Research on Semantic Judgment of Key Words in Ontology - Based Dynamic Requirements Traceability

MengNi Rao
School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
1353943697@qq.com

YongHua Li
School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
251812804@qq.com

***Abstract*—The accuracy of trace links can be improved by using ontology-based dynamic requirements traceability method, but it is a complicated and tedious process to establish a reasonable and effective ontology. Because the size of the ontology affects the experimental results directly, this paper shows a method to combine the modifiers with the general ontology. Firstly, the method makes semantic choices for the keywords through the collocation rules of the modifier and the keyword. Then the similarity of the keywords are calculated by the semantic distance in the WordNet and adjusted through the modifier ontology. The semantic choices of the keywords are reflected by the similarity scores. Finally, the similarity between the source artifacts and target artifacts are calculated by the similarity of the keywords. The number of modifiers is small in the requirements documents, design documents and etc. so this can reduce time and labor cost brought by the construction of domain ontology. In order to verify the effectiveness of this method, the method in this paper will be compared with VSM based methods and domain based ontology methods.**

Keywords- dynamic requirements traceability; modifier; ontology; polysemy; semantic analysis

I. INTRODUCTION

At present, semantic issues are the pivotal problem in dynamic requirements traceability. With the intensive study and the widely use of ontology, more and more scholars adopt ontology to solve the semantic problems in dynamic requirements traceability. Zhiwei Chen[1] proposed a method for assessing semantic mining. Firstly, he obtained a standard dataset by combining nine semantic relationships in WordNet and synonyms in UMLS (Unified Medical Language System). Then, he evaluated embedded words in this dataset. Sujata R. Kolhe [2] used Latent Semantic Indexing (LSI) to cluster and create tags to facilitate the retrieval and management of large-scale text databases, and calculated similarity by combining extended queries and cosine similarity. Liviu Sebastian Matei[3] calculated the semantic distance between words firstly and then calculate the similarity between the texts based on the dynamic timing. Compared with the traditional vector space model, he taked the influence of word order on the semantics into account and proposed the time series model to improve the accuracy of the results. Chalitha Kulathunga[4] identified ambiguous words in financial texts by integrating ontology and clustering methods. Although this method eliminated the semantic ambiguity of the text and improved the performance of the clustering algorithm, it did not use the financial dataset and the validity of this method cannot be verified. Gong Li[5] used Weibo essays as material to construct an ontology knowledge base for the security domain. Then he used ontology knowledge to expand the initial query words, and combined local query feedback to filter candidate extended words. Finally, he got the results by performing two queries and iterations. Because Weibo were mainly short and keywords and information are sparse, the accuracy of this method would decrease with the increasing number of the query results.

According to relevant research, 78% of words in the requirements document are nouns or verbs[6]. So the nouns and verbs are the main research objects in dynamic requirements traceability. It is easy to cause errors of semantic differences because of the ambiguous nouns. The method based on domain ontology is to solve the phenomena of "word polysemy" that cannot be solved based on IR method. However, this method must build related domain ontology. It is a complicated and tedious process to construct domain ontology. To solve this problem, this paper proposes a method which combines WordNet and the ontologies of modifiers to determine the semantics of nouns in texts. This method can reduce experimental errors and improve accuracy. The number of modifiers in the requirements document is relatively small. Compared with the construction of domain ontology, the construction of the modifier ontology has an advantages of smaller workload, and has a higher applicability to the entire domain.

In this paper, we study ontology-based keyword retrieval method. The structure is organized as follows: Section 2 presents the method used in this paper. Section 3 is mainly the dataset, experimental steps and experimental analysis of the experiment in this paper. Finally, Section 4 concludes the paper and points out further work.

II. ONTOLOGY-BASE KEYWORD RETRIEVAL METHOD

A. Ontology-based keyword retrieval method framework

This paper proposes an ontology-based keyword retrieval method. The workflow is shown in Figure 1. The basic idea of this method is as follows: Firstly, the collocation types of modifiers and keywords is determined by keywords and modifiers .Secondly, the method reads the semantic attribute table of keywords in WordNet and extracts modified attributes of modifiers by using modifiers ontology. The modifiers ontology in this paper is established by manual. Thirdly, the semantics of keywords are determined by utilizing the keyword semantic attributes and modified attributes of modifiers through a set of polysemy semantic selection rules. Finally, the method calculates the similarity of keywords based on the

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00055

distance-based similarity method in WordNet and the semantic similarity of keywords are modified by exploiting the semantic similarity of modifiers. This correction can make it more accurate to reflect the actual semantic relationship between keywords.

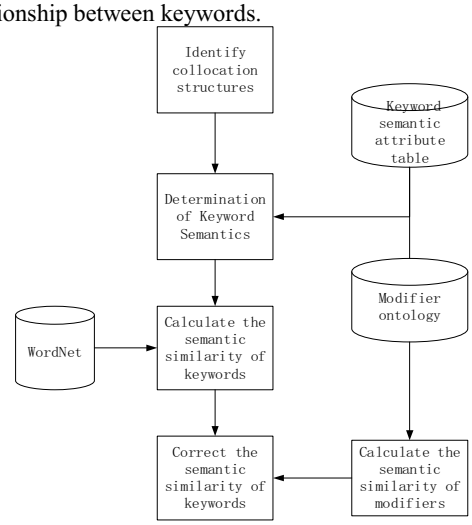

Figure 1. Ontology-based keyword retrieval method framework.

B. Modifiers for Selective Rules of Polysemy

Definition 2.1 Semantic attribute set. Each word has a different meaning. WordNet groups the meaning of the English words. Each meaning is called a semantic attribute, which is denoted as att . Then for any words w , there is a semantic attribute set $Att=\{att_1,att_2,\cdots,att_n\}$ .Then at the att semantics, the semantics of the word are denoted as $S(att_i)_w$.

Definition 2.2 Semantic attribute relationships. The semantic relationship between hyponymy and hypernymy can be divided into four categories:

(1) If the word A is part of the word *B*, the relationship between *A* and *B* are denoted by *A apo B*.
(2) If word *A* is a member of word *B*, the relationship between *A* and *B* is recorded as *A amo B*.
(3) If word *A* is the constituent material of word *B*, the relationship between *A* and *B* is denoted as *A amf B*.
(4) If the word *A* is the refinement of the word *B*, the relationship between *A* and *B* is recorded as *A ako B*.

In the four relationships mentioned above, the relations of *amo, amf*, and *ako* satisfy transitivity, while *apo* does not satisfy transitivity. For example, the sentences "the branch is a part of the tree" and "the tree is a part of the forest" cannot deduce to "the branch is a part of the forest" because there are two different relationships between "branch/tree" and "tree/forest".

Definition 2.3 There is a semantic attribute set $Att=\{att_1,att_2,\cdots,att_n\}$ for the word w . If there is att_i, there is word w which satisfies *w amo W* or *w amf W* or *w ako W* . Then W is the parent class of w under the semantic attribute att_i , which denoted as $f(att_i)_w=W$. $f(att_i)_w$ is a collection containing all parent classes under the att_i semantics.

Definition 2.4 Collocation relationship. For the keyword w , its semantic attribute set is $Att=\{att_1,att_2,\cdots,att_n\}$, and its modified attribute set of mw is $MAtt=\{matt_1,matt_2,\cdots,matt_n\}$,then there is a relation R , so that if $\exists\, att_i\in Att,\ matt_j\in Matt$, s.t. $R(matt_j)=att_i$. If $\exists att_i\in Att,\ \forall matt_j\in Matt,$ s.t. $R(matt_j)=att_i,$ then $R(matt_j)=0.$

Relationships are used to describe the collocations between modifiers and keywords. Different modifiers define different semantic properties. This relationship is a many-to-many relationship rather than a one to one relationship.

According to the above definition, the words w_1 and w_2 are assumed to have attributive attributes $att=\{att_1,att_2,\cdots,att_n\}$ and $Att=\{Att_1,Att_2,\cdots,Att_m\}$. The w_1 and w_2 modifiers are mod_1 and mod_2 respectively, and the modifier modifiers are $matt=\{matt_1,matt_2,\cdots,matt_p\}$ and $MAtt=\{MAtt_1,MAtt_2,\cdots,MAtt_q\}$ respectively. There are the following semantic selection rules:

Rule 2.1 When modifiers exist for the same semantic attribute and the corresponding words w_1 and w_2 have the same semantic attributes, the words w_1 and w_2 select the same semantic attribute.

$$\exists\, matt_i\in matt,\ Matt_j\in MAtt,\ (att\cap Att\neq\varnothing)\cap(R(matt_i)=R(MAtt_j))$$
$$\cap(R(matt_i)\neq 0)\cap(R(MAtt_j)\neq 0)\cap(R(matt_i)\in(att\cap Att))$$
$$\cap(R(MAtt_j)\in(att\cap Att))\rightarrow\exists\, att_h\in(att\cap Att),Att_k\in(att\cap Att)$$
$$,\text{s.t. } S(att_h)_{w1}=S(Att_k)_{w2}.$$

Rule 2.2 When modifiers exist for the same semantic attribute and there is one of the four relations *apo*, *amo*, *amf*, and *ako* for one of semantic attributes of the corresponding words w_1 and w_2 , the words w_1 and w_2 select this semantic attribute.

$$\exists\, matt_i\in matt,\ MAtt_j\in MAtt,\ (R(matt_i)\neq 0)\cap(R(MAtt_j)\neq 0)$$
$$\cap((R(matt_i)\ apo\ R(MAtt_j))\cup(R(matt_i)\ amo\ R(MAtt_j))\cup$$
$$(R(matt_i)\ amf\ R(MAtt_j))\cup(R(matt_i)\ ako\ R(MAtt_j)))))\rightarrow$$
$$\exists\, att_h\in att,\ Att_k\in Att,\ \text{s.t. } S(att_h)_{w1}=S(Att_k)_{w2}.$$

Rule 2.3 When modifiers exist for the same semantic attribute and there is one of the three relationships *amo*, *amf*, and *ako* for the word w_1 and the parent semantic attributes of the word w_2, the words w_1 and w_2 select the semantic attribute.

$$\exists\ matt_i\in matt,\ MAtt_j\in MAtt,((R(matt_i)\ amo\ f(R(MAtt_j)))\cup$$
$$(R(matt_i)\ amf\ f(R(MAtt_j)))\cup(R(matt_i)\ ako\ f(R(MAtt_j)))))$$
$$\cap(R(matt_i)\neq 0)\cap(R(MAtt_j)\neq 0)\cap(f(R(MAtt_j))\neq\varnothing)\rightarrow$$
$$\exists\, att_h\in att,Att_k\in Att,\text{s.t. } S(att_h)_{w1}=S(Att_k)_{w2}.$$

Rule 2.4 When the modifier exists in the same semantic attribute and there is one of the three relationships *amo*, *amf*, and *ako* between the parent class of the word w_1 and the parent semantic attribute of the word w_2 , the word w_1 and w_2 Select this semantic attribute.

$\exists\, matt_i \in matt,\ MAtt_j \in MAtt\ ,\ \Big(\big(f\big(R(\ matt_i\ )\big)\ amo\ f\big(R(\ MAtt_j)\big)\big)\cup$
$\big(f\big(R(\ matt_i\ )\big)\ amf\ f\big(R(\ MAtt_j)\big)\big)\cup\big(f\big(R(\ matt_i\ )\big)\ ako\ f\big(R(\ MAtt_j)\big)\big)$
$\cup\big(f\big(R(\ matt_i\ )\big)\ = f\big(R(\ MAtt_j)\big)\big)\big)\cap\big(f\big(R(\ matt_i\ )\big)\neq\varnothing\ \big)\cap$
$\big(f\big(R(\ MAtt_j)\big)\neq\varnothing\big)\rightarrow\exists att_h\in att,\ Att_k\in Att,$ s.t. $S(att_h)_{w1}=S(Att_k)_{w2}$.

For example "short time" and "several second", there are 10 nominal semantic attributes in "time" and "an indefinite period" in semantic attributes that can be modified by "short". There are also 10 nominal semantic attributes in "second", and the two semantic attributes "1/60 of a minute" and "an indefinitely short time" can be modified by "several". According to the above rules, "time" has an intersection of the semantic attributes of "second", that is, a semantic attribute of a certain period of time, and "second" is the semantic meaning of "an indefinitely short time".

C. Direct match based keyword search method

Through Stanford Parser's analysis of English sentences, there are two kinds of keyword modification and collocation. They are recorded as left collocation $(mod_1,\cdots,mod_m,N)$ and right collocation $(N,mod_1,\cdots,mod_m)$, and the modifier vector is $Mod=\{\mathrm{mod}_1,mod_2,\cdots,mod_m\}$ According to the research results of English linguistics, the closer the distance between the modifiers and the key words is, the greater the weight is to calculate the similarity degree of the modifiers. Therefore, we can assign different weights for the modifiers according to the distance between the central word and the central word. The weights assigned to this paper are as follows:

$$w(t)\ =\ \frac{1}{dis(t)} \tag{1}$$

$dis(t)$ represents the distance between the modifier and the keyword, and $w(t)$ represents the weight of the word.

The degree of closeness between the meanings of the modified words is reflected by the similarity calculated by the modifier ontology. The similarity is recorded as sim, and the semantic correlation coefficient matrix of the modified words is expressed as follows:

$$MSim=\begin{pmatrix} sim_{11} & sim_{12} & \cdots & sim_{1n} \\ sim_{21} & sim_{22} & \cdots & sim_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ sim_{m1} & sim_{m2} & \cdots & sim_{mn} \end{pmatrix} \tag{2}$$

The process based on the direct collocation method is as follows:

Input: Keywords n_i、n_j, two modifier vectors：

$Mod_i=\{\mathrm{mod}_{i1},mod_{i2},\cdots,mod_{im}\},Mod_j=\{\mathrm{mod}_{j1},mod_{j2},\cdots,mod_{jn}\}$

Output: the similarity of n_i and n_j after adjustment based on direct collocation.

1. The semantic attribute tables of the keyword n_i and n_j are read through WordNet, and the modified attributes of the modifier words $Mod_i=\{\mathrm{mod}_{i1},mod_{i2},\cdots,mod_{im}\}$ and $Mod_j=\{\mathrm{mod}_{j1},mod_{j2},\cdots,mod_{jn}\}$ are read through the modifier ontology.
2. Through the proposed rules of semantic meaning selection, the semantics of keywords are determined by the keyword semantic attribute table and modified attributes of modifiers that were read in step 1, and the similarity of keywords is calculated by the distance method.
3. Calculate the semantic correlation coefficient matrix $MSim$ of the modifier words $Mod_i=\{\mathrm{mod}_{i1},mod_{i2},\cdots,mod_{im}\}$ and $Mod_j=\{\mathrm{mod}_{j1},mod_{j2},\cdots,mod_{jn}\}$ based on the modifier ontology, and calculate the modifier word weight according to Equation (1).
4. The correlation coefficient vector $msim_i=(sim_{i1},sim_{i2},\cdots,sim_{in})$ for each modifier is obtained through the correlation coefficient matrix $MSim$ obtained in step 3, and the semantic correlation coefficient $mSim_k$ of mod_{ik} in $Mod_i=\{\mathrm{mod}_{i1},mod_{i2},\cdots,mod_{im}\}$ is calculated by the correlation coefficient vector:

$$mSim_k=\max\left(sim_{k1}*w_{k1},sim_{k2}*w_{k2},\cdots,sim_{kn}*w_{kn}\right) \tag{3}$$

5. Calculate the semantic correlation coefficient $MSim_{ij}$ of vector Mod_{ij}:

$$MSim_{ij}=\frac{\sum mSim_k}{\sum w_k} \tag{4}$$

6. Correct sim（n_i, n_j）similarity via $MSim_{ij}$:

$$\mathrm{sim}(n_i,n_j)=\begin{cases} MSim_{ij} & sim(n_i,n_j)=0 \\ sim(n_i,n_j)\times\left(\alpha+\dfrac{MSim_{ij}-sim(n_i,n_j)}{sim(n_i,n_j)}\right) & sim(n_i,n_j)\neq 0 \end{cases} \tag{5}$$

The advantage of this method is that keyword semantics can be clearly selected. In the actual process, the semantic relevance of keywords is adjusted through different collocation words. In the English text, there are also situations where both left and right collocations exist at the same time. Therefore, this chapter further proposes a method based on mixing and collocation.

D. Keyword Search Method Based On Mixed Collocation

In actual English documents, the situation based solely on left and right collocations is relatively unusual, and in most cases, the situation of left and right collocations exists. Therefore, this paper proposes the method of mixed collocation to judge the true choice of semantic modifiers. The algorithm 2 is based on the algorithm 1 and takes the presence of both left and right collocations into account.

The process based on the mixed collocation method is as follows:

Input: Keywords n_i、n_j, two modifier vectors:

$Mod_{i_L}=\{\mathrm{mod}_{i_L1},mod_{i_L2},\cdots,mod_{i_Lm_1}\},Mod_{i_R}=\{\mathrm{mod}_{i_R1},mod_{i_R2},\cdots,mod_{i_Rm_2}\}$

,

$Mod_{j_L}=\{\mathrm{mod}_{j_L1},mod_{j_L2},\cdots,mod_{j_Ln_1}\},Mod_{j_R}=\{\mathrm{mod}_{j_R1},mod_{j_R2},\cdots,mod_{j_Rn_2}\}$

Output: the similarity of n_i and n_j after adjustment based on mixed collocation.

1. Based on steps 1-4 of algorithm 1, the left and right collocation semantic correlation coefficients are

calculated:

$$mSim_{i_{LL}} = \left\{ mSim_{i_{LL}1}, mSim_{i_{LL}2}, \cdots; mSim_{i_{LL}m_1} \right\}, mSim_{i_{LR}} = \left\{ mSim_{i_{LR}1}, mSim_{i_{LR}2}, \cdots, mSim_{i_{LR}m_2} \right\},$$

$$mSim_{i_{RL}} = \left\{ mSim_{i_{RL}1}, mSim_{i_{RL}2}, \cdots; mSim_{i_{RL}m_2} \right\}, mSim_{i_{RR}} = \left\{ mSim_{i_{RR}1}, mSim_{i_{RR}2}, \cdots, mSim_{i_{RR}m_2} \right\}$$

2. According to step 1, calculate the semantic correlation coefficient of the left and right collocation according to Equation (4).
3. Get $mSim_{i_{LL}}$, $mSim_{i_{LR}}$, $mSim_{i_{RL}}$ and $mSim_{i_{RR}}$ according to step 2, and calculate the semantic correlation coefficient of vector Mod_{ij} :

$$MSim_{ij} = \max\left(mSim_{i_{LL}}, mSim_{i_{LR}}, mSim_{i_{RL}}, mSim_{i_{RR}} \right) \quad (6)$$

4. Correct the semantic similarity of keywords the similarity $sim(n_i, n_j)$ according to the fifth step of algorithm 1.

The method of the mixed collocation determines the semantics of the modified words by combining the right and left modified words. However, when the semantic similarities of the two collocations have high degree and the semantic range of the keywords cannot be determined by modifiers, the semantics of selection cannot be achieved.

III. EVALUATION

The dataset used in this experiment is a requirements traceability test set from a port production business management system. The dataset included a total of 346 source artifacts and 1264 target artifacts, including 429 source artifacts with correct trace links. In the experiment, due to the limitation of manpower, 56 source materials and 107 target materials were selected, among which 40 correct tracking chains were established manually. In this experiment, the calculation of the similarity between the source artifacts and the target artifacts uses the method of Yonghua Li[7] to obtain the results. And the port ontology uses the ontology built in my laboratory. After many experiments, the value of parameter in ontology-based keyword retrieval method (OKR) is 0.75.

TABLE I. EXPERIMENTAL RESULTS

results	VSM	Domain	OKR
recall	0.996	0.948	0.938
precision	0.193	0.312	0.299
F2	0.543	0.672	0.640

It can be seen from Table 1 that the domain ontology-based method and OKR method are lower in recall rate than the VSM-based method because the overall similarity score between the source artifacts and the target artifacts of these method is generally low, and the candidate links once If the match is successful, the similarity score will be high. Compared with ontology-based dynamic requirement traceability method, the OKR method without domain ontology can achieve almost the same effect in recall and precision. Compared with the traditional vector space model method, the experiment results shows that the OKR method can effectively improve the accuracy of the requirement traceability. The result is also depicted in fig 3.

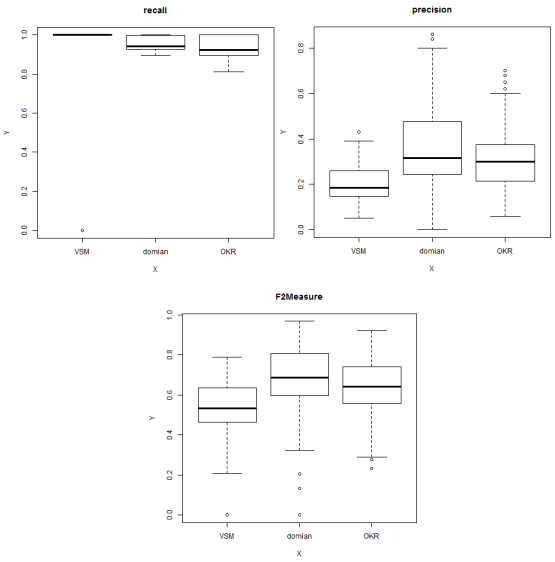

Figure 2. Results of the experiments.

IV. CONCLUSION

This paper presented a method by combining the modifiers with the universal ontology to solve semantic differences. The experiment results shows that this method can improve the accuracy of the requirement traceability effectively.The next step will focus on how to combine sentence structure and modifiers, and concentrate on understanding the meaning of sentence semantics on the level of sentence structure to improve the accuracy of the recommended trace links.

REFERENCES

[1] Chen Z, He Z, Liu X, et al. “An exploration of semantic relations in neural word embeddings using extrinsic knowledge,” IEEE International Conference on Bioinformatics and Biomedicine. IEEE Computer Society, pp. 1246-1251, 2017.

[2] Kolhe S R, Sawarkar S D. “A concept driven document clustering using WordNet,” International Conference on Nascent Technologies in Engineering, pp. 1-5, 2017.

[3] Matei L S, Matu S T. “Document semantic distance based on the time series model,” Roedunet Conference: NETWORKING in Education and Research, pp. 1-4,2016.

[4] Chalitha Kulathunga, D.D. Karunaratne. “An Ontology-based and Domain Specific Clustering Methodology for Financial Documents,” 2017 International Conference on Advances in ICT for Emerging Regions, pp. 209‑216, 2017.

[5] Gong Li, Du Junping, Lai Jincai, et al. “Microblog query expansion algorithm based on ontology and local query feedback,” Journal of Nanjing University: Natural Science Edition, vol. 53, issue 6, pp.1004-1011, 2017.

[6] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani and et.al.Developing Language Processing Components with GATE Version 7 (a User Guide). （1995-01）http://gate.ac.uk/sale/tao/splitch14.html#sec:ontologies:vr.

[7] Li Y, Cleland-Huang J. “Ontology-based trace retrieval,” International Workshop on Traceabilityin Emerging Forms of Software Engineering, pp. 30-36, 2013.