

Study on distributed generation algorithm of variable precision concept lattice based on ontology heterogeneous database

Qingrong WANG^{1, a}, Changfeng ZHU^{2, b}

¹School of Electronic and Information Engineering, Lanzhou Jiaotong University
Lanzhou 730070, China

²School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou ,
730070, China

Email: ^acfzhu003@163.com, ^bwangqr003@163.com

Abstract. Integration of distributed heterogeneous data sources is the key issues under the big data applications. In this paper the strategy of variable precision is introduced to the concept lattice, and the one-to-one mapping mode of variable precision concept lattice and ontology concept lattice is constructed to produce the local ontology by constructing the variable precision concept lattice for each subsystem, and the distributed generation algorithm of variable precision concept lattice based on ontology heterogeneous database is proposed to draw support from the special relationship between concept lattice and ontology construction. Finally, based on the standard of main concept lattice of the existing heterogeneous database generated, a case study has been carried out in order to testify the feasibility and validity of this algorithm, and the differences between the main concept lattice and the standard concept lattice are compared. Analysis results show that this algorithm above-mentioned can automatically process the construction process of distributed concept lattice under the heterogeneous data sources.

1 Introduction

Heterogeneous information system is generally divided into four kinds, namely semantic heterogeneity, platform heterogeneity, grammar heterogeneity and construction heterogeneity. In recent years, domestic and overseas scholars have carried out research in heterogeneous information system [1]. CORBA, federated database, virtual database DCOM have been used to handle different types of heterogeneous problem. However, because of the particularity of semantic heterogeneous, the problem has not been solved satisfactorily [2]. Firstly, different systems use different terminology to express the same attribute or concept; secondly, the same terminology express different concepts in different environments; thirdly, the terms which express the same concepts are used in different types of data in different data sources; fourthly, the concept of different data sources are incomplete parity. In the paper [3,4], relevant scholars put forward many algorithms which with ontology as the analysis basic, they can clearly describe the concept, but also can accurately record the relationship of concepts. The paper [5] use the full concept lattice theory and by means of the theory of formal concept analysis to get ontology. However, these algorithms are based on the ontology already exists or obtained by manual construction, causing the algorithm has certain limitation. Therefore, ontology



of the automatic of semi-automatic construction has become an important work of heterogeneous database data mining.

Based on the study of related theory and algorithm, introduce variable threshold concept lattice, distributing constructing algorithm BLDCL (based on ontology distribute concept lattice generation algorithm) for solving the heterogeneous database concept lattice have been proposed in the paper. The variable threshold concept lattice was constructed for each subsystem with algorithm. According to mapping relationship between variable threshold concept lattice and Ontology concept lattice, local Ontology was generated automatically or semi-automatically. Then use the method mentioned in the document [6,7] to get the global ontology.

2 The basic concept

Definition 1: Formal context $K = (E, A, R)$ form the lattice with the partial order relation is concept lattice [8]. E is instance set. A is attribute set. R is relationship set. Even for a concept $C(E_c, A_c)$, $f(E_c) = A_c$ and $g(A_c) = E_c$ are satisfied.

Definition 2: Variable threshold concept lattice is a concept lattice satisfies the following condition based on definition 1:

$\delta(K, A_c)$ used to measured the potential of the set in concept $C(E_c, A_c)$.

Definition 3: Sets $O(C, E, R, F, I)$ is ontology. C is concept set. E is instance set. R is the relationship of concept set. F is concept hierarchical relation set. I is axiom (regulation) set.

Definition 4: The indivisible concept is atomic concept, denoted as C^{atom} . Atomic concept is the concept which has only one attribute, then $a(i) = h(C_i^{atom})$. h is the mapping relationship between atomic concept and attribute.

Definition 5: Local ontology is the ontology that meets the following conditions:

① Any one of concept in ontology O has limited sub-concepts, namely $\forall C \in O$. There are $N(C)$ sub-concepts such that $C = \bigcup_{i=1}^{N(C)} C_i$ any sub-concepts C_i can be expressed as $C_i = \bigcup_{j=1}^{N(C_i)} C_j$.

② Any one of concept in ontology can be decomposed into atomic concepts.

Definition 6: A number of local ontology from the same field are merged into one ontology called the global ontology.

Definition 7: The attribute set of all concepts in ontology is ontology attribute set, denoted by $A(O)$.

Variable threshold concept lattice can generate an isomorphic with ontology. Therefore, we have theorem 1.

Theorem 1: Exist a ontology and the variable threshold concept lattice isomorphism.

Proof: Set any one of concept in variable threshold concept lattice meet the following condition

$$A_c = \{a \mid |g(A_c)| \geq \delta(K, A_c)\}$$

This concept lattice construct concept set is

$$C_{lattice} = \{C \mid C(E_c, A_c)\}$$

Construct ontology $O(C_{lattice}, E, \{f, g\}, \leq, I)$, and $I = \{A_c = \{a \mid |g(A_c)| \geq \delta(K, A_c)\}\}$.

Therefore, we obtain the following two conclusions:

① There is a one-to-one mapping between the concept in ontology O and the concept of variable threshold concept lattice.

② Any one of relationship for any one of concept in ontology O can be found in relation set which corresponding to variable threshold concept lattice and vice versa.

Integrated ① and ② indicated that O and L is isomorphism. QED.

When the accuracy requirements are very low, namely, $\delta(K, A) = 0$. Then we get corollary 1, the concept lattice can generate an ontology which isomorphic with it.

Corollary 1: There is an ontology and concept lattice isomorphism.

Proof: By Theorem 1 know that there is an ontology and variable threshold concept lattice isomorphism. Set the condition $\delta(K, A) = 0$ in this variable threshold concept lattice, and then this variable threshold concept lattice becomes a concept lattice which meets Definition 1. Namely, there is an ontology and concept lattice isomorphism. QED.

Theorem 2: Ontology $O(C, E, R, F, I)$ can generate a concept lattice $L(E, A(L), f, g, \leq)$ which is isomorphic with it.

Proof: Any one of concept $\forall C \in O$ in Ontology $O(C, E, R, F, I)$ is composed of $N(C)$ numbers of sub-concepts, namely

$$C = \bigcup_{i=1}^{N(C)} C_i \quad (1)$$

And any sub-concept C_i can be expressed as

$$C_i = \bigcup_{j=1}^{N(C_i)} C_j \quad C_i = \bigcup_{j=1}^{N(C_i)} C_j \quad (2)$$

By condition ② in Definition 5 know that the finite iteration (2) to (1) can obtain the concept set which is composed of atomic concept, namely

$$C = \bigcup C_i^{atom}$$

By Definition 4 know that each atomic concept corresponding to a attribute, namely

$$a(i) = h(C_i^{atom})$$

Where h is the mapping relationship between atomic concepts and attributes.

Set $\forall C \in O$ is composed of N numbers atomic concepts, these atomic concepts are mapping into a attribute set is

$$A(C) = \{a \mid a = a(i), i = 1, L, N\}$$

And set all concepts in ontology O corresponding to attribute set is

$$A(O) = \{a \mid a \in A(C), \forall C \in O\} \quad (3)$$

Formal context $K(E, A(O), R)$ generate any one of concept $C^*(E^*, A^*)$, $E^* \subseteq E, A^* \subseteq A(O)$ among that. And existence of the following relationship

$$E^* = f(A^*) = f(C^*) \quad (4)$$

If $C^* \notin O$, then by condition ③ in Definition 5 and formula (4) know that $E^* = \Phi$, so

$$C^* \notin L(E, A(L), R(L)) \quad (5)$$

Conversely, if $C^* \in O$, then by condition ③ in Definition 5 and formula (4) know that $E^* \neq \Phi$, so

$$C^* \in L(E, A(L), R(L)) \quad (6)$$

Thus $\forall C^\diamond \in L(E, A(L), R(L))$, then

$$C^\diamond \in O \quad (7)$$

By formula (7) know that any one of concept in $L(E, A(L), f, g, \leq)$ have parent concept or sub concept, namely has the partial relation \leq . Thus formal context $K(E, A(O), R)$ generate the lattice $L(E, A(L), f, g, \leq)$ is concept lattice. QED.

By Theorem 2 know that the global ontology can construct the main concept lattice.

Theorem 3: Set variable threshold concept lattice generated from formal context $K(E, A, R)$ as $L(E, A, f, g, \leq)$. There is a ontology $O(C^\diamond, E^\diamond, R^\diamond, F^\diamond, I^\diamond)$ generated from L meet the condition $A \subseteq A(O)$, $E \subseteq E^\diamond$, then for $\forall C_1(A_1, E_1) \in L(E, A, f, g, \leq)$, if there is δ makes $|E_1| \geq \delta$, then $C_1 \subseteq C^\diamond$ and $C_1 \in O$.

Proof: By known from problem set

$$A \subseteq A(O) \quad (8)$$

Set any one of concept in L as $C_1(A_1, E_1)$, and then get relationship from problem set and formula (8)

$$A_1 \subseteq A \subseteq A(O) \quad , \quad E_1 \subseteq E \subseteq E^\diamond \quad (9)$$

By Definition 1 know that there is $f \in R$

$$E_1 = g(A_1) \quad (10)$$

According to conditions of problem

$$|E_1| \geq \delta \quad (11)$$

According to formula (11), (12) and Definition 2, set $\delta = \delta(K, A_1)$, then get $C \subseteq C^\diamond$ and $C \in O$. QED.

3 Algorithm descriptions

According to mapping relationship between variable threshold concept lattice and Ontology concept lattice, local Ontology was generated automatically or semi-automatically. Then use the method mentioned to get the global ontology[9]. Finally, use global ontology to construct main concept lattice. The basic procedure of the algorithm is shown as Figure1.

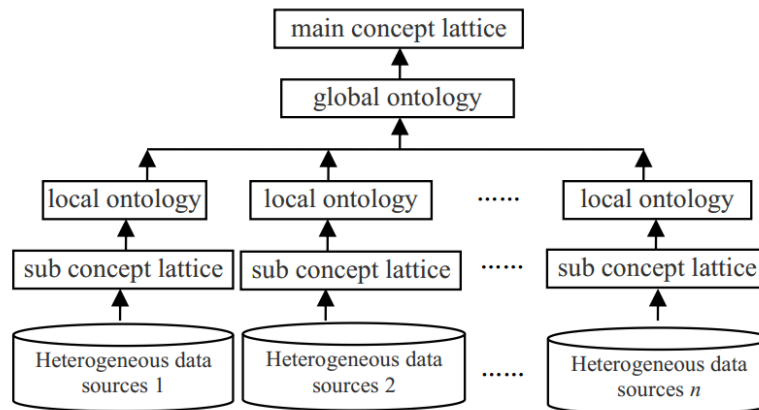


Figure 1: Flowchart of algorithm

Algorithm BODCL: (Based on ontology distribute concept lattice generation algorithm)

Input Φ

Output L

Begin /* Subsystem algorithm */

Step1: The heterogeneous data source i generate variable threshold concept lattice (sub concept lattice);

Step2: The sub-concept lattice automatically generate local ontology by using the method which mentioned in Theorem 1;

Step3: Send local ontology to main system;

End

Begin /*Main system algorithm*/

Step1: Local ontology merging into global ontology by using the method which mentioned in document[10];

Step2: Transform the global ontology into main system and generate concept lattice (main concept lattice) by using the method which mentioned in Theorem 2;

End

4 Experiments and Analysis

4.1 Standard of experiment

The experiment reference the gold standard mentioned [10] as criteria. Select an existing main concept lattice generated from heterogeneous database as standard.

The main measurement index:

(1) Concept difference $diffC(L_{test}, L_{st})$:

$$diffC(L_{test}, L_{st}) = \left(\frac{|C_{test} / C_{st}|}{|C_{st}|} + \frac{|C_{st} / C_{test}|}{|C_{st}|} \right) / 2$$

(2) Concept hierarchy similarity $sim(L_{test}, L_{st})$:

$$sim(L_{test}, L_{st}) = \sum_{c \in C_{test} \cap C_{st}} simL(c, L_{test}, L_{st})$$

Where $simL(c, L_{test}, L_{st}) = (simP + simC) / 2$

$$simP = \frac{|parents(c, L_{st}) \cap parents(c, L_{test})|}{|parents(c, L_{st})|} \quad simC = \frac{|children(c, L_{st}) \cap children(c, L_{test})|}{|children(c, L_{st})|}$$

Where $parents(c, L)$ is the parent concept numbers of concept c in L . $children(c, L)$ is the sub concept number of concept c in L . $diffC(L_{test}, L_{st})$ is the extent of the deviation between concept set generated automatically by algorithm and standard concept set. $sim(L_{test}, L_{st})$ is the total deviation between concept relationships of corresponding position.

The test sets use the data of human resource management system in two branch schools of Beijing Longwen Education Centre. All data size set as 10000. The information of test sets is shown as Table 1.

	branch schools 1		branch schools 2	
	TN	AN	TN	AN
T1	E	18	YG	16
T2	S	15	GZ	12
T3	W	18	JX	23

Table 1: The information of test set

In Table1, TN is table name, AN is number of attributes, T is test set, E is employee, YG is yangong, S is salary, GZ is gongzi, W is performance, JX is jixiao.

4.2 Experimental results and analysis

To verify the feasibility of this algorithm, the testing environment is set as following: in the Windows XP operating system, using C language and MPI parallel language to realize BODCL.

	Generated concept	Standard concept
T1	1189	980
T2	786	452
T3	2982	2801

Table2: Variation between generated concept and standard concept

	$diffC$	Sim
T1	17.86%	50.23%
T2	5.77%	87.08%
T3	28.90%	79.71%

Table 3: Important index value

The experimental results show that the algorithm proposed in this paper can automatically process the construction process of distributed concept lattice under heterogeneous data source. And the results that compare the main concept lattice generated automatically with the standard main concept lattice are ideal.

5 Conclusions

Based on the existing research, the strategy of variable precision is introduced to the concept lattice, and the one-to-one mapping mode of variable precision concept lattice and ontology concept lattice is constructed to produce the local ontology by constructing the variable precision concept lattice for each subsystem, and the distributed generation algorithm of variable precision concept lattice based on ontology heterogeneous database is proposed to draw support from the special relationship between concept lattice and ontology construction in this paper. Analysis results show that this algorithm above-mentioned can automatically process the construction process of distributed concept lattice under the heterogeneous data sources, and can get a better result.

Acknowledgment

This work is supported by The National Nature Science Foundation of China (No. 61364028) , The Chinese Ministry of Education of Humanities and Social Science Project (No. 15XJAZH002). The Nature Science Foundation of Gansu Province (No. 1508RJZA061). The science and technology planning project of Lanzhou (No. 2015-2-94).

References

- [1] Amit P. Sheth. ,“Changing focus on interoperability in information systems: Form system , syntax , structure to semantics”, *Interoperating Geographic Information Systems*, Boston: Kluwer Academic Publishers, PP. 1-25, (2005).
- [2] Noy N. F., Musen M. A. PROMPT. “Algorithm and tool for automated ontology merging and alignment”, Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth conference on Innovative Applications of Artificial Intelligence, AAAI Press, PP. 450-455, (2000).
- [3] LI Xing-yi, GAO Wen-hao, SHI Hua-ji. “Heterogeneous databases integration approach based on ontology”, *Computer Engineering and Design*, volume 30, PP. 1931-1935,(2009).
- [4] LIU Qiang, ZHAO Di, ZHONG Hua, et al.“Ontology-Aided Automatic Schema Matching”, *Journal of Software*, volume 20, PP. 234-245, (2009).
- [5] Tao Chun, Zhang Liang, Shi Baile. “Query processing for ontology-based XML data integration”, *Journal of Computer Research and Development*, volume 28, PP. 68-77, (2005).
- [6] HU Wei, QU Yuzhong, Falcon AO. “A Practical Ontology Matching System”. *Web Semantics: Science, Services and Agents on the Worlds Wide Web*, volume 6, PP. 237-239, (2008).
- [7] LIU Chunchen , LIU Dayou, WANG Shengsheng, et a. “Improved Semantic Similarity Calculating Model and Application” , *Journal of Jilin University*, volume39, PP. 119-123, (2009).
- [8] TANG Jie, LIANG Bang-Yong, LI Juan-Zi. “Automatic Ontology Mapping in Semantic Web”,*Chinese Journal of Computers*, volume29, PP. 1956-1974, (2008).
- [9] ZHANG Peng, YANG Feng, LU Shuai. “Clustering-Based Ontology Block Matching Approach”,*Journal of Jilin University*, volume49, PP. 493-497, (2011).
- [10] MA Chi, WANG Nian-bin, ZHANG Hai-yan. “Similarity Calculating-based Ontology Mapping Framework”, *Computer Engineering*, volume35, PP.61-63, (2009).