [8,9].

Ranking summaries generated from databases is useful in the context of descriptive data mining tasks where a single data set can be generalized in many different ways and to many levels of granularity. Our approach to generating summaries is based upon a data structure called a *domain generalization graph* (DGG) [7,10]. A DGG for an attribute is a directed graph where each node represents a domain of values created by partitioning the original domain for the attribute, and each edge represents a generalization relation between these domains. Given a set of DGGs corresponding to a set of attributes, a *generalization space* can be defined as all possible combinations of domains, where one

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 232–241, 1999.

domain is selected from each DGG for each combination. This generalization space describes, then, all possible summaries consistent with the DGGs that can be generated from the selected attributes. When the number of attributes to be generalized is large or the DGGs associated with the attributes are complex, the generalization space can be very large, resulting in the generation of many summaries. If the user must manually evaluate each summary to determine whether it contains an interesting result, inefficiency results. Thus, techniques are needed to assist the user in identifying the most interesting summaries.

In this paper, we introduce and evaluate twelve new heuristics based upon measures from economics, ecology, and information theory, in addition to the four previously mentioned in [8] and [9], and present additional experimental results describing the behaviour of these heuristics when used to rank the interestingness of summaries. Together, we refer to these sixteen measures as the *HMI set* (i.e., heuristic measures of interestingness).

Although our measures were developed and utilized for ranking the interestingness of generalized relations using DGGs, they are more generally applicable to other problem domains. For example, alternative methods could be used to guide the generation of summaries, such as Galois lattices [6], conceptual graphs [3], or formal concept analysis [19]. Also, summaries could more generally include views generated from databases or summary tables generated from data cubes. However, we do not dwell here on the methods or technical aspects of deriving summaries, views, or summary tables. Instead, we simply refer collectively to these objects as summaries, and assume that some collection of them is available for ranking.

The heuristics in the HMI set were chosen for evaluation because they are well-known measures of diversity, dispersion, dominance, and inequality that have previously been successfully applied in several areas of the physical, social, ecological, management, information, and computer sciences. They share three important properties. First, each heuristic depends only on the probability distribution of the data to which it is being applied. Second, each heuristic allows a value to be generated with at most one pass through the data. And third, each heuristic is independent of any specific units of measure. Since the tuples in a summary are unique, they can be considered to be a population with a structure that can be described by some probability distribution. Thus, utilizing the heuristics in the HMI set for ranking the interestingness of summaries generated from databases is a natural and useful extension into a new application domain.

2 The HMI Set

A number of variables will be used in describing the HMI set, which we define as follows. Let m be the total number of tuples in a summary. Let n_i be the value contained in the *Count* attribute for tuple t_i (all summaries contain a derived attribute called *Count*; see [8] or [9] for more details). 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. So, given the sample summary shown in Table 1, for example, we have $m = 4$, $n_1 = 3$, $n_2 = 1$, $n_3 = 1$, $n_4 = 2$, $N = 7$, $p_1 = 0.429$, $p_2 = 0.143$, $p_3 = 0.143$, $p_4 = 0.286$, $\bar{u} = 1.75$, $\bar{q} = 0.25$, $r_1 = 0.339$, $r_2 = 0.196$, $r_3 = 0.196$, and $r_4 = 0.268$.

Table 1. A sample summary

Tuple ID	*Colour*	*Shape*	*Count*
t_1	red	round	3
t_2	red	square	1
t_3	blue	square	1
t_4	green	round	2

We now describe the sixteen heuristics in the HMI set. Examples showing the calculation of each heuristic are not provided due to space limitations.

***$I_{Variance}$*.** Based upon sample variance from classical statistics [15], $I_{Variance}$ 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 $1/(m-1)$.

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

***$I_{Simpson}$*.** A variance-like measure based upon the Simpson index [18], $I_{Simpson}$ 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.

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

***$I_{Shannon}$*.** Based upon a relative entropy measure from information theory (known as the *Shannon index*) [17], $I_{Shannon}$ measures the average information content in the tuples of a summary.

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

***I_{Total}*.** Based upon the Shannon index from information theory [23], I_{Total} measures the total information content in a summary.

$$I_{Total} = m * I_{Shannon}$$

***I_{Max}*.** Based upon the Shannon index from information theory [23], I_{Max} measures the maximum possible information content in a summary.

$$I_{Max} = \log_2 m$$

***$I_{McIntosh}$*.** Based upon a heterogeneity index from ecology [14], $I_{McIntosh}$ 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.

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

***I_{Lorenz}*.** Based upon the Lorenz curve from statistics, economics, and social science [20], I_{Lorenz} measures the average value of the Lorenz curve derived from the probabilities p_i associated with the tuples in a summary. 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), $(p_1 + p_2, q_1 + q_2)$, $\ldots$. 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.

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

***I_{Gini}*.** Based upon the Gini coefficient [20] which is defined in terms of the Lorenz curve, I_{Gini} 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.

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

***I_{Berger}*.** Based upon a dominance index from ecology [2], I_{Berger} measures the proportional dominance of the tuple in a summary with the highest probability p_i.

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

***I_{Schutz}*.** Based upon an inequality measure from economics and social science [16], I_{Schutz} measures the relative mean deviation of the actual distribution of the counts in a summary from a uniform distribution of the counts.

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

***I_{Bray}*.** Based upon a community similarity index from ecology [4], I_{Bray} measures the percentage of similarity between the actual distribution of the counts in a summary and a uniform distribution of the counts.

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

***$I_{Whittaker}$*.** Based upon a community similarity index from ecology [21], $I_{Whittaker}$ measures the percentage of similarity between the actual distribution of the counts in a summary and a uniform distribution of the counts.

$$I_{Whittaker} = 1 - \left(0.5 \sum_{i=1}^{m} |p_i - \bar{q}|\right)$$

$\boldsymbol{I_{Kullback}}$. Based upon a distance measure from information theory [11], $I_{Kullback}$ measures the distance between the actual distribution of the counts in a summary and a uniform distribution of the counts.

$$I_{Kullback} = \log_2 m - \left(\sum_{i=1}^{m} p_i \log_2 \frac{p_i}{\bar{q}} \right)$$

$\boldsymbol{I_{MacArthur}}$. Based upon the Shannon index from information theory [13], $I_{MacArthur}$ 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.

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

$\boldsymbol{I_{Theil}}$. Based upon a distance measure from information theory [20], I_{Theil} measures the distance between the actual distribution of the counts in a summary and a uniform distribution of the counts.

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

$\boldsymbol{I_{Atkinson}}$. Based upon a measure of inequality from economics [1], $I_{Atkinson}$ 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.

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

3 Experimental Results

To generate summaries, a series of seven discovery tasks were run: three on the NSERC Research Awards Database (a database available in the public domain) and four on the Customer Database (a confidential database supplied by an industrial partner). These databases have been frequently used in previous data mining research [8,9,12] and will not be described again here. We present the results of the three NSERC discovery tasks, which we refer to as *N-2*, *N-3*, and *N-4*, where 2, 3, and 4 correspond to the number of attributes selected in each discovery task. Similar results were obtained from the Customer Database.

Typical results are shown in Tables 2 through 5, where the 22 summaries generated from the *N-2* discovery task are ranked by the various measures. In Tables 2 through 5, the *Summary ID* column describes a unique summary identifier (for reference purposes), the *Non-ANY Attributes* column describes the number of non-ANY attributes in the summary (i.e., attributes that have not

Table 2. Ranks assigned by $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, and I_{Total} from *N-2*

Summary	*Non-ANY*	*No. of*	$I_{Variance}$		$I_{Simpson}$		$I_{Shannon}$		I_{Total}	
ID	*Attributes*	*Tuples*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
1	1	2	0.377595	1.5	0.877595	1.5	0.348869	1.5	0.697738	1.5
2	1	3	0.128641	5.0	0.590615	5.0	0.866330	5.0	2.598990	5.0
3	1	4	0.208346	3.5	0.875039	3.5	0.443306	3.5	1.773225	3.5
4	1	5	0.024569	10.0	0.298277	10.0	1.846288	10.0	9.231440	7.0
5	1	6	0.018374	12.0	0.258539	14.0	2.125994	11.0	12.755962	9.0
6	1	9	0.017788	13.0	0.253419	15.0	2.268893	13.0	20.420033	13.0
7	1	10	0.041606	8.5	0.474451	8.5	1.419260	8.5	14.192604	10.5
8	2	2	0.377595	1.5	0.877595	1.5	0.348869	1.5	0.697738	1.5
9	2	4	0.208346	3.5	0.875039	3.5	0.443306	3.5	1.773225	3.5
10	2	5	0.079693	6.0	0.518772	6.0	1.215166	6.0	6.075830	6.0
11	2	9	0.018715	11.0	0.260833	12.0	2.194598	12.0	19.751385	12.0
12	2	9	0.050770	7.0	0.517271	7.0	1.309049	7.0	11.781437	8.0
13	2	10	0.041606	8.5	0.474451	8.5	1.419260	8.5	14.192604	10.5
14	2	11	0.013534	14.0	0.226253	16.0	2.473949	16.0	27.213436	14.0
15	2	16	0.010611	17.0	0.221664	18.0	2.616697	18.0	41.867161	16.0
16	2	17	0.012575	15.0	0.260017	13.0	2.288068	15.0	38.897160	15.0
17	2	21	0.008896	18.0	0.225542	17.0	2.567410	17.0	53.915619	18.0
18	2	21	0.011547	16.0	0.278568	11.0	2.282864	14.0	47.940136	17.0
19	2	30	0.006470	19.0	0.220962	19.0	2.710100	19.0	81.302986	19.0
20	2	40	0.002986	20.0	0.141445	20.0	3.259974	20.0	130.39897	20.0
21	2	50	0.002078	21.0	0.121836	21.0	3.538550	21.0	176.92749	21.0
22	2	67	0.001582	22.0	0.119351	22.0	3.679394	22.0	246.51939	22.0

Table 3. Ranks assigned by I_{Max}, $I_{McIntosh}$, I_{Lorenz}, and I_{Berger} from *N-2*

Summary	*Non-ANY*	*No. of*	I_{Max}		$I_{McIntosh}$		I_{Lorenz}		I_{Berger}	
ID	*Attributes*	*Tuples*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
1	1	2	1.000000	1.5	0.063874	1.5	0.532746	1.5	0.934509	2.5
2	1	3	1.584963	3.0	0.233956	5.0	0.429060	3.0	0.712931	5.0
3	1	4	2.000000	4.5	0.065254	3.5	0.277279	7.5	0.934509	2.5
4	1	5	2.321928	6.5	0.458697	10.0	0.402945	4.0	0.393841	12.0
5	1	6	2.584963	8.0	0.496780	14.0	0.379616	5.0	0.393841	12.0
6	1	9	3.169925	10.0	0.501894	15.0	0.261123	9.0	0.393841	12.0
7	1	10	3.321928	12.5	0.314518	8.5	0.165982	14.5	0.603704	8.5
8	2	2	1.000000	1.5	0.063874	1.5	0.532746	1.5	0.934509	2.5
9	2	4	2.000000	4.5	0.065254	3.5	0.277279	7.5	0.934509	2.5
10	2	5	2.321928	6.5	0.282728	6.0	0.283677	6.0	0.666853	6.5
11	2	9	3.169925	10.0	0.494505	12.0	0.253015	10.0	0.365614	16.5
12	2	9	3.169925	10.0	0.283782	7.0	0.166537	13.0	0.666853	6.5
13	2	10	3.321928	12.5	0.314518	8.5	0.165982	14.5	0.603704	8.5
14	2	11	3.459432	14.0	0.529937	16.0	0.236883	11.0	0.365614	16.5
15	2	16	4.000000	15.0	0.534837	18.0	0.175297	12.0	0.365614	16.5
16	2	17	4.087463	16.0	0.495313	13.0	0.142521	16.0	0.365614	16.5
17	2	21	4.392317	17.5	0.530693	17.0	0.132651	17.0	0.365614	16.5
18	2	21	4.392317	17.5	0.477246	11.0	0.118036	18.0	0.420841	10.0
19	2	30	4.906891	19.0	0.535592	19.0	0.100625	21.0	0.365614	16.5
20	2	40	5.321928	20.0	0.630569	20.0	0.108058	19.0	0.234297	21.0
21	2	50	5.643856	21.0	0.657900	21.0	0.102211	20.0	0.234297	21.0
22	2	67	6.066089	22.0	0.661515	22.0	0.083496	22.0	0.234297	21.0

been generalized to the level of the most general node in the associated DGG that contains the default description "ANY"), the *No. of Tuples* column describes the number of tuples in the summary, and the *Score* and *Rank* columns describe the calculated interestingness and the assigned rank, respectively, as determined by the corresponding measure. Some measures are ranked by score in descending order and some in ascending order (this is easily determined by examining the ranks assigned in Tables 2 through 5). This is done so that each measure ranks the less complex summaries (i.e., those with few tuples and/or few non-ANY attributes) as more interesting. Tables 2 through 5 do not show any single-tuple summaries (e.g., a single-tuple summary where both attributes are generalized to ANY and a single-tuple summary that was an artifact of the DGGs used), as these summaries are considered to contain no information and are, therefore, uninteresting by definition. The summaries in Tables 2 through 5 are shown in increasing order of the number of non-ANY attributes and the number of tuples in each summary, respectively.

Table 4. Ranks assigned by I_{Schutz}, I_{Bray}, $I_{Whittaker}$, and $I_{Kullback}$ from *N-2*

Summary	*Non-ANY*	*No. of*	I_{Schutz}		I_{Bray}		$I_{Whittaker}$		$I_{Kullback}$	
ID	*Attributes*	*Tuples*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
1	1	2	0.434509	4.5	0.565491	4.5	0.565491	4.5	0.348869	1.5
2	1	3	0.379598	3.0	0.620402	3.0	0.620402	3.0	0.866330	5.0
3	1	4	0.684509	11.5	0.315491	11.5	0.315491	11.5	0.443306	3.5
4	1	5	0.310744	2.0	0.689256	2.0	0.689256	2.0	1.846288	10.0
5	1	6	0.294042	1.0	0.705958	1.0	0.705958	1.0	2.125994	11.0
6	1	9	0.466300	6.0	0.533700	6.0	0.533700	6.0	2.268893	13.0
7	1	10	0.734509	19.5	0.265491	19.5	0.265491	19.5	1.419260	8.5
8	2	2	0.434509	4.5	0.565491	4.5	0.565491	4.5	0.348869	1.5
9	2	4	0.684509	11.5	0.315491	11.5	0.315491	11.5	0.443306	3.5
10	2	5	0.534397	9.0	0.465603	9.0	0.465603	9.0	1.215166	6.0
11	2	9	0.516940	8.0	0.483060	8.0	0.483060	8.0	2.194598	12.0
12	2	9	0.712175	15.0	0.287825	15.0	0.287825	15.0	1.309049	7.0
13	2	10	0.734509	19.5	0.265491	19.5	0.265491	19.5	1.419260	8.5
14	2	11	0.486637	7.0	0.513363	7.0	0.513363	7.0	2.473949	16.0
15	2	16	0.600273	10.0	0.399727	10.0	0.399727	10.0	2.616697	18.0
16	2	17	0.699103	14.0	0.300897	14.0	0.300897	14.0	2.288068	15.0
17	2	21	0.696302	13.0	0.303698	13.0	0.303698	13.0	2.567410	17.0
18	2	21	0.743921	22.0	0.256079	22.0	0.256079	22.0	2.282864	14.0
19	2	30	0.723102	16.0	0.276898	16.0	0.276898	16.0	2.710100	19.0
20	2	40	0.734397	17.5	0.265603	17.5	0.265603	17.5	3.259974	20.0
21	2	50	0.734397	17.5	0.265603	17.5	0.265603	17.5	3.538550	21.0
22	2	67	0.742610	21.0	0.25739	21.0	0.257390	21.0	3.679394	22.0

Table 5. Ranks assigned by $I_{MacArthur}$, I_{Theil}, $I_{Atkinson}$, and I_{Gini} from *N-2*

Summary	*Non-ANY*	*No. of*	$I_{MacArthur}$		I_{Theil}		$I_{Atkinson}$		I_{Gini}	
ID	*Attributes*	*Tuples*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
1	1	2	0.184731	3.5	0.651131	1.5	0.505218	1.5	0.217254	1.5
2	1	3	0.218074	5.0	0.718633	3.0	0.914901	22.0	0.158404	5.0
3	1	4	0.399511	11.5	1.556694	7.5	0.792127	8.5	0.173861	3.5
4	1	5	0.144729	2.0	0.757153	4.0	0.759314	6.0	0.078822	8.0
5	1	6	0.132377	1.0	0.777902	5.0	0.693136	3.0	0.067906	11.0
6	1	9	0.243857	6.0	1.710559	9.0	0.765973	7.0	0.065429	13.0
7	1	10	0.457814	16.5	2.508888	13.5	0.821439	11.5	0.076804	9.5
8	2	2	0.184731	3.5	0.651131	1.5	0.505218	1.5	0.217254	1.5
9	2	4	0.399511	11.5	1.556694	7.5	0.792127	8.5	0.173861	3.5
10	2	5	0.298402	9.0	1.195810	6.0	0.859044	16.0	0.126529	6.0
11	2	9	0.264620	8.0	1.898130	10.0	0.759162	5.0	0.067231	12.0
12	2	9	0.452998	15.0	2.249471	12.0	0.884562	19.0	0.086449	7.0
13	2	10	0.457814	16.5	2.508888	13.5	0.821439	11.5	0.076804	9.5
14	2	11	0.260255	7.0	2.025527	11.0	0.727091	4.0	0.056104	14.0
15	2	16	0.342143	10.0	2.939297	15.0	0.797472	10.0	0.044494	16.0
16	2	17	0.441534	14.0	3.512838	16.0	0.860465	17.0	0.045517	15.0
17	2	21	0.440642	13.0	3.890191	17.0	0.852812	13.0	0.037253	18.0
18	2	21	0.487441	20.0	3.982314	18.0	0.862917	18.0	0.038645	17.0
19	2	30	0.494412	21.0	4.485426	19.0	0.894697	21.0	0.027736	19.0
20	2	40	0.479347	18.0	5.317662	20.0	0.854864	15.0	0.020222	20.0
21	2	50	0.482560	19.0	5.751495	21.0	0.854329	14.0	0.016312	21.0
22	2	67	0.515363	22.0	6.181546	22.0	0.885877	20.0	0.012656	22.0

Tables 2 through 5 show similarities in how some of the sixteen measures rank summaries. For example, the six most interesting summaries (i.e., 1, 2, 3, 8, 9, and 10) are ranked identically by $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, I_{Total}, $I_{McIntosh}$, and $I_{Kullback}$, while the four least interesting summaries (i.e., 19, 20, 21, and 22) are ranked identically by $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, I_{Total}, I_{Max}, $I_{McIntosh}$, $I_{Kullback}$, I_{Theil}, and I_{Gini}.

To quantify the extent of the ranking similarities between the sixteen measures across all seven discovery tasks, we calculated the Gamma correlation coefficient for each pair of measures and found that 86.4% of the coefficients are highly significant with a *p-value* below 0.005. We also found the ranks assigned to the summaries have a high positive correlation for some pairs of measures. For the purpose of this discussion, we considered a pair of measures to be highly correlated when the average coefficient is greater than 0.85. Thus, 35% of the pairs (i.e., 42 of 120 pairs) are highly correlated using the 0.85 threshold. Following careful examination of the 42 highly correlated pairs, we found two distinct groups of measures within which summaries are ranked similarly. One group consists of the measures $I_{Variance}$, $I_{Simpson}$, $I_{Shannon}$, I_{Total}, I_{Max}, $I_{McIntosh}$, I_{Berger}, $I_{Kullback}$, and I_{Gini}. The other group consists of the measures I_{Schutz}, I_{Bray}, $I_{Whittaker}$, and $I_{MacArthur}$. There are no similarities (i.e., no high positive correlations) shared between the two groups. Of the remaining three measures, I_{Theil}, I_{Lorenz}, and $I_{Atkinson}$, I_{Theil} is only highly correlated with I_{Max}, while I_{Lorenz} and $I_{Atkinson}$ are not highly correlated with any of the other measures. There were no highly negative correlations between any of the pairs of measures.

One way to analyze the measures is to determine the complexity of summaries considered to be of high, moderate, and low interest (i.e., the relative interestingness). These results are shown in Table 6. In Table 6, the values in the *H*, *M*, and *L* columns describe the complexity index for a group of summaries considered to be of high, moderate, and low interest, respectively. The *complexity index* for a group of summaries is defined as the product of the average number of tuples and the average number of non-ANY attributes contained in the group of summaries. For example, the complexity index for summaries determined to be of high interest by the $I_{Variance}$ index for discovery task *N-2*, is 4.5 (i.e., 3×1.5, where 3 and 1.5 are the average number of tuples and average number of non-ANY attributes, respectively). High, moderate, and low interest summaries were considered to be the top, middle, and bottom 20%, respectively, of summaries. The *N-2*, *N-3*, and *N-4* discovery tasks generated sets containing 22, 70, and 214 summaries, respectively. Thus, the complexity index of the summaries from the *N-2*, *N-3*, and *N-4* discovery tasks is based upon the averages for four, 14, and 43 summaries, respectively.

Table 6. Relative interestingness of summaries from the NSERC discovery tasks

Interestingness Measure	*Relative Interestingness*								
	N-2			*N-3*			*N-4*		
	H	*M*	*L*	*H*	*M*	*L*	*H*	*M*	*L*
$I_{Variance}$	4.5	11.3	93.6	9.0	64.7	520.3	34.6	430.5	3212.9
$I_{Simpson}$	4.5	20.3	93.6	9.0	72.9	477.4	38.0	447.8	3163.1
$I_{Shannon}$	4.5	11.3	93.6	9.0	72.9	520.3	29.8	430.2	3210.2
I_{Total}	4.5	13.2	93.6	8.1	65.8	545.5	27.2	423.6	3220.5
I_{Max}	3.6	14.0	93.6	8.3	63.7	545.5	27.0	424.2	3221.6
$I_{McIntosh}$	4.5	20.3	93.6	9.0	72.9	477.4	38.0	447.8	3163.1
I_{Lorenz}	3.9	20.3	93.6	21.1	104.8	249.3	133.6	1373.9	482.6
I_{Berger}	4.5	15.8	93.6	9.6	86.6	457.5	48.8	587.8	2807.2
I_{Schutz}	4.0	13.1	48.6	23.4	367.9	146.7	289.8	1242.2	227.0
I_{Bray}	4.0	13.1	48.6	23.4	367.9	146.7	289.8	1242.2	227.0
$I_{Whittaker}$	4.0	13.1	48.6	23.4	367.9	146.7	289.8	1242.2	227.0
$I_{Kullback}$	4.5	11.3	93.6	9.0	72.9	520.3	29.8	430.2	3210.2
$I_{MacArthur}$	4.9	13.1	84.0	23.2	251.4	220.8	249.5	1210.3	233.2
I_{Theil}	3.9	17.1	93.6	9.1	66.2	533.3	33.8	558.9	2668.4
$I_{Atkinson}$	8.0	18.0	49.1	31.5	270.5	103.7	531.1	555.6	1611.1
I_{Gini}	4.5	13.2	93.6	9.0	60.5	537.7	27.9	425.1	3220.5

Table 6 shows that in most cases the complexity index is lowest for the most interesting summaries and highest for the least interesting summaries. For example, the complexity index for summaries determined by the $I_{Variance}$ index to be of high, moderate, and low interest are 4.5, 11.3, and 93.6 from *N-2*, respectively, 9.0, 64.7, and 520.3 from *N-3*, respectively, and 34.6, 430.5, and 3212.9 from *N-4*, respectively. The only exceptions occurred in the results for the I_{Lorenz}, I_{Schutz}, I_{Bray}, $I_{Whittaker}$, $I_{MacArthur}$, and $I_{Atkinson}$ indexes from the *N-3* and *N-4* discovery tasks.

A comparison of the summaries with high relative interestingness from the *N-2*, *N-3*, and *N-4* discovery tasks is shown in the graph of Figure 1. In Figure 1, the horizontal and vertical axes describe the measures and the complexity indexes, respectively. Horizontal rows of bars correspond to the complexity indexes of summaries from a particular discovery task. The back most horizontal row of bars corresponds to the average complexity index for a particular measure. Figure 1 shows a maximum complexity index on the vertical axes of 60.0 (although the complexity indexes for I_{Lorenz}, I_{Schutz}, I_{Bray}, $I_{Whittaker}$, $I_{MacArthur}$, and $I_{Atkinson}$ from the *N-4* discovery task each exceed this value by a minimum of 189.5). The measures, listed in ascending order of the complexity index, are (position 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}, I_{Bray}, and $I_{Whittaker}$ (11), and $I_{Atkinson}$ (12).

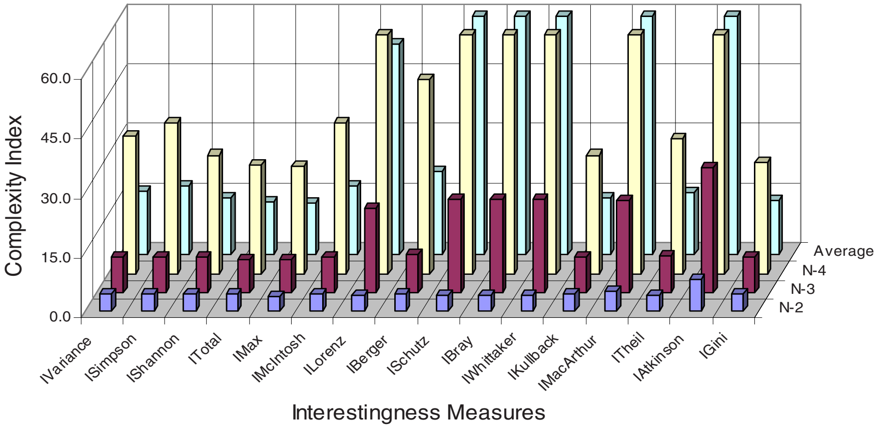

Fig. 1. Relative complexity of summaries from the NSERC discovery tasks

4 Conclusion and Future Research

We described the HMI set of heuristics for ranking the interestingness of summaries generated from databases. Although the heuristics have previously been applied in several areas of the physical, social, ecological, management, information, and computer sciences, their use for ranking summaries generated from databases is a new application area. The preliminary results presented here show that the order in which some of the measures rank summaries is highly correlated, resulting in two distinct groups of measures in which summaries are ranked similarly. Highly ranked, concise summaries provide a reasonable starting point for further analysis of discovered knowledge. That is, other highly ranked summaries that are nearby in the generalization space will probably contain information at useful and appropriate levels of detail. Future research will focus on determining the specific response of each measure to different population structures.

References

1. A.B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.
2. W.H. Berger and F.L. Parker. Diversity of planktonic forminifera in deep-sea sediments. *Science*, 168:1345–1347, 1970.
3. 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.
4. J.R. Bray and J.T. Curtis. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27:325–349, 1957.
5. A.A. Freitas. On objective measures of rule surprisingness. 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 1–9, Nantes, France, September 1998.
6. R. Godin, R. Missaoui, and H. Alaoui. Incremental concept formation algorithms based on galois (concept) lattices. *Computational Intelligence*, 11(2):246–267, 1995.
7. H.J. Hamilton, R.J. Hilderman, L. Li, and D.J. Randall. Generalization lattices. 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 328–336, Nantes, France, September 1998.
8. 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.
9. R.J. Hilderman, H.J. Hamilton, and B. Barber. Ranking the interestingness of summaries from data mining systems. In *Proceedings of the 12th International Florida Artificial Intelligence Research Symposium (FLAIRS'99)*, pages 100–106, Orlando, Florida, May 1999.
10. 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.
11. S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
12. 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.
13. R.H. MacArthur. Patterns of species diversity. *Biological Review*, 40:510–533, 1965.
14. R.P. McIntosh. An index of diversity and the relation of certain concepts to diveristy. *Ecology*, 48(3):392–404, 1967.
15. W.A. Rosenkrantz. *Introduction to Probability and Statistics for Scientists and Engineers.* McGraw-Hill, 1997.
16. R.R. Schutz. On the measurement of income inequality. *American Economic Review*, 41:107–122, March 1951.
17. C.E. Shannon and W. Weaver. *The mathematical theory of communication.* University of Illinois Press, 1949.
18. E.H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.
19. 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.
20. H. Theil. *Economics and information theory.* Rand McNally, 1970.
21. R.H. Whittaker. Evolution and measurement of species diversity. *Taxon*, 21 (2/3):213–251, May 1972.
22. Y.Y. Yao, S.K.M. Wong, and C.J. Butz. On information-theoretic measures of attribute importance. In N. Zhong and L. Zhou, editors, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, pages 133–137, Beijing, China, April 1999.
23. J.F. Young. *Information theory.* John Wiley & Sons, 1971.