Mining of Implicit Association Relationship of Text based on Audit Knowledge Base

Gang Liu*, TengFei Li*, ZhenYu Zhou
Harbin Engineering University, Harbin 150001, China

*Corresponding author e-mail: liugang@hrbeu.edu.cn, *1397159025@qq.com, b18334143@qq.com

Abstract. In the knowledge base, the concept is refined, and it is difficult to realize the analysis of correlation between concepts through statistical means. However, a large number of domain business relationships stored in the knowledge base provide the basis for the inter-concept relationship. By analyzing the characteristics of the direct correlation between concepts, we can find the indirect relation between the concepts implicit in the knowledge base. To measure the degree of the indirect correlation between the concepts, we can select the strong association relation that meets the requirement as the association rule with potential significance in the knowledge base.

1. Introduction
Knowledge database is an effective way to organize information. Taking the form of knowledge database, the management of surveyed information can unify the management of knowledge conception, manage the evolution of text knowledge and deal with various concepts and connections in the field of social security.

2. The Acquisition and Analysis of Domain Concept Relationship
The description of domain knowledge through ontology enables the computer system to operate on domain knowledge in the same form.

3. The Acquisition of Conceptual Relationships
At present, most ontology construction methods are based on certain application fields, and their application scope is limited. The common methods of ontology construction are as follows: IDEF5 method, skeleton method, METHONLOGY method, TOVE method, KACTUS method, seven-step method, SENSUS method.

3.1. The Type of Relationship Between Concepts
The relationship between concepts can be divided into similarity and correlation. Semantic similarity refers to the similarity between two concepts, usually referring to the degree of similarity between two concepts with some common characteristics. Semantic relevance refers to the degree of correlation between two concepts. There may be no similarity between the two concepts, but it can be related through other relationships. The basic similarity relationships are: inheritance, instance and part-
whole. The correlation relationship between concepts is divided into dependency, operational, causal, calculation, comparison and syntagmatic.

3.2. The Implicit Relationship in the Knowledge Database

In the domain knowledge database, the relationship between concepts is divided into two categories: explicit relationships and implicit relationships. Explicit relationships generated by the concept of semantic relations in information field. Implicit relationship generated by a number of explicit relationships between concepts through combination and semantic transmission.

4. The Strong Correlation of Concept Recessive Relation

The discovery of the knowable recessive relationship is based on the structure of the transmission chain composed of the dominant relationship, and the analysis of whether the whole transmission chain is related or not.

In this paper, factor matching theory is used to solve two problems: how to judge whether a relationship is implicit or not and how to determine its relationship type if it is implicit.

4.1. Determination of Standard Relational Factors and Characterization of Relational Factors

In this paper, the standard relational factors to be used are selected as follows: Composition, Connection, Function, Hierarchy, Space, Time, Similarity, Inherent, Physical, Separation, Contains and State. "Semantic relationships can be decomposed into factors" thoughts make relations can be expressed as the determining factor as the base of a vector space, to coordinate system to abstract the multidimensional space, factor is the dimension of the generalized coordinate system. The determinants of nine existing relationships in the knowledge base are shown in Fig.1.

![Fig.1 The factor distribution matrix of the relationship](image)

In Fig.1, the factor value "1" indicates that a relationship has the characteristics of this factor. The value of "0" indicates that some relationship does not have the characteristics of this factor. The value "α" indicates that it is difficult to determine whether a relationship has the characteristics of this factor, and the value is the interval (0, 1).

4.2. Relationship Factor Operation and Result Determination

In the transmission of the explicit relationship, it is necessary to judge which characteristics of the relationship can be passed on.

4.2.1. Vector Matching Calculation. According to the vector matching theory, common factors between each pair of concept vector can be determined, then the main task is to use common factors vector concept between the common values to calculate. In this paper, three calculation methods are proposed, which are as follows:

In this method, the common value is equal to the total number of dimensions with a value of 1 on the common factor, as shown in formula (1), $C_1$ represents the number of dimensions of 1:
In this method, the generic value is equal to the sum of all non-zero value dimension weights in the generic factor vector, as shown in formula (1):

$$x = C_1$$  \hspace{1cm} (1)

In this method, the generic value is equal to the sum of all non-zero value dimension weights in the generic factor vector, as shown in formula (2):

$$x = c_1 \frac{D}{2} + c_\beta + c_\alpha \ast \theta$$  \hspace{1cm} (2)

Which $c_1$ represents the number of dimensions of 1 in the common factor vector. $c_\beta$ represents the number of dimensions of $\beta$ in the common factor vector. $c_\alpha$ represents the number of dimensions of $\alpha$ in the common factor vector. $D$ represents the total number of the common factor vector dimensions, and $\theta$ represents the weight of the corresponding factors of $\alpha$, which is between 0 and D/2.

In this method, the generic value is equal to the sum of all the non-zero value dimensionality weights in the common factor vector and the sum of the dimensionality of the dimensionality of zero value, as shown in formula (3):

$$x = c_1 \frac{D}{2} + c_\beta + c_\alpha \ast \theta - c_0$$  \hspace{1cm} (3)

Which $c_1$ represents the number of dimensions of 1 in the common factor vector. $c_\beta$ represents the number of dimensions of $\beta$ in the common factor vector. $c_\alpha$ represents the number of dimensions in the common factor vector for $\alpha$. $c_0$ represents the number of dimensions in the common factor vector that are 0. $D$ represents the total number of the common factor vector dimensions, and $\theta$ represents the weight of the corresponding factors of $\alpha$, which is between 0 and D/2.

In the calculation of the common values, the calculated values obtained from the above three methods are all values greater than 1, while the common values between the concepts should be evaluated between 0 and 1. Therefore, it is necessary to normalize the common values obtained by the normalized formula for the three methods, so as to guarantee the value between 0 and 1. The normalized formula is as follows:

$$com = 2 \ast \left(\frac{1}{1 + k^{-x}} - 0.5\right)$$  \hspace{1cm} (4)

In this formula, $com$ represents the common value between concepts. $x$ is calculated as the result of formula (1), formula (2) and formula (3). Formula (4) is the formalized formula, which can normalize the value of the interval not in [0, 1] to [0, 1], where $k = 1.01$.

In order to evaluate the accuracy of the above three methods, 30 concept pairs were selected from the results of decomposition of conceptual factors. For these 30 pairs of concepts, artificial evaluation is adopted and the proposed method is used to calculate. Comparing the results calculated by the three methods with those obtained by manual evaluation, a more accurate calculation method is selected. The comparison analysis diagram of three methods and artificial evaluation are shown in Fig.2.
Fig.2 Comparison of various methods.

As can be seen from Fig.2, the second of the three methods is the closest to manual evaluation, and the curve of the second method is smooth and relatively stable. Therefore, this paper adopts method two as the calculation method and the formula (2) and (4) to calculate.

4.2.2. Relational Factor Operation. For the transfer effect of two dominant relationships, the effects of all the relationship factors that determine the characteristics of the relationship are obtained. In the factor space, this process can be transformed into operations of the corresponding dimensions of two relational vectors. To define the transfer operation of the relationship vector $\otimes$:

In the factor space of the relationship, there are any two relations $S_1$ and $S_2$, corresponding to: Vector$_1(x_1,x_2,\ldots,x_8)$ and Vector$_2(y_1,y_2,\ldots,y_8)$. And defining the Vector$_1 \otimes$ Vector$_2 = \text{Vector}_c(x_1 \times y_1, x_2 \times y_2,\ldots,x_8 \times y_8)$.

For how to discover implicit relationships, this paper proposes HAAMA (Hidden Attribute Association Mining Algorithms). The HAAMA is described as follows:

1. for each $C_i \in KD$
2. RM[i][0]←$C_i$ and RM[i][0]←$C_i$
3. for each $C_i \in KD$
4. if $\exists R_{\text{direct}}(C_i,C_j) \in KD$ and $i<j$
5. RM[i][j]←$R_{\text{direct}}(C_i,C_j)$
6. for each RM[i][j] = null and $i<j$
7. GetList(RM[i][j])
8. for each $R_{\text{list}} \in \text{RM[i][j]}$ and $i<j$
9. $R_{\text{temp}}\leftarrow R_{\text{list}}$
10. for ($m=1;m<\text{count}(R_{\text{temp}})-1;m++$
11. Vector←$R_{\text{temp}}[m] \otimes R_{\text{temp}}[m+1];$
12. if ((Vector $\in \text{AAM}$) or (Vector $\approx R$ and $R \in \text{AAM}$))
13. $R_{\text{temp}}[m+1]←\text{Vector}$
14. if (Vector is a new relation)
15. $R_{\text{temp}}[m+1]←\text{Vector}$ and AAM←Vector;
(16) if(Vector is not a relation)
(17) \[ R_{\text{temp}}[1] \leftarrow \text{null and break}; \]
(18) if(R_{\text{temp}}[1]=\text{null})
(19) delete R_list;

The description of the function GetList() is as follows (The number i and j of the two concepts of indirect connection need to be obtained):

(1) for each RM[i][k]\neq \text{null} and i<k and k<j;
(2) \quad \text{if} (RM[k][j] = \text{null and } k<j);
(3) \quad \text{if} (RM[k][j]=\text{null and } k<j);
(4) \quad \text{if} (RM[k][j]\neq \text{null and } k<j);
(5) \quad \text{RM}[i][j]\leftarrow \text{R_list}(C_i,C_k,C_j);

The time complexity of the whole algorithm is \[ O(n^2) = O(n^2 / 2) + O(n^3 / 4) + O(n^4 / 4) = O(n^4). \] Because the algorithm mainly constructs the relational matrix RM as the structure of data storage and analysis, its space complexity is \( O(n^2) \).

4.3. Extraction of Association Rules

After the discovery of the implicit relationship, the implicit relation with strong relation is selected as the association rule in the knowledge base.

4.3.1. The measure of the attribute association. Combined with relevant methods measuring the degree of knowledge base of ontology semantic relations, the above analysis, this paper analyses the concept of node in the network density, the relationship between depth and the connection type factors, such as dominant relationship measurement formula is given as follows:

\[
\text{Rel}(a,b) = \frac{\lambda \times \sqrt{\frac{1}{2} (e(a) + e(b))}}{\max[d(a),d(b)]}
\]

(5)

In this formula, \( \text{Rel}(a,b) \) represents the relation weight between concept \( a \) and \( b \). \( e(a) \) is the number of edges derived from concept \( a \), representing the local network density of concept \( a \). \( e(b) \) is the number of edges derived from concept \( b \), representing the local network density of concept \( b \). \( d(a) \) represents the depth of concept \( a \). \( d(b) \) represents the depth of concept \( b \). The lambda represents the weight of the different relationship types.

On the basis of explicit relationship measurement, this paper considers the influence of transmission factors, and gives the following formula of implicit relationship measurement:

\[
\text{Rel}(a,b) = \frac{\text{Rel}(a,m_i) \times \alpha_i + \cdots + \text{Rel}(m_i,b) \times \alpha_i}{i+1}
\]

(6)

\( m_i \) is the intermediate concept of the relationship between \( A \) and \( B \), where \( i \) is the number of intermediate concepts, \( \alpha_i \) is the attenuation factor of the explicit relationship measurement under the increment of the relationship transfer length.

4.3.2. The generation of association rules. The recessive relation of all correlated intensity relation Relation > Rel was selected after a given correlation strength threshold Rel. The "concept1=>concept2\|relation|relationship strength " output is used as the association rule in the knowledge base.

Comparing the traditional methods of finding association rules with statistical methods, we can find that the method of discovering association rules in this paper deviates from the traditional method, but it still follows the general rules of discovering association rules and successfully finds the implicit relationship hidden in a large number of semantic data.
5. Experiment and Result Analysis

5.1. Experimental Environment and Data
In this paper, the relevant literature data in the field of social basic endowment insurance were used as experimental data. These materials were divided into 126 small documents as the source information of audit information in the experiment. At the same time, Chengdu special office of audit office - computer audit experience compilation and the audit business analysis report Social security business analysis manual in the development process of "social security net audit system" by the audit office of Heilongjiang province and Harbin engineering university are referred to, and form the audit business process and description information of the social endowment insurance fund audit.

5.2. Results Analysis
On the same text data, this paper compares the classification mining effect of traditional association rule mining algorithm on text feature representation and the mining effect of knowledge base ontology representation of HAAMA algorithm text. The classification mining method of traditional association rule mining algorithm and the number of association rules obtained by HAAMA algorithm are shown in Fig.4. It can be found that the number of association rules mined by traditional association rules mining algorithm is far less than that mined by HAAMA algorithm. Through content analysis, can be found on each category HAAMA algorithm to obtain the number of association rules are classification mining methods than the traditional association rule mining algorithm. This is due to the effect of mining association rules based on statistical law is often influenced by the emergence of the concept of quantity, the text of some concepts such as calculation relationship between relationship, these in the policy text information cannot meet the requirements of statistics, the number of occurrences of the classification of the traditional association rule mining algorithm mining method will be ignored. The ontology representation of knowledge base will save this kind of relationship.

![Fig.4 Number of association rule mining](image-url)
Two kinds of methods in the classification of the traditional association rule mining algorithm on the mining efficiency mining method is better than the HAAMA algorithm, the experiment get two algorithm method of association rules mining in the number of the same information time, the comparison is shown in Fig.5.

Under the same number of social media information, the classification of the traditional association rule mining algorithm mining efficiency is higher than HAAMA algorithm of mining method, and with the increase of amount of time needed for HAAMA algorithm of incremental method is greater than the traditional classification of algorithm for mining association rules mining method. This is because the HAAMA algorithm when found the implicit relationship between the concept of text, it needs to determine whether all related pathways have correlation between concepts, the recessive relationship of the required processing compared with the traditional classification of algorithm for mining association rules mining method to deal with the relationship between more, along with an increase in the number of text, including the concept of quantity is increasing, and the number of possible relationships between concepts will be more, thus HAAMA algorithm of overall efficiency decline.

6. Conclusion
In the knowledge base in association rules mining, the influence factors of relationship still need manual analysis of determining factor matching theory in the judgment of the recessive relationship still exist error, need artificial differential adjustment. Under the influence of such relations semantic judgment accuracy, this paper puts forward method of mining association rules in the knowledge base of certain deficiencies, in the future work also need to further optimize the relationship between factors of space dimension, each the effect of improving relationship, through the work to further improve in the field of audit in the knowledge base is the audit information correlation analysis result.

Acknowledgments
This work is sponsored by the Humanities and Social Sciences Research Planning Fund Project in Ministry of Education under grant number 19YJAZH053 and the Special Fund Project of Basic Scientific Research Business fee in Central Colleges and Universities under grant number 3072019CF0601.

References
[1] Yang B, Wang X. Construction of Logistics Financial Risk Ontology Model Based on Risk Correlation [C]/2019 International Conference on Modeling, Simulation, Optimization and Numerical Techniques (SMONT 2019). Atlantis Press, 2019.
[2] Lan W, Ming C. A New Model to Compute Semantic Similarity from Multi-ontology [J]. Communications in Computer & Information Science, 2013, 406: 1-10.
[3] Shajalal, Md, Aono, Masaki. Semantic textual similarity between sentences using bilingual
word semantics [J]. Progress in Artificial Intelligence, 2019: 1-10.

[4] David J Russomanno. A plausible inference prototype for the Semantic Web. Journal of Intelligent Information Systems, 2006, 26(3): 227-246P

[5] Zhang X, Sun, Shouqian, Zhang, Kejun. An information Content-Based Approach for Measuring Concept Semantic Similarity in WordNet [J]. Wireless Personal Communications, 2018 (3): 1-16.