A Concept Similarity Method in Structural and Semantic Levels

Sheng Yan   Li Yun   Luan Luan

(Institute of Information Engineering, Yangzhou University, Jiansu Yangzhou 225009)

E-mail: liyun@yzu.edu.cn

Abstract: Based on the formal concept analysis, combining the rough set and semantic ontology, this paper proposes a formal concept similarity measurement by improving the original Tversky similarity model. The method used the hierarchical information of concept lattice in the lattice structure and its own semantic meaning to measure the similarity of from the structure information of the concepts and semantic information respectively, which reflect the true similarity of formal concept to some extent.

Keywords: formal concept analysis, hierarchical information, semantic, concept similarity

1. Introduction

As the basic unit of human's thinking and knowledge, the concept has been deeply focused by philosophers, scientists and become an important subject of artificial intelligence. Concept lattice is the core structure of Formal Concept Analysis(FCA), which describes the concepts and their relationship. It is a highly simplified description of objective world to a certain extent. The advantage of the simplification is its good mathematical nature. Prof Wille creates the FCA based on this simplification. However, it is precisely because of this simplification makes the concept lattice can not simply be understood as part of the real world model, while is seen as an artificial data sets derived from a number of indications.

Based on the concept lattice, combined with rough set theory[2] and semantic ontology theory[3] , we improved the basic Tversky[4] similarity model and put forward a method of computing concept similarity. Measuring the concept similarity by structure and semantic can reflect the true similarity of concept from different levels.

2 Backgrounds

2.1 Formal Concept Analysis

Formal Concept Analysis (FCA)[1] is an useful data analysis technique based on number theory, whose content is formal context, formal concept and the relation among the formal concepts. Formal Context is defined as a triple K=(G, M, I), where G is the set of objects, M is the set of attributes and I is the binary relationship between G and M, that is I⊆G×M, gIm. In the formal context K, two mapping function f and g is defined as fellow:

$$A' = \{m \in M | gRm, \forall g \in A\} \tag{1}$$

$$B' = \{g \in G | gRm, \forall m \in B\} \tag{2}$$

They are called the Galois connection between G and M. If the tuple(A, B) from P(G)×P(M) satisfied two conditions: A=g(B) and B=f(A), or A=B' and B=A', then, we called (A, B) an formal concept from formal context K, denoted C=(A, B), which B and A is called the Intent and Extent of formal concept C separately. Assumed that $C_1=(A_1, B_1)$ and $C_2=(A_2, B_2)$ are two formal concepts, the order relation"≤"is defined as $C_1 \le C_2 \Leftrightarrow B_2 \le B_1$. C_1 is the sub concept of C_2, C_2 is super concept of C_1. All formal concepts and their relations consist of a concept lattice.

Formal concept lattice is a complete lattice, because any concepts have only and unique suprema and infima which are given by:

$$\wedge_{t\in T}(A_t, B_t) = \left(\bigcap_{t\in T} A_t, \left(\bigcup_{t\in T} B_t\right)''\right) \qquad (3)$$

$$\vee_{t\in T}(A_t, B_t) = \left(\left(\bigcup_{t\in T} A_t\right)'', \bigcap_{t\in T} B_t\right) \qquad (4)$$

They are also called Meet and Join operation.

2.2 Rough Set

Rough set theory[2] is characterized by that we don't need to give out the quantity of certain characteristics or attributes and from a given set of the problem description, confirm the approximate domain by the indiscernibility classes directly and then find out the inherent laws of problems.

Def 1: Assume that U denotes a non-empty finite universe, R denotes a set of equivalence relations on U, U/R is all of the equivalence classes on R, for any $X \subseteq U$, the lower approximated set and upper approximate set of X on R can be defined respectively as below:

$$\underline{R}(X) = \{Y \in U/R \mid Y \subseteq [x]_R\} \qquad (5)$$

$$\overline{R}(X) = \{Y \in U/R \mid Y \bigcap [x]_R != \Phi\} \qquad (6)$$

$\underline{R}(X)$ is the maximum set of objects which belong to X on the basis of known knowledge; $\overline{R}(X)$ is the union set of the equivalence classes $[X]_R$ whose intersection with X is not empty and also is the minimum set of object which maybe belong to X. The boundary regions of set X is defined as $BN(X) = \overline{R}(X) - \underline{R}(X)$. Sets have uncertainty owing to the existence of the boundary regions. The bigger the boundary regions are, the lower their accuracy is, and the greater their roughness is. When condition $\overline{R}(X) = \underline{R}(X)$ is satisfied, set X is called the accuracy set of R; Otherwise when $\overline{R}(X) \neq \underline{R}(X)$, set X is called the rough set of R. The rough set can be depicted by the upper and lower approximations of precise set.

2.3 Tversky Similarity Model

Tversky[4] put forward an concept similarity method based on the set theory:

$$S(a,b) = \frac{f(B_1 \cap B_2)}{f(B_1 \cap B_2) + \alpha f(B_1 - B_2) + \beta f(B_2 - B_1)} \qquad (7)$$

There into, $\alpha, \beta \geq 0$, f is an function for measure the similarity between B_1 and B_2, the attribute set of B_1 and B_2 is a and b.

After this, Rodriguez and Egenhofer[5] have proposed an assessment of semantic similarity among entity classes in different ontologies based on the normalization of Tversky's similarity model:

$$S(a,b) = \frac{|B_1 \cap B_2|}{|B_1 \cap B_2| + \alpha(a,b)|B_1 - B_2| + (1-\alpha(a,b))|B_2 - B_1|} \qquad (8)$$

Where the function |.| represents the cardinality of a set, $depth(a), depth(b)$ represents the depth of a,b in ontologies. If $depth(a) \leq depth(b)$, then $\alpha(a,b) = \frac{depth(a)}{depth(a) + depth(b)}$, else $\alpha(a,b) = 1 - \frac{depth(a)}{depth(a) + depth(b)}$.

Zhao[3,6] proposed extension of the measure with the employment of rough set theory:

$$S(a,b)=\frac{|(B_1\vee B_2)_{LA}|}{|(B_1\vee B_2)_{LA}|+\alpha|B_{1LA}-B_{2LA}|+(1-\alpha)|B_{2LA}-B_{1LA}|} \tag{9}$$

where $B_{LA}=\text{intent}(\wedge\{(x,y)\in\zeta\mid y\subseteq B\}),\zeta$ is the set of all concepts of lattice.

Wang[7]advanced a more complex method to compute the similarity of formal concepts:

$$\begin{aligned}S^{\wedge}_{LA}((A_1,B_1),(A_2,B_2))&=\omega\frac{|(A_1\cap A_2)^{\wedge}_{LA}|}{|(A_1\cap A_2)^{\wedge}_{LA}|+\frac{1}{2}|A^{\wedge}_{1LA}-A^{\wedge}_{2LA}|+\frac{1}{2}|A^{\wedge}_{2LA}-A^{\wedge}_{1LA}|}\\&+(1-\omega)\frac{|(B_1\cap B_2)^{\wedge}_{LA}|}{|(B_1\cap B_2)^{\wedge}_{LA}|+\frac{1}{2}|B^{\wedge}_{1LA}-B^{\wedge}_{2LA}|+\frac{1}{2}|B^{\wedge}_{2LA}-B^{\wedge}_{1LA}|}\end{aligned} \tag{10}$$

This formula is very complex because it contains the part of the intent and extent of concepts. The paper improved the basic Tversky similarity model combined the Lattice and Rough set theory to obtain more accurate results.

2.4 Information Entropy

The main idea of information entropy is: There are many events in system S=(E1, E2, …, En), each probability of events P=(P1, P2, …,Pn), then each events' information content is $I(E)=-\log Pi$, the average information content of the whole system is:

$$H(S)=\sum_{i=1}^{n}PiI(E)=-\sum_{i=1}^{n}Pi\log Pi \tag{11}$$

The average information content is called as information entropy. The paper compute the information content of each attribute using entropy to obtain the semantic similarity.

3 Concept Similarity Method

3.1 Structure Similarity

Concept lattice is a complete lattice, which has lots of available information, such as that the upper the node is, the more abstract it is and the less information it has while the more downer the node is, the more specificity it is and the more information it has. When we compared formal concept C_1 to other concepts C_2 and C_3, the level number of two concepts C_1 and C_2 is the closer than the one of C_1 and C_3, then, obviously, C_2 should be better than C_3 because of closer hierarchy which the true similarity of C_1 and C_2 is bigger than the similarity of C_1 and C_3. Thus, the use of hierarchy information of formal concept is necessary in calculating the concept similarity.

Secondly, considering the hierarchy structure of formal concept and combining with rough set theory, using the approximate set of intent instead of intent self to calculate the similarity helps to contain more specific information. Based on two points above, this paper improves the Tversky similarity model and adds information of hierarchy and approximates to obtain a more effective result.

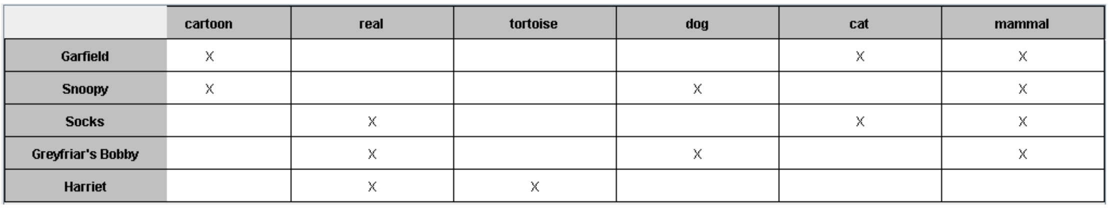
Def 2 The structure similarity among concepts：Formal concepts C_1 (A_1, B_1) and C_2 (A_2, B_2) are two formal concepts in concept lattice, the structure similarity is defined as follows:

$$StructSim(C_1,C_2)=\frac{|(B_1\vee B_2)_{LA}|}{|(B_1\vee B_2)_{LA}|+\alpha(C_1,C_2)|B_{1LA}-B_{2LA}|+(1-\alpha(C_1,C_2))|B_{2LA}-B_{1LA}|} \tag{12}$$

$(B_1\vee B_2)_{LA}=\text{intent}(\wedge\{(x,y)\in\zeta\mid y\subseteq B\}),\zeta$ stands for all formal concepts, | | represents the cardinality of a set, $|B_{1_{LA}} - B_{2_{LA}}|$ shows the number of intents which belong to the lower approximate set of B$_1$ but don't belong to the lower approximate set of B$_2$, $|B_{2_{LA}} - B_{1_{LA}}|$ shows the number of intents which belong to the lower approximate set of B$_2$ but don't belong to the lower approximate set of B$_1$; $\alpha(C_1, C_2) = \begin{cases} \dfrac{depth(C_1)}{depth(C_1) + depth(C_2)}, & \text{if } depth(C_1) \le depth(C_2) \\ 1 - \dfrac{depth(C_1)}{depth(C_1) + depth(C_2)}, & \text{if } depth(C_2) \le depth(C_1) \end{cases}$,

$depth(C_1)$ represents the level number of C$_1$ in concept lattice, which the number of top node is noted as 0. The coefficient α represents the hierarchical relationship among the concepts.

The formal context (Figure 1) is from literature[8] to illustrate how to compute the structure similarity of concepts. In the formal context, we used Godin[9] to construct the formal lattice below (Figure 2):



	cartoon	real	tortoise	dog	cat	mammal
Garfield	X				X	X
Snoopy	X			X		X
Socks		X			X	X
Greyfriar's Bobby		X		X		X
Harriet		X	X			

Figure 1 The formal context

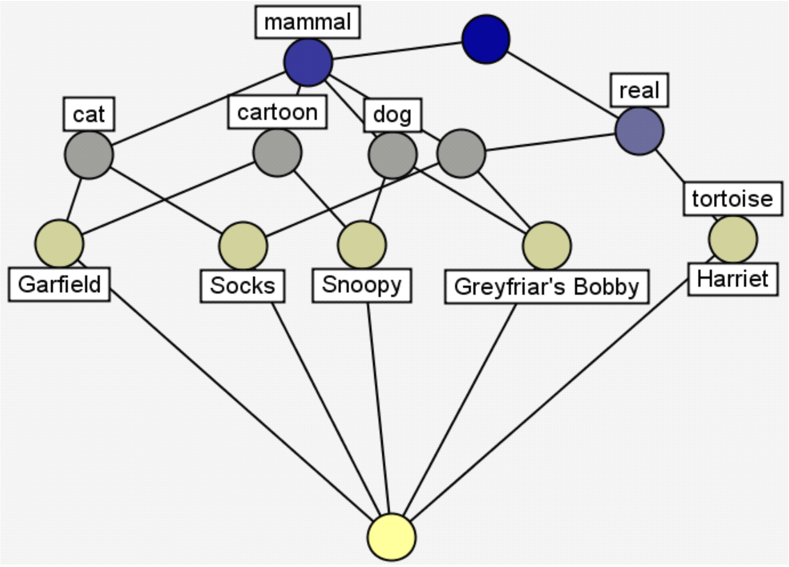


Figure 2 The relative concept lattice

Considering the formal concepts C$_1$((Garfield, Socks), (mammal, cat)), C$_2$((Grayfriar's Bobby, snoopy), (dog, mammal)), C$_3$((Grayfriar's Bobby), (real, dog, mammal)), the structure similarity between C$_1$ and C$_2$, C$_1$ and C$_3$ according to formula (12):

$$StructSim(C_1, C_2) = \frac{1}{1 + \frac{1}{2} * 1 + (1 - \frac{1}{2}) * 1} = \frac{1}{2}$$

$$StructSim(C_1, C_3) = \frac{1}{1 + \frac{2}{5} * 1 + (1 - \frac{2}{5}) * 2} = \frac{5}{13}$$

The similarity between C_1 and C_2 is bigger than the one between C_1 and C_3.

3.2 Semantic Similarity

In order to measure the semantic similarity, we adopt WordNet[10] as the basic copus to extract the IS-A relations among the words and then add the probabilty of each word to obtain a weighted IS-A relations. At last, we get a semantic similarity table by computing the information content of each word.

WordNet is a broad coverage of English vocabulary semantic network. The paper mainly uses the IS-A relation of words and SynSet to obtain a weighted IS-A hierarchical relationship.

Def 3 The weighted IS-A relationships[11]: Given an english corpus ε , the weighted IS-A relationships $H(\varepsilon)$ can be generated by each word's probability and their IS-A relationship. The probability of each word is defined as:

$$p(n)=\frac{freq(n)}{M}+\sum_{i\in sub(n)}p(i) \qquad (13)$$

$freq(n)$ represents the number of word n in the whole corpus; M represents the number of all words in the whole corpus; $\sum_{i\in sub(n)}p(i)$ represents the sum of probability of words which are the children nodes of word n. If word n is in bottom and it has no child, then the relative value is 0. At the same time, the top node is defined as Top and its $p(Top)=1$.

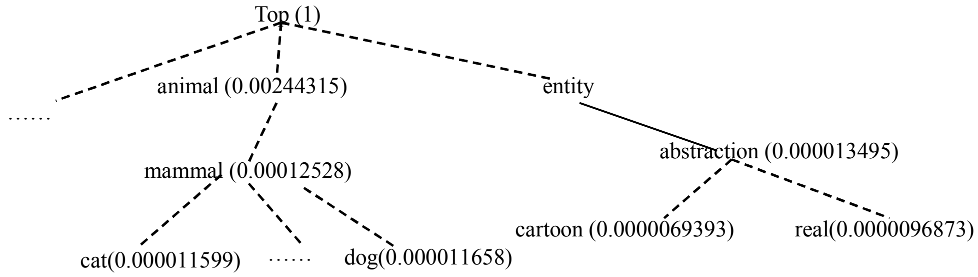


Figure 3 A weighted IS-A hierarchical relationship

From the IS-A relationships, we can get the information content of each word according to the information content formula $I(E)=-\log Pi$, for example:

$I(cat)=-\log 0.000011599{=}4.935579$,

$I(dog)=-\log 0.000011658{=}4.933376$,

$I(mammal)=-\log 0.00012528=3.902118$,

From the weighted IS-A hierarchical relationship, we can see that the word is upper, its information content is less and if the parent node of two words has more information content, they share more information and more similar. The paper adopt the similarity of information content advanced in the literature[11] to improve the semantic similarity of formal concepts.

Def 4 Information content similarity, ics(n_1, n_2) [12]: Given an English corpus ε and a weighted IS-A hierarchical relationship $H(\varepsilon)$, any two words n_1, n_2 in corpus, if n_1= n_2 or n_1 belongs to the Synset of n_2, then $ics(n_1,n_2)=1$; else $ics(n_1,n_2)=\frac{2I(n')}{I(n_1)+I(n_2)}$. n' is the maximum common parent node of n_1 and n_2, $I(n')=\max_{n\in S(n_1,n_2)}\{I(n)\}$, $S(n_1,n_2)$ is all parent node of n_1 and n_2.

According to this definition, the similarity of any two words can be obtained. For example:

$$ics(cat,dog)=\frac{2I(mammal)}{I(cat)+I(dog)}=\frac{2*3.902118}{4.935579+4.933376}=0.790786$$

Because the intent of a formal concept reflects its nature, we can just consider the intent to measure the similarity of formal concepts.

Def 5 Semantic similarity among concepts[11]**:** The semantic similarity of two formal concepts $C_1(A_1, B_1)$ and $C_2(A_2, B_2)$ is defined as:

$$SemanticSim(C_1,C_2)=\frac{M(B_1,B_2)}{\max\{|B_1|,|B_2|\}} \tag{14}$$

$M(B_1,B_2)=\max\{\sum_{b_1\in B_1,b_2\in B_2} ics(b_1,b_2)\}$, b_1 and b_2 can only occur once in every combination.

For example, the semantic similarity between formal concept C_1((Garfield, Socks), (mammal, cat)) and C_2((Grayfriar's Bobby, snoopy), (dog, mammal)) is:

$$SemanticSim(C_1,C_2)=\frac{M((mammal,cat),(dog,mammal))}{\max\{|(mammal,cat)|,|(dog,mammal)|\}}=0.895393$$

While the semantic similarity between formal concept C_1((Garfield, Socks), (mammal, cat)) and C_3((Grayfriar's Bobby), (real, dog, mammal)) is:

$$SemanticSim(C_1,C_3)=\frac{M((mammal,cat),(real,dog,mammal))}{\max\{|(mammal,cat)|,|(real,dog,mammal)|\}}=0.596929$$

3.3 The formula of similarity among concepts

After obtaining the structure and semantic similarity of two formal concepts, we can combine with the two similarities to compute the final similarity.

Def 6 Sim(C_1, C_2)：The final similarity between formal concepts $C_1(A_1, B_1)$ and $C_2(A_2, B_2)$ is:

$$Sim(C_1,C_2)=\omega StructSim(C_1,C_2)+(1-\omega)SemanticSim(C_1,C_2) \tag{15}$$

ω is a weighted coefficient and adjust the importance of structure and semantics similarity, whose value is 0~1.

For example, we compute the similarity between C_1((Garfield, Socks), (mammal, cat)) and C_2((Grayfriar's Bobby, snoopy), (dog, mammal)) is:

$$Sim(C_1,C_2)=0.5*\frac{1}{2}+0.5*0.895393=0.697697$$

While the similarity between C_1((Garfield, Socks), (mammal, cat)) and C_3((Grayfriar's Bobby), (real, dog, mammal)) is:

$$Sim(C_1,C_3)=0.5*\frac{5}{13}+0.5*0.596929=0.490772$$

It shows that the similarity of formal concepts C_1 and C_2 is higher than the one of C_1 and C_3, which is in line with the actual situation.

4 Conclusions and Future Work

The paper proposes a concept similarity method which improved the basic Tversky model and measure the similarity among the concepts in structural and semantic levels and tried best to find the true similarity of different concepts. If the method is applied to ontology engineering, such as ontology mapping, ontology merging, we can measure the similarity between ontology concepts; And in information retrieval, it can be used for matching keywords as a basic tool to measure similarity. The future work is to verify effects and accuracy.

Acknowledgment

This research was supported in part by the Chinese National Natural Science Foundation under grant No. 60673060, Natural Science Foundation of Jiangsu Province under contract BK2008206 and Natural Science Foundation of Education Department of Jiangsu Province under contract 08KJB520012.

References

[1] B.Ganter , R.Wille. Formal Concept Analysis: Mathematical Foundations. Springer-Verlag, 1999.

[2] Pawlak Z. Rough set[J]. International of Computer and Information Science, 1982, 11: 341-356

[3] Y. Zhao, X. Wang, W.A. Halang, Ontology mapping based on rough formal concept analysis, in: Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, 2006, February 19-25, p.180

[4] A. Tversky, Features of similarity, Psychological Review ,84 (1977) 327-352.

[5] M.A. Rodriguez, M.J. Egenhofer, Determining semantic similarity among entity classes from different Ontologies, IEEE Transactions on Knowledge and Data Engineering 15 (2003) 442-456.

[6] Y. Zhao, W.A. Halang, Rough concept lattice based ontology similarity measure, in: Proceedings of the First International Conference on Scalable Information Systems, Hong Kong, 2006, May 30-June 01.Vol. 152, p.15

[7] Lidong Wang, Xiaodong Liu. A new model of evaluating concept similarity. Knowledge-Based Systems 21(2008), 842-846.

[8] Uta Priss. Formal Concept Analysis in Information Science. Knowledge-Based Systems archive, 2008, Vol. 21 , Issue 1 : 80-87

[9] Godinr, Missaouir, Alaouih. Incremental concept formation algorithms based on galois (concept) lattices[J]. Computational Intelligence, 1995, 11 (2) : 246-267.

[10] ftp://ftp.cogsci.princeton.edu/pub/wordnet/2.0/WordNet-2.0.exe

[11] Shengyan, Liyun, Lituo, LuanLuan. Ontology Mapping Based on the Concept Lattice Model, Journal of Nanjing Normal University(Engineering and technology), 2008, Vol.8 ,No.4, 91-94.

[12] Anna Formica. Concept similarity in Formal Concept Analysis: An information content approach. Knowledge-Based Systems archive, 2008, Vol. 21, Issue 1: 80-87