

Study on association rules mining based on semantic relativity

Zhang Lei Xia Shixiong Zhou Yong Xia Zhanguo

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: An association rules mining method based on semantic relativity is proposed to solve the problem that there are more candidate item sets and higher time complexity in traditional association rules mining. Semantic relativity of ontology concepts is used to describe complicated relationships of domains in the method. Candidate item sets with less semantic relativity are filtered to reduce the number of candidate item sets in association rules mining. An ontology hierarchy relationship is regarded as a directed acyclic graph rather than a hierarchy tree in the semantic relativity computation. Not only direct hierarchy relationships, but also non-direct hierarchy relationships and other typical semantic relationships are taken into account. Experimental results show that the proposed method can reduce the number of candidate item sets effectively and improve the efficiency of association rules mining.

Key words: ontology; association rules mining; semantic relativity

So many rules are obtained through the existing association rules mining algorithms, but most of the rules are of little use. In order to solve the problem, domain knowledge and background knowledge are imported into the process of association rules mining. Xie et al. proposed an optimized *a priori* mining algorithm based on ontology^[1]. Češpivová et al.^[2] introduced medical ontology and other background knowledge into the process of association mining. Kuo et al.^[3] described an approach to categorize attributes in preparation for mining association rules in the data. Farzanfar et al.^[4] proposed a new approach that improved efficiency in the classical fuzzy association rules mining problem by providing the capability to handle domain ontology relationships. Wu et al.^[5] showed an ontology-based system framework for multi-dimensional association rules mining. Tseng et al.^[6] devised two efficient algorithms for mining association rules with ontological information. Won et al.^[7] proposed a method to allow smaller and more relevant search space compared to the original data sets. Only a direct hierarchy relationship was considered in the above researches. In fact, there are so many complicated relationships in the ontology, such as direct hierarchy relationships, indirect hierarchy relationships and other typical semantic relationships. How to use the most relationships in the ontology to improve association rules mining is a problem.

In this paper, an association rules mining method based on semantic relativity is proposed by combining ontology with

association rules mining. Not only direct hierarchy relationships and indirect hierarchy relationships, but also other typical semantic relationships are considered in the method.

1 Problem Statements

Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of items, $i_j (1 \leq j \leq n)$ is an item of the dataset. Given dataset $DB = \{t_1, t_2, \dots, t_n\}$, DB denotes the set of transactions. Each transaction $t_i = \{tid, A\}$ has a unique identifier tid and a set of items $A, A \subseteq I$.

Suppose that $X, Y \subseteq I$ and $X \cap Y = \emptyset$, then $X \Rightarrow Y$ is an association rule. The support of this rule is denoted as $S(X \Rightarrow Y)$ and the confidence of the rule is denoted as $C(X \Rightarrow Y)$.

$$S(X \Rightarrow Y) = |\{t \mid t \text{ includes } X \text{ and } Y\}| / |DB|$$

$$C(X \Rightarrow Y) = |\{t \mid t \text{ includes } X \text{ and } Y\}| / |\{t \mid t \text{ includes } X\}|$$

Association rules mining means finding all association rules which support the rules and whose confidences are greater than minimum support s_{\min} and minimum confidence c_{\min} , respectively.

Definition 1 Ontology is defined as $O = (C, RT, R, H)$, where C is concept set, $c \in C$ expresses one concept; RT is the set of semantic relationship types, $RT = \{\text{SameAs}, \text{DisjointWith}, \text{Equivalent}\}$; R expresses the set of non-hierarchy relationships. A relation $r \in R$ is expressed as $r = (c_i, c_j, RT)$. Among them, $c_i, c_j \in C, rt \in RT$. H expresses the concept hierarchy set; $H \subseteq C \times C$; $(c_i, c_j) \in H$ expresses that c_i is the sub-concept of c_j .

Each item i_j of a dataset is the concept in an ontology.

2 Semantic Relativity

2.1 Hierarchy relativity

The existing calculation methods of concept hierarchy relativity are classified into the side calculation method^[8], the information contents method^[9] and the mixture method^[10] etc. The concept hierarchy is regarded as the hierarchy tree in the above methods. But, a concept may have several parent concepts in application. So, the concept hierarchy in an ontology should be taken as a directed acyclic graph rather than a hierarchy tree.

Definition 2 The hierarchy path is the sequence of concepts associated with hierarchy relationships. $HPath = \{c_1, \dots, c_i, \dots, c_m\}$, where $\forall c_i, 1 \leq i \leq m-1; (c_i, c_{i+1}) \in H$; $|HPath|$ is the number of concepts in the hierarchy path; $HPath(i)$ is the i -th concept in the hierarchy path.

Definition 3 Concept hierarchy paths for concepts c_1 and c_2 are the set of hierarchy paths connecting both concepts. $CPath(c_1, c_2) = \{HPath_1, \dots, HPath_i, \dots, HPath_m\}$, where $HPath_i(1) = c_1, HPath_i(|HPath|) = c_2$.

In concept hierarchy paths, $|CPath(c_1, c_2)|$ is the number of different hierarchy paths between c_1 and c_2 .

Definition 4 The ancestry of concept c is the set of con-

Received 2008-04-15.

Biographies: Zhang Lei (1977—), male, doctor, lecturer, zhanglei_zyx@163.com; Xia Shixiong (1961—), male, professor, xiasx@cumt.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 50674086), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20060290508), the Science and Technology Fund of China University of Mining and Technology (No. 2007B016).

Citation: Zhang Lei, Xia Shixiong, Zhou Yong, et al. Study on association rules mining based on semantic relativity[J]. Journal of Southeast University (English Edition), 2008, 24(3): 358 – 360.

cepts co-existing with the hierarchy path and is denoted as $\text{Ascent}(c) = \{c_i \mid \exists \text{CPath}(c_i, c), \text{CPath}(c_i, c) \neq \emptyset\}$.

Definition 5 (disjunctive ancestry) For concept c , $\forall c_1, c_2 \in \text{Ascent}(c)$. If the following conditions are satisfied:

- $\exists \text{HPath}, \text{HPath} \in \text{CPath}(c_i, c) \wedge c_j \notin \text{HPath}$;
- $\exists \text{HPath}, \text{HPath} \in \text{CPath}(c_j, c) \wedge c_i \notin \text{HPath}$.

Then, c_i or c_j is called disjunctive ancestry with each other for c . It is denoted as $\text{DisAscent}(c, c_i, c_j)$.

Definition 6 (common ancestry) The ancestry with two common concepts are called common ancestry. Common ancestry of concept c_1 and c_2 is denoted as $\text{CAscent}(c_1, c_2) = \{c_i \mid c_i \in \text{Ascent}(c_1), c_i \in \text{Ascent}(c_2)\}$.

Definition 7 (common disjunctive ancestry) For $\forall c_1, c_2$, their common disjunctive ancestry, $\text{CDAscent}(c_1, c_2)$, must satisfy the following conditions:

- $\text{CDAscent}(c_1, c_2) \subseteq \text{CAscent}(c_1, c_2)$;
- $\forall c_i, c_j \in \text{CDAscent}(c_1, c_2), \text{DisAscent}(c_i, c_1, c_2), \text{DisAscent}(c_j, c_1, c_2)$;
- $\forall c_i, c_j \in \text{CDAscent}(c_1, c_2), \text{CPath}(c_i, c_j) = \emptyset \wedge \text{CPath}(c_j, c_i) = \emptyset$.

For concept $c \in C$, its information quantity is expressed as $\text{IC}(c)$. $\text{IC}(c) = -\log(p(c))$, where $p(c)$ is the probability of the concept c and its sub-concept.

$$p(c) = \frac{|I(c)| + \sum_{c', c \in \text{Ascent}(c')} I(c')}{|I(C)|}$$

Average information content of common disjunctive ancestry is taken as the shared information of concept c_1, c_2 .

$$\text{Share}(c_1, c_2) = \frac{\sum_{c \in \text{CDAscent}(c_1, c_2)} \text{IC}(c)}{|\text{CDAscent}(c_1, c_2)|}$$

Concept hierarchy relativity is

$$\text{CHR}(c_1, c_2) = \frac{2\text{Share}(c_1, c_2)}{\text{IC}(c_1) + \text{IC}(c_2)}$$

2.2 Typical semantic relativity

Semantic relationship weight, $\text{rw}(r_i)$, is associated with each type of semantic relationship to express the weight of semantic relationship.

Definition 8 Semantic path is denoted as SRPath . $\text{SRPath} = \{c_1, c_2, \dots, c_i, \dots, c_m\}$, where $\forall c_i, 1 \leq i \leq m-1, \exists r_k(c_i, c_{i+1}) \in R, r_k \in \text{RT}$. $|\text{SRPath}|$ is the number of concepts in the semantic path. $\text{SRPath}(i)$ is the i -th concept in the semantic path.

Definition 9 Concept semantic path of concept c_1 and c_2 is $\text{SAss}(c_1, c_2)$. $\text{SAss}(c_1, c_2) = \{\text{SRPath}_1, \dots, \text{SRPath}_i, \dots, \text{SRPath}_j\}$. Where $\text{SRPath}_i(1) = c_1, \text{SRPath}_i(|\text{SRPath}|) = c_2$.

Semantic path strength, $\text{Stren}(\text{SRPath})$, is used to quantify the semantic path.

$$\text{Stren}(\text{SRPath}) = \prod_{c_i, i=1, |\text{SRPath}|-1, r_i(c_i, c_{i+1})} \text{rw}(r_i)$$

For concept c_1 and c_2 , their typical semantic relativity is denoted as $\text{CSR}(c_1, c_2)$.

$$\text{CSR}(c_1, c_2) = \max_{\text{SRPath}_i \in \text{SAss}(c_1, c_2)} \text{Stren}(\text{SRPath}_i)$$

Concept semantic relativity is denoted as $\text{CR}(c_1, c_2)$.

$$\text{CR}(c_1, c_2) = \frac{\omega_H \times \text{CHR}(c_1, c_2) + \omega_R \times \text{CSR}(c_1, c_2)}{\omega_H + \omega_R}$$

where ω_H, ω_R are weights.

3 Association Rules Mining Based on Semantic Relativity

The algorithm is based on semantic relativity. The candidate item sets are pruned as follows:

1) Only items with semantic relativity greater than user-defined minimum semantic relativity(sr_{\min}) are considered in candidate item sets;

2) Any candidate item set that contains both an item and its ancestry concept can be pruned.

The proposed algorithm is described as follows:

Input: Dataset DB, Ontology O , minimum support s_{\min} , minimum confidence c_{\min} , and minimum semantic relativity (sr_{\min});

Output: The set of association set L , according to support, confidence and semantic relativity.

1) Compute the semantic relativity of items in DB;

2) Generate frequent fuzzy item sets.

Repeat

If $k=1$ Find the frequent 1-itemsets L_1

Else Produce the candidate k item sets C_k from L_{k-1} .

Delete the candidate k -item sets in which the semantic relativity between the k -th item and any of the form $k-1$ items is less than sr_{\min} .

Delete any candidate in C_k that includes an item and its ancestry concept.

Compute the support degree and confidence degree for each item set in C_k ;

Delete the k -item sets of which support is less than s_{\min} or confidence is less than c_{\min} . The remaining item sets are frequent item sets L_k .

Until $L_k = \emptyset$

$L = \bigcup_k L_k$

3) Make association rules and output the set of association set L with according support, confidence and semantic relativity.

4 Experimental Results

Synthetic datasets are used in the experiments to evaluate the performance of the proposed algorithms. The parameters are that: the number of items is 328; the average size of transactions is 30 items; the ontology has 395 concepts, and the max concept hierarchy is 5. On average, one concept has three relationships. The execution times in different transaction numbers of the approach in this paper and the approach in Ref. [4] are compared. The results are shown in Fig. 1. In the approach in this paper, as the number of the transactions increases, the time of pruning candidates is longer than the time of scanning the item sets. The execution times and candidate item sets in different numbers of items without considering the support and the confidence of the approach in this paper and the approach in Ref. [4] are compared. It shows that the semantic relativity of items plays a more im-

portant role in pruning candidates as the number of items increases (see Fig. 2 and Fig. 3).

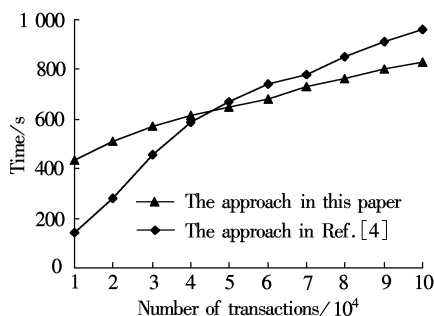


Fig. 1 Execution times for different numbers of transactions

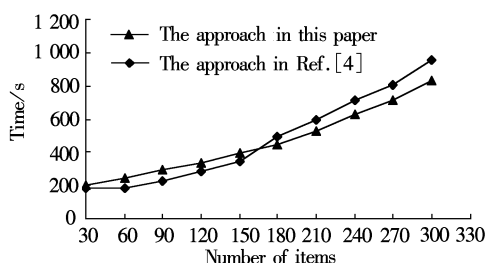


Fig. 2 Execution times for different numbers of items

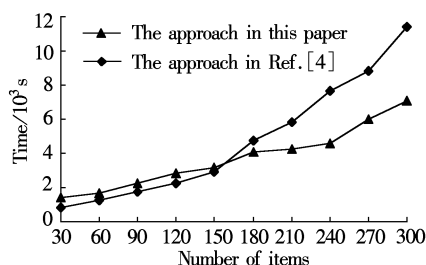


Fig. 3 Candidate item sets for different numbers of items

5 Conclusion

An association rules mining based on semantic relativity is proposed. Not only the hierarchy relationship, but also the other semantic relationships are considered in computing semantic relativity. Candidate item sets are pruned based on semantic relativity. It leads to shortening the execution time of the algorithm and makes association rules easier to understand with according semantic relativity. The experimental

results show that the approach in this paper is superior to the approach in Ref. [4].

References

- [1] Xie Hongwei, Yu Xueli, Li Juanli, et al. Study on ontology-based *a priori* algorithm applying to emergency decision system[C]//*Proceedings of Fuzzy Systems and Knowledge Discovery*. Nevada, USA, 2007: 669 – 673.
- [2] Češpivová H, Rauch J, Svátek V, et al. Roles of medical ontology in association mining CRISP-DM cycle [C]//*Proceedings of Knowledge Discovery and Ontologies at 15th European Conference on Machine Learning/8th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Pisa, Italy, 2004: 217 – 229.
- [3] Kuo Yen-Ting, Lonie Andrew, Sonenberg Liz, et al. Domain ontology driven data mining: a medical case study[C]//*Proceedings of the 2007 International Workshop on Domain Driven Data Mining*. San Jose, CA, USA, 2007: 11 – 17.
- [4] Farzanyar Zahra, Kangavari Moharrnadreza, Hashemi Sattar. A new algorithm for mining fuzzy association rules in the large databases based on ontology[C]//*Proc of Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'OG)*. Hong Kong, China, 2006: 65 – 69.
- [5] Wu Chin-Ang, Lin Wen-Yang, Wu Chuan-Chun. Ontology-assisted query formulation in multidimensional association rules mining [C]//*Proceedings of 2007 IEEE International Conference on Granular Computing*. Washington, DC, USA: IEEE Computer Society, 2007: 358 – 361.
- [6] Tseng Ming-Cheng, Lin Wen-Yang, Jeng Rong. Incremental maintenance of ontology-exploiting association rules [C]//*Proc of 2007 International Conference on Machine Learning and Cybernetics*. Hong Kong, China, 2007: 2280 – 2285.
- [7] Won Dongwoo, McLeod Dennis. Ontology-driven rules generalization and categorization for market data[C]//*Proceedings of the 23rd ICDE Workshops on Data Mining and Business Intelligence*. Istanbul, Turkey, 2007: 917 – 923.
- [8] Li Y, Bandar Z A, Mclean D. An approach for measuring semantic similarity between words using multiple information sources[J]. *IEEE Trans on Knowledge and Data Engineering*, 2003, **15**(4): 871 – 882.
- [9] Lord P W, Stevens R D. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation[J]. *Bioinformatics*, 2003, **19**(10): 1275 – 1283.
- [10] Rodríguez M Andrea, Egenhofer Max J. Determining semantic similarity among entity classes from different ontologies [J]. *IEEE Trans on Knowledge and Data Engineering*, 2003, **15**(2): 442 – 456.

基于语义相关性的关联规则挖掘研究

张磊 夏士雄 周勇 夏战国

(中国矿业大学计算机科学与技术学院, 徐州 221116)

摘要:为了解决传统关联规则挖掘中候选集数量过多, 计算时间复杂度过高的问题, 提出了基于语义相关性的关联规则挖掘方法. 该方法采用本体概念之间的语义相关性描述领域中的复杂关系, 通过语义相关度过滤掉领域中相关性较小的候选集, 以减少关联规则挖掘中候选集的数量. 计算语义相关性时, 将本体层次关系看作有向无环图而不是层次树, 不仅考虑直接层次关系, 还考虑非直接层次关系和其他典型语义关系. 实验结果表明, 该方法能有效减少候选集数量, 提高关联规则挖掘的效率.

关键词:本体; 关联规则挖掘; 语义相关性

中图分类号: TP311.5