Measuring the Similarity for Heterogenous Data: An Ordered Probability-Based Approach

SiQuang Le and TuBao Ho

Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292 Japan
{quang, bao}@jaist.ac.jp

Abstract. In this paper we propose a solution to the similarity measuring for heterogenous data. The key idea is to consider the similarity of a given attribute-value pair as the probability of picking randomly a value pair that is less similar than or equally similar in terms of order relations defined appropriately for data types. Similarities of attribute value pairs are then integrated into similarities between data objects using a statistical method. Applying our method in combination with distance-based clustering to real data shows the merit of our proposed method.

Key words: data mining, similarity measures, heterogenous data, order relations, probability integration.

1 Introduction

Measuring similarities between data objects is one of primary objectives in data mining systems. It is essential for many tasks, such as finding patterns or ranking data objects in databases with respect to queries. Moreover, measuring similarities between data objects affects significantly effectiveness of distance-based data mining methods, e.g. distance-based clustering methods and nearest neighbor techniques.

To measure the similarity for heterogenous data that comprise different data types, quantitative data, qualitative data, item set data, etc., has been a challenging problem in data mining due to natural differences among data types. Two primary tasks of the problem are (1) determining the same essential (dis)similarity measures for different data types and (2) integrating properly (dis)similarities of attribute value pairs into similarities between data objects. The first task is rather difficult because each data type has its own natural properties that leads to itself particular similarity measures. For example, the similarity between two continuous value are often considered as their absolute difference meanwhile the similarity between two categorical values is simply identity or non-identity of these two values. Thus, it is hard to define one proper similarity measure

for all data types. Further, similarity measures for some data types are so poor due to the poorness of data structure (e.g., categorial, item set).

To date, a few similarity measuring methods for heterogeneous data have been proposed [**?**,**?**,**?**,**?**,**?**,**?**]. Their approach is to apply the same similarity measure scheme for different data types, and then, dissimilarity between two data objects is assigned by adding linearly dissimilarities of their attribute value pairs or by distance Minkowski. In [**?**,**?**,**?**], the authors measure the similarity between two values based on three factors: *position*, *span*, and *content* where *position* indicates the relative position of two attribute values, *span* indicates the relative sizes of attribute values without referring to common parts, and *content* is a measure of the common parts between two values. It is obvious that similarities of value pairs of different data types have the same meaning since they are all based on these three factors. However, it is not hard to see that these three factors are not always applicable or suitable for all data types. For example, *position* arises only when the values are quantitative. Similar to that, methods of [**?**,**?**,**?**] take sizes of union (the joint operation $\otimes$) and intersection (the meet operation $\oplus$) of two values into account of measuring their similarity. Obviously, these two operators are not always suitable for all data types. For example, the intersection is not suitable for continuous data since continuous values are often different and therefore, the intersection of two continuous values seems to be always empty. In short, the methods [**?**,**?**,**?**,**?**,**?**,**?**] are based on factors or operators that are required to be suitable for all data types. However, due to the nature difference of data types, the factors or operators are hard to exist or not discovered yet.

In this paper, we address the similarity measuring problem for heterogeneous data by a probability-based approach. Our intuition in similarity for a value pair is that the more number of pairs that are less than or equally similar, the greater their similarity. Based on the idea, we define the similarity of one value pair as the probability of picking randomly a value pair that is less similar than or equally similar in terms of order relations defined appropriately for each data types. By this way, we can obtain the same meaning similarities between values of different data types meanwhile each of the similarities is still based on particular properties of the corresponding data type. After that, similarities of attribute value pairs of two objects are then integrated using a statistical method to assign the similarity between them.

This paper is organized as follows: the ordered probability-based similarity measuring method for single attributes is described in Section 2.

The integration methods and an example are given in Section 3 and Section 4. In section 5, we investigate characteristics of the proposed method. Next, complexity evaluation and experiment evaluations in combination with clustering methods to real data sets is shown in Section 6. Conclusions and further works are discussed in the last section.

2 Ordered probability-based similarity measure for an attribute

Let $A_1, \ldots, A_m$ be m data attributes where A_i can be any data type such as quantitative data, qualitative data, interval data, item set data, etc. Let $D \subseteq A_1 \times \ldots \times A_m$ be a data set and $\mathbf{x} = (x_1, \ldots, x_m), x_i \in A_i$ be a data object of D. For each attribute A_i, denote $\preceq_i$ an order relation on A_i^2 where $(x_i', y_i') \preceq_i (x_i, y_i)$ implies that value pair (x_i', y_i') is less similar than or equally similar to value pair (x_i, y_i).

2.1 Ordered probability-based similarity measure

The first task of measuring similarity for heterogeneous data is to determine similarity measures for value pairs of each attribute. We define the ordered probability-based similarity for value pair (x_i, y_i) of attribute A_i as follows:

Definition 1. *The ordered probability-based similarity between two values x_i and y_i of attribute A_i with respect to order relation $\preceq_i$, denoted by $S_{\preceq_i}(x_i, y_i)$, is the probability of picking randomly a value pair of A_i that is less similar than or equally similar to (x_i, y_i)*

$$S_{\preceq_i}(x_i, y_i) = \sum_{(x_i', y_i') \preceq_i (x_i, y_i)} p(x_i', y_i')$$

where $p(x_i', y_i')$ is the probability of picking value pair (x_i', y_i') of A_i.

Definition 1 implies that the similarity of one value pair depends on both the number of value pairs that are less similar than or equally similar and probabilities of picking them. It is obvious that the more number of less than or equally similar value pair one pair has, the more similar they are.

As it can be induced from Definition 1, similarities of value pairs do not depend on data types. They are based only on order relations and probability distributions of value pairs. Hence, similarities of value pairs have the same meaning regardless of their data types. In other hand, each

similarity is based on an order relation built properly for each data type. Thus, the similarity measure still reserved particular properties of this data type.

2.2 Order relations for real data

In the following we define order relations of some common real data types, e.g. continuous data, interval data, ordinal data, categorical data, and item set data.

- **Continuous data:** A value pair is less similar or equally similar to another value pair if and only if the absolute difference of the first pair is greater than or equal to that of the second pair.

$$(x', y') \preceq (x, y) \Leftrightarrow |x' - y'| \geq |x - y|$$

- **Interval data:** A value pair is less similar than or equally similar to another value pair if and only if the proportion between the intersection interval and the union interval of the first pair is smaller than or equal to that of the second pair.

$$(x', y') \preceq (x, y) \Leftrightarrow \frac{|x' \cap y'|}{|x' \cup y'|} \leq \frac{|x \cap y|}{|x \cup y|}$$

- **Ordinal data:** A value pair is less similar than or equally similar to another value pair if and only if the interval between two values of the first pair contains that of the second pair:

$$(x', y') \preceq (x, y) \Leftrightarrow [x'..y'] \supseteq [x..y]$$

- **Categorical data:** A value pair is less similar than or equally similar to another value pair if and only if either they are identical or values of the first pair are not identical meanwhile those of the second pair are:

$$(x', y') \preceq (x, y) \Leftrightarrow \begin{cases} x' = x, y' = y \\ x' \neq y', x = y \end{cases}$$

- **Item set data:** Following the idea of Geist [**?**], the order relation for item set value pairs that come from item set M is defined as follows:

$$(X, Y), (X', Y') \in M^2 : (X', Y') \preceq (X, Y) \Leftrightarrow \begin{cases} X' \cap Y' \subseteq X \cap Y \\ \overline{X}' \cap \overline{Y}' \subseteq \overline{X} \cap \overline{Y} \\ X' \cap \overline{Y} \supseteq X \cap \overline{Y} \\ \overline{X}' \cap Y' \supseteq \overline{X} \cap Y \end{cases}$$

It is easy to see that these order relations are transitive.

2.3 Probability approximation

Now we present a simple method to estimate the probability of picking randomly a value pair. Assuming that values of each attribute are independent, the probability of picking a value pair (x_i, y_i) of A_i is approximately estimated as:

$$p(x_i, y_i) = \frac{\delta(x_i)\delta(y_i)}{n^2}$$

where $\delta(x_i)$ and $\delta(y_i)$ are the numbers of objects that have attribute value x_i, y_i respectively, and n is the number of data objects.

3 Integration methods

The similarity between two data objects consisting of m attributes is measured by a combination of m similarities of their attribute value pairs. Taking advantage of measuring similarities of attribute value pairs in terms of probability, we consider integrating similarities of m attribute value pairs as the problem of integrating m probabilities.

Denote $S(\mathbf{x}, \mathbf{y}) = f(S_1, \ldots, S_m)$ the similarity between two data objects $\mathbf{x}$ and $\mathbf{y}$ where S_i is the similarity between values x_i and y_i of attribute A_i, and $f(.)$ is a function for integrating m probabilities $S_1, \ldots, S_m$. Here we describe some popular methods for integrating probabilities [**?**,**?**,**?**].

The most popular method is due to Fisher's transformation [**?**], which uses the test statistic

$$T_F = -2 \sum_{i=1}^{m} \ln S_i$$

and compares this to the χ^2 distribution with $2m$ degrees of freedom.

In [**?**], Stouffer et. al. defined

$$T_s = \sum_{i=1}^{m} \frac{\Phi^{-1}(1 - S_i)}{\sqrt{m}}$$

where Φ^{-1} is the inverse normal cumulative distribution function. The value T_s is compared to the standard normal distribution.

Another P-value method was proposed by Mudholkar and Geore [**?**]

$$T_M = -c \sum_{i=1}^{m} \log \frac{S_i}{1 - S_i}$$

where

$$c = \sqrt{\frac{3(5m+4)}{m\pi^2(5m+2)}}$$

The combination value of $S_1, \ldots, S_m$ is referenced to the t distribution with $5m + 4$ degrees of freedom.

In practice, probability integrating functions are often non-decreasing functions. It means that the greater $S_1, \ldots, S_m$ are, the greater $S(\mathbf{x}, \mathbf{y})$ is. In particular, it is easy to prove that the mentioned probability integrating functions are non-decreasing functions.

4 Example

To illustrate how the similarity between two data objects is measured using our method, consider the simple data set given in Table 1 that was obtained from an user internet survey. This data set contains 10 data objects comprising 3 different attributes e.g. age (continuous data), connecting speed (ordinal data), and time on internet (interval data). Consider the first data object ($\{26, 128k, [6..10]\}$ and the second one $\{55, 56k, [7..15]\}$, the similarity between them is measured as follows:

$$\begin{aligned} S_{age}(26, 55) &= p(23, 55) + p(55, 23) + p(25, 55) + p(55, 25) + \ldots + p(57, 26) \\ &= \frac{1 \times 1}{10^2} + \frac{1 \times 1}{10^2} + \frac{1 \times 1}{10^2} + \frac{1 \times 1}{10^2} + \ldots + \frac{1 \times 1}{10^2} \\ &= 0.18 \end{aligned}$$

$$\begin{aligned} S_{speed}(128k, 56k) &= p(14k, 128k) + p(128k, 14k) + p(28k, 128k) \\ &\quad + p(128k, 28k) + \ldots + p(56k, > 128k) + p(> 128k, 56k) \\ &= \frac{2 \times 1}{10^2} + \frac{2 \times 2}{10^2} + \frac{2 \times 1}{10^2} + \frac{2 \times 2}{10^2} + \ldots + \frac{2 \times 1}{10^2} + \frac{2 \times 2}{10^2} \\ &= 0.42 \end{aligned}$$

$$\begin{aligned} S_{time}([6..10], [7..15]) &= p([5..10], [20..30]) + p([20..30], [5..10]) + p([5..10], [12..20]) \\ &\quad + p([12..20], [5..10]) + \ldots + p([3..7], [5..12]) \\ &= \frac{1 \times 1}{10^2} + \frac{1 \times 1}{10^2} + \frac{1 \times 1}{10^2} + \frac{1 \times 1}{10^2} + \ldots + \frac{1 \times 1}{10^2} \\ &= 0.76 \end{aligned}$$

Table 1. An example: a data set obtained from an user internet survey includes 10 data objects, comprising 3 different attributes e.g., age (continuous data), connecting speed (ordinal data) and time on internet (interval data)

No.	Age (year)	Connecting Speed (k)	Time on Internet (hour)
1	26	128	[6..10]
2	55	56	[7..15]
3	23	14	[5..10]
4	25	36	[20..30]
5	56	> 128	[12..20]
6	45	56	[15..18]
7	34	28	[3..4]
8	57	28	[3..7]
9	48	14	[8..12]
10	34	> 128	[5..10]

Now we use Fisher's transformation test statistic [**?**] to integrate S_{age}, S_{speed} and S_{time} :

$$\begin{aligned} T_F &= -2(\ln S_{age} + \ln S_{speed} + \ln S_{time}) \\ &= -2(\ln(0.18) + \ln(0.42) + \ln(0.76)) \\ &= 5.71 \end{aligned}$$

The value of the χ^2 distribution with 6 degrees of freedom at point 5.71 is 0.456. Thus, the similarity between the first and the second objects, $S(\{26, 128k, [6..10]\}, \{55, 56k, [7..15]\})$, is 0.456.

5 Characteristics

In this subsection, we investigate characteristics and properties of our proposed method. For convenience let us recall an important required property of similarity measures that was proposed by Geist et. al. [**?**].

Definition 2. *Similarity measure $\rho : \Gamma^2 \to R^+$ is called an order-preserving similarity measure with respect to order relation $\preceq$ if and only if it holds true for:*

$$\forall(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \Gamma^2, (\mathbf{x}', \mathbf{y}') \preceq (\mathbf{x}, \mathbf{y}) \Rightarrow \rho(\mathbf{x}', \mathbf{y}') \leq \rho(\mathbf{x}, \mathbf{y})$$

Since order-preserving measures play important roles in practice, most common similarity measures (e.g., Euclidean, Hamming, Russel and Rao, Jaccard and Needham) possess the property with respect to reasonable order relations.

Theorem 1. *Similarity measure $S_{\preceq_i} : A_i^2 \to R^+$ is an order-preserving similarity measure with respect to order relation $\preceq_i$ if order relation $\preceq_i$ is transitive.*

Proof. Denote $\Lambda(x_i, y_i)$ the set of pairs which are smaller than or equally (x_i, y_i)

$$\Lambda(x_i, y_i) = \{(x_i', y_i') : (x_i', y_i') \preceq_i (x_i, y_i)\}$$

Since $\preceq_i$ is a transitive relation, for any two value pairs (x_{i_1}, y_{i_1}) and (x_{i_2}, y_{i_2}), when $(x_{i_1}, y_{i_1}) \preceq_i (x_2, y_2)$ we have $\forall (x_i, y_i) \in \Lambda(x_{i_1}, y_{i_1}) : (x_i, y_i) \preceq (x_{i_1}, y_{i_1})$ implies $(x_i, y_i) \preceq_i (x_{i_2}, y_{i_2})$. This means $(x_i, y_i) \in \Lambda(x_{i_2}, y_{i_2})$, and thus

$$\Lambda(x_{i_1}, y_{i_1}) \subseteq \Lambda(x_{i_2}, y_{i_2}) \tag{1}$$

In other hand, we have

$$S_{\preceq_i}(x_i, y_i) = \sum_{(x_i', y_i') \preceq_i (x_i, y_i)} p(x_i', y_i') = \sum_{(x_i', y_i') \in \Lambda(x_i, y_i)} p(x_i', y_i') \tag{2}$$

From (1) and (2),

$$S_{\preceq_i}(x_{i_1}, y_{i_1}) = \sum_{(x_i, y_i) \in \Lambda(x_{i_1}, y_{i_1})} p(x_i, y_i) \leq \sum_{(x_i, y_i) \in \Lambda(x_{i_2}, y_{i_2})} p(x_i, y_i) = S_{\preceq_i}(x_{i_1}, y_{i_1})$$

Thus, $S_{\preceq_i}(.,.)$ is an order-preserving measure. □

In practice, order relation $\preceq_i$ are often transitive. Thus, the ordered probability-based similarity measures for attributes are also order-preserving similarity measures.

Denote $\mathbb{A} = A_1 \times \ldots \times A_m$ the product space of m attributes $A_1, \ldots, A_m$. We define the product of order relation $\preceq_1, \ldots, \preceq_m$ as follows:

Definition 3. *The product of order relations $\preceq_1, \ldots, \preceq_m$, denoted by $\prod_{i=1}^m \preceq_i$, is an order relation $\preceq$ on $\mathbb{A}^2$, for which one data object pair is said to be less similar than or equally similar to another data object pair with respect to $\prod_{i=1}^m \preceq_i$ if and only if attribute value pairs of the first data object pair are less similar than or equally similar to those of the second data object pair*

$$\forall (\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \mathbb{A}^2 : (\mathbf{x}', \mathbf{y}') \preceq (\mathbf{x}, \mathbf{y}) \Leftrightarrow (x_i', y_i') \preceq_i (x_i, y_i), i = 1, \ldots, m$$

Proposition 1. *The product of order relations $\preceq_1, \ldots, \preceq_m$ is transitive when order relations $\preceq_1, \ldots, \preceq_m$ are transitive.*

Proof. Denote $\preceq = \prod_i^m \preceq_i$. For any triple data object pairs $(\mathbf{x_1}, \mathbf{y_1})$, $(\mathbf{x_2}, \mathbf{y_2})$, and $(\mathbf{x_3}, \mathbf{y_3})$. if $(\mathbf{x_1}, \mathbf{y_1}) \preceq (\mathbf{x_2}, \mathbf{y_2})$, and $(\mathbf{x_2}, \mathbf{y_2}) \preceq (\mathbf{x_3}, \mathbf{y_3})$, we have

$$(\mathbf{x_1}, \mathbf{y_1}) \preceq (\mathbf{x_2}, \mathbf{y_2}) \Leftrightarrow (x_{i_1}, y_{i_1}) \preceq_i (x_{i_2}, y_{i_2}) \ \ \forall i = 1 \ldots m$$
$$(\mathbf{x_2}, \mathbf{y_2}) \preceq (\mathbf{x_3}, \mathbf{y_3}) \Leftrightarrow (x_{i_2}, y_{i_2}) \preceq_i (x_{i_3}, y_{i_3}) \ \ \forall i = 1 \ldots m$$

Since $\preceq_i$ is transitive for $i = 1 \ldots m$, $(x_{i_1}, y_{i_1}) \preceq_i (x_{i_2}, y_{i_2})$ and $(x_{i_2}, y_{i_2}) \preceq_i (x_{i_3}, y_{i_3})$ implies $(x_{i_1}, y_{i_1}) \preceq_i (x_{i_3}, y_{i_3})$. Hence $(\mathbf{x_1}, \mathbf{y_1}) \preceq (\mathbf{x_3}, \mathbf{y_3})$.

Thus, $\prod_i^m \preceq_i$ is transitive. □

Theorem 2. *Similarity measure $S : \mathbb{A}^2 \rightarrow R^+$ is an order-preserving similarity measure with respect to $\prod_{i=1}^m \preceq_i$ when order relations $\preceq_1, \ldots, \preceq_m$ are transitive and probability integrating function f is non-decreasing.*

Proof. Denote $\preceq = \prod_i^m \preceq_i$, $(\mathbf{x}', \mathbf{y}')$ and $(\mathbf{x_2}, \mathbf{y_2})$ two data object pairs. We have

$$(\mathbf{x}', \mathbf{y}') \preceq (\mathbf{x}, \mathbf{y}) \Leftrightarrow (x'_i, y'_i) \preceq (x_i, y_i) \ \forall i = 1, \ldots, m;$$

Since $\preceq_i$ is transitive for $i = 1, \ldots, m$, following Theorem 1,

$$S'_i = S_{\preceq_i}(x'_i, y'_i) \leq S_{\preceq_i}(x_i, y_i) = S_i \ \forall i = 1, \ldots, m$$

Since f is a non-decreasing function,

$$S(\mathbf{x}', \mathbf{y}') = f(S'_1, \ldots, S'_m) \leq f(S_1, \ldots, S_m) = S(\mathbf{x}, \mathbf{y})$$

Since $(\mathbf{x}', \mathbf{y}') \preceq (\mathbf{x}, \mathbf{y}) \Rightarrow S(\mathbf{x}', \mathbf{y}') \leq S(\mathbf{x}, \mathbf{y})$, $S(.,.)$ is an order-preserving similarity measure with respect to $\prod_i^m \preceq_i$. □

Theorem 3 says that if attribute value pairs of a object pair are less similar than or equally similar to those of another object pair, the similarity of the first object pair is smaller than or equally the similarity of the second object pair in conditions that order-relations $\preceq_1, \ldots, \preceq_m$ are transitive and probability integrating function f is non-decreasing.

6 Evaluation

6.1 Complexity evaluation

In this subsection, we analysis the complexity for computing similarity for a value pair and for a two data objects described by m attributes.

The simplest way two measure the similarity between two values of attribute A_k is to scan all value pairs of this attribute. By this way, the

complexity of measuring similarity for a value pair is obviously $O(n_k^2)$ where n_k is the number of values of attribute A_k. In practice, n_k is often small (from dozens to a hundred) and therefore the complexity is absolutely acceptable. However, n_k may be large (up to n) when A_k is continuous data. In this case, we design two especial methods for computing similarity between two continuous values in $O(\log_2 n_k)$ or $O(n_k)$ depending on memory space requirements.

Let denote A_k a continuous attribute with n_k values $a_1, ..., a_{n_k}$. Assuming that $a_1 < ... < a_{n_k}$.

Computing similarity for continuous data in $O(\log_2 n_k)$ In this methods, we first sort n_k^2 value pairs. Then the similarity of value pair (v, v') at index i is simply the similarity of pair (u, u') at index $i-1$ plus the probability of getting (v, v') and stored in vector S. After that the similarity between any value pair can be referred from vector S in $O(\log_2 n_k)$ by the binary search technique [**?**]. The method is rather convenient in sense of complexity since $O(\log_2 n_k)$ is so small even when n_k is very large. However, it requires $O(n_k^2)$ memory space.

Computing similarity for continuous data in $O(\log n_k)$ Since $O(n_k^2)$ memory requirement is out of today computer's ability when the number of values is up to hundred thousands or millions, the method of computing similarity for value pairs in $O(\log_2 n_k)$ seems to be unrealistic when facing with data sets describing by continuous attributes with large numbers of values. In this part, we introduce the method required $O(n_k)$ memory space and gives the similarity between two values in $O(n_k)$.

Theorem 3. *Given $a_1, ..., a_{n_k}$ be a ordered values. For any value pair (v, v') and (a_i, a_j) with $i \leq j$, it holds true*

1. *if $(a_i, a_j) \preceq (v, v')$, then $(a_i, a_t) \preceq (v, v')$when $t \geq j$.*
2. *if $(a_i, a_j) \not\preceq (v, v')$, then $(a_i, a_t) \not\preceq (v, v')$when $i \leq t \leq j$.*

Proof. 1. We have $(a_i, a_j) \preceq (v, v') \Leftrightarrow a_j - a_i \geq |v - v'|$. Since $a_t \geq a_j$ when $t \geq j$, $a_t - a_i \geq |v - v'|$. Thus $(a_i, a_j) \preceq (v, v')$.□

2. We have $(a_i, a_j) \not\preceq (v, v') \Leftrightarrow a_j - a_i \not\geq |v - v'|$. Since $a_j \leq a_t \leq a_i$ when $i \leq t \leq j$, $a_t - a_i \not\geq |v - v'|$. Thus, $(a_i, a_t) \not\preceq (v, v')$.□

From Theorem 3, it is easy to see that the similarity between two values v and v' can be computed as

$$Sim(v, v') = \sum_{i=1}^{n_k} p(a_i) \sum_{j=t_i}^{n_k} p(a_j) \tag{3}$$

where t_i is the smallest number that is greater than i and satisfies $(a_i, a_{l_i}) \preceq (v, v')$.

Based on equation 3, we build an algorithm for determining similarity between two value (v, v') (see Figure 1). It is not hard to prove that the complexity for computing similarity between two values is $O(n_k)$ and required memory store is $O(n_k)$.

Procedure Sim_Determine
IN: two values v and v'.
OUT: Similarity of (v, v')
BEGIN
1: $i = 1, j = 1, sp = 1, Sim = 0$
2: **for** $i = 1$ to n_k **do**
3: **while** $((v, v') \preceq (a_i, a_j))$ and $(j \leq n_i)$ **do**
4: $sp = sp - p(a_j)$
5: $j = j + 1$
6: **end while**
7: $Sim = Sim + p(a_i) * sp$
8: **end for**
9: return Sim
END

Fig. 1. Algorithm for computing similarity between two continuous values in $O(n_k)$

After obtaining similarities for m attribute value pairs of two data objects, it is not hard to prove that integrating these m similarities requires $O(m)$.

It is obvious that the complexity to measure a value pair of the proposed measure is higher than $O(m)$. However, in real applications, the complexity is acceptable as the value number of each attribute is often small or using the especial methods to reduce the complexity.

6.2 Evaluation with real data

In the following we analyze real data sets using our similarity measuring approach in conjunction with clustering methods. We try to mine group of users with particular properties from internet survey data.

Data set The Cultural Issues in Web Design data set was obtained from the GVU's 8th WWW User Survey (http://www.cc.gatech.edu/gvu/user_surveys/survey-1997-10/). The data set is a collection of users's opinions on influences of languages, colors, culture, etc. on web designs. The data set includes 1097

respondents, which are described by 3 item set attributes, 10 categorical attributes, and 41 ordinal attributes.

Methodology

- *Similarity measure method*
  We apply the proposed method to measure similarities between respondents of the Cultural Issues in Web Design data set. We use the order relations and the probability approximation method as mentioned in Section 2. We choose the Fisher's transformation to integrate similarities of attribute value pairs.
- *Clustering method*
  A clustering method can be categorized into either partitioning approaches (e.g., K-means [**?**], Kmedoid [**?**]) or hierarchical approaches (e.g., single linkage [**?**], complete linkage [**?**], group average linkage [**?**,**?**]). Since partitioning approaches are not proper for noncontinuous data, we choose agglomeration hierarchical average linkage clustering method, which overcomes the *chain* problem of single linkage methods and discover more *balanced* clusters than complete linkage methods do.

Clustering results The Cultural Issues in Web Design data set was clustered into 10 clusters. However, we present characteristics of only three clusters due to space limitation, see Table 2. A characteristic of a cluster is presented as an attribute value that majority of respondents of the cluster answered. For example, value *can't write* of attribute *Unfamiliar site* is considered as a characteristic of the first cluster because 92% respondents of this cluster answered the value.

Discussion As it can be seen from Table 2, the clusters have many characteristics, e.g. the first cluster has 13 characteristics, the third has 15 characteristics. Moreover, characteristics are different from cluster to cluster. In particular, when visiting an *unfamiliar site*, the problem of 92% respondents of the first cluster is *cannot write*, while 81% respondents of the second cluster is *cannot translate*, and 84% respondents of the third cluster is *cannot read*. Moreover, answers of respondents in the same clusters are somehow similar. For example, all respondents of the first cluster can neither read *Rabic* and *Hebrew* nor speak *Bengari* and *Hebrew*. In short, almost respondents in the same cluster have the same answers but they are different from answers of respondents of different clusters. The analysis of characteristics from these clusters shows that

our similarity measuring method in combination with the agglomeration hierarchical average linkage clustering method discovers valuable clusters of real data sets.

7 Conclusions and further works

We introduced a method to measure the similarity for heterogeneous data in the statistics and probability framework. The main idea is to define the similarity of one value pair as the probability of picking randomly a value pair that is less similar than or equally similar in terms of order relations defined appropriately for data types. Similarities of attribute value pairs of two objects are then integrated using a statistical method to assign the similarity between them.

The measure possess the order-preserving similarity property. Moreover, applying our approach in combination with clustering methods to real data shows the merit of our proposed method.

However, the proposed method is designed for data sets whose data objects have the same number of attribute values. In future works, we will adapt this method for more complex data sets whose data objects may have different numbers of attribute values.

Acknowledgments

We appreciate professor Gunter Weiss and Le Sy Vinh at the Heinrich-Heine University of Duesseldorf, Germany for helpful comments on the manuscript.

References

1. Gowda K. C. and Diday E. Symbolic clustering using a new dissimilarity measure. *In Pattern Recognition*, 24(6):567–578, 1991.
2. Gowda K. C. and Diday E. Unsuppervised learning throught symbolic clustering. *In Pattern Recognition lett.*, 12:259–264, 1991.
3. Gowda K. C. and Diday E. Symbolic clustering using a new similarity measure. *IEEE Trans. Syst. Man Cybernet*, 22(2):368–378, 1992.
4. Ichino M. and Yaguchi H. Generalized minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems Man, and Cybernetics*, 24(4), 1994.
5. de Carvalho F.A.T. Proximity coefficients between boolean symbolic objects. In E. et al Diday, editor, *New Approaches in Classification and Data Analysis*, volume 5 of *Studies in Classification, DataAnalysis, and Knowledge Organisation*, pages 387–394, Berlin, 1994. Springer-Verlag.

6. de Carvalho F.A.T. Extension based proximity coefficients between constrained boolean symbolicobjects. In Hayashi C. et al., editor, *IFCS96*, pages 370–378, Berlin, 1998. Springer.
7. Geist S., Lengnink K., and Wille R. An order-theoretic foundation for similarity measures. In Diday E. and Lechevallier Y., editors, *Ordinal and symbolic data analysis, studies in classification, data analysis, and knowledge organization*, volume 8, pages 225–237, Berlin, Heidelberg, 1996. Springer.
8. Fisher R.A. *Statistical methods for research workers.* Oliver and Boyd, 11th edition, 1950.
9. Stouffer S.A, Suchman E.A, Devinney L.C, and Williams R.M. Adjustment during army life. *The American Solder*, 1, 1949.
10. Mudholkar G.s and George E.O. The logit method for combining probabilities. In J. Rustagi, editor, *Symposium on Optimizing methods in statistics*, pages 345–366. Academic press, NewYork, 1979.
11. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms.* MIT Press and McGraw-Hill., the third edition, 2002.
12. MacQueen J. Some methods for classification and analysis of multivariate observation. In *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
13. Kaufmann L. and Rousseeuw P.J. Clustering by means of medoids. *Statistical Data Analysis based on the L1 Norm*, pages 405–416, 1987.
14. Sneath P.H.A. The application of computers to taxonomy. *Journal of general microbiology*, 17:201–226, 1957.
15. McQuitty L.L. Hierarchical linkage analysis for the isolation of types. *Education and Psychological measurements*, 20:55–67, 1960.
16. Sokal R.R. and Michener C.D. Statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438, 1958.
17. McQuitty L.L. Expansion of similarity analysis by reciprocal pairs for discrete and continuous data. *Education and Psychological measurements*, 27:253–255, 1967.

Table 2. Characteristics of three discovered clusters

Cluster 1

No.	Att. Names	Value	P_a	Value	P_a	Value	P_a	Value	P_a
1	Unfamiliar sites	Can't write	92					Other	8
2	Read Arabic	None	100						
3	Read Hebrew	None	100						
4	Speak Bengali	None	100						
5	Speak Hebrew	None	100						
6	Primary same as Native	Yes	98	No	2				
7	Important problem	Can't write	92	None	2			Other	6
8	American images	None	79					Other	21
9	Native Language	English	79	Chinese	4	German	4	Other	12
10	Read German	None	71	Basic phrases	19	Native	8	Other	2
11	Software	Yes both	73	Yes get	25	No	2		
12	Speak English	Native	69	Conver.	17	None	14		
13	Provide native sites	Agree strongly	69	Agree somewhat	23	Disag. somewhat	4	Other	4

Cluster 2

No.	Att. Names	Value	P_a	Value	P_a	Value	P_a	Value	P_a
1	Unfamiliar sites	Can't translate	81					Other	19
2	Read Chinese	None	100						
3	Read Hindi	None	100						
4	Read Japanese	None	100						
5	Speak Hindi	None	100						
6	Due to culture	No	93	Yes-both	8				
7	Sites in non-fluent	Few	89	None	9	Most	2		
8	Non-English sites	Few	89	None	8	Half	4		
9	Translations	Yes-useful	87					Other	13
10	Read German	None	83	Basic phrases	9	Literate	8		
11	Native Language	English	81	Spanish	8	Arabic	2	Other	9
12	Speak German	None	81	Basic phrases	11	Conver.	8		
13	Designed culture	Yes	70	No	28	Don't know	2		

Cluster 3

No.	Att. Names	Value	P_a	Value	P_a	Value	P_a	Value	P_a
1	Unfamiliar sites	Can't read	84					Other	16
2	Read Arabic	None	100						
3	Read Chinese	None	100						
4	Read Hindi	None	100						
5	Speak Arabic	None	100						
6	Speak Bengali	None	100						
7	Speak Hindi	None	100						
8	Read Italian	None	93	Basic phrases	4	Native	2	Other	2
9	Speak Italian	None	93	Basic phrases	7				
10	Speak Spanish	None	84	Basic phrases	14	Conver.	2		
11	Read Spanish	None	82	Basic phrases	18				
12	Sites designed for culture	Yes	68	No	29	Dontknow	4		
13	Sites in non-fluent	Few	77	All	11	None	7	Other	5
14	Software	Yes get	77	Yesboth	18	No	5		
15	Non-English sites	Few	68	None	21	Half	9	Other	2