University of Helsinki
Department of Computer Science
Series of Publications C, No. C-1997-66

Similarity of Attributes by External Probes

Gautam Das, Heikki Mannila, and Pirjo Ronkainen

Helsinki, October 1997

Report C-1997-66

University of Helsinki
Department of Computer Science
P. O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, FINLAND

Similarity of Attributes by External Probes

Gautam Das[†], Heikki Mannila[‡], and Pirjo Ronkainen[§]

University of Helsinki, Department of Computer Science, Report C-1997-66

Helsinki, October 1997, 22 pages

Abstract

In data mining, similarity between attributes is one of the central notions. Such a notion can be used to build attribute hierarchies etc. Similarity metrics can be user-defined, but an important problem is defining similarity on the basis of data. Several methods based on statistical techniques exist, but for defining the similarity between A and B they typically consider only the values of A and B, not the other attributes. We describe how a similarity or distance between attributes can be defined by considering the other attributes. The basic idea is that in a 0/1 relation r, two attributes A and B are similar if the subrelations $\sigma_{A=1}(r)$ and $\sigma_{B=1}(r)$ are similar. Similarity between these relations is defined by considering the marginal frequencies of a selected subset of other attributes. Thus for example in a market basket database two products A and B would be deemed similar if the customers buying A and B have similar buying behavior with respect to the other products. We show that the framework produces natural notions of similarity. Empirical results on the Reuters-21578 document dataset show, for example, how natural classifications for countries can be discovered from keyword distributions in documents. The similarity notion is easily computable with scalable algorithms. We also show how the notion can be generalized to arbitrary relations.

Key Words:

Similarity, Distance, Hierarchies, Data mining, Knowledge Discovery

CR Classification:

H.3.3 Information Search and Retrieval

I.2.6 Artificial Intelligence: Learning

[†]University of Memphis, Department of Mathematical Sciences Memphis, TN 38152, USA, dasg@msci.memphis.edu.

[‡]University of Helsinki, Department of Computer Science, P.O. Box 26, FIN-00014 Helsinki, Finland, Heikki.Mannila@cs.helsinki.fi.

[§]University of Helsinki, Department of Computer Science, P.O. Box 26, FIN-00014 Helsinki, Finland, Pirjo.Ronkainen@cs.helsinki.fi.

Contents

1 Introduction **1**

2 Desired properties of distance notions between attributes **3**

3 Internal measures of similarity **5**

4 External measures **7**
4.1 Basic measure 7
4.2 Variations 9
4.3 Constructing external measures from internal measures 10
4.4 Complexity considerations 10
4.5 Introducing new attributes 11

5 Experimental results **11**
5.1 Documents and keywords 11
5.2 Course enrollment data 15
5.3 Telecommunications alarm data 15
5.4 Random data 17

6 Selection of probes **17**

7 General relations **19**

8 Conclusions **20**

References **20**

1 Introduction

Similarity of objects is one of the central concepts in data mining and knowledge discovery: in order to look for patterns or regularities in the data we have to be able to quantify how far from each other two objects in the database are. Recently, there has been considerable interest into defining intuitive and easily computable measures of similarity between complex objects and into using abstract similarity notions in querying databases [1, 3, 8, 11, 14, 19, 21].

Two typical examples of data sets are shown in Figure 1.

Row ID	Chips	Mustard	Sausage	Pepsi	Coca-Cola	Miller	Bud
t_1	1	0	0	0	0	1	0
t_2	1	1	1	1	0	1	0
t_3	1	0	1	0	1	0	0
t_4	0	0	1	0	0	1	0
t_5	0	1	1	1	0	0	1
t_6	1	1	1	0	0	1	0
t_7	1	0	1	1	0	1	0
t_8	0	1	1	0	1	0	1
t_9	1	0	0	1	0	0	1
t_{10}	0	1	1	0	1	1	1

Name	Sex	Age	Occupation	Car Make	Car Year	Accidents
Smith	M	34	engineer	BMW	1990	1
Jones	F	35	doctor	Volvo	1995	0
Edwards	M	45	manager	Cadillac	1992	0
Hill	M	24	engineer	Volvo	1989	2
Stone	F	40	consultant	BMW	1995	1
Cooper	M	39	writer	Ford	1988	0

Figure 1: Two example tables.

In the first example, market basket data, the data objects represent customers in the supermarket, and the columns represent different products. Similar data sets occur in, e.g., information retrieval: there the rows are documents and the columns are (key)words occurring in the documents. In fact, one of our experimental data sets is from this setting. In the second example, the objects are customers of an insurance company and the attributes describe different properties of the customers. The supermarket basket data set consists of binary attributes only, while in the insurance company example there are also attributes

with larger domains (e.g., attribute Age).

When discussing similarity for cases like these one typically talks about similarity of the objects stored in the database, i.e,, similarity between customers. Such a notion can be used in customer segmentation, prediction, and other applications. There is, however, another class of similarity notions. We can discuss similarity between *binary attributes*, or between *attribute values*. Consider the supermarket basket data example. We can define different notions of similarity between *products* by looking at how the customers buy these products. In the insurance company example we could look at the different values of attribute Age and try to find out how similar two values of Age (say 24 and 25) are, when viewed from the point of view of the other data in the table.

Such similarity notions between attributes or attribute values can be used to form hierarchies or clusters of the attributes or their values. A hierarchy can itself give useful insight into the structure of the data, and hierarchies can also be used to produce more abstract rules etc. [20]. Typically, one assumes that the hierarchy is given by a domain expert. This is indeed a good solution, but in several cases the data can be such that no domain expert is available, and hence there is no ready source for the hierarchy.

Given a notion of attribute similarity we could in the supermarket example produce a hierarchy for the products that is not based on existing notions of similarity, but rather is derived from the buying patterns of the customers. For the insurance data example with attribute Age we could derive a discretization of the Age attribute based on the other data in the table.

In this paper we consider the problem of defining similarity notions between attributes or attribute values in large data sets. For simplicity, we consider mostly the case of similarity of binary attributes, as the extension to similarity of attribute values is straightforward.

We discuss two basic approaches for attribute similarity, *internal* and *external* measures. An internal measure of similarity between two attributes A and B is defined purely in terms of the values in the A and B columns, whereas an external measure takes into account also the data in other columns. We contend that external measures can in several cases give more accurate and useful results.

The basic idea in external measures is that in a 0/1 relation r, two attributes A and B are similar if the subrelations $\sigma_{A=1}(r)$ and $\sigma_{B=1}(r)$ are similar. Similarity between these relations is defined by considering the marginal frequencies of a selected subset of other attributes. Thus for example in a market basket database two products A and B would be deemed similar if the customers buying A and B have similar buying behavior with respect to the other products.

The rest of this paper is organized as follows. In Section 2 we describe the properties we require of the measures. Section 3 studies internal measures of similarity, and shows that they fail to satisfy these requirements. In Section 4

we present the definition of external measures, describe some of their properties, and discuss their computation. Section 5 gives experimental results on three real-life datasets: the Reuters-21578 collection [16] of newswire articles, a course enrollment data set, and a telecommunications alarm sequence. In Section 6 we discuss the selection of the probes used in the external measures, while in Section 7 we outline how the same approach can be used for nonbinary data to compute similarity measures between attributes values. Section 8 is a short conclusion.

We conclude this section by introducing some notation. Given a 0/1 relation r with n rows over attributes R, and a selection condition θ on the rows of r, we denote by $f(r, \theta)$ the fraction of rows of r that satisfy θ. For example, $f(r, A = 1 \wedge B = 0)$ gives the fraction of rows with $A = 1$ and $B = 0$. If r is obvious from the context, we just write $f(\theta)$. We use the abbreviation $f(A)$ for $f(A = 1)$, and $f(ABC)$ for $f(A = 1 \wedge B = 1 \wedge C = 1)$. An *association rule* [2] on the relation r is an expression $X \Rightarrow B$, where $X \subseteq R$ and $B \in R$. The *frequency* or *support* of the rule is $f(r, X \cup \{B\})$, and the *confidence* of the rule is $f(r, X \cup \{B\})/f(r, X)$.

2 Desired properties of distance notions between attributes

In this section we outline the intuitions guiding our development of similarity notions. We develop the theory in terms of the complementary notion of distance.

Given a relation schema R with p attributes, denote by $\mathcal{D}$ the class of all 0/1 relations r with over R. We wish to define a *distance* function $d : \mathcal{D} \times R \times R \rightarrow \mathbb{R}$. That is, $d(r, A, B)$ denotes the distance between attributes A and B for relation r. We write just $d(A, B)$ if no confusion is possible.

Preferably, a distance function should have the following properties.

1. It should be *symmetric*, i.e., for all attributes A and B we have $d(A, B) = d(B, A)$.

2. The distance should be 0 if and only if the attributes are identical (from the point of view of the application).

3. It should satisfy the *triangle inequality*, i.e., for all attributes A, B, and C we have $d(A, B) + d(B, C) \geq d(A, C)$.

4. It should be easy and efficient to compute.

5. The function should (in an inverse sense) capture the notion of the *similarity* between attributes. Thus, if two attributes are very similar, then the distance between them should be small (i.e. tend to zero), and vice versa.

The first three requirements are the familiar conditions for a metric. Such a distance measure will be useful in further computations, such as in forming clusters of similar attributes. The requirement of symmetry is quite natural. The same holds for the condition that $d(A, B) = 0$ if and only if A and B are identical from the application's point of view.

The need for triangle inequality is not immediately as obvious. Consider, however, the problem of searching for objects similar to object A from a large class of objects. If we know that A is close to B and B is far from C, then the triangle inequality tells us that A must be far from C, too. This makes it possible to discard C from the search without actually looking at the distance between A and B, a crucial property for efficient access to large collections of objects.[1]

The fourth requirement, efficient computability, should be obvious. For our application of determining similarity or distance between attributes, a quadratic algorithm in the number p of attributes is more or less acceptable, at least provided there are not more than a couple of thousands of attributes. After all, the set of all pairwise distances is quadratic in p. However, a cubic algorithm might already be too slow: we can easily have thousands of products.

The fifth requirement, that the distance function really measures distance or similarity is much harder to formalize. Obviously, the exact similarity notion depends often on the application, and hence it is not even possible to write down a unique set of requirements that would apply to all cases. In the remainder of this section we describe some intuitive properties that a similarity notion should in our opinion have, and formalize these properties.

For the case of market basket data like 0/1 relations, we might require that adding a row t with $t(A) = t(B) = 0$ should not influence the distance between attributes A and B. That is, if $s = r \cup \{t\}$, then $d(r, A, B) = d(s, A, B)$. The intuition here is that rows with A and B equal to zero tell us nothing about the products A and B, and hence there is no change in the distance. We call a distance function d satisfying this condition *oblivious to double zeroes*. Denoting $r_{AB} = \{t \in r \mid t(A) = 1 \vee t(B) = 1\}$, by applying the above condition repeatedly gives us that $d(r, A, B) = d(r_{AB}, A, B)$.

In the next condition, using the terminology of the market basket example, we concentrate on the effect of a choice of the selection of a random subset of buyers. We require that if we choose a random subset of the buyers of a product A and say that all these buy a new product B, then A and B should be similar products. After all, all the differences between A and B are caused by random fluctuations. Formally, consider a relation r with attributes R, and let $A \in R$.

[1] Recently, Fagin and Stockmeyer have shown that for this purpose, it typically suffices that the triangle inequality holds to within a constant multiplicative factor. There also are some interesting applications in which the normal triangle inequality does not hold, but the weakened form does hold [7].

Let $0 < e < 1$. Generate a new attribute B_e, and for a row $t \in r$ assign a value to $t(B_e)$ as follows. If $t(A) = 1$, then $t(B_e) = 1$ with probability e, and $t(B_e) = 0$ otherwise; if $t(A) = 0$, then $t(B_e) = 0$ always.

We require that the attributes A and B_e should be similar. Obviously, random fluctuations make it possible that A and B_e differ significantly, but as the size of the relation r grows, the probability should get smaller. To formalize this, denote by $r \cdot k$ the relation obtained by taking k copies of relation r and applying the above construction to it. We say that the distance function is *asymptotically zero with respect to random subsets*, if

$$\lim_{k \to \infty} d(r \cdot k, A, B_e) = 0$$

in probability, for all e.

Our final property is partially influenced by the above condition, but is more loosely defined. We want the distance between A and B to be small, if the behavior of the buyers of A and B with respect to other products is about the same. What this exactly means will be formalized later; the key issue is that the values of the attributes A and B alone do not suffice to determine the similarity of A and B.

3 Internal measures of similarity

Given a 0/1 relation r with n rows over attributes R, an internal measure of similarity d is a measure whose value $d(A, B)$ for attributes A and B depends only on the values on the A and B columns of r. As there are only 4 possible value combinations,[2] we can express the sufficient statistics for any internal measure by the familiar 2-by-2 contingency table:

	$B = 1$	$B = 0$	sum
$A = 1$	n_{11}	n_{10}	$n_{1\cdot}$
$A = 0$	n_{01}	n_{00}	$n_{0\cdot}$
sum	$n_{\cdot 1}$	$n_{\cdot 0}$	n

Here $n_{\cdot 1} = n_{11} + n_{01}$ etc.

We can measure the strength of association between A and B in numerous ways; see [9] for a compendium of methods. Possibilities include the χ^2 test statistic, measuring the deviation of the observed values n_{ij} from the expected values under the assumption of independence. Note that this measure is not oblivious to double zeroes, for obvious reasons. There exist several modifications of this measure, and some of them are oblivious to double zeroes. Alternatively,

[2] and assuming the order of the rows in r does not make any difference

we could just apply the χ^2 measure to the relation r_{AB} consisting of those rows that have a 1 in at least one of A and B.

If one would like to focus on the positive information, an alternative way would be to use the (relative) size of the symmetric difference of the rows with $A = 1$ and $B = 1$:

$$I_{sd}(A, B) = \frac{f((A = 1 \wedge B = 0) \vee (A = 0 \wedge B = 1))}{f(A = 1 \vee B = 1)} = \frac{f(A) + f(B) - 2f(AB)}{f(A) + f(B) - f(AB)}.$$

This measure is oblivious to double zeroes, as the coefficients in the formula do not include any terms influenced by rows with A and B equal to 0.

In data mining contexts, one might use be tempted to use the confidences of the association rules [2] $A \Rightarrow B$ and $B \Rightarrow A$. For example, one could use the distance function

$$I_c(A, B) = (1 - c(A \Rightarrow B)) + (1 - c(B \Rightarrow A)).$$

We have the following.

Proposition 1 *The functions I_{sd} are I_c are metrics on the set of all attributes. and they are oblivious to double zeroes. They are not asymptotically zero with respect to random subsets.*

The drawback with all the internal measures is that they do not necessarily measure similarity in the desired way. Consider two products A and B, and assume that the same type of customers buy A and B, i.e., it does not make any difference to a customers purchase behavior whether he/she buys A or B, given that he/she buys at least one of them. In this case our intuition is A and B are similar products. However, it might be that A and B are (1) complementary: nobody buys both, (2) independent: buyers of A buy B with the same probability as the nonbuyers of A, or (3) totally correlated: everybody who buys A also buys B, and vice versa. Examples of these situations are shown below.

	$B = 1$	$B = 0$	sum
$A = 1$	0	100	100
$A = 0$	100	0	100
sum	100	100	200

	$B = 1$	$B = 0$	sum
$A = 1$	50	50	100
$A = 0$	50	50	100
sum	100	100	200

	$B=1$	$B=0$	sum
$A=1$	100	0	100
$A=0$	0	100	100
sum	100	100	200

No internal measure can view all these three situations as showing that A and B are similar. There simply is not enough information in the columns of A and B to estimate the type of similarity that we are looking for.

What about the behavior of internal measures with respect to the selection of a random subset? Recall from Section 2 the definition of the new attribute B_e: $B_e = 1$ with probability e if $A = 1$, otherwise $B_e = 0$. Suppose n_A is the number of rows with $A = 1$. Then the contingency table looks as follows.[3]

	$B_e=1$	$B_e=0$	sum
$A=1$	en_A	$(1-e)n_A$	n_A
$A=0$	0	$n-n_A$	$n-n_A$
sum	en_A	$n-en_A$	n

No typical measure of association for 2-by-2 contingency tables will indicate that A and B_e are similar to each other.

Thus it seems that internal measures will not capture the type of similarity of attributes we want. In the next section, we move to considering external measures. For them, in addition to the values of A and B, the values of other attributes also have an influence on the similarity notion.

4 External measures

4.1 Basic measure

Given a 0/1 relation r over attributes R, we aim to measure the similarity of attributes A and B by the similarity of the relations $r_A = \sigma_{A=1}(r)$ and $r_B = \sigma_{B=1}(r)$. This corresponds to our basic intuition: products are similar if the buying behavior of the customers buying them are similar.

Defining similarity between attributes by similarity between relations might seem a step backwards. We wanted to define similarity between two objects of size $n \times 1$ and reduce this to similarity between objects of dimensions $n_A \times p$ and $n_B \times p$, where $n_A = |r_A|$ and $n_B = |r_B|$, and that p is the number of attributes. However, there is a fairly easy way out of this.

Let $\{f_i\}$ be a collection of functions measuring the properties of relations that we are interested in. The value $f_i(r_A)$ of function f_i on relation r_A might measure the frequency of buying Pepsi in the customers in r_A, the average total

[3] Assuming that for exactly en_A rows of the relation we set $B_e = 1$.

volume of sales for the customers in r_A, frequency of visiting the shop, etc. We want to measure the similarity between r_A and r_B by investigating the values $f_i(r_A)$ and $f_i(r_B)$, for all i. As $r_A \neq r_B$, they will probably be different. How do we quantify the differences in such a way that is independent of the actual values of the functions f_i?

We consider relations r_A and r_B to be similar, if the differences between them might arise by chance. That is, we use as a criterion for the difference between $f_i(r_A)$ and $f_i(r_B)$ the probability that this difference would arise also in the case when r_A and r_B are arbitrary random subsets of the relation $r_{AB} = r_A \cup r_B$ (or sizes n_A and n_B). If f_i is a simple function, such as the average of a quantity, we can estimate this probability directly. If f is more complicated,[4] we can use randomization testing: simply generate a sufficiently large number of subsets s_A and s_B of sizes n_A and n_B from r_{AB}, and test how often the difference between $f_i(s_A)$ and $f_i(s_B)$ is at least as big as the difference between $f_i(r_A)$ and $f_i(r_B)$. Such randomization methodology is currently fairly widely used in statistics (see [6] for a very readable introduction).

Here we choose the functions f_i to measure the frequency of a certain probe variable. That is, we select a set $P \subseteq R$ of *probe attributes*, and measure for each $D \in P$ how different the frequency of D is in relations r_A and r_B. A simple test for this is to use the χ^2 test statistic for two proportions, as is widely done in, e.g., epidemiology [18]. Given a probe variable D, the value of the test statistic is after some simplifications

$$F(A, B, D) = \frac{(f(r_A, D) - f(r_B, D))^2 f(r, A) f(r, B)(n-1)}{f(r, D)(1 - f(r, D))(f(r, A) + f(r, B))},$$

where n is again the number of rows in the whole dataset. Note that this is quadratic in the difference in the frequencies of the probe attribute D in the relations r_A and r_B. Further, we have $c(A \Rightarrow D) = f(r_A, D)$, i.e, the confidence of the association rule $A \Rightarrow D$ is the frequency of D in the relation r_A. The other terms in the test statistic take into account the sizes of the relations r_A and r_B.

To obtain the distance measure we sum over all the probes D:

$$F_P(A, B) = \sum_{D \in P} F(A, B, D),$$

and this will be χ^2 distributed with $|P|$ degrees of freedom. If the value $F_P(A, B)$ is large, then A and B are not behaving in the same way with respect to the attributes in P.

One might be tempted to use F_P or some similar notion as an measure of similarity. However, as [9] puts it, "The fact that an excellent test of independence may be based on χ^2 does not at all meant that χ^2, or some simple function

[4] or we have forgotten the test statistics

of it, is an appropriate *measure* of degree of association." [9] goes on to use different models based on predictive ability etc., and we follow this advice.

Therefore we use the measure

$$E_P(A,B) = \sum_{D \in P} |f(r_A, D) - f(r_B, D)|$$

for the distance between attributes A and B. That is, we just take the sum of the absolute values of the differences in the frequencies of D in r_A and r_B. The following proposition is immediate.

Proposition 2 *The measure E_P is a pseudometric on the set R of attributes; i.e., it is symmetric, and satisfies the triangle inequality, but the value of the distance can be 0 even if two attributes are not identical.*

An intuitive interpretation of the function E_P in terms of predictive activity is as follows. Suppose we have to predict the purchase patterns of a customer t on the products in P. That is, for each product $D \in P$ we have to give a probability that customer t will buy product P. Assume that we know only that customer t bought A (or B). Then the distance $E_P(A,B)$ measures the expected total number of errors we make in predicting the purchase pattern of t on the basis of full information about the purchase patterns of persons who bought A (or people who bought B).

A reformulation of the definition of E_P in terms of confidences of rules is

$$E_P(A,B) = \sum_{D \in P} |c(A \Rightarrow D) - c(B \Rightarrow D)|.$$

This turns out to be crucial for the efficient computation of E_P for all pairs of attributes A and B.

Note that for the internal distance I_c defined on the basis of confidences we have $I_c(A,B) = E_{\{A,B\}}(A,B)$.

Our definition of E_P can be considered as an approximation of the distance between the multivariate distributions f and g on P defined by r_A and r_B. Instead of considering the full distributions, we look at the marginal distributions on single variables. There are some connections to the notions of Kullbach-Leibler distance (also known as relative entropy or cross entropy) and the Hellerstein distance between distributions [15, 5].

4.2 Variations

The definition of E_P is by no means the only possible one. Denote $P = \{D_1, \ldots, D_k\}$, the vector $v_{A,P} = [f(r_A, D_1), \ldots, f(r_A, D_k)]$, and similarly the vector $v_{B,P}$. We have at least the following alternative ways of defining a measure of similarity.

1. Instead of using the L_1 metric, we could use the more general L_p metric and define $E_P(A, B)$ as the L_p distance between $v_{A,P}$ and $v_{B,P}$.

2. We can generalize the probe set P to be a set of boolean formulas θ_i, where θ_i is constructed from atomic formulae of the form "$A = 0$" and "$A = 1$" for $A \in R$ by using standard boolean operators. The the distance function is $\sum_i |f(r_A, \theta_i) - f(r_B, \theta_i)|$.

3. We could give a *weight* $w(D)$ to each probe D, and define the distance as $\sum_{D \in P} w(D)(c(A \Rightarrow D) - c(B \Rightarrow D))$. Intuitively, the weight of a probe denotes its significance in the distance calculations. For example, we could set $w(D) = f(r, D)$. With market basket data, this implies that high volume products are given a higher significance.

The use of the L_p metric does not seem to have a large effect on the distances. The importance of the other variations is not immediately obvious and is left for further study.

4.3 Constructing external measures from internal measures

Suppose $I(A, B)$ is an internal distance measure between attributes. Given a collection $P = \{D_1, \ldots, D_k\}$ of probes, we can use I do define an external distance notion as follows. Given attributes A and B, denote $v_{A,P} = [I(A, P_1), \ldots, I(A, P_n)]$, i.e., the vector of internal distances of A to each of the probes. Similarly, let $v_{B,P} = [I(B, P_1), \ldots, I(B, P_n)]$. Then, we can define the external distance between A and B by using any suitable distance notion between the vectors $v_{A,P}$ and $v_{B,P}$: $E_P(A, B) = d(v_{A,P}, v_{B,P})$.

An example of this would be to use the symmetric difference internal distance in building the external distance. However, this measure is/seems to be quite sensitive to actual sizes of the relations r_A and r_B, and hence it does not have the desired properties.

4.4 Complexity considerations

Given a 0/1 relation r with n rows over relation schema R with p attributes, we determine the complexity of computing the distance $E_P(A, B)$ for a fixed set $P \subseteq R$ and all pairs $A, B \in R$. To compute these quantities, we need the frequencies of D in r_A and r_B, for each $D \in P$. That is, we have to know the confidence of the association rules $A \Rightarrow D$ and $B \Rightarrow D$ for each triplet (A, N, D). There are $p^2|P|$ of these triplets. For moderate values of p and for reasonably small probe sets P we can keep all these in memory, and one pass through the database suffices to compute all the necessary counts.

In fact, computing these counts is a special case of the problem of computing all the frequent sets that arises in association rule discovery [2, 4]. If we are not interested in probe sets of small frequency, we can use variations of the Apriori [4] algorithm for the computations needed. This method is fast and scales nicely to very large data sets.

4.5 Introducing new attributes

One of the reasons for considering attribute similarity was the ability to build attribute hierarchies, i.e., to do hierarchical clustering on attributes. In this subsection we sketch how this can be done extremely efficiently using our distance measures.

Suppose we have computed the distances between all pairs of attributes, and assume A and B are the closest pair. Then we can form a new attribute E as the combination of A and B. This new attribute is interpreted as the union of A and B in the sense that $f(E) = f(A) + f(B) - f(AB)$.

Suppose we then want to continue the clustering of attributes. The new attribute E represents a new cluster, so we have to be able to compute the distance of E from the other attributes. For this, we need to be able to compute the confidence of the rules $E \Rightarrow D$ for all probes $D \in P$. This confidence is defined as

$$c(E \Rightarrow D) = \frac{s((A \vee B)D)}{s(A \vee B)} = \frac{s(AD) + s(BD) + s(ABD)}{s(A) + s(B) - s(AB)}.$$

Note that if we have computed the frequent set information for all subsets of R with sufficiently high frequency, then all the terms in the above formula are known. Thus we can continue the clustering without having to look at the original data again.

5 Experimental results

We have used three data sets in our experiments: the so-called Reuters-21578 collection [16] of newswire articles, a database about students and courses at the Computer Science Department of the University of Helsinki, and telecommunication alarm sequence data.

5.1 Documents and keywords

The data set consists of 21578 articles from the Reuters newswire in 1987. Each article has been tagged with keywords. There are altogether 445 different keywords. Over 1800 articles have no keywords at all, one article has 29

keywords, and the average number of keywords per article is slightly over 2. Examples of the keywords are given in Table 1. Thus, in the association rule framework the keywords are the attributes and the articles are the rows. The frequency of a keyword is the fraction of the articles in which the keyword appears.

To test the intuitiveness of the resulting measures, we chose 14 names of countries as our test set:

> Argentina, Brazil, Canada, China, Colombia, Ecuador, France, Japan, Mexico, Venezuela, United Kingdom, USA, USSR, West Germany.

(Recall that the articles are from 1987, so the USSR and West Germany both existed at that time.)

We have lots of background information about the similarities and dissimilarities between these keywords, so testing the naturalness of the results should be relatively easy.

As probe sets, we used several sets of related keywords: see Table 1.

Probe set	Probes
Economic terms	earn, trade, interest
Organizations	ec, opec, worldbank, oecd
Mixed terms	earn, acq, money-fx, crude, grain, trade, interest, wheat, ship, corn, rice

Table 1: The probe sets used for the Reuters data.

Internal vs. external measures We start by comparing the two notions of internal distances I_{sd} and I_c with the external distances E_P for different probe sets P.

Obviously, the actual values of a distance function are irrelevant; we can multiply or divide the distance values by any constant without modifying the properties of the metric. In several applications (such as some types of clustering) what actually matters is only the relative order of the values. That is, as long as for all A, B, C, and D we have $d(A,B) < d(C,D)$ if and only if $d'(A,B) < d'(C,D)$, the measures d and d' behave in the same way.

Figure 2 (top left) shows the distribution of points $(I_{sd}(A,B), I_c(A,B))$ for all pairs (A,B) of countries. We see that the values of the I_c measure tend to be quite close to 2, indicating that the confidences of the rules $A \Rightarrow B$ and $B \Rightarrow A$

are both quite low. Similarly, a large fraction of the values of the I_{sd} measure are close to 1. These phenomena are to be expected, as few of the pairs of countries occur in the same articles.

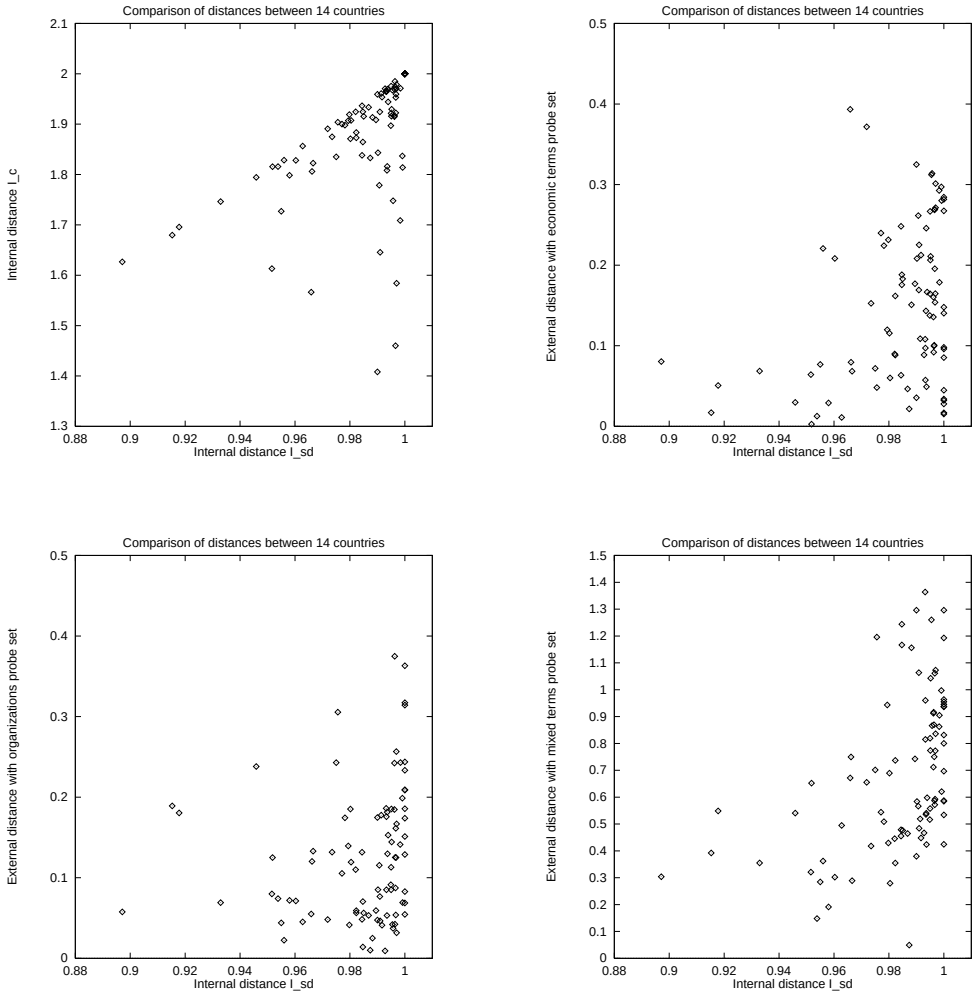

Figure 2: Relationships between internal distances and external distances between the 14 countries for the Reuters data.

The other three plots in Figure 2 show how the internal distance I_{sd} is related to the external distances E_P for three different selections of P. We note that the point clouds are fairly wide, indicating that the measures truly measure different things.

The effect of the probe set P How does the choice of the probe set P effect the measure E_P? Given two sets of probes P and Q which have no relation to each other, there is no reason to assume that $E_P(A, B)$ and $E_Q(A, B)$ would have any specific relationship. Actually, the whole point of constructing external

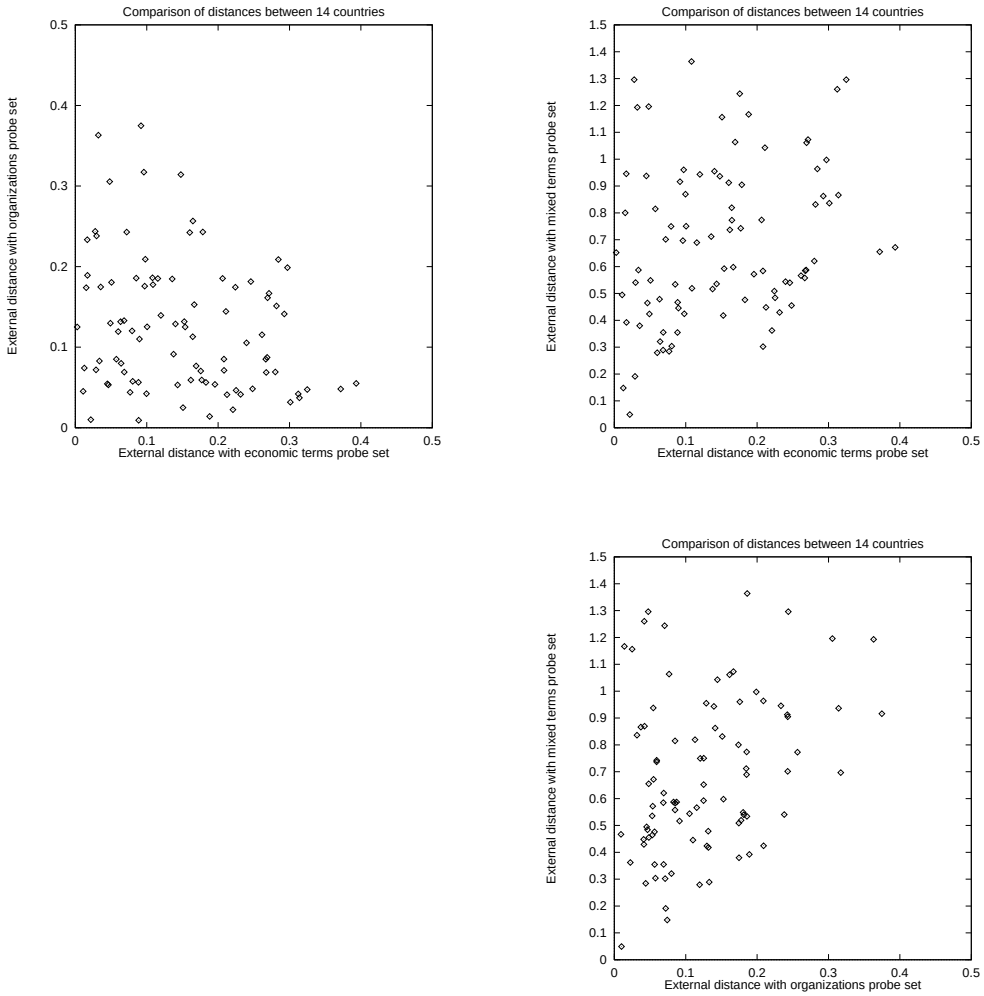

Figure 3: Relationships between external distances for various probe sets between the 14 countries for the Reuters data.

measures between attributes was to let the choice of the probe sets affect the distance!

The plots in Figure 3 show scatterplots of the measures computed using different probe sets. Again, we see great variation, as is to be expected.

Clustering using the external distances To further illustrate the behavior of the external and internal distance functions, we clustered the 14 countries using a standard agglomerative hierarchical clustering algorithm [12, 13]. As distance between clusters, we used the minimum distance between the elements of the clusters. Figure 4 shows the result.

It can be seen that the resulting clustering is quite natural, and corresponds to our views of the geopolitical relationships between the countries.

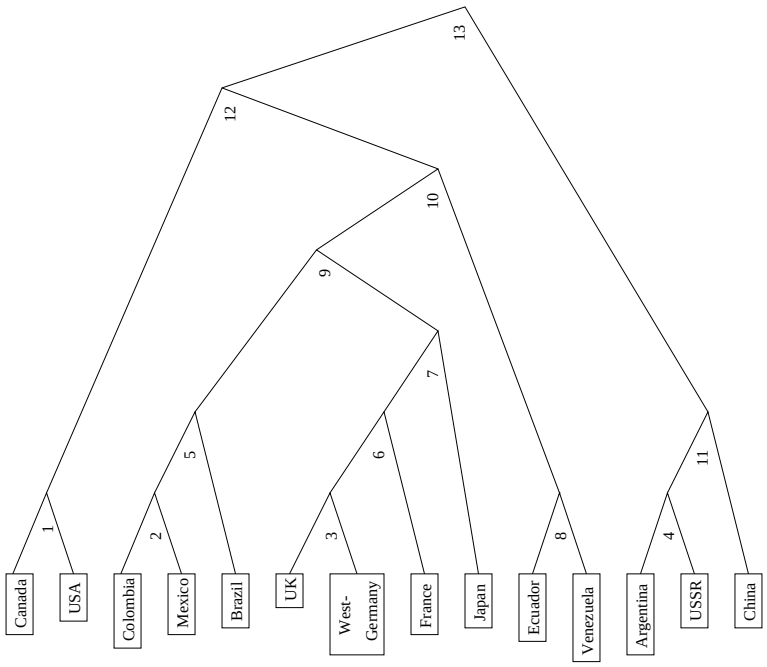

Figure 4: Clustering of countries produced by using the mixed probe set and the minimum distance clustering criterion.

5.2 Course enrollment data

We used course enrollment data from the Department of Computer Science at the University of Helsinki. The data consists of 1050 students and 128 courses; the rows represent students and the columns represent courses.

We made several experiments using with data. Here we report only on a few. We computed the distances between 8 courses from the 3rd and 4th year: Computer Networks; Logic programming; Computer-aided instruction; Object-oriented databases; User interfaces; String algorithms; Design and analysis of algorithms; and Database Systems. As probes we used the 2nd year courses Computer communications; Graphics; Operating systems; and Artificial intelligence.

Computing the distances and using the minimum distance hierarchical clustering methods yields the clustering tree in Figure 5. Again, the results are quite natural.

5.3 Telecommunications alarm data

In telecommunication network management the handling of so called alarm sequences is quite important. In this application, data mining methods have been shown to be useful [10]. Here we describe how our external distance measures

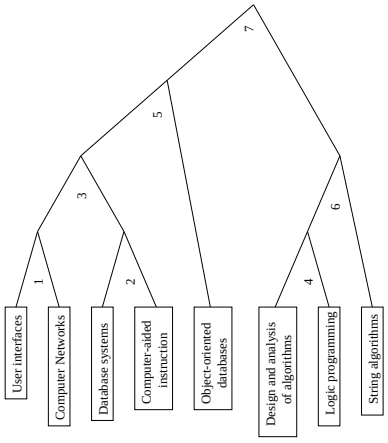

Figure 5: Clustering of the courses produced by the minimum distance clustering criterion.

can be used to detect similarities between alarms.

We analyzed a sequence of 58616 alarms with associated occurrence times from a telephone exchange, collected during a time period of 12 days. There are 247 different alarms; the most frequent alarm occurs more the 8000 times, while some alarms occur only once or twice. We used only the 108 most frequent of the alarms in our experiments.

We transformed this event sequence into a binary matrix as follows. Each of the 247 alarms corresponds to a column. We look at the sequence through windows of width 60 seconds, and slide the window by increments of 1 second through the data. For each row there is a 1 in the column of an alarm, if that alarm occurred within the time window corresponding to the row. There are about 10^6 windows, i.e., rows, as the time period was about 10^6 seconds. (These rows do not have to be constructed explicitly, and one can also use the fact that subsequent windows overlap to get quite efficient algorithms for finding frequent sets in this application [17].)

The alarms are identified by their code numbers, which are assigned by the software developers. The code numbers for the alarms have been assigned so that alarms having the same prefix typically have something to do with each other, at least in the mind of the designer.

We computed the pairwise distances between the 108 alarms, using as probes all the 108 alarms themselves. Below is the list of the 10 smallest pairwise distances.

A	B	E_P
7316	7317	0.130
2241	2478	0.307
7421	7422	0.377
7132	7139	0.407
2064	2241	0.534
7801	7807	0.611
2064	2478	0.688
7410	7411	1.233
7414	7415	1.336
7001	7030	1.421

Even with no knowledge about the actual application, it is obvious that the similarity metric captures some aspects of the underlying structure of the application: the pairs of alarms that are deemed similar by the E_P measure have in most cases numbers that are close to each other. Note that the distance computation uses no information about the actual structure of the network nor about the alarm numbers. On the other hand, two alarms can have short distance for two reasons: either they occur closely together in time (and hence in similar contexts), or they just appear in similar contexts, not necessarily closely together in time. In the above list there are examples of both cases. We omit the details for brevity.

5.4 Random data

As an artificial case, we considered random relations r over attributes $R = \{A_1, \ldots, A_p\}$, where $t(A_i) = 1$ with probability c for some $c < 1/p$ independently of the other entries. That is, all the attributes of r are random and independent from each other, but the density of 1's varies. We computed pairwise external distances for such relations for various probe sets. The results show that the distance for all pairs A_i, A_j is approximately the same. Again, this is the expected behavior.

6 Selection of probes

A proper selection of the probes set is crucial for the usefulness of the distance measure. Our goal in developing the external measure of similarity was that the probes describe the facets of subrelations that the user thinks are important. Optimally, the user should have sufficient domain knowledge to determine which attributes should be used as probes and which are not.

In our experiments it was clear that different probe sets produce different similarity notions. This is as it should be.

We first consider the effect of adding one probe D to the probe set P; denote $Q = P \cup \{D\}$. Recall the formula

$$E_P(A, B) = \sum_{D \in P} |c(A \Rightarrow D) - c(B \Rightarrow D)|.$$

Thus $E_Q(A, B) = E_P(A, B) + |c(A \Rightarrow D) - c(B \Rightarrow D)|$. Assume the frequencies of A and B are relatively high compared to the frequency of D, say $f(D) \leq cf(A)$ and $f(D) \leq cf(B)$ for some $c < 1$. Then the confidence $c(A \Rightarrow D)$ is between 0 and c, and thus the added term in E_Q is below $2c$. Hence adding a probe with a small frequency is not going to produce large changes in the distance measure. Generalizing, if $Q = P \cup \{D_i \mid i = 1, \ldots, k\}$, let

$$c_i = \frac{f(D_i)}{\min(f(A), f(B)}.$$

Then $E_P(A, B) \leq E_Q(A, B) \leq E_P(A, B) + \sum_i 2c_i$.

(If the additional probe D is very frequent, it will also have little effect on the distance. In the extreme case, if $f(D) = 1$, then $E_P(A, B) = E_Q(A, B)$ for $Q = P \cup \{D\}$.)

Thus a very low-frequency probe (compared to other probes) cannot have much effect on the distance. Thus small variations in the choice of probes do not typically produce dramatic changes in the distances.

In the rest of this section we consider a formal version of the probe selection problem. Suppose a domain expert decides that certain relationships between various attributes should hold (for example, between attributes A, B and C, we may want $E_P(A, B) < E_P(A, C)$). It is then interesting to look for a probe set (if one exists) which satisfies such constraints.

Formally, there are several versions of the problem:

1. Given a single user-defined constraint such as $E_P(A, B) < E_P(A, C)$, compute all single probes D that satisfies it.

2. Given a collection of constraints (such as $E_P(A, B) < E_P(A, C)$ AND $E_P(E, F) >= E_P(A, F)$ AND, etc), compute all single probes D that satisfy them.

3. Compute the largest set of probes P which satisfies a single constraint $E_P(A, B) < E_P(A, C)$.

4. Given a collection of constraints, compute all probe sets that satisfy them.

The first problem can be algorithmically solved in an obvious manner as follows: for all possible probes D, check if $|f(r_A, D) - f(r_B, D)| < f(r_A, D) - f(r_B, D)|$.

Consider an analogy from market-basket databases. Coke, Pepsi and beer are all "beverages". However, a domain expert might insist that the distance between coke and Pepsi be smaller than the distance between coke and beer (perhaps because the former are "soft drinks"). A probe such as "pastries" may satisfy this constraint; customers are more likely to eat pastries with Coke/Pepsi than with beer.

Thus, we can envision interactive experimentations with different probe sets that proceed as follows: first "coarse" distances and clusters using a large probe set are computed, then each cluster is refined to obtain finer differences by using different, smaller probe sets. In the above market-basket example, a first coarse clustering may group Coke, Pepsi and beer into one cluster, and later a more refined analysis would reveal that Coke and Pepsi are closer to each other than to beer.

We may use the algorithm for the first problem to solve the second problem. For each constraint, use the first algorithm to compute the set of satisfying single probes. Then compute the intersection of these sets.

The third problem requires different techniques, but surprisingly, the algorithm is quite easy. We start with the entire set of possible probes, $R = \{D_1, D_2, \ldots, D_p\}$. Given attribute A, recall the vector $v_{A,R} = [f(D_1, r_A), \ldots, f(D_p, r_A)]$. Clearly $E_R(A, B)$ is the L_1 distance between $v_{A,R}$ and $v_{B,R}$.

Let $v_{AB} = [|f(D_1, r_A) - f(D_1, r_B)|, \ldots, |f(D_p, r_A) - f(D_p, r_B)|]$. Note that the sum of the components (which are all positive) of this vector is $E_R(A, B)$. Similarly let $v_{AC} = [|f(D_1, r_A) - f(D_1, r_C)|, \ldots, |f(D_p, r_A) - f(D_p, r_C)|]$.

Now compute the vector $v = v_{AB} - v_{AC}$. The components of v may be either positive or negative. It is easy to see that the largest probe set that we are seeking is the largest subset of the components of v whose sum is negative. This may be computed by sorting v and taking the longest prefix whose sum is negative.

As for the fourth problem, the naive way is to compute all possible satisfying probe sets for each constraint, and then computing the intersection of these families. However we are investigating more efficient techniques (such as incremental techniques, constraint graph techniques, etc) for this problem.

7 General relations

In this section we briefly outline how the idea of external distances can be used to define similarity notions between the values in the domain of an attribute in an arbitrary relation.

Given a relation schema S and a relation s over S, consider a fixed attribute $A \in S$ and two values a and b from the domain of A. The basic idea is again that a and b are similar, if the subrelations $s_a = \sigma_{A=a}(s)$ and $s_b = \sigma_{A=b}(s)$ are

similar.

To determine the similarity of the subrelations s_a and s_b, we apply a set of probe functions f_i, and combine the answers by using, e.g., the L_1 metric: $E(a,b) = \sum_i |f_i(s_a) - f_i(s_b)|$. We omit the details for brevity.

8 Conclusions

Similarity is an important concept for advanced retrieval and data mining applications. In this paper we considered the problem of defining an intuitive similarity or distance notion between attributes of a 0/1 relation. We listed certain natural conditions that such a measure should satisfy, and showed that the typical internal similarity measures do not suffice to capture these intuitions.

We then introduced the notion of an external measure between attributes A and B, defined by looking at the values of probe functions on subrelations defined by A and B. We showed that this measure has the desired properties, and that it can be computed fast. We also outlined how the use of association rule algorithms can help in building hierarchies based on this notion.

We gave experimental results on three different real-life data sets and showed that the similarity notion indeed captures some of the true similarities between the attributes.

There are several open problems. One is semiautomatic probe selection: how can we provide guidance to the user in selecting the probe sets. The other is the use of hierarchies generated by this method in rule discovery: what properties will the discovered rules have. Also, the connection to statistical tests sketched in the beginning of Section 4 needs to be strengthened, and the relationships to mutual entropy and the Hellerstein distance are worth studying. Further experimentation is also needed to determine the usability of external distances in various application domains.

References

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficiency similarity search in sequence databases. In D. B. Lomet, editor, *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO'93)*, pages 69 – 84, Chicago, Illinois, USA, Oct. 1993. Lecture Notes in Computer Science, Vol. 730, Springer-Verlag.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'93)*, pages 207 – 216, Washington, D.C., USA, May 1993. ACM.

[3] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In U. Dayal, P. M. D. Gray, and S. Nishio, editors, *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, pages 490 – 501, Zürich, Switzerland, 1995.

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307 – 328. AAAI Press, Menlo Park, CA, 1996.

[5] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, Dec. 1989.

[6] P. R. Cohen. *Empirical Methods for Artificial Intelligence*. MIT Press, 1995.

[7] R. Fagin and L. Stockmeyer. Relaxing the triangle inequality in pattern matching. Report RJ 10031, IBM Research Division, Almaden Research Center, San Jose, California, USA, June 1996.

[8] D. Q. Goldin and P. C. Kanellakis. On similarity queries for time-series data: Constraint specification and implementation. In U. Montanari and F. Rossi, editors, *Proceedings of the 1st International Conference on Principles and Practice of Constraint Programming (CP'95)*, Cassis, France, Sept. 1995. Lecture Notes in Computer Science, Vol. 976, Springer-Verlag.

[9] L. A. Goodman and W. H. Kruskal. *Measures of Association for Cross Classifications*. Springer-Verlag, 1979.

[10] K. Hätönen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen. Knowledge discovery from telecommunication network alarm databases. In *Proceedings of the 12th International Conference on Data Engineering (ICDE'96)*, pages 115 – 122, New Orleans, Louisiana, Feb. 1996.

[11] H. Jagadish, A. O. Mendelzon, and T. Milo. Similarity-based queries. In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'95)*, pages 36 – 45, San Jose, California, USA, May 1995. ACM.

[12] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[13] L. Kaufman and P. Rousseauw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley Inc., New York, NY, 1990.

[14] A. J. Knobbe and P. W. Adriaans. Analysing binary associations. In E. Simoudis, J. W. Han, and U. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 311 – 314, Portland, Oregon, USA, Aug. 1996. AAAI Press.

[15] Kullbach and Leibler. On information theory and sufficiency. In *Annals of Mathematical Statistics, Volume 22*. 1951.

[16] D. Lewis. The reuters-21578, distribution 1.0. http://www.research.att.com/~lewis/reuters21578.html, 1997.

[17] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3), 1997.

[18] O. S. Miettinen. *Theoretical Epidemiology*. John Wiley Inc., New York, NY, 1985.

[19] D. Rafiei and A. Mendelzon. Similarity-based queries for time series data. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 26(2):13–25, May 1997.

[20] R. Srikant and R. Agrawal. Mining generalized association rules. In U. Dayal, P. M. D. Gray, and S. Nishio, editors, *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, pages 407 – 419, Zürich, Switzerland, 1995. Morgan Kaufmann.

[21] D. A. White and R. Jain. Algorithms and strategies for similarity retrieval. Technical Report VCL-96-101, Visual Computing Laboratory, University of California, San Diego, 9500 Gilman Drive, Mail Code 0407, La Jolla, CA 92093-0407, July 1996.