Similarity of Attributes by External Probes

Gautam Das
University of Memphis
Department of Mathematical Sciences
Memphis TN 38152, USA
dasg@msci.memphis.edu

Heikki Mannila and **Pirjo Ronkainen**
University of Helsinki
Department of Computer Science
P.O. Box 26, FIN-00014 Helsinki, Finland
{*Heikki.Mannila,Pirjo.Ronkainen*}*@cs.helsinki.fi*

Abstract

In data mining, similarity or distance 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. For defining the similarity between two attributes A and B they typically consider only the values of A and B, not the other attributes. We describe how a similarity notion between attributes can be defined by considering the values of 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 the two relations is defined by considering the marginal frequencies of a selected subset of other attributes. 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.

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 (Agrawal, Faloutsos, & Swami 1993; Agrawal *et al.* 1995; Goldin & Kanellakis 1995; Jagadish, Mendelzon, & Milo 1995; Knobbe & Adriaans 1996; Rafiei & Mendelzon 1997; White & Jain 1996).

A typical data set is shown in Figure 1. In this 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. (Due to the lack of space, we use as only one example relation, containing just binary attributes, and do not discuss how the ideas can be adapted to attributes with larger domains.)

When discussing similarity one typically talks about similarity of the objects stored in the database, e.g., similarity between customers. Such a notion can be used in customer segmentation, prediction, and other applications. There is, however, another class of similarity notions, *similarity between (binary) attributes.* For example, in the supermarket basket data we can define different notions of similarity between products by looking at how the customers buy these products. A simple example is that Coke and Pepsi can be deemed similar attributes, if the buying behaviour of buyers of Coke and buyers of Pepsi is similar.

Similarity notions between attributes can be used to form hierarchies or clusters of attributes. 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. (Han, Cai, & Cercone 1992; Srikant & Agrawal 1995). 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.

In this paper we consider the problem of defining similarity between attributes in large data sets. 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 certain other columns, the *probe* columns. We contend that external measures can in several cases give additional insight into the data. Note that there is no single correct notion of similarity, and varying the probe sets makes it possible to obtain similarity notions reflecting

Row ID	Chips	Mustard	Sausage	Pepsi	Coke	Miller	Bud
t_1	1	0	0	0	1	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	1	1	1	1	0	0	1
t_6	1	1	1	0	0	1	0
t_7	1	0	1	1	0	1	0

Figure 1: An example data set.

different viewpoints.

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 $fr(r,\theta)$ the fraction of rows of r that satisfy θ. For example, $fr(r, A = 1 \wedge B = 0)$ gives the fraction of rows with $A = 1$ and $B = 0$. If r is obvious from the context, we just write $fr(\theta)$. We use the abbreviation $fr(A)$ for $fr(A = 1)$, and $fr(ABC)$ for $fr(A = 1 \wedge B = 1 \wedge C = 1)$. An *association rule* (Agrawal, Imielinski, & Swami 1993) 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 $fr(r, X \cup \{B\})$, and the *confidence* of the rule is $conf(X \Rightarrow B) = fr(r, X \cup \{B\})/fr(r, X)$.

Internal measures of similarity

Given a 0/1 relation r with n rows over attributes R, an internal measure of similarity d_I is a measure whose value $d_I(A, B)$ 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,[1] we can express the sufficient statistics for any internal measure by the familiar 2-by-2 contingency table.

We can measure the strength of association between A and B in numerous ways; see (Goodman & Kruskal 1979) for a compendium of methods. Possibilities include the χ^2 test statistic, which measures the deviation of the observed values from the expected values under the assumption of independence. There exist several modifications of this measure.

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 $B = 1$:

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

In data mining contexts, one might use be tempted to use the confidences of the association rules $A \Rightarrow B$ and $B \Rightarrow A$. For example, one could use the distance function $d_{I_{conf}}(A, B) = (1 - conf(A \Rightarrow B)) + (1 - conf(B \Rightarrow A))$. The functions $d_{I_{sd}}$ are $d_{I_{conf}}$ are both metrics on the set of all attributes.

Internal measures are useful in several applications; however, as the similarity between A and B is based on solely the values in columns A and B, they cannot reflect certain types of similarity.

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

External measures of similarity

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 $r_B = \sigma_{B=1}(r)$. 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, Pepsi and Coke could be deemed similar if the customers buying them have similar buying behavior with respect to the other products.

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 $n_A \times m$ and $n_B \times m$, where $n_A = |r_A|$ and $n_B = |r_B|$, and that m is the number of attributes. However, we will see in the sequel that for the similarity of relations we can rely on some well-established notions.

Consider a set of attributes P, the *probe attributes*, and assume that the relations r_A and r_B have been projected to this set. These relations can be viewed as defining two multivariate distributions g_A and g_B on $\{0,1\}^P$: given an element $\bar{x} \in \{0,1\}^P$, the value $g_A(x)$ is the relative frequency of $\bar{x}$ in the relation r_A.

One widely used distance notion between distributions is the Kullbach-Leibler distance (also known as relative entropy or cross entropy) (Kullbach & Leibler 1951; Basseville 1989):

$$re(g_A, g_B) = \sum_{\bar{x}} g_A(\bar{x}) \log \frac{g_A(\bar{x})}{g_B(\bar{x})},$$

or the symmetrized version of it: $re(g_A, g_B) + re(g_B, g_A)$. The problem with this measure is that the sum has $2^{|P|}$ elements, so direct computation of the measure is not feasible for larger sets P. Therefore, we look for simpler measures that would still somehow reflect the distance between g_A and g_B.

One way to remove the exponential dependency on $|P|$ is to look at only a single attribute $D \in P$ at a time. That is, we define the distance $d_{F,P}(A, B)$ as

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

where $F(A, B, D)$ is some measure of how closely A and B agree with respect to the probe attribute D. Of course, this simplification looses power compared to the full relative entropy measure. Still, we suggest the sum above as the external distance between A and B, given the set P of probe attributes and a measure of distance of A and B with respect to an attribute D. If the value $d_{F,P}(A, B)$ is large, then A and B are not behaving in the same way with respect to the attributes in P.

There are several possibilities for the choice of the function $F(A, B, D)$. We can measure 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 (Miettinen 1985), and also in data mining (Brin, Motwani, & Silverstein 1997). Given a probe variable D, the value of the test statistic $F_\chi(A, B, D)$ is after some simplifications

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

where n is the number of rows in the whole dataset r. To obtain the distance measure we sum over all the probes D: $d_{\chi,P}(A, B) = \sum_{D \in P} F_\chi(A, B, D)$. This measure is χ^2 distributed with $|P|$ degrees of freedom.

One might be tempted to use $d_{\chi,P}$ or some similar notion as a measure of similarity. However, as (Goodman & Kruskal 1979) 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 of it, is an appropriate *measure* of degree of association." One well-known problem with the χ^2 measure is that it is very sensitive to cells with small counts; see (Guo 1997) for a discussion of the same problems in the context of medical genetics.

An alternative is to use the term from the relative entropy formula:

$$F(A, B, D) = fr(r_A, D) \log(fr(r_A, D)/fr(r_B, D)).$$

However, experiments show that this is also quite sensitive to small fluctuations in the frequency of the attribute D in relations r_A and r_B.

A more robust measure is $F_{fr}(A, B, D) = |fr(r_A, D) - fr(r_B, D)|$: if D is a rare attribute, then $F_{fr}(A, B, D)$ cannot obtain a large value. In our experiments we used this variant, and the resulting measure for the distance between attributes A and B is

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

The measure $d_{fr,P}$ 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. A reformulation of the definition of $d_{fr,P}$ in terms of confidences of rules is $d_{fr,P}(A, B) = \sum_{D \in P} |conf(A \Rightarrow D) - conf(B \Rightarrow D)|$. This turns out to be crucial for the efficient computation of $d_{fr,P}$ for all pairs of attributes A and B. Note that for the internal distance $d_{I_{conf}}$ defined on the basis of confidences we have $d_{I_{conf}}(A, B) = d_{fr,\{A,B\}}(A, B)$.

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

1. Instead of using the L_1 metric, we could use the more general L_p metric and define $d_{fr,P}(A, B)$ 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 formulae 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. Then the distance function is $\sum_i |fr(r_A, \theta_i) - fr(r_B, \theta_i)|$.

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

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

Complexity considerations Given a 0/1 relation r with n rows over relation schema R with m attributes, we determine the complexity of computing the distance $d_{fr,P}(A, B)$ 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, B, D). There are $m^2|P|$ of these triplets. For moderate values of m 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 computing all the frequent sets that arises in association rule discovery (Agrawal, Imielinski, & Swami 1993; Agrawal *et al.* 1996). If we are not interested in probe attributes of small frequency, we can use variations of the Apriori (Agrawal *et al.* 1996) algorithm. This method is fast and scales nicely to very large data sets.

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. Now we sketch how this can be done 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 $fr(E) = fr(A) + fr(B) - fr(AB)$.

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

$$\begin{aligned} conf(E \Rightarrow D) &= \frac{fr((A \vee B)D)}{fr(A \vee B)} \\ &= \frac{fr(AD)+fr(BD)-fr(ABD)}{fr(A)+fr(B)-fr(AB)}. \end{aligned}$$

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.

Experimental results

We have used three data sets in our experiments: the so-called Reuters-21578 collection (Lewis 1997) of newswire articles, a database about students and courses at the Computer Science Department of the University of Helsinki, and telecommunication alarm sequence data.

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. 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.[2] 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: economic terms (earn, trade, interest), organizations (ec, opec, worldbank, oecd), and mixed terms (earn, acq, money-fx, crude, grain, trade, interest, wheat, ship, corn, rice).

[2] In 1987, both the USSR and West Germany still existed.

Internal vs. external measures We start by comparing the two notions of internal distances $d_{I_{sd}}$ and $d_{I_{conf}}$ with the external distances $d_{fr,P}$ 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 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 $(d_{I_{sd}}(A,B), d_{I_{conf}}(A,B))$ for all pairs (A,B) of countries. We see that the values of the $d_{I_{conf}}$ measure tend to be quite close to 2, indicating that the confidences of the rules $A \Rightarrow B$ and $B \Rightarrow A$ are both low. Similarly, a large fraction of the values of the $d_{I_{sd}}$ measure are close to 1. These phenomena are to be expected, as few of the pairs of countries occur in the same articles.

The other three plots in Figure 2 show how the internal distance $d_{I_{sd}}$ is related to the external distances $d_{fr,P}$ 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 $d_{fr,P}$? Given two sets of probes P and Q which have no relation to each other, there is no reason to assume that $d_{fr,P}(A,B)$ and $d_{fr,Q}(A,B)$ would have any specific relationship. Actually, the whole point of constructing external measures between attributes was to let the choice of the probe sets affect the distance!

Figure 3 shows scatter plots of the measures computed using different probe sets. Again, we see great variation, as is to be expected.

Clustering using the internal and 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 (Jain & Dubes 1988; Kaufman & Rousseauw 1990). As a distance between clusters, we used the minimum distance between the elements of the clusters. Figure 4 shows two clusterings produced by using $d_{fr,P}$, as well as a clustering produced by using $d_{I_{sd}}$.

The clusterings resulting from external distances are quite natural, and correspond mainly to our views of the geopolitical relationships between the countries. The clusterings are different, reflecting the different probe sets. The flexibility of the external measure is that the probes can be used to define the viewpoint. A slightly suprising feature in the leftmost clustering on Figure 4 is that Argentina and the USSR are relatively close to each other. In the data set most of the articles about Argentina and the USSR were about grain, explaining

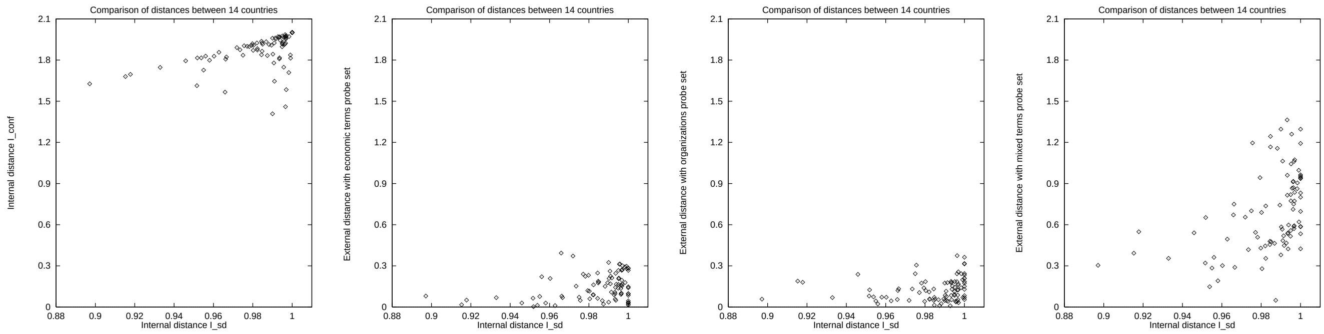

Figure 2: Relationships between internal distances and external distances between the 14 countries for the Reuters data.

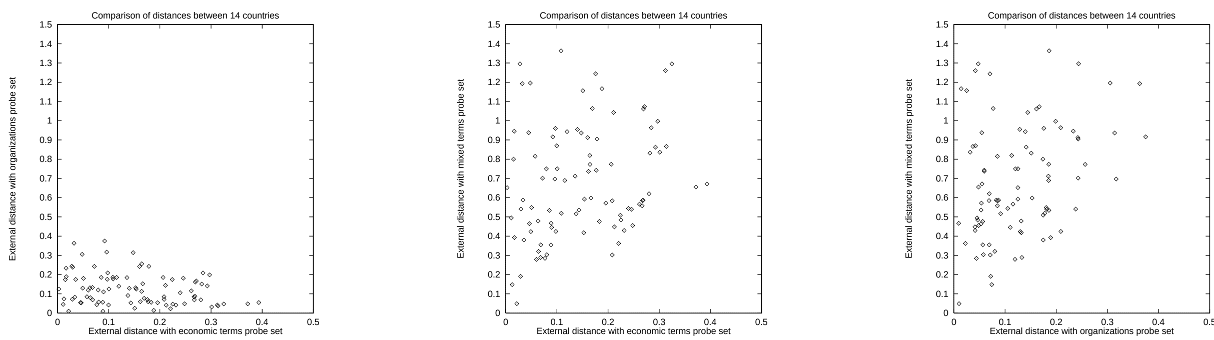

Figure 3: Relationships between external distances for various probe sets between the 14 countries for the Reuters data.

the phenomenon. The third clustering based on internal distance reflects mainly the number of co-occurrence of the keywords in the articles, whereas the clusterings based on external measures weigh the co-occurrence with the probe attributes. Lack of space prevents us from discussing the qualities of clusterings in detail in this version.

Course enrollment data As the second data set 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 with the data. We computed, for example, 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.

Telecommunications alarm data In telecommunication network management the handling of so called

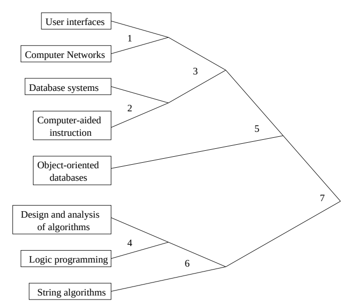

Figure 5: Clustering of the courses produced by the minimum distance clustering criterion.

alarm sequences is quite important. In this application, data mining methods have been shown to be useful (Hätönen *et al.* 1996). Here we describe how our external distance measures 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, and transformed the event sequence into a binary matrix as follows. Each

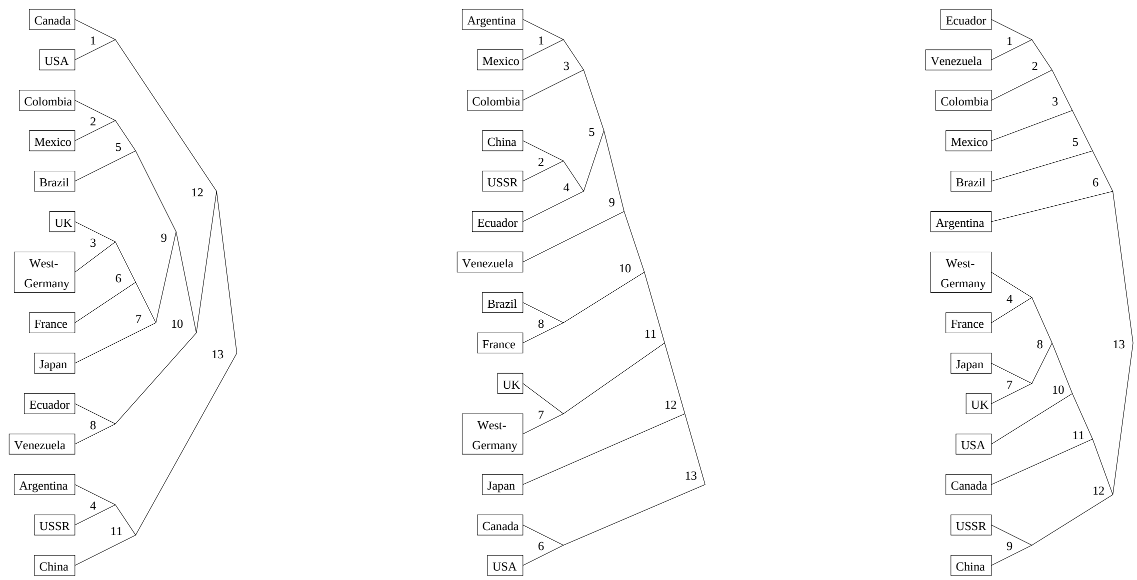

Figure 4: Clustering of countries produced with the minimum distance clustering criterion by using $d_{fr,P}$ with the mixed probe set (left) and the economic terms probe set (middle), as well as by using $d_{I_{sd}}$ (right).

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. See (Mannila, Toivonen, & Verkamo 1997) for efficient algorithms for finding frequent sets in this application.

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. Following is the list of the 10 smallest distances.

A	B	$d_{fr,P}$
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 $d_{fr,P}$ 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. 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 list there are examples of both cases. We omit the details for brevity.

Random data As an artificial case, we considered random relations r over attributes $R = \{A_1, \ldots, A_m\}$, where $t(A_i) = 1$ with probability c independently of the other entries. That is, all the attributes of r are random and independent from each other. We computed pairwise external distances for such relations for various probe sets. The results show that the distance for all pairs is approximately the same. Again, this is the expected behavior.

Selection of probes 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.

The experiments showed clearly that different probe sets produce different similarity notions. This is as it should be: the probe set defines the point of view from which similarity is judged, and thus different selections produces different measures. There is no single optimal solution to the probe selection problem. In the full paper we describe some strategies that can be used to help the user in the selection of probes.

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 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 also outlined how the use of association rule algorithms can help in building hierarchies based on this notion. After that 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 needs to be strengthened, and the relationships to mutual entropy and the Hellerstein distance are worth studying. Moreover, it needs to be shown how good an approximation of the Kullbach-Leibler distance our measure is. Further experimentation is also needed to determine the usability of external distances in various application domains. Finally, extending the method for distances between attribute values is worth investigating.

Acknowledgments We thank Usama Fayyad for useful comments on an earlier version of this paper.

References

Agrawal, R.; Lin, K.-I.; Sawhney, H. S.; and Shim, K. 1995. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, 490 – 501. Zürich, Switzerland: Morgan Kaufmann.

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press. 307 – 328.

Agrawal, R.; Faloutsos, C.; and Swami, A. 1993. Efficiency similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO'93)*, 69 – 84. Chicago, Illinois: Lecture Notes in Computer Science, Vol. 730.

Agrawal, R.; Imielinski, T.; and Swami, A. 1993. Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'93)*, 207 – 216. Washington, D.C.: ACM.

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

Brin, S.; Motwani, R.; and Silverstein, C. 1997. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'97)*, 265 – 276. Tucson, Arizona: ACM.

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

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

Guo, S.-W. 1997. Linkage disequilibrium measures for fine-scale mapping: A comparison. *Human Heredity* 47(6):301 – 314.

Han, J.; Cai, Y.; and Cercone, N. 1992. Knowledge discovery in databases: an attribute-oriented approach. In *Proceedings of the 18th International Conference on Very Large Data Bases (VLDB'92)*, 547 – 559. Vancouver, Canada: Morgan Kaufmann.

Hätönen, K.; Klemettinen, M.; Mannila, H.; Ronkainen, P.; and Toivonen, H. 1996. Knowledge discovery from telecommunication network alarm databases. In *Proceedings of the 12th International Conference on Data Engineering (ICDE'96)*, 115 – 122. New Orleans, Louisiana: IEEE Computer Society Press.

Jagadish, H.; Mendelzon, A. O.; and Milo, T. 1995. Similarity-based queries. In *Proceedings of the 14th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'95)*, 36 – 45. San Jose, California: ACM.

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

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

Knobbe, A. J., and Adriaans, P. W. 1996. Analysing binary associations. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, 311 – 314. Portland, Oregon: AAAI Press.

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

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

Mannila, H.; Toivonen, H.; and Verkamo, A. I. 1997. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery* 1(3):259 – 289.

Miettinen, O. S. 1985. *Theoretical Epidemiology*. New York, NY: John Wiley Inc.

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

Srikant, R., and Agrawal, R. 1995. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, 407 – 419. Zürich, Switzerland: Morgan Kaufmann.

White, D. A., and Jain, R. 1996. Algorithms and strategies for similarity retrieval. Technical Report VCL-96-101, Visual Computing Laboratory, University of California, San Diego.