Research on the Theory and Methods for Similarity Calculation of Rough Formal Concept in Missing-Value Context

Wang kai , LI Shao-Wen , Zhang You-Hua , Liu Chao

School of Information and Computer Science, Anhui Agricultural University, Hefei, China
wangkai1113@yeah.net

Abstract. In this paper, for the low similarity computation accuracy of concept in the field of agriculture ontology mapping, formal concept analysis theory and rough set theory are introduced to similarity computation. Jointly considering attribute hierarchies in concept lattice, the semantic hierarchy of the concepts are weighted differently, and the theory and methods of similarity calculation of rough formal concept in missing-value context is given. Finally, similarity computing model is prospected. Experimental results show the model has a high computational accuracy.

1. Introduction

With the explosive growth of knowledge, representation, sharing and exchanging of which has become urgent to solve. Ontology is a shared concept of a clear explanation standardization, which makes it possible to resolve all these issues. It has increasingly become important component of knowledge engineering, knowledge management, information retrieval and semantic Web. Formal concept analysis theory is a mathematical method put by Professor Wille R, which comes from the understanding of concept related to the areas of philosophy, and reflects the hierarchy between the concepts. Rough set theory is a theory of data analysis proposed by Z. Pawlak, which uses the approximate relationship between the upper and lower use of data to describe the uncertainty of information. Now has been widely used in decision analysis, pattern recognition, machine learning and knowledge discovery.

Wu Qiang broadens the scope of formal concept in the paper [1], applying rough set to Formal concept analysis area, to definite the concept which can not be studied. Yang wenping discusses the rough approximation of formal concept, proposing to use the rough set method to solve the upper and lower approximation of rough concept, which theoretically proves that the result of this theory is equivalent to the approximate extension of other ones. Kent gives the analysis of the approximate operator model of concept lattice, illustrates the relationship between the approximation operators. Yu-Ping studies the disorder duality relations between the concept pairs, mainly for the dependence among attributes, modeling for the concept lattice in the tolerance approximation space. Shao [2] synthesizes the basic rough model and reduction concept lattice model, getting the corresponding functional dependency algorithms, by the theory of rough set operations. All the research above mainly focused on the discovery of the common characteristics between rough set model and concept lattice model, and theoretical feasibility, ignoring the

*LI Shao-Wen is the Corresponding author.

*Fund projects: National High Technology Research and Development Program of China (No. 2006AA10Z249); National Natural Science Foundation of China (No. 30971091, 30800663)

characteristics of the concept of ontology structure, lack of the measures of similarity between the concept of ontology in rough environment.

The phenomenon of missing values is a widespread problem in knowledge engineering, which mainly related to the expression and processing of uncertain concepts. With the rapid development of the Semantic Web, the number of domain ontology gradually increases, to some extent, causes a sharp increase of uncertain concepts, which seriously affects the sharing and reuse of knowledge between domain ontologies.

Ontology mapping is an effective method to solve the problem foregoing, the key of which is to get conceptual similarity. Due to the reasons above-mentioned, this article uses upper and lower approximation of rough set theory, proposes rough formal concept in missing-value context, put forward the improved theory and methods for similarity calculation of rough formal concept in missing-value context.

2. Related Concepts

2.1 Rough Set Theory

Rough Set Theory is some kind of mathematical tool dealing with fuzzy and uncertain knowledge. The main idea of this theory is to access decision-making or classification rules, in the premise of maintaining the same classification, by knowledge reduction. Rough Set Theory generally refers to some undefined subset, regularly be approximation defined by two precision sets (upper approximation and lower approximation).

Definition 1. For a given knowledge base K= (U, R), U refers to non-empty set of objects; R is a family of equivalence relations based on U. If $P \subseteq R$, and P is not empty, then the intersection of all equivalence relations in P set is also an equivalence relation, called indiscernibility relations. For each subset $X \subseteq U$ and equivalence relation R, we can get these as follows:

$$\underline{R}\ X = \bigcup \{Y \in U / R \mid Y \subseteq X\} \tag{1}$$

$$\overline{R}\ X = \bigcup \{Y \in U / R \mid Y \cap X \neq \varnothing\} \tag{2}$$

They are called the upper approximation and lower approximation of relation P respectively. If the upper and lower approximation are not equal, X is called the rough set of R, otherwise, called the defined or precise set of R.

2.2 Theory of Formal Concept Analysis

Formal Concept Analysis are widely used in many area, such as data analysis and rule extraction, the core of which is the concept lattice, that is the concept hierarchy based on binary relation. Concepts exist in the form of relations of posets in the lattice. For each concept, they are composed of extensions and intensions of their own. The relationship of the concept of upper and lower nodes is the one of father and son. The concept hierarchy between concepts can be clearly seen by the Hasse map. The visualization of data could be easily got.

Definition 2. There is a set of L, and a，b，c∈L, the prerequisites that binary relation $\leq$ is the poset of L are:1）a$\leq$a (reflexivity); 2）a$\leq$b and b$\leq$a$\Rightarrow$a=b（anti-symmetry）; 3）a$\leq$b and b$\leq$c$\Rightarrow$a$\leq$c（transitivity）.

Definition 3. (Infimum and Supremum) There is a subset S∈L in the poset （L，$\leq$）, and then arbitrary element in set L is called the lower bound of subset S. And if it existing a largest element, we call it Infimum; similarly, the definition of Supremum could be got.

Definition 4. (Lattice) The poset （L，$\leq$）can be called Lattice only if it could Satisfy the following requirements: 1）For any a, b∈L, a$\wedge$b and a$\vee$b are both exist; 2）Infimum and Supremum are presence for any a, b∈L.

2.3 Missing-Value Context and Rough Formal Concept

Ordinary formal context K=（G, M, I） is a triple context, G is a set of objects, M is a set of attributes, and I is a binary relation such that $I \subseteq G \times M$, and it is identified. However, in real life due to lack of information or unpredictable cases happening, knowledge we need can not be obtained in the normal way, which makes it hard to express. Based on the reasons above it is necessary to take specially steps to deal with these problems. First of all, the definition of missing-value context is given. When the relationship between certain object g and property m is uncertain, I $\subseteq$(g, m) not being judged, it could be called being missing-value [3].

Definition 5. (Missing-Value Context) Missing-Value Context T=（U，A，R） is a triple context, U={o1,o2, ……on} is a set of objects, A={a1,a2, ……an} is set of Attributes. R is an uncertain relationship between U and A.

Better to explain the definition of rough formal concept, formal concept is given firstly.

Definition 6. (Formal Concept) For a given formal context K=（G, M, I）, concept (A , B) is called Formal Concept, if it satisfies the conditions that 1）A∈G，B∈M；2）（A ，B）∈I；3）A'=B, B'=A, among which A' represents the shared attribute sets of objects A, B' is on behalf of the shared object sets of attributes B.

Definition 7. (Rough Formal Concept) For a given Missing-Value Context T=（U，A，R）, only if the concept meets the conditions as follows, it could be named Rough Formal Concept: 1) A∈G, B∈M; 2) A$\times$B is a rough set based on relation I.

3. Traditional Similarity Calculation Model

3.1 Tversky Ratio Model

Tversky measured the degree of similarity between concepts by using the shared feature sets of entities. The computational model is as follows:

$$Sin(m,n)=\frac{f(M \cap N)}{f(M \cap N)+a \cdot f(M-N)+b \cdot f(N-M)} \quad (3)$$

Among which Sin (m, n) denotes the similarity between concept m and n; M and N are the feature sets of m and n; f is the metric function of feature sets; (M-N) indicates the feature sets

which lies in M rather than in N; Similarly, (N-M) indicates the feature sets which lies in N rather than in M; Parameter a、b adjust the cases dealing with the asymmetric feature sets.

3.2 The Similarity Calculation Model Based on Formal Concept Analysis

Based on the Tversky Ratio Model, the similarity calculation model to the basis of Formal Concept Analysis is proposed by Souza and Davis, which uses the mathematical operation $\wedge$ and $\vee$ to calculate the irreducible infimum.

$$Sin(m,n)=\frac{|(m\vee n)^{\wedge}|}{|(m\vee n)^{\wedge}|+a\,|(m-n)^{\wedge}|+(1-a)\,|(n-m)^{\wedge}|} \qquad (4)$$

$m\vee n$ represents the supremum of formal concept m and n; $(m\vee n)^{\wedge}$ means the element sets of irreducible infimum of the supremum features; $(m\text{-}n)^{\wedge}$ denotes the irreducible infimum element sets which lies in m instead of n; and vice versa.

3.3 The Similarity Calculation Model Based on Information

Based on the Tversky Ratio Model, Formica put forward the similarity calculation model that based on information, which makes use of the similar map of concepts in domain knowledge.

$$Sin((M1,N1),(M2,N2))=\frac{|M1\ \mathbf{I}\ M2|}{l}\times w+[\frac{1}{q}\bullet \max(\sum_{<m,n>\in P} f(m,n))]\times(1-w) \qquad (5)$$

M1 $\mathbf{I}$ N1 is the number of the same objects in the pairs of the concepts; l is the bigger value to be compared with the numerical objects; q is the larger value to be compared with the numerical attributes; $\sum_{<m,n>\in P} f(m,n)$ is the summation of the concepts whose attributes matches one another in the concept similarity map; w is the weighting factor adjusting different cases.

It is not difficult to draw conclusions from the model above that many scholars just improved the model raised by Tversky. The semantic meaning of similarity model is enriched by pulling in the Rough Set Theory and Formal Concept Analysis. But there are still limitations, specifically expressed in two aspects: 1) simply counting the numbers of upper concept nodes, lacking of accuracy measurement, 2) no consideration about the differences between the feature properties of different concept level, only depending on computing the semantic distance between concepts to determine the value of similarity.

4. Rough Formal Concept Similarity Calculation Model

4.1 Equivalence Relation, Irreducible Supremum and Infimum in Formal Concept Analysis

Formal concept is formed of objects and attributes, which can determine the equivalence relation of the object G and attribute M by the formal context K=（G, M, I）. For every non-empty set of formal concept, there always exists the sole largest sub-concept and smallest parent concept, which is called Supremum and Infimum. If existing the only element not expressed by the largest sub-concept of others, it is named Irreducible Supremum element; Similarly, it is easy to get the definition of Irreducible Infimum element.

4.2 Improved Rough Concept lattice Similarity Calculation Model

On the one hand, by observing the structure of the Hasse map, we know that the included attributes of the upper parent node in concept lattice is the minimal subset attributes of the lower sub-class nodes. And if two nodes have the same feature properties, the conclusion that they must have the same upper parent node can be got. On the other hand, from the taxonomic point of view, the similarity degree between underlying object is higher than the one between upper layer object. Based on the above analysis, the semantic parameters of the upper layer is greater than the one of the lower layer. The improved rough concept lattice similarity calculation model is given below.

$$f_{RSIM} = ((A_{1\_}, B_{1\_}), (A_{2\_}, B_{2\_})) = \frac{|A_{1\_} \mathbf{I} A_{2\_}|}{g} \times a + \frac{\sum_{i=1}^{n} X_i W_i}{\sum_{i=1}^{n} Y_i W_i} \times (1-a) \qquad \textbf{(6)}$$

Specific parameters are defined as follows:

Xi=fi(B1_ $\mathbf{I}$ B2_) represents the shared property features of the rough concept lattice in level i;

Yi=fi(B1_ $\mathbf{I}$ B2_)+fi(B1_ − B2_)+fi(B2_ − B1_) denotes the property features of the rough concept lattice in level i; Wi means the weight of the conceptual elements in level i.

A1_ is the lower approximation concept (A1_， B1_)'s object, while (A1_， B1_) is the rough formal concept of (A1，B1), so as the A2_; B1_ is the upper approximation concept (A1_，B1_)'s attribute; Parameter a is the weighting factor adjusting accuracy of the model.

The weight of the different level is determined by 1/2i-1, known through the literature [4], among which i stands for the number of the level.

In order to better explain the model above, a formal context of the domain ontology modeling of tea pests is given below, shown as table 1.

Table 1. Formal Context of the Domain Ontology Modeling of Tea Pests

Dark compo-und eyes	*Light skin color*	*Short reproductive cycle*	*Serious harm of*	*Severely destruction*

					nymph
Pest 1	×		×		
Pest 2	×		×		
Pest 3		×	×		×
Pest 4		×	×		
Pest 5				×	
	Easy to control	Widely distribute-d	Regional	Lots of feet	Harm for shoots
Pest 1	×				×
Pest 2	×				×
Pest 3		×			
Pest 4			×		×
Pest 5				×	×

Based on the literature[4], the generating algorithm for building concept lattice to the basis of matrix column rank with attribute priority, by which uses the matrix column rank of the concept and the union operation of the concept pairs to generate rough formal concepts having hierarchical structure. The hierarchical concepts corresponding with table1 are just as follows.

Layer 1 ： C1 $(G,\varnothing)$

Layer 2 ： C2{{ Pest 1，Pest 2，Pest 4，Pest 5}，{ Harm for shoots }}，C3{{ Pest 1，Pest 2，Pest 3，Pest 4}，{ Short reproductive cycle }}；

Layer 3 ： C4{{ Pest 1，Pest 2，Pest 4}，{ Harm for shoots，Short reproductive cycle }}，C5{{ Pest 1，Pest 2，Pest 5}，{ Easy to control }}；

Layer 4 ：C6{{ Pest 1，Pest 2}，{ Harm for shoots，Short reproductive cycle，Easy to control，Dark compound eyes }}，C7{{ Pest 3，Pest 4}，{ Light skin color，Short reproductive cycle }}；

Layer 5 ： C8{{ Pest 5}，{ Lots of feet，Harm for shoots，Easy to control，Serious harm of nymph }}，C9{{ Pest 3}，{ Light skin color，Severely destruction，Widely distributed，Short reproductive cycle }}，C10{{ Pest 4}，{ Light skin color，Harm for shoots，Regional，Short reproductive cycle }}；

Layer 6 ： C11 $(\varnothing,M)$

The rough formal concept with the hierarchical structure is generated just as the map 1. The weight of different layers is given in table 2. Using the formula above, the similarity between the concept nodes can be calculated. For example, the similarity value between concept C2 and C6 is got by setting the parameter a =0.25.

$$f_{RSIM(C_2,\ C_6)}=\frac{2}{4}\times 0.25+\frac{1\times 1+0\times\frac{1}{2}+0\times\frac{1}{4}+0\times\frac{1}{8}}{(1+3)\times 1+(0+1)\times\frac{1}{2}+(0+1)\times\frac{1}{4}+(0+0)\times\frac{1}{8}}=0.34 \quad (7)$$

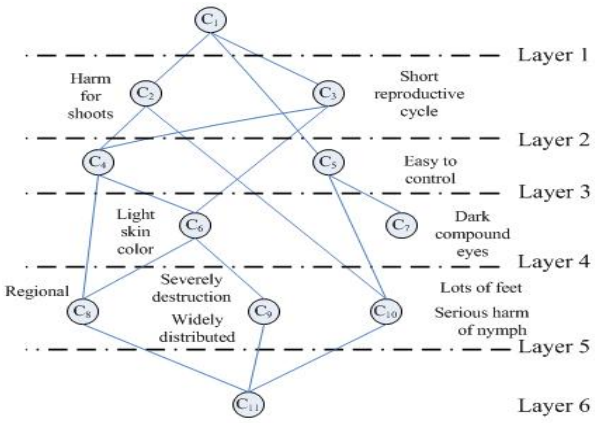

Fig. 1. Concept lattice with hierarchical structure

By using the similarity way, all the objects and attributes which are non-empty could be calculated. In order to analyze these data we get, the result of the improved model and the one of the Souza model are put together to make comparition shown as the table 3.The table is divided into two parts by the diagonal of the number one. The value of the upper triangular is the result of the improved similarity of rough formal concepts. The value of the lower triangular is the result of the similarity of the Souza model.

Fig. 2. Property values of the related levels

Layer 1	Harm for shoots，Short reproductive cycle	Weight：1
Layer 2	Easy to control	Weight：1/2
Layer 3	Light skin color，Dark compound eyes	Weight：1/4
Layer 4	Regional, Severely destruction, Widely distributed，Lots of feet，Serious harm of nymph	Weight：1/8

Fig. 3.Values of the similarity between concepts

	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_2	1	0.19	0.69	0.19	0.34	0.06	0.63	0.00	0.4
C_3	0.00	1	0.69	0.13	0.49	0.92	0.00	0.46	0.4
C_4	0.67	0.67	1	0.17	0.90	0.52	0.36	0.40	0.9
C_5	0.00	0.00	0.00	1	0.35	0.00	0.52	0.00	0.0
C_6	0.40	0.40	0.67	0.40	1	0.33	0.57	0.36	0.8
C_7	0.00	0.67	0.50	0.00	0.33	1	0.00	0.95	0.8
C_8	0.40	0.00	0.33	0.40	0.50	0.00	1	0.00	0.3
C_9	0.00	0.40	0.33	0.00	0.25	0.67	0.00	1	0.4
C_{10}	0.40	0.40	0.67	0.00	0.50	0.67	0.25	0.50	1

5. Model Analysis

For the reason that precision and recall always changes with the mutative thresholds, they can not be used to analyze the result of the similarity accurately. Based on the values in table 3, the concept node in the high-level such as C2 and its child node C4 are chosen to analyze the relationship between other concept nodes. Compared to the Souza model, the conclusion that the improved model is better could be got at two aspects. On the one hand, the result of accuracy is improved. The values of the improved model increase to the different degrees, for jointly considering attribute hierarchies in concept lattice and the relations between objects and attributes. On the other hand, irrelevant concept pairs decrease. Due to the concepts in the domain area have certain similar characteristics, all the values of the concept pairs could not be zero. The improved model cuts down the number of the pairs with zero effectively, enhancing the measurement accuracy between concept pairs.

6. Conclusions and future work

The paper puts forward the theory and methods for similarity calculation of rough formal concept in missing-value content, in which formal concept analysis theory and rough set theory are introduced to similarity computation. Jointly considering attribute hierarchies in concept lattice, the semantic hierarchies of the concepts are weighted differently. Experimental results show the model has a high computational accuracy. The model above proposes a practical theory and methods to merge domain ontology, helping to raising the accuracy of ontology integration.

References

1. Wu Qiang, Liu Zong-tian. The Rough Concept in FCA, Journal of Chinese Computer Systems, Beijing, vol. 26, pp. 1563-1565, April 2005.
2. Shao, M. Set approximations in fuzzy formal concept analysis, Fuzzy Sets and Systems, Beijing, 2007, pp.2627-2640.
3. Xie Zhi-peng, Liu Zong-tian. Concept Analysis and Knowledge Acquisition in Missing-Context, Computer Science, Beijing, 2000, pp. 36-39.
4. Mao Hua, Dou Lin-li. An Algorithm of Concept Lattice Based on Matrix Column Rank with Attribute Priority. Journal of Hebei University (Natural Science Edition), Shi jiazhuang, vol. 29, Dec. 2009, pp. 130-132.