Application of an Improved Synthetical Semantic Similarity Method in Water Knowledge Graph

Kaili Gao, Jianzhuo Yan and Yongchuan Yu*

Information Department, Beijing University of Technology, Beijing 100124, China
Email: yuyongchuan@bjut.edu.cn

Abstract. In order to integrate different kinds of information in the water affairs field, this paper proposes an improved comprehensive semantic similarity algorithm and applies it to the construction of water affairs ontology. In this article, the algorithm is to give the concept instance similarity algorithm and concept definition similarity algorithm appropriate weight. First, we apply the three algorithms to the concept of the local ontology of smart water, and then calculate the similarity between the concepts to judge the similarity between the concepts to merge the smart water data, and finally compare the similarities of the three algorithms Degree value and calculation and comparison of three similarity variance values prove the effectiveness of the method. It can be seen from the experimental results that the variance of the comprehensive similarity calculation method is smaller than that of the other two methods, and it is easier to determine the similarity between two concepts.

Keywords. Integration; ontology construction; similarity calculation; water affairs.

1. Introduction

Water information plays an important role in urban development and national security. However, smart water is confronted with big data from multiple types and sources. In order to achieve the deep integration of complex characteristic big data, a knowledge architecture recognized in the business scope should be implemented first, in which the objective entities involved in the business are preserved in the semantic form, and the semantic attributes of the entities are recorded, and can be parsed and understood by the computer. The water authority has built a Shared exchange platform before, which can achieve the visualization effect to the water service data. But there are many problems in the information system, such as the information can not be interconnected, information does not contain semantic relations, can not achieve complex reasoning can only achieve simple query effect and so on. At present, the level of water information is relatively backward. Due to the large amount of time data and spatial data, semantic heterogeneity exists between information, so the mapping between ontologies cannot be realized normally. Therefore, how to effectively integrate smart water information has become an urgent problem to be solved.

At present, the data fusion of the surveyed smart water service mainly includes two aspects, among which the technologies are mainly divided into: (1) data mode layer fusion: concept combination, concept hyponym relation combination, and concept attribute combination. Domestic and foreign researchers have made a lot of contributions in these aspects. Domestic ontology mapping technology mainly includes dictionary wordnet, structure and instance-based method, as well as using background knowledge and previous knowledge to find the matching result output. In general, the mapping of two related ontologies, and the output is the ontology mapping composed of the corresponding relations between the matching ontology concepts. Ontology mapping is useful for ontology evolution and
different information integration, such as ontology integration [1]. Lily ontology mapping system of central south university of China [2] has developed well in recent years. Many famous universities abroad and laboratory has carried on the research of ontology mapping, have developed some specific mapping system and implementation method, such as the Glue of research at the University of Washington system [3] based on the concept of instance method, ontology algebra method [4] at Stanford University, University of Maryland, the semantics of the digestion method [5], Macedche et al. put forward the method of using concept definition of [6], the university of Karlsruhe KAON engineering of ontology mapping framework MAFRA [7]. The Rimom framework proposed by Li et al. [8] for ontology alignment improves the effect by automatically combining multiple strategies. The AML ontology mapping system entered by Faria in 2015 OAEI is an ontology mapping system based on ontology concept [9, 10]. Forsati et al. proposed a method based on ontology mapping (HSOMAP) et al. [11] (2) data layer fusion: entity combination, entity attribute combination, conflict detection and resolution. Hogan et al. [12] discussed the scalable and distributed methods of entity integration for large-scale and static predictive database of associated data. However, these methods are not effective in intelligent water information fusion.

This paper studies the data in the field of smart water. Through the investigation, we find that we have two kinds of different data, one is the real-time monitoring data of the comprehensive database of the sharing and exchange platform of Beijing water authority, and the other is the basic information data of the first water census of Beijing water authority. These two kinds of data are quite different in organizational form and content, which can be regarded as two different ontologies, many of which describe the same entities. In order to build a comprehensive knowledge map of water utilities, ontology mapping plays a crucial role. Therefore, in this paper, we refer to relevant literatures and propose an improved comprehensive semantic similarity calculation method, so as to effectively realize the mapping in the body and find out the semantic association between entities in different ontologies.

2. Text

2.1. Comprehensive Semantic Similarity Computing Model
In view of the limitations of the current ontology mapping method, the current method is not universal and the mapping is one-sided. Through the research, it is found that the instance of the concept, the definition of the concept and the hierarchical structure of the concept all affect the similarity between concepts to some extent. For example, the method in Ref. [6] shows that the concept definition and structure are more important in influencing similarity, and the instance information of the concept is not mentioned. In Ref. [13], the similarity is only calculated by considering the instance between two concepts. This method is not comprehensive. At this stage of the water in the field of two ontology, ontology structure at build time difference is bigger, so can be ignored in the field of structure similarity, integrated the advantages and disadvantages of the above methods, comprehensive consideration of various kinds of feature information in ontology, this chapter put forward a kind of strategy choice more comprehensive similarity method, this method is through the concept instances and defines the similarity of comprehensive comparison between the concept of ontology, the following respectively based on the concept instances and based on the concept definition of similarity are introduced.

2.1.1. Similarity Based on Concept Instances. To some extent, the instances of concepts reflect the semantic relationship between concepts. If the instances of two concepts are mostly the same, the semantic relationship between the two concepts is very strong.

Based on the concept example, the similarity $S_{\text{ins}}(A, B)$ is calculated, where $A$ and $B$ respectively represent one concept in the two ontologies. We use the method in machine learning to calculate the joint probability of the concept’s instance $S_{\text{ins}}(A, B)$, so as to obtain the matrix of the concept’s instance similarity. For a concept, Jaccard coefficient is used to calculate the instance similarity of the concept [14]. The formula is as follows.
Calculate $P(A, B)$, $P(A, \overline{B})$, and $P(\overline{A}, B)$ according to the concrete examples of concept A and concept B in ontology. Where $P(A, B)$ represents the probability that one of the instances belongs to both concept A and concept B, that is, the proportion of the instances belonging to both concepts A and B in all the instances of concept A and concept B, the formula is as follows.

$$P(A, B) = \frac{U_1^{A,B} + U_2^{A,B}}{U_1 + U_2}$$

(2)

$P(A, \overline{B})$ represents the probability that one of the instances belongs to concept A but does not belong to concept B. The formula is as follows

$$P(A, \overline{B}) = \frac{U_1^{A,\overline{B}} + U_2^{A,\overline{B}}}{U_1 + U_2}$$

(3)

Similarly, $P(\overline{A}, B)$ represents the probability that one of these instances is not a member of concept A but is a member of concept B. The formula is as follows:

$$P(\overline{A}, B) = \frac{U_1^{\overline{A},B} + U_2^{\overline{A},B}}{U_1 + U_2}$$

(4)

Then, equation (1) can be used to calculate the instance-based similarity values of concepts A and B.

2.1.2. Similarity Based on Concept Definition. Only instance of the concept of computing, there are certain limitations, such as the concept and the concept of "graduate" B "master's" represented by the same entity, but in different schools of ontology, the name is different, between instances also does not have the same part, so we should consider the similarity based on the concept definition of wordnet, contains the names of the computing concept and attribute of the concept of similarity calculation.

For similarity based concept names, we can look at the hierarchy of concepts A and B in wordnet to determine their relationship. If they are on the same node, $=1$; If they have no concept of a common superclass at a limited level, then $=0$, the specific formula is shown in (5):

$$Sim_{name}(A, B) = \frac{2 \cdot \log_2 p(A, B)}{\log_2 p(A) + \log_2 p(B)} \cdot x$$

(5)

where the coefficient $x$,

$$x = k \sqrt{1 - \frac{|Dep(A) - Dep(B)|}{Dep(A) + Dep(B)}}$$

(6)

$\log_2 p (A)$, $\log_2 p (B)$ is the information entropy of synonyms of concepts A and B in wordnet.

For similarity based on concept attributes, if concept A and concept B have the same attributes, the two concepts are highly similar. In ontology, attributes are also concepts, and the similarity between attributes depends on whether the name and data type are the same or related. ai is the attribute of concept A, bi is the attribute of concept B, then the similarity calculation formula between the two is

$$Sim_{attr}(a_i, b_j) = Sim_{name}(a_i, b_j) \times h$$

(7)

where $h$ is the data type matching value of $a_i$ and $b_j$, we set the relevant matching value as shown in table 1, and $S_{name}(a_i, b_j)$ can be calculated by using equation (5).
Table 1. Matching value.

| Matching | Float | Int | Char | Date |
|----------|-------|-----|------|------|
| Float    | 1     | 0.9 | 0.1  | 0.7  |
| Int      | 0.9   | 1   | 0.1  | 0.8  |
| Char     | 0.1   | 0.1 | 1    | 0.1  |
| Date     | 0.7   | 0.8 | 0.1  | 1    |
| Float    | 1     | 0.9 | 0.1  | 0.7  |

We set a weight of $W^k_{attribute}$ for each pair of attributes, and then the similarity calculation formula based on the attributes

$$Sim_{pro}(A, B) = \sum_{k=1}^{m} W^k_{property} Sim(a_i, b_j)$$  \hspace{1cm} (8)

Finally, the similarity formula based on the concept definition is

$$Sim_{def}(A, B) = \alpha \times Sim_{name}(A, B) + \beta \times Sim_{pro}(A, B)$$  \hspace{1cm} (9)

where $\alpha$ and $\beta$ are set by the domain expert.

2.1.3. Comprehensive Semantic Similarity. Combined with the calculation formula based on concept instance and concept definition, the comprehensive semantic similarity calculation formula is as follows:

$$Sim(A, B) = \chi Sim_{instance}(A, B) + \delta Sim_{def}(A, B)$$  \hspace{1cm} (10)

Two of these coefficients are set by domain experts.

2.2. Framework

Step 1: input two ontologies (Ontology1, Ontology2);

Step 2: type selector: instance--instance to calculate the instance-based similarity between concepts;

Step 3: type selector: concept definition--concept definition, respectively calculate the similarity based on concept name and concept attribute in wordnet.

Step 4: the domain expert sets the weight, and outputs the final comprehensive similarity calculation result according to the comprehensive semantic similarity calculation formula, so as to judge the final similarity between concepts.

2.3. Experiment

2.3.1. Knowledge Map Construction. At present, we have both internal data and external data. The comprehensive database of the exchange and sharing platform and the data of the basic water information platform are analyzed. The comprehensive database data are all real-time monitoring data, and the basic water information platform is the basic description data of infrastructure. For these two kinds of data, we use the seven-step method and build two ontologies with reference to wordnet dictionary and relevant industry standards. We select the concepts of rivers, sluice gates, lakes, Wells and rural areas for ontology mapping experiments, and some ontologies are shown as follows.

Comprehensive library ontology, ontology 1: the concept of River contains 16 examples (qingshui River, yongding River, white River, tidal River, shahe, etc.), the concept of Gate_dam contains 44 instance (Ann penstock, north commissioner flood diversion sluice, the grand palace gate lock, dongzhimen brake, etc.), the concept of Lake has 10 examples, concept Reservoir has 15 examples, the concept of Rural has 20 examples, as shown in figure 1.
Figure 1. Ontology 1.

Ontology 2: basic water information platform ontology, in which concept river contains 180 instances, concept gate 1061 instances, concept lake 41 instances, concept RS 88 instances and concept CWS 125 instances, as shown in figure 2.

Figure 2. Ontology 2.

In the process of constructing knowledge map, we currently have the above two completely different forms of ontology, a concept of rivers, gate, lakes and Wells, the countryside, the concepts of the instance, some of the same instance, in terms of concept definition, also has the similar concepts, but in terms of structure, can be intuitive to see, the above four concepts of the parent of each are not identical. Therefore, following the mapping process based on multi-strategy selection, we performed the following relevant calculations.

2.3.2. Comprehensive Semantic Similarity Calculation. (1) Similarity Calculation Based on Concept Instances

In order to calculate the similarity based on the concept instance, the instance distribution of the concept is described as shown in table 2 below. Where A represents the number of concept instances of comprehensive library, B represents the number of concept instances of foundation level platform, A \( \cap \) B represents the intersection of two ontology concept instances.
Table 2. Example distribution of the concept.

| A ∩ B | A | Gate_Dam | Lake | Reservoir | Rural |
|-------|---|----------|------|-----------|-------|
| B     | River | 180 | 13 | 0 | 0 | 0 |
|       | Gate | 1061 | 0 | 31 | 0 | 0 |
|       | Lake | 41 | 0 | 0 | 10 | 0 |
|       | RS | 88 | 0 | 0 | 0 | 13 |
|       | CWS | 125 | 0 | 0 | 0 | 18 |

According to the similarity calculation method based on concept instances in chapter 2 and by referring to table 2 above, we can get the instance similarity of the two concepts as shown in table 3.

Table 3. The similarity based on examples of concept.

| Sim_{estance}(a_i,b_j) | River | Gate_Dam | Lake | Reservoir | Rural |
|------------------------|-------|----------|------|-----------|-------|
| river                  | 0.13  | 0        | 0    | 0         | 0     |
| gate                   | 0     | 0.028    | 0    | 0         | 0     |
| lake                   | 0     | 0        | 0.235| 0         | 0     |
| RS                     | 0     | 0        | 0    | 0.14      | 0     |
| CWS                    | 0     | 0        | 0    | 0.15      |       |

Analysis: from the results in table 3, we can conclude that the instance similarity of each pair of concepts basically reflects the semantic relationship between concepts. At the same time, it can be seen that Sim(River, river), Sim(Gate_dam, gate) and other five pairs are similar. Due to the large difference in the instances containing the concepts in the two ontologies, the similarity value is relatively small. In fact, these two pairs are very similar.

2) Similarity Calculation Based on Concept Definition

Then we consider similarity calculation based on concept definition, which is divided into similarity based on concept name and similarity based on concept attribute in wordnet.

According to the two ontologies we initially constructed and wordnet dictionary, Depth(River)=Depth(river)=5, Depth(Gate_dam)=8, Depth(gate)=9, Depth(Lake)=Depth(lake)=3, Depth(Reservoir)=4, Depth(RS)=4, Depth(Rural)=Depth(CW)=3 can be obtained. First, we refer to equations (5) and (6) in chapter 2, and get the results as shown in table 4.

Table 4. Similarity based on conceptual names.

| Sim_{name}(a_i,b_j) | River | Gate_Dam | Lake | Reservoir | Rural |
|---------------------|-------|----------|------|-----------|-------|
| river               | 1     | 0.121    | 0.221| 0.101     | 0.118 |
| gate                | 0.113 | 0.754    | 0.105| 0.11      | 0.09  |
| lake                | 0.105 | 0.11     | 1    | 0.13      | 0.113 |
| RS                  | 0.132 | 0.118    | 0.105| 0.575     | 0.052 |
| CWS                 | 0.108 | 0.115    | 0.092| 0.089     | 0.174 |

Then the attribute-based similarity between two concepts is calculated. Similarly, we calculate according to equations (7) and (8) in chapter 2, and get the results as shown in table 5.
Table 5. Similarity based on conceptual properties.

| Sim_{prop}(a_i,b_j) | River | Gate_Dam | Lake | Reservoir | Rural |
|---------------------|-------|----------|------|-----------|-------|
| river               | 0.625 | 0        | 0.27 | 0         | 0     |
| gate                | 0     | 0.043    | 0    | 0         | 0     |
| lake                | 0.248 | 0        | 0.745| 0         | 0     |
| RS                  | 0     | 0        | 0    | 0.589     | 0     |
| CWS                 | 0     | 0        | 0    | 0.668     | 0     |

Integrating the similarity based on concept instances and concept attributes we calculated, we set $\alpha=\beta=0.5$ according to the opinions of domain experts, and calculated the similarity based on concept definitions according to equation (9), as shown in table 6.

Table 6. The similarity based on concept definition.

| Sim_{def}(A,B) | River | Gate_Dam | Lake | Reservoir | Rural |
|----------------|-------|----------|------|-----------|-------|
| river          | 0.813 | 0.061    | 0.248| 0.051     | 0.059 |
| gate           | 0.057 | 0.599    | 0.053| 0.052     | 0.045 |
| lake           | 0.176 | 0.052    | 0.873| 0.065     | 0.057 |
| RS             | 0.067 | 0.059    | 0.052| 0.582     | 0.026 |
| CWS            | 0.054 | 0.058    | 0.046| 0.049     | 0.421 |

Finally, according to the Suggestions of the field experts, we set the weight as $\gamma=0.2$, $\eta=0.8$, and the equation (10) for calculating the comprehensive similarity is shown in table 7.

Table 7. Comprehensive similarity calculation.

| S(A,B) | River | Gate_Dam | Lake | Reservoir | Rural |
|--------|-------|----------|------|-----------|-------|
| river  | 0.676 | 0.049    | 0.198| 0.041     | 0.047 |
| gate   | 0.046 | 0.485    | 0.042| 0.042     | 0.036 |
| lake   | 0.141 | 0.042    | 0.745| 0.052     | 0.046 |
| RS     | 0.053 | 0.047    | 0.042| 0.494     | 0.023 |
| CWS    | 0.043 | 0.046    | 0.037| 0.039     | 0.367 |

Analysis: from table 7, we can clearly find the concepts River and river, Gate_dam and gate, Lake and lake. Concept Reservoir and concept RS, concept Rural and concept CW have a large degree of similarity. It can be preliminarily concluded that the five pairs of concepts are similar and can be mapped one by one to prepare for the subsequent operation of knowledge mapping.

2.3.3. Evaluate. In the following, we present the line graph of the similarity of the three algorithms of the concept pair selected from the two ontologies as shown in figure 3.

We consulted ten domain experts in the water industry who used domain knowledge to evaluate the semantic relevance between the concepts selected from the two ontologies. We will average the value as a threshold, finally calculated the threshold for: $\Phi = 0.30$. We used 25 sets of data from the three algorithms and the threshold calculated by the domain experts to make the variance, and the final results are shown in table 8.
Figure 3. Similarity comparison chart.

Table 8. Variance comparison chart.

| Algorithm                                  | Variance |
|--------------------------------------------|----------|
| Algorithm based on concept instance        | 0.0782   |
| Algorithm based on concept definition      | 0.0755   |
| Comprehensive Similarity Algorithm         | 0.0656   |

By comparing the variance of the three algorithms, it can be seen that the variance value of the synthetic similarity algorithm is the minimum, indicating that the accuracy is the highest. Experimental results further demonstrate the effectiveness of the proposed method.

The comprehensive experimental results show that, although the comprehensive semantic similarity algorithm is not as high as the semantic similarity calculated by the algorithm based on the concept definition, the accuracy and effectiveness of the comprehensive semantic similarity algorithm are the highest as can be seen from the final variance calculation results. Therefore, the multi-strategy comprehensive semantic similarity computing model proposed in this paper can be applied to the water service ontology mapping. From the experimental results, it is one-sided and insufficient to infer the semantic relationship between concepts based on a single similarity. The comprehensive semantic similarity takes into account all aspects of the ontology concept. In the future, we can also adjust the weight to better support the mapping of different domains.

3. Conclusion
The comprehensive similarity algorithm proposed in this paper combines the instance-based method and the define-based method to generate an improved similarity calculation method, which can better realize the mapping between different water ontology and lay a good foundation for the future construction of water knowledge Graph. Ontologies are generally developed independently by different organizations, so the same or related industries have different experience, which will affect the role of ontologies in information interaction, easy to cause semantic heterogeneity, and cause obstacles to the interoperability of ontologies. Ontology mapping can exactly solve this problem. It is an important link of knowledge map development and evolution.

Due to the limitations of time and conditions, the study in this paper has the following problems that need further study: (1) the ontology mapping mechanism proposed in this paper is more suitable for the ontology mapping of current water utilities, and there may be better ontology mapping methods that need further study and application; (2) the comprehensive semantic similarity computing model proposed can well complete the ontology mapping task. In the future, we should further improve the model and study more in-depth and rich semantic similarity computing problems. (3) at present, the water ontology we have built is not comprehensive enough. We need to continue to build water ontology at other levels to lay a good foundation for the construction of the final water knowledge map.
Acknowledgement
This research was funded by [Water Pollution Control and Treatment Science and Technology Major Project] grant number [2018ZX07111005]” and “The APC was funded by [Water Pollution Control and Treatment Science and Technology Major Project].

References
[1] Wang P, Zhou Y and Xu B 2011 Matching large ontologies based on reduction anchors. *International Joint Conference on IJCAI* pp 2343-2348.
[2] Doan A H, Madhavan J, Dominings P and Halevy A 2002 Learning to map between ontologies on the semantic web *International Conference on World Wide Web* pp 662-673.
[3] Mitra P, Wiederhold G and Kersten M L 1999 A graph-oriented model for articulation of ontology Interdependencies Lecture Notes in Computer Science 99 86-100.
[4] Pang Y, Zou Y and Luan X 2003 Semantic Resolution for E-Commerce. *Springer Berlin Heidelberg*, 2564 355-366.
[5] Odr R, Guez M A and Egenhfer M J 2003 Determining semantic similarity among entity classes from different ontologies *IEEE Transactions on Knowledge and Data Engineering* 15 442-456.
[6] Macedche A, Motik B, Silva N and Volz R 2002 MAFRA-A Mapping FRAmework for Distributed Ontologies (Berlin Heidelberg: Springer) pp 235-250.
[7] Li J, Tang J, Li Y and Luo Q 2009 RiMOM: A dynamic multistrategy ontology alignment framework *IEEE Transactions on Services Computing*.
[8] Faria D, Pesquita C, Santos E, Palmonari M and Sowkarthiga B 2013 AML results for OAEI 2015 *ODBASE-OTM 2013 Conference* pp 527-541.
[9] Faria D, Martins C, Nanavaty A and Oliveira D 2015 MAFRA-A mapping framework for distributed ontologies *S ISWC International Workshop on ontology Matching (OM)* pp 235-250.
[10] Forsati R and Shamsfard M 2016 Symbiosis of evolutionary and combinatorial ontology mapping approaches *Information Sciences* 342 53-80.
[11] Hogan A, Zimmermann A, Umbrich J, Polleres A and Decker S 2012 Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora *Web Semantics & Science Services and Agents on the World Wide Web* pp 76-110.
[12] Doan A H, Madhavan J, Dhamankar R, Domingosl P and Halevy A 2002 Learning to map ontologies on the semantic web *Proceeding of the Eleventh International World Wide Web Conference* pp 303-319.
[13] Van R 1979 *Information Retrieval* (London: Butter Worths).
[14] Ehring M and Staab S 2004 QOM-quick ontology mapping *Proc of ISWC 2004* (Springer-Verlag) pp 683-697.