Ranking the Interestingness of Summaries from Data Mining Systems

Robert J. Hilderman, Howard J. Hamilton and Brock Barber

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

Abstract

We study data mining where the task is description by summarization, the representation language is generalized relations, the evaluation criteria are based on heuristic measures of interestingness, and the method for searching is the Multi-Attribute Generalization algorithm for domain generalization graphs. We present and empirically compare four heuristics for ranking the interestingness of generalized relations (or *summaries*). The measures are based on common measures of the diversity of a population, statistical variance, the Simpson index, and the Shannon index. All four measures rank less complex summaries (i.e., those with few tuples and/or non-ANY attributes) as most interesting. Highly ranked summaries provide a reasonable starting point for further analysis of discovered knowledge.

Introduction

The process of knowledge discovery from databases includes these steps: data selection, cleaning and other preprocessing, reduction and transformation, data mining to identify interesting patterns, interpretation and evaluation, and application [7]. The goal is to identify valid, previously unknown, potentially useful patterns in data [7; 9]. The data mining step requires the choice of four items: a data mining task (such as prediction, description, or anomaly detection), a representation language for patterns, evaluation criteria for patterns, and a method for searching for patterns to be evaluated. Within the category of descriptive tasks, summarization has received considerable attention and several fast, effective algorithms have been developed. The task of performing *attribute-oriented generalization* (AOG) requires the creation of a generalized relation (or *summary*) where specific attribute values in a relation are replaced with more general concepts according to user-defined *concept hierarchies* (CHs) [5]. If the original relation is the result of a database query, the generalized relation is a summary of these results, where, for example, names of particular laundry soaps might be replaced by the general concepts "laundry soap" or "aisle 9" depending on the concept hierarchy. The GDBR and FIGR algorithms perform attribute oriented generalization in $O(n)$ time (proven to be optimal) while requiring $O(p)$ space, where n is the number of tuples in the input relation, and p is the number of tuples in the summaries (typically $p \ll n$) [5].

Until recently, AOG methods were limited in their ability to efficiently generate summaries when multiple CHs were associated with an attribute. To resolve this problem, we previously introduced new serial and parallel AOG algorithms [12; 16] and a data structure called a *domain generalization graph* (DGG) [12; 13; 16; 21]. 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 among 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 such that one domain is selected from each DGG. Our algorithms generate the summaries by traversing the generalization space in a time- and space-efficient manner. 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.

We study data mining where the data mining task is description by summarization, the representation language is generalized relations, the evaluation criteria are based on heuristic measures of interestingness, and the method for searching is the Multi-Attribute Generalization algorithm [12] for domain generalization graphs. In [15], we proposed four heuristics, based upon information theory and statistics, for ranking the interestingness of summaries generated from a database. Preliminary results suggested that the order in which the summaries are ranked is highly correlated among these measures. In this paper, we present additional experimental results describing the behaviour of these heuristics when used to rank the interestingness of summaries.

Techniques for determining the interestingness of discovered knowledge have previously received some attention in the literature. A rule-interest function is

proposed in [20] which prunes uninteresting implication rules based upon a statistical correlation threshold. In [2], two interestingness functions are proposed. The first function measures the difference between the number of tuples containing an attribute value and the number that would be expected if the values for the attribute were uniformly distributed. The second function measures the difference between the proportion of records that contain specified values in a pair of attributes and the proportion that would be expected if the values were statistically independent. A measure from information theory, called KL-distance, is proposed in [8] which measures the distance of the actual distribution of terms in text files from that of the expected distribution. KL-distance is also proposed in [12] for measuring the distance between the actual distribution of tuples in a summary to that of a uniform distribution of the tuples. In [25], another measure from information theory is proposed which measures the average information content of a probabilistic rule. In [19], deviations are proposed which compare the difference between measured values and some previously known or normative values. In [11], two interestingness measures are proposed that measure the potential for knowledge discovery based upon the complexity of concept hierarchies associated with attributes in a database. A variety of interestingness measures are proposed in [18] that evaluate the coverage and certainty of a set of discovered implication rules that have previously been identified as potentially interesting. In [1], transaction support, confidence, and syntactic constraints are proposed to construct rules from databases containing binary-valued attributes. A measure is proposed in [10] which determines the interestingness (called surprise there) of discovered knowledge via the explicit detection of occurrences of Simpson's paradox. Finally, an excellent survey of information-theoretic measures for evaluating the importance of attributes is described in [26].

Although our measures were developed and utilized for ranking the interestingness of generalized relations as described earlier in this section, they are more generally applicable to other problem domains, such as ranking *views* (i.e., precomputed, virtual tables derived from a relational database) or *summary tables* (i.e., materialized, aggregate views derived from a data cube). However, we do not dwell here on the technical aspects of deriving generalized relations, 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 remainder of this paper is organized as follows. In the next section, we describe heuristics for ranking the interestingness of summaries and provide a detailed example for each heuristic. In the third section, we present experimental results and compare the four interestingness measures. In the last section, we conclude with a summary of our work and suggestions for future research.

Interestingness

We now formally define the problem of ranking the interestingness of summaries, 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 each v_{ij} is a value from the domain D_j associated with attribute A_j. One attribute A_n is a derived attribute, called *Count*, whose domain D_n is the set of positive integers, and whose value v_{in} for each attribute-value pair (A_n, v_{in}) is equal to the number of tuples which have been aggregated from the base relation (i.e., the unconditioned data present in the original relational database). The *interestingness* I of a summary S is given by $I = f(S)$, where f is typically a function of the cardinality and degree of S, the complexity of its associated CHs, or the probability distributions of the tuples in S.

A sample summary is shown below in Table 1. In Table 1, there are $n = 3$ attribute-domain pairs (i.e., (A_1, D_1) = (*Colour*,{*red, blue, green*}), (A_2, D_2) = (*Shape*,{*round, square*}), and (A_3, D_3) = (*Count*,{*positive integers*})) and $m = 4$ sets of unique attribute-value pairs. The *Colour* and *Shape* attributes describe the colour and shape, respectively, of some arbitrary object, and the *Count* attribute describes the number of objects aggregated from the base relation which possess the corresponding colour and shape characteristics. A *Tuple ID* attribute is being shown for demonstration purposes only (to simplify the presentation that follows) and is not actually part of the summary. For example, the tuple {(*Colour, blue*),(*Shape, square*),(*Count*, 1)} is simply referred to as t_3. Table 1 will be used as the basis for all calculations in the examples which follow.

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 four heuristics for ranking the interestingness of summaries and provide a detailed example for each heuristic. These heuristics have been selected for evaluation as interestingness measures because they are common measures of diversity of a population. The well-known, domain-independent formulae, upon which these heuristics are based, have previously seen extensive application in several areas of the physical, social, management, and computer sciences. The I_{var} measure is based upon variance, which is the most common measure of variability used in statistics [22]. The I_{avg} and I_{tot} measures, based upon a relative entropy measure (also known as the Shannon index) from information theory [23; 27], measure the average information

content in a single tuple in a summary and the total information content in a summary, respectively. The I_{con} measure, a variance-like measure based upon the Simpson index [24], 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 tuples in a summary are unique, and therefore, can be considered to be a population described by some probability distribution. In the discussion that follows, the probability of each t_i occurring in the actual probability distribution of S is given by:

$$p(t_i) = \frac{v_{in}}{(v_{1n} + v_{2n} + \ldots + v_{mn})},$$

and the probability distribution of each t_i occurring in a summary where the tuples have a uniform probability distribution is given by:

$$q(t_i) = \frac{\frac{(v_{1n}+v_{2n}+\ldots+v_{mn})}{m}}{(v_{1n} + v_{2n} + \ldots + v_{mn})} = \frac{1}{m},$$

where v_{in} is the value associated with the *Count* attribute A_n in tuple t_i.

The I_{var} Measure

Given a summary S, we can measure how far the actual probability distribution (hereafter called simply a distribution) of the counts for the t_i's in S varies from that of a uniform distribution. The *variance* of the distribution in S from that of a uniform distribution is given by:

$$I_{var} = \frac{\sum_{i=1}^{m}(p(t_i) - q(t_i))^2}{m-1},$$

where higher values of I_{var} are considered more interesting. For example, given the summaries S_1, S_2, S_3, and S_4 with variance of 0.08, 0.02, 0.05, and 0.03, respectively, we order the summaries, when ranked from most interesting to least interesting, as S_1, S_3, S_4, and S_2. Our calculation for variance uses $m-1$ because we assume the summary may not contain all possible combinations of attributes, meaning we are not observing all possible tuples.

Example 1: From the actual distribution of the tuples in Table 1, we have $p(t_1) = 0.429, p(t_2) = 0.143, p(t_3) = 0.143$, $p(t_4) = 0.286$, and from a uniform distribution of the tuples, we have $q(t_i) = 0.25$, for all i. So, the interestingness of the summary using the I_{var} measure is:

$$\begin{aligned} I_{var} &= ((0.429 - 0.25)^2 + (0.143 - 0.25)^2 + \\ &\quad (0.143 - 0.25)^2 + (0.286 - 0.25)^2)/3 \\ &= 0.018. \end{aligned}$$

The I_{avg} Measure

Given a summary S, we can determine the average information content in each tuple. The *average information content*, in bits per tuple, is given by:

$$I_{avg} = -\sum_{i=1}^{m} p(t_i) \log_2 p(t_i),$$

where lower values of I_{avg} are considered more interesting. For example, given the summaries S_1, S_2, S_3, and S_4 with average information content of 1.74, 0.95, 3.18, and 2.21 bits, respectively, we order the summaries, when ranked from most interesting to least interesting, as S_2, S_1, S_4, and S_3.

Example 2: The actual distribution for p is given in Example 1. So, the interestingness of the summary using the I_{avg} measure is:

$$\begin{aligned} I_{avg} &= -(0.429 \log_2 0.429 + 0.143 \log_2 0.143 + \\ &\quad 0.143 \log_2 0.143 + 0.286 \log_2 0.286) \\ &= 1.842 \textit{ bits.} \end{aligned}$$

The I_{tot} Measure

Given a summary S, we can determine its total information content. The *total information content*, in bits, is given by:

$$I_{tot} = m * I_{avg},$$

where lower values of I_{tot} are considered more interesting. For example, given the summaries S_1, S_2, S_3, and S_4 with total information content of 31.8, 6.96, 3.83, and 15.5 bits, respectively, we order the summaries, when ranked from most interesting to least interesting, as S_3, S_2, S_4, and S_1.

Example 3: The average information content for the summary is given in Example 2. So, the interestingness of the summary using the I_{tot} measure is:

$$\begin{aligned} I_{tot} &= m * I_{avg} \\ &= 4 * 1.842 \\ &= 7.368 \textit{ bits.} \end{aligned}$$

The I_{con} Measure

Given a summary S, we can measure the extent to which the counts are distributed over the tuples in the summary. The *concentration* of the distribution in S is given by:

$$I_{con} = \sum_{i=1}^{m} p(t_i)^2,$$

where higher values of I_{con} are considered more interesting. For example, given the summaries S_1, S_2, S_3, and S_4 with concentration of 0.57, 0.24, 0.11, and 0.32, respectively, we order the summaries, when ranked from most interesting to least interesting as S_1, S_4, S_2, and S_3.

Example 4: The actual distribution for p is given in Example 1. So, the interestingness of the summary using the I_{con} measure is:

$$\begin{aligned} I_{con} &= 0.429^2 + 0.143^2 + 0.143^2 + 0.286^2 \\ &= 0.306. \end{aligned}$$

Table 2: Ranks assigned by each interestingness measure for *N-2*

Summary ID	*Non-ANY Attributes*	*No. of Tuples*	I_{var}		I_{avg}		I_{tot}		I_{con}	
			Score	*Rank*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
1	1	2	0.18880	1.5	0.34887	1.5	0.69774	1.5	0.87760	1.5
2	1	3	0.08576	5	0.86633	5	2.59899	5	0.59062	5
3	1	4	0.15626	3.5	0.44331	3.5	1.77323	3.5	0.87504	3.5
4	1	5	0.01966	10	1.84629	10	9.23144	7	0.29828	10
5	1	6	0.01531	13	2.12599	11	12.75600	9	0.25854	14
6	1	9	0.01581	12	2.26899	13	20.42000	13	0.25342	15
7	1	10	0.03745	8.5	1.41926	8.5	14.19260	10.5	0.47445	8.5
8	2	2	0.18880	1.5	0.34887	1.5	0.69774	1.5	0.87760	1.5
9	2	4	0.15626	3.5	0.44331	3.5	1.77323	3.5	0.87504	3.5
10	2	5	0.06375	6	1.21517	6	6.07583	6	0.51877	6
11	2	9	0.04513	7	1.30905	7	11.78140	8	0.51727	7
12	2	9	0.01664	11	2.19460	12	19.75140	12	0.26083	12
13	2	10	0.03745	8.5	1.41926	8.5	14.19260	10.5	0.47445	8.5
14	2	11	0.01230	14	2.47495	16	27.21340	14	0.22625	16
15	2	16	0.00995	17	2.61670	18	41.86720	16	0.22166	18
16	2	17	0.01184	15	2.28807	15	38.89720	15	0.26002	13
17	2	21	0.01100	16	2.28286	14	47.94010	17	0.27857	11
18	2	21	0.00847	18	2.56741	17	53.91560	18	0.22554	17
19	2	30	0.00625	19	2.71010	19	81.30300	19	0.22096	19
20	2	40	0.00291	20	3.25997	20	130.39900	20	0.14145	20
21	2	50	0.00204	21	3.53855	21	176.92700	21	0.12184	21
22	2	67	0.00156	22	3.67939	22	246.51900	22	0.11935	22

Experimental Results

In this section, we present experimental results which contrast the various interestingness measures. All summaries in our experiments were generated using DB-Discover [5; 6], a software tool which uses AOG for KDD. DB-Discover was run on a Silicon Graphics Challenge M, with twelve 150 MHz MIPS R4400 CPUs, using Oracle Release 7.3 for database management.

Description of Databases

To generate summaries, series of discovery tasks were run on the NSERC Research Awards Database (a database available in the public domain) and the Customer Database (a confidential database supplied by an industrial partner). The NSERC Research Awards Database, frequently used in previous data mining research [3; 4; 11; 17], consists of 10,000 tuples in six tables describing a total of 22 attributes. The Customer Database, also frequently used in previous data mining research [6; 14; 16], consists of 8,000,000 tuples in 22 tables describing a total of 56 attributes. The largest table contains over 3,300,000 tuples representing account activity for over 500,000 customer accounts and over 2,200 products and services.

Our previous experience in applying AOG data mining techniques to the databases of our industrial partners has shown that domain experts typically perform discovery tasks on a few attributes that have been determined to be relevant. Consequently, we present results for experiments containing a maximum of four relevant attributes. Discovery tasks were run against the NSERC database, where two, three, and four attributes were selected for discovery, and against the Customer database, where two and three attributes were selected for discovery. We refer to the NSERC discovery tasks as as *N-2*, *N-3*, and *N-4*, respectively, and the Customer discovery tasks as *C-2* and *C-3*, respectively. Since similar results were obtained from the NSERC and Customer discovery tasks, we focus on the NSERC tasks.

Comparative Results

We now compare the ranks assigned to the summaries by each interestingness measure. A typical result is shown in Table 2, where 22 summaries, generated from the two-attribute NSERC discovery task, are ranked. In Table 2, 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 been generalized to the level of the root node in the associated DGG), 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 interestingness measure. Table 2 does not show any single-tuple summaries (e.g., the single-tuple summary where both attributes are generalized to ANY and a single-tuple summary that was an artifact of the concept hierarchies used), as these summaries are considered to contain no information and are, therefore, uninteresting by definition. The summaries in Table 2 are shown in increasing order of the number of non-ANY attributes and the number of tuples in each summary, respectively.

The *Rank* column for each interestingness measure uses a ranking scheme that breaks ties in the interestingness scores by averaging the ranks and assigning the same rank to each summary involved in the tie, even though the resulting rank may be fractional. For example, if two summaries are tied when attempting to rank the fourth summary, each is given a rank of $(4+5)/2 = 4.5$, with the next summary ranked sixth. If instead, three summaries are tied, each is given a rank of $(4+5+6)/3 = 5.0$, with the next summary ranked seventh. The general procedure should now be clear. This ranking scheme was adopted to conform to the requirements of the Gamma correlation coefficient used

to analyze the ranking similarities of the interestingness measures and described later in this section.

Table 2 shows there are numerous ties in the ranks assigned by each interestingness measure. For example, Summaries 1 and 8, the most interesting one- and two-attribute summaries, respectively, have a rank of 1.5. This tie is also an artifact of the concept hierarchies used in the discovery task. Summary 1 is shown in Table 3. In the concept hierarchy associated with the *Province* attribute, there is a one-to-one correspondence between the concept *Canada* in Summary 8 and the concept *ANY* in Summary 1. Consequently, this results in a summary containing two non-ANY attributes being assigned the same interestingness score as a summary containing one non-ANY attribute. All ties in Table 2 result from a similar one-to-one correspondence between concepts in the concept hierarchies used.

Table 3: Summary 1 from *N-2*

Province	*DiscCode*	*Count*
ANY	Other	8376
ANY	Computer	587

Table 2 also shows some similarities in how the four interestingness measures rank summaries. For example, the six most interesting summaries (i.e., Summaries 1, 8, 2, 10, 3, and 9) are ranked identically by the four interestingness measures, as are the four least interesting summaries (i.e., Summaries 19, 20, 21, and 22). There are also similarities among the moderately interesting summaries. For example, Summary 11 is ranked seventh by the I_{var}, I_{avg}, and I_{con} measures, and Summary 12 is ranked twelfth by the I_{avg}, I_{tot}, and I_{con} measures.

To determine the extent of the ranking similarities between the four interestingness measures, we can calculate the Gamma correlation coefficient for each pair of interestingness measures. The Gamma statistic assumes that the summaries under consideration are assigned ranks according to an ordinal (i.e., rank order) scale, and is a probability computed as the difference between the probability that the rank ordering of two interestingness measures agree minus the probability that they disagree, divided by 1 minus the probability of ties. The value of the Gamma statistic varies in the interval $[-1, 1]$, where values near 1, 0, and -1 represent significant positive, no, and significant negative correlation, respectively.

The Gamma correlation coefficients (hereafter called the coefficients) for the two-, three-, and four-attribute discovery tasks are shown in Table 4. In Table 4, the *Interestingness Measures* column describes the pairs of interestingness measures being compared and the *N-2*, *N-3*, and *N-4* columns describe the coefficients corresponding to the pairs of interestingness measures in the two-, three-, and four-attribute discovery tasks, respectively. Table 4 shows that the ranks assigned to the summaries by all pairs of interestingness measures are similar, as indicated by the high coefficients. The coefficients vary from a low of 0.82862 for the pair containing the I_{tot} and I_{con} measures in the three-attribute discovery task, to a high of 0.96506 for the pair containing the I_{avg} and I_{con} measures in the same discovery task. The ranks assigned by the pair containing the I_{var} and I_{avg} measures are most similar, as indicated by the average coefficient of 0.95494 for the two-, three-, and four-attribute discovery tasks, followed closely by the ranks assigned to the pair containing the I_{avg} and I_{con} measures with an average coefficient of 0.95253. The ranks assigned by the pairs of interestingness measures in the three-attribute discovery task have the least similarity, as indicated by the average coefficient of 0.90166, although this is not significantly lower than the two- and four-attribute average coefficients of 0.91813 and 0.92555, respectively. Given the overall average coefficient is 0.91511, we conclude that the ranks assigned by the four interestingness measures are highly correlated.

Table 4: Comparison of ranking similarities

Interestingness Measures	*Gamma Correlation Coefficient*			
	N-2	*N-3*	*N-4*	*Average*
I_{var} & I_{avg}	0.94737	0.95670	0.96076	0.95494
I_{var} & I_{tot}	0.92983	0.86428	0.91904	0.90438
I_{var} & I_{con}	0.91228	0.93172	0.94929	0.93110
I_{avg} & I_{tot}	0.91228	0.86356	0.90947	0.89510
I_{avg} & I_{con}	0.94737	0.96506	0.94516	0.95253
I_{tot} & I_{con}	0.85965	0.82862	0.86957	0.85261
Average	0.91813	0.90166	0.92555	0.91511

We now discuss the complexity of the summaries ranked by the various interestingness measures. We define the *complexity index* of a summary as the product of the number of tuples and the number of non-ANY attributes contained in the summary. A desirable property of any ranking function is that it rank summaries with a low complexity index as most interesting. However, although we want to rank summaries with a low complexity index as most interesting, we do not want to lose the meaning or context of the data by presenting summaries that are too concise. Indeed, in previous work, domain experts agreed that more information is better than less, provided that the most interesting summaries are not too concise and remain relatively easy to understand [11]. Although the most interesting summaries ranked by our interestingness measures are concise, they are generally in accordance with the low complexity property and provide a reasonable starting point for further analysis of more complex summaries.

One way to analyze the interestingness measures and evaluate whether they satisfy the guidelines of our domain experts, is to determine the complexity indexes of summaries considered to be of high, moderate, and low interest, as shown in Table 5. In Table 5, the *Task ID* column describes the discovery task, the *Relative Interest* column describes the relative degree of interest of the corresponding group of summaries on a three-tier scale (i.e., H=High, M=Moderate, L=Low), the *Tuples* and *NA* columns describe the average number of tuples and the average number of non-ANY attributes,

Table 5: Relative interestingness of summaries versus the complexity index

Task ID	Relative Interest	I_{var}			I_{avg}			I_{tot}			I_{con}		
		Tuples	NA	CI	Tuples	NA	CI	Tuples	NA	CI	Tuples	NA	CI
N-2	H	2.0	1.5	3.0	1.5	1.0	1.5	1.5	1.0	1.5	1.5	1.0	1.5
	M	9.0	1.5	13.5	5.5	1.0	5.5	10.0	1.5	15.0	13.0	1.5	19.5
	L	34.0	1.5	51.0	58.5	2.0	117.0	58.5	2.0	117.0	58.5	2.0	117.0
N-3	H	4.1	1.6	6.6	3.0	1.4	4.2	3.0	1.4	4.2	3.0	1.4	4.2
	M	26.7	2.2	58.7	24.9	2.2	54.8	23.4	2.4	56.2	33.9	2.1	58.6
	L	215.4	2.7	589.7	232.4	3.0	697.2	253.1	3.0	759.3	232.4	3.0	697.2
N-4	H	8.3	1.7	14.1	7.9	1.7	13.4	6.5	1.6	10.4	7.9	1.7	13.4
	M	139.5	2.9	404.6	136.1	2.8	381.1	146.2	3.0	438.6	171.9	2.9	498.5
	L	1014.2	3.7	3752.5	1044.5	3.8	3969.1	1075.8	3.9	4195.6	1044.5	3.8	3969.1
C-2	H	5.0	1.2	6.0	4.4	1.1	4.8	4.4	1.1	4.8	4.5	1.2	5.4
	M	19.3	1.6	30.9	21.4	1.7	36.4	22.3	1.7	37.9	80.5	1.3	104.7
	L	101.7	1.8	183.1	101.7	1.8	183.1	101.7	1.8	183.1	101.7	1.8	183.1
C-3	H	9.2	1.8	16.6	8.7	1.7	14.8	8.3	1.8	14.9	9.1	1.7	15.5
	M	82.8	2.4	198.7	79.4	2.4	190.1	100.8	2.2	221.8	108.4	2.4	260.2
	L	361.3	2.7	975.5	361.1	2.7	975.0	361.3	2.7	975.5	343.1	2.8	960.7

respectively, in the corresponding group of summaries, and the *CI* column describes the complexity index of the corresponding group of summaries. High, moderate, and low interest summaries were considered to be the top, middle, and bottom 10%, respectively, of summaries as ranked by each interestingness measure. The two-, three-, and four-attribute NSERC discovery tasks generated sets containing 22, 70, and 214 summaries, respectively, and the two- and three-attribute Customer discovery tasks generated sets containing 196 and 1016 summaries, respectively. Thus, the complexity index of the summaries in the two-, three-, and four-attribute NSERC tasks is based upon two, seven, and 22 summaries, respectively, and the complexity index of the summaries in the two- and three-attribute Customer tasks is based upon 20 and 102 summaries, respectively.

Table 5 shows that each interestingness measure ranks summaries with a low complexity index as most interesting, and vice versa. For example, the complexity index of the I_{var} measure for the two-attribute task shows a typical result. Summaries of high, moderate, and low interest have complexity indexes of 3.0, 13.5, and 51.0, respectively. This result is consistent for all interestingness measures in all discovery tasks.

A comparison of the complexity indexes of the summaries ranked by the four interestingness measures for the two-, three-, and four-attribute discovery tasks are shown in the graph of Figure 1. In Figure 1, the first, second, and third row of bars, where the first row is at the front of the graph, correspond to the two-, three-, and four-attribute discovery tasks, respectively. For the two- and three-attribute discovery tasks, the summaries ranked as most interesting by the I_{avg}, I_{tot}, and I_{con} measures have the lowest complexity indexes, followed by the I_{var} measure. For the four-attribute discovery task, the summaries ranked as most interesting by the I_{tot} measure have the lowest complexity index followed by the I_{avg} and I_{con} measures, and the I_{var} measure.

The fourth row of bars in Figure 1 shows the average complexity index of summaries derived from the two-, three-, and four-attribute discovery tasks for each interestingness measure. For example, the average for the I_{var} measure was derived from the complexity indexes

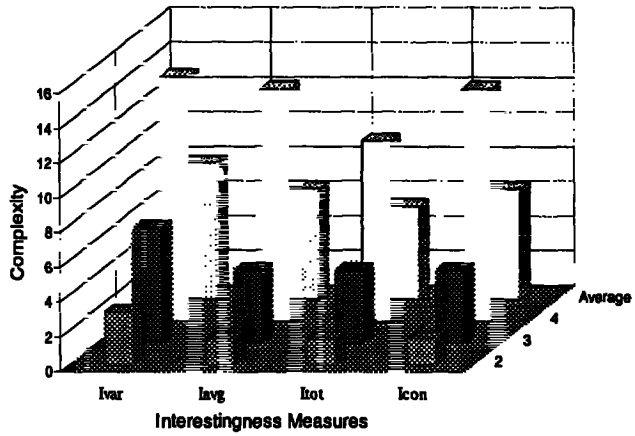

Figure 1: Relative complexity of summaries

3.0, 6.6, and 14.1 for the two-, three-, and four-attribute tasks, respectively, for an average complexity index of 7.9. The summaries ranked as most interesting by the I_{tot} measure have the lowest complexity index, followed by the I_{avg} and I_{con} measures, and the I_{var} measure.

Conclusion and Future Research

We described heuristics for ranking the interestingness of summaries generated from databases. The four interestingness measures evaluated rank summaries in a similar manner, as indicated by the high Gamma correlation coefficients for all possible pairs of interestingness measures. Although all four interestingness measures rank summaries with a low complexity index (i.e., those with few tuples and/or non-ANY attributes) as most interesting, summaries ranked by the I_{tot} measure have the lowest complexity index. Domain experts agree that a low complexity index is a desirable property, and that summaries with a low complexity index provide a reasonable starting point for further analysis of discovered knowledge.

Future research will focus on developing new heuris-

tics. KL-distance will be further studied as an interestingness measure. Additional diversity measures from information theory and statistics will be evaluated. Finally, techniques for attaching domain knowledge to the measures will be investigated, to allow closer mimicking of domain experts' rankings.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 207–216, Montreal, August 1995.

[2] L. Bhandari. Attribute focusing: Machine-assisted knowledge discovery applied to software production process control. In *Knowledge Discovery in Databases: Papers from the 1993 Workshop*, pages 61–69, Menlo Park, CA, 1993. AAAI Press. WS-93-02.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Adavances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI/MIT Press, 1996.

[8] R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 112–117, Montreal, August 1995.

[9] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases: An overview. In *Knowledge Discovery in Databases*, pages 1–27. AAAI/MIT Press, 1991.

[10] 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.

[11] H.J. Hamilton and D.F. Fudger. Measuring the potential for knowledge discovery in databases with DBLearn. *Computational Intelligence*, 11(2):280–296, 1995.

[12] H.J. Hamilton, R.J. Hilderman, and N. Cercone. Attribute-oriented induction using domain generalization graphs. In *Proceedings of the Eighth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'96)*, pages 246–253, Toulouse, France, November 1996.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] J.A. Major and J.J Mangano. Selecting among rules induced from a hurricane database. In *Knowledge Discovery in Databases: Papers from the 1993 Workshop*, pages 28–41, Menlo Park, CA, 1993. AAAI Press. WS-93-02.

[19] C.J. Matheus and G. Piatetsky-Shapiro. Selecting and reporting what is interesting: The kefir application to healthcare data. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 401–419, Menlo Park, CA, 1996. AAAI Press/MIT Press.

[20] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[21] D.J. Randall, H.J. Hamilton, and R.J. Hilderman. Temporal generalization with domain generalization graphs. *International Journal of Pattern Recognition and Artificial Intelligence*. To appear.

[22] W.A. Rosenkrantz. *Introduction to Probability and Statistics for Scientists and Engineers*. McGraw-Hill, 1997.

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

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

[25] P. Smyth and R.M. Goodman. Rule induction using information theory. In *Knowledge Discovery in Databases*, pages 159–176. AAAI/MIT Press, 1991.

[26] Y.Y. Yao, S.K.M. Wong, and C.J. Butz. On information-theoretic measures of attribute importance. In N. Zhong, editor, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, Beijing, China, April 1999.

[27] J.F. Young. *Information theory*. John Wiley & Sons, 1971.