Similarity Relations Based on Uniqueness Measure
Masashi Emoto\textsuperscript{1} Rolly Intan\textsuperscript{2,3} Manami Matsumoto\textsuperscript{2} Masao Mukaidono\textsuperscript{2}

\textsuperscript{1}School of Business Administration, Meiji University
\textsuperscript{2}Department of Computer Science, Meiji University
\textsuperscript{3}Department of Informatics Engineering,
Petra Christian University, Surabaya, Indonesia
E-mail: \{masashi, rosty, ca-tail, masao\}@cs.meiji.ac.jp

Abstract: Various concepts of similarity relations used in clustering have been proposed until now dealing with number of coincided attributes, Euclid distance, etc. Generally, their degree of similarity between two objects is calculated by using only attribute values of the objects. However, as usually happen in human perception, degree of similarity between two objects may be different depending on by whom the perception is performed. For example, two blond hair persons may have different degrees of similarity depending on the perceptions performed by, for instance, a Northern European and a Japanese. In this paper, we propose a new concept of similarity, which is closer to human perception than traditional concepts of similarity. The concept of similarity is based on rarity of attribute values in a subset of objects regarded as human knowledge. That is, we use not only attribute values of two objects, but also a subset of objects. Then, we clarify mathematical properties of the concept of similarity and show its differences from a traditional concept of similarity.

Keywords: Similarity Relation, Probability, Uniqueness Measure

1 Introduction

Various concepts of similarity relations have been proposed until now. Mostly those concepts of similarity are dealing with number of coincided attributes, Euclid distance, etc. Generally, degree of similarity between two objects is calculated only by using attribute values of the objects. However, as usually happen in human perception, degree of similarity between two objects may be subjectively different depending on by whom the perception is performed. For example, a Northern European and a Japanese may have different perceptions in determining degree of similarity between two blond hair persons in which two blond hair person may be looked more similar by Japanese than by Northern European.

Here, in this paper, we need to propose a new concept of similarity that is closer to human perception. The concept of similarity is not only based on similar value of attributes between two objects, but also based on rarity of information (value of attributes) by introducing a uniqueness measure. In the concept of similarity, a rarity is obtained from entropy of attribute value. Basically, our proposed concept of similarity has the same idea with a concept of similarity proposed in[7]. However, the proposed concept differs from[7] in that our proposed concept is based on human’s perception by using rarity of attribute value in a subset of all objects regarded as knowledge. The structure of paper is the following. In Section 2, we discuss and explain the proposed concept of the similarity by giving some examples. Based on discussion in Section 2, Section 3 formally defines the concept of similarity based on uniqueness measure. Then, we clarify mathematical properties of the similarity based on uniqueness measure in Section 4. Finally, a conclusion is given in Section 5.

2 Human Perception

The concept of uniqueness-based similarity proposed in this paper is based on human perception. In this section, first, we show an illustrative example, and discuss how human perception works in a process of recognition. Then, based on human perception, we propose a new concept of similarity, called uniqueness-based similarity.

2.1 Monkey and Human

Let us suppose that there are two photographs, a photograph of two different monkeys and a photograph of two different humans. As human, obviously we cannot recognize differences between two monkeys in the photograph well, but we can recognize differences between two humans in their photograph (if they are not identical twin). This is because we have much knowledge related to the attributes of humans by having many opportunities (experiences) to see them, but we have few knowledge related to the monkeys. Hence, it can be said that it is difficult for a human to recognize a monkey as a particular monkey compared with another monkey, because of our limited knowledge about...
the monkeys. In other words, we do not know any important features or attributes of monkey to be recognized. It could be said that it may be difficult for monkeys to recognize a human as a particular human compared to the other human. In our point of view, humans and monkeys have different knowledge.

2.2 Human and Their Features

Let us suppose that we have two pictures, a picture of two Japanese men (Fig.1) and a picture of the same persons with mustache (Fig.2). First, we consider a case when Japanese look at the pictures. Generally, Japanese could understand some important features needed to recognize men without mustache. So it is easy for Japanese to distinguish men without mustache. Furthermore, in many cases, Japanese have a rare opportunity to see men with mustache, so they tend to have a strong impression for a mustached image. Consequently, their attention will gather for the mustache. As a result, men with mustache are looked more similar for Japanese than men without mustache.

Next, we consider another case when Arabs look at a picture of two different Arabs with a mustache and a picture of the same persons without mustache. Since Arabs have plenty of knowledge and experience to see men (Arabs) with mustache, they could easily recognize the persons with mustache without being confused by the mustache. Moreover, it is well known that almost all men of Arabs have mustache. As a result, Arabs without mustache may be looked more similar for Arabs than Arabs with a mustache.

Finally, we consider Arabs look at the pictures of Japanese in Fig.1 and Fig.2. It seems that Arabs will do the same evaluation as Japanese. We can guess that Arabs' knowledge has a concept that mostly Japanese do not have mustache.

2.3 Summary of Discussion

From the discussion above, it can be summarized that human can recognize anything if he has enough knowledge or experience about it, where knowledge plays important role in human perception. Here, knowledge is constituted based on subjectivity for every human, and the subjectivity is made from what it has been experienced until now. When an element and its combination (feature) are unique in knowledge, human has a tendency to react to the unique thing sensitively. For instance, it is applied to the features such as, monkeys are covered with red hair, humans (Japanese) have mustache, etc. Therefore, in this paper, we propose a new concept of similarity based on uniqueness (measure) in which the concept of similarity naturally represents human perception. That is, human will pay much attention to a rare (unique) case in the whole data.

3 Similarity based on Uniqueness Measure

This section formally defines the concept of similarity based on the uniqueness measure. Also, we define a simple concept of traditional similarities in order to make a comparison with the proposed concept of similarity.

First, we need to define a data table called information system as usually used in knowledge representation. Simply, we use a binary information system as the following definition.

**Definition 1** Information system contains data about objects of interest characterized by some attributes. A binary information system is defined as a quadruple \( I = (U, A, V, \rho) \), where \( U \) is a non-empty finite set of objects called the universe, \( A \) is a non-empty finite set of attributes, \( V \) is a set of \( \{0, 1\} \) and \( \rho : (U \times A) \rightarrow V \).

**Example 1** Given a binary information system in Table 1 consists of ten objects, \( \{u_1, u_2, \cdots, u_{10}\} \), and five attributes, \( \{a_1, a_2, \cdots, a_5\} \). For example in Table 1, an attribute value of object \( u_7 \) in attribute \( a_2 \) is given by \( \rho(u_7, a_2) = 1 \). Also, an attribute value of object \( u_7 \) in attribute \( a_1 \) is given by \( \rho(u_7, a_1) = 0 \).

Next, we define a degree of coincidence by comparing attribute values between two objects in a given attribute.
Definition 2 Degree of coincidence between two objects \( u_i, u_j \in U \) in attribute \( a_k \in A \) is denoted by \( M(u_i, u_j, a_k) \), and defined as follows.

\[
M(u_i, u_j, a_k) = \begin{cases} 
1 & \text{if } \rho(u_i, a_k) = \rho(u_j, a_k) \\
0 & \text{if } \rho(u_i, a_k) \neq \rho(u_j, a_k).
\end{cases}
\]

(1)

Next, we define a simple concept of traditional similarities by using degree of coincidence as the following definition.

Definition 3 Traditional similarity between two objects \( u_i, u_j \in U \) is denoted by \( S_{tra}(u_i, u_j) \), and defined by:

\[
S_{tra}(u_i, u_j) = \frac{\sum_{k=1}^{\left|A\right|} M(u_i, u_j, a_k)}{\left|A\right|},
\]

(2)

where \( a_k \in A \), and \( \left|A\right| \) is cardinality (number of elements) of set of attributes \( A \).

The above similarity is obtained by dividing a number of coincided attributes of two objects by a number of all attributes.

Example 2 Using Definition 3, similarity between \( u_2 \) and \( u_5 \), and similarity between \( u_8 \) and \( u_{10} \) in Table 1 are calculated by:

\[
S_{tra}(u_2, u_5) = \frac{2}{5} = 0.4,
\]

\[
S_{tra}(u_8, u_{10}) = \frac{5}{5} = 1.
\]

In Section 2, we discuss that degree of similarity may be changed by knowledge in which knowledge is regarded as a subset of \( U \). Next, we need to define some functions based on the knowledge in order to define the concept of uniqueness-based similarity.

Definition 4 Let \( u_i \in U \) be an object, \( X \subseteq U \) is knowledge, and \( a_k \in A \) is an attribute. \( P(u_i, a_k, X) \) is defined as probability of an attribute value in \( X \) as given by:

\[
P(u_i, a_k, X) = \frac{\sum_{u \in X} M(u_i, u_j, a_k)}{|X|}.
\]

(3)

Example 3 Let \( X = \{u_1, u_2, u_3, u_4, u_5\} \) be knowledge. \( P(u_2, a_1, U), P(u_5, a_3, U), P(u_2, a_1, X) \) and \( P(u_5, a_3, X) \) are calculated as follows.

\[
P(u_2, a_1, U) = \frac{2}{10} = 0.2,
\]

\[
P(u_5, a_3, U) = \frac{5}{10} = 0.5,
\]

\[
P(u_2, a_1, X) = \frac{2}{5} = 0.4,
\]

\[
P(u_5, a_3, X) = \frac{5}{5} = 1.
\]

Using the probability of an attribute value in a given subset of objects as defined in Definition 4, a concept of uniqueness measure is characterized by a function to calculate uniqueness of relationship between attribute values of two objects in a certain attribute.

Definition 5 Given \( X \subseteq U \), uniqueness measure relationship between two objects \( u_i, u_j \in U \) in an attribute \( a_k \in A \) is characterized by a function \( C(u_i, u_j, a_k, X) \).

\[
C(u_i, u_j, a_k, X) = \begin{cases} 
1 - P(u_i, a_k, X)^2 & \text{if } M(u_i, u_j, a_k) = 1 \\
1 - 2 \times P(u_i, a_k, X) \times P(u_j, a_k, X) & \text{if } M(u_i, u_j, a_k) = 0.
\end{cases}
\]

(4)

Generally, this function calculates subtraction a probability, which attribute values of two objects occur at once from one. Since all attribute values take 0 or 1, occurrences of attribute values of two objects are divided into four cases. They are a case of 0 and 0, a case of 0 and 1, a case of 1 and 0, and a case of 1 and 1. The case, which attribute values are different is two cases (0 and 1, and 1 and 0). Hence, probability is twice when \( M(u_i, u_j, a_k) = 0 \).

Example 4 We show two examples of a case when two attribute values are different, \( C(u_2, u_5, a_2, X) \), and a case when two attribute values are the same, \( C(u_2, u_5, a_3, X) \).

Let knowledge be \( X = \{u_1, u_2, u_3, u_4, u_5\} \).

\[
C(u_2, u_5, a_2, X) = 1 - 2 \times P(u_2, a_2, X) \times P(u_5, a_2, X),
\]

\[
= 1 - 2 \times 0.2 \times 0.8,
\]

\[
= 1 - 0.32,
\]

\[
= 0.68.
\]
In the second example, attribute values of \( a_3 \) in \( X \) are all 0's. Consequently, attribute values in \( a_3 \) are not rare. Therefore, \( C(u_2, u_5, a_3, X) = 0 \).

Now, by using uniqueness measure as defined in Definition 5, we define a concept of similarity between two objects as follows.

**Definition 6** Degree of similarity based on uniqueness measure between two objects

\[
S_{uni}(u_i, u_j, X) = \frac{\sum_{k=1}^{[A]} (C(u_i, u_j, a_k, X) \times M(u_i, u_j, a_k))}{\sum_{k=1}^{[A]} C(u_i, u_j, a_k, X)}.
\]

**Example 5** Given the information system in Table 1 and let \( X = \{u_1, u_2, u_3, u_4, u_5\} \). By using Eq. 5, we calculate degree of similarity between \( u_2 \) and \( u_5 \) as follow.

\[
S_{uni}(u_2, u_5, U) = 0.96 + 0.75 = 1.71
\]

\[
S_{uni}(u_2, u_5, X) = 3.47
\]

\[
S_{uni}(u_2, u_8, U) = 0.84 + 0 = 0.84
\]

\[
S_{uni}(u_2, u_8, X) = 2.56
\]

\[
S_{uni}(u_2, u_{10}, U) = 0.84 + 0.68 + 0 + 0.52 + 0.52 = 2.56
\]

\[
S_{uni}(u_2, u_{10}, X) = 3.48
\]

Similarly, degree of similarity between \( u_8 \) and \( u_{10} \) is given by:

\[
S_{uni}(u_8, u_9, U) = 0.36 + 0.36 + 0.75 + 0.75 + 0.51 = 2.73
\]

\[
S_{uni}(u_8, u_9, X) = 1
\]

\[
S_{uni}(u_8, u_{10}, U) = 0.64 + 0.36 + 1 + 0.84 + 0.64 = 2.89
\]

\[
S_{uni}(u_8, u_{10}, X) = 3.48
\]

In knowledge \( X \), all attribute values of \( a_3 \) are equal to 0. Consequently, uniqueness of 0 in attribute \( a_3 \) disappeared. Hence, it can be verified that

\[
S_{uni}(u_2, u_5, U) \geq S_{uni}(u_2, u_5, X).
\]

Also, in the case of \( S_{uni}(u_8, u_{10}, U) \) and \( S_{uni}(u_8, u_{10}, X) \), since attribute values of two objects in all attributes are the same, their similarity is equal to 1 for any subsets.

In these two cases, when a subset of objects consists of only one object and when all objects in the subset have exactly the same attribute values for all attributes, results of calculation by using Eq. 5 are equal to \( \frac{0}{0} \). Here, we consider \( \frac{0}{0} \) as 1 in this research.

**4 Mathematical Properties of Similarity Based on Uniqueness Measure**

As a primary goal in this paper, this section discusses mathematical properties of similarity based on the uniqueness measure. There are several well known mathematical properties of binary relations such as equivalence relation, fuzzy similarity relation [8], weak fuzzy similarity relation [5], resemblance (proximity) relation [1] and etc. Equivalence relation is considered as the strongest binary relation, and it can be verified that similarity between crisp data satisfies the equivalence relation as given by the following definition.

**Definition 7** It is called equivalence relation, if binary relation satisfies the following properties.

- Reflexivity: \( \forall x \ R(x, x) = 1 \),
- Symmetry: \( \forall x, y \ R(x, y) = R(y, x) \),
- Transitivity: \( \forall x, y, z \ R(x, z) \geq R(y, z) \).

Fuzzy similarity relation has a weaker transitive property than equivalence relation for dealing with fuzzy data.

**Definition 8** It is called fuzzy similarity relation, if binary relation satisfies the following properties.

- Reflexivity: \( \forall x \ R(x, x) = 1 \),
- Symmetry: \( \forall x, y \ R(x, y) = R(y, x) \),
- Max-min Transitivity: \( \forall x, y, z \ R(x, z) \geq \max_{y \in U} \min \{R(x, y), R(y, z)\} \).
The weak fuzzy similarity relation is proposed based on conditional probability relation [6]. In conditional probability relation, similarity relationship between two data is assumed similar to the relationship between two events in conditional probability.

**Definition 9** It is called weak fuzzy similarity relation, if binary relation satisfies the following properties.

- **Reflexivity:**
  \[ R(x, x) = 1, \]

- **Conditional Symmetry:**
  \[ R(x, y) > 0 \text{ if } R(y, x) > 0, \]

- **Conditional Transitivity:**
  \[ R(x, y) > 0 \text{ and } R(y, z) > 0 \text{ if } R(x, z) > 0. \]

**Definition 10** The binary relation that satisfies only reflexivity and symmetry properties is called resemblance (proximity) relation.

**Theorem 1** Similarity based on uniqueness measure satisfies resemblance relation.

It can be verified easily.

In order to explore more properties, it is necessary to define number of coincidence attributes.

**Definition 11** Number of coincidence attributes is given by the following equation:

\[
N(u_i, u_j) = \sum_{k=1}^{\vert A \vert} M(u_i, u_j, a_k). \tag{6}
\]

**Theorem 2** In the traditional similarity, all \( u_i, u_j \in U \) satisfy the following property.

\[
N(u_i, u_j) \leq N(u_k, u_l)
\]

if and only if

\[
S_{\text{tra}}(u_i, u_j) \leq S_{\text{tra}}(u_k, u_l). \tag{7}
\]

However, in the similarity based on uniqueness measure, the property is not always satisfied.

**Proof:**

In the traditional similarity, it is clear to satisfy the above property from Definition 3.

In the similarity based on uniqueness measure, we show a counter example. In the binary information system in Table 1, let knowledge be \( X = \{u_1, u_2, u_3, u_4, u_5\} \). Then we show degrees of similarity based on uniqueness measure between \( u_2 \) and \( u_6, u_4 \) and \( u_5 \) as follows.

\[
S_{\text{uni}}(u_2, u_6, X) = \frac{0.84}{0.52 + 0.68 + 1 + 0.84 + 0.52} = 0.84
\]

\[= 3.56, \]

\[= 0.236. \]

\[
S_{\text{uni}}(u_4, u_6, X) = 0.36 + 0.52 + 0.52 + 0.52 = 1.92 = 0.188.
\]

**Definition 12** Related to the next theorem, let \( u \in U \) has \( \rho(u, a) \) for \( a \in A \). We define that \( \sim \) has \( 1 - \rho(u, a) \) for \( a \in A \).

**Theorem 3** In the traditional similarity, all \( u_i, u_j \in U \) satisfy the following property.

\[
S_{\text{tra}}(u_i, u_j) + S_{\text{tra}}(u_i, \sim u_j) = 1. \tag{8}
\]

However, in the similarity based on uniqueness measure, all \( u_i, u_j \in U \) do not always satisfy that \( S_{\text{uni}}(u_i, u_j, X) + S_{\text{uni}}(u_i, \sim u_j, X) \) is equal to 1.

**Proof:**

First, we prove that Eq.9 is satisfied in the traditional similarity.

\[
S_{\text{tra}}(u_i, u_j) + S_{\text{tra}}(u_i, \sim u_j) = \frac{\sum_{k=1}^{\vert A \vert} M(u_i, u_j, a_k)}{\vert A \vert} + \frac{\sum_{k=1}^{\vert A \vert} 1 - M(u_i, u_j, a_k)}{\vert A \vert},
\]

\[
= \frac{\sum_{k=1}^{\vert A \vert} M(u_i, u_j, a_k) + 1 - M(u_i, u_j, a_k)}{\vert A \vert},
\]

\[
= \frac{\sum_{k=1}^{\vert A \vert} 1}{\vert A \vert} = \frac{\vert A \vert}{\vert A \vert} = 1.
\]

In the similarity based on uniqueness measure, we show a counter example, which does not satisfy \( S_{\text{uni}}(u_i, u_j, X) + S_{\text{uni}}(u_i, \sim u_j, X) = 1 \). We use the
information system in Table 1. Let knowledge be \( X = \{ u_1, u_2, u_3, u_4, u_5 \} \). We consider a similarity of \( u_2 \) and \( u_5 \).

\[
S_{\text{uni}}(u_2, u_5, X) = 0.84 + 0 + 0.84 + 0.68 + 0.52 + 0.52' = 2.56', \quad S_{\text{uni}}(\sim u_2, \sim u_5, X) = 0.96 + 1 + 0.84 + 0.68 + 1 = 3.36', \quad \text{From } S_{\text{uni}}(u_2, u_5, X) + S_{\text{uni}}(\sim u_2, \sim u_5, X) = 0.944 \neq 1, \text{ it can be proved that the similarity based on uniqueness measure does not satisfy } S_{\text{uni}}(u_i, u_j, X) + S_{\text{uni}}(\sim u_i, \sim u_j, X) = 1. \text{ Q.E.D.}
\]

**Theorem 4** In the traditional similarity, all \( u_i, u_j \in U \) satisfy the following property.

\[
S_{\text{tra}}(u_i, u_j) = S_{\text{tra}}(\sim u_i, \sim u_j). \quad (9)
\]

However, in the similarity based on uniqueness measure, the above property is not always satisfied.

**Proof:** First, we prove that Eq.10 is satisfied in the traditional similarity.

\[
S_{\text{tra}}(\sim u_i, \sim u_j) = \sum_{k=1}^{|A|} M(\sim u_i, \sim u_j, a_k) / |A| = \sum_{k=1}^{|A|} 1 - M(u_i, u_j, a_k) / |A| = \sum_{k=1}^{|A|} M(u_i, u_j, a_k) / |A| = S_{\text{stra}}(u_i, u_j).
\]

In the similarity based on uniqueness measure, we show a counter example which does not satisfy Eq.10.

Again, we use the information system in Table 1. Let knowledge be \( X = \{ u_1, u_2, u_3, u_4, u_5 \} \). We consider a similarity of \( u_2 \) and \( u_5 \).

\[
S_{\text{uni}}(u_2, u_5, X) = 0.84 + 0 + 0.84 + 0.68 + 0 + 0.52 + 0.52' = 2.56', \quad S_{\text{uni}}(\sim u_2, \sim u_5, X) = 0.96 + 0.84 + 1 + 0.84 + 0.68 + 1 = 3.36', \quad \text{From } S_{\text{uni}}(u_2, u_5, X) + S_{\text{uni}}(\sim u_2, \sim u_5, X) = 0.944 \neq 1, \text{ it can be proved that the similarity based on uniqueness measure does not satisfy } S_{\text{uni}}(u_i, u_j, X) + S_{\text{uni}}(\sim u_i, \sim u_j, X) = 1. \quad \text{Q.E.D.}
\]

**Theorem 5** For any knowledge, \( X_1, X_2 \subseteq U \), similarities among objects, \( u_i, u_j, u_k, u_l \in U \), do not satisfy: \( S_{\text{uni}}(u_i, u_j, X_1) \leq S_{\text{uni}}(u_k, u_l, X_1) \) if and only if \( S_{\text{uni}}(u_i, u_j, X_2) \leq S_{\text{uni}}(u_k, u_l, X_2) \).

**Proof:** We show a counter example. In the information system of Table 1, let be \( X_1 = \{ u_1, u_2, u_3, u_4, u_5 \} \) and \( X_2 = \{ u_6, u_7, u_8, u_9, u_{10} \} \). Then,

\[
S_{\text{uni}}(u_4, u_5, X_1) = 0.188, \quad S_{\text{uni}}(u_1, u_5, X_1) = 0.329, \quad S_{\text{uni}}(u_4, u_5, X_2) = 0.382, \quad S_{\text{uni}}(u_1, u_5, X_2) = 0.140.
\]

Therefore,

\[
S_{\text{uni}}(u_4, u_5, X_1) \leq S_{\text{uni}}(u_1, u_5, X_1), \quad S_{\text{uni}}(u_4, u_5, X_2) \geq S_{\text{uni}}(u_1, u_5, X_2).
\]

The similarity based on uniqueness measure does not satisfy that \( S_{\text{uni}}(u_i, u_j, X_1) \leq S_{\text{uni}}(u_k, u_l, X_1) \) if and only if \( S_{\text{uni}}(u_i, u_j, X_2) \leq S_{\text{uni}}(u_k, u_l, X_2) \). \quad \text{Q.E.D.}

**Example 6** Theorem 5 proved that order of similarity could be changed. We explain by using the following example.

In the information system of Table 1, let be \( X = \{ u_1, u_2, u_3, u_4, u_5 \} \) and \( U = \{ u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10} \} \). Then,
\[ S_{\text{uni}}(u_2, u_3, X) = 0, \]
\[ S_{\text{uni}}(u_2, u_7, X) = 0.190, \]
\[ S_{\text{uni}}(u_2, u_3, U) = 0.235, \]
\[ S_{\text{uni}}(u_2, u_7, U) = 0.177, \]

\[ S_{\text{uni}}(u_2, u_3, X) \leq S_{\text{uni}}(u_2, u_7, X), \]
\[ S_{\text{uni}}(u_2, u_3, U) \geq S_{\text{uni}}(u_2, u_7, U). \]

In knowledge \( X \), \( u_7 \) is more similar to \( u_2 \) than \( u_3 \). Also, in knowledge \( U \), \( u_3 \) is more similar to \( u_2 \) than \( u_7 \). In knowledge \( X \) and \( U \), order of similarity is changed.

5 Conclusion

In this paper, we discussed and introduced a new concept of similarity dealing with uniqueness measure, called uniqueness-based similarity. We compared the concept of uniqueness-based similarity with a concept of traditional similarity by which the concept of uniqueness-based similarity is closer to human perception than the concept of traditional similarity. We then showed mathematical properties of uniqueness-based similarity. In the future, we would like to clarify more properties of the uniqueness-based similarity, and apply the concept to the real world application.

References


