A method of computing conceptual semantic similarity based on part-whole relationship

Xinyu Cao¹, Jing Zhao¹, Fan Zhang¹, Gang Wu¹, Chao Zhao¹ and Haitao Wang¹*

¹China National Institute of Standardization, Beijing, Beijing, 100191, China
*Corresponding author’s e-mail: wanght@cnis.ac.cn

Abstract. Conceptual semantic similarity calculation is an important content in research fields such as natural language processing, machine translation and is widely used in information retrieval, knowledge acquisition, text classification, and clustering. It’s very important to development of natural language processing by improving the accuracy of semantic similarity calculation. This paper proposes a method for calculating the similarity of concepts based on the part-whole relationship between concepts. This method is based on the partial overall relationship diagram and uses the hierarchical features and positions of the concepts in the diagram to calculate the conceptual similarity.

1. Introduction
Semantic similarity calculation is an important research content in the field of natural language processing. Especially with the rapid development of the Internet, the processing of textual information has become more complex and important. Due to the greater freedom of web text, the calculation of similarity between two concepts has become an important content in research fields such as sentence analysis and text clustering. It improves the accuracy of the calculation of semantic similarity and has a tremendous effect on text information processing. Conceptual similarity calculation methods can be roughly divided into two categories: one is to use large-scale corpora for statistics and to calculate based on the probability distribution of vocabulary context information. This method is more accurate, but it depends on the corpus used for training. Another type of calculation is based on some kind of world knowledge, mainly based on the hierarchical relationship of a knowledge-complete semantic dictionary. This method is relatively intuitive and susceptible to people's subjective consciousness. [1, 2].

Part-whole relationship is an important semantic relationship [3], which represents a special association between an entity and its parts, and is a basic semantic relationship. The understanding of part-whole relations is the basis for people to understand some phenomena in the world [4]. If concept X and concept Y can be expressed in similar language expressions such as “X is part of Y”, “Y includes parts such as X”, then it is considered that there is a part-whole relationship between concept X and concept Y [5]. For example, there is a part-whole relationship between “bear” and “bear paw”.

Part-whole relationship is a kind of complex semantic relationship. Many researchers think that part-whole relationship should be regarded as a collection of relationships, rather than a single relationship [6]. In 1987, Winston et.al. divided some of the overall relationships into six categories based on three relational elements: Functional, Homeomerous, and Separable. Classify WCH [5]. In this paper, WCH classification is used to classify part-whole relationships. Functional in WCH classification refers to whether the part has a fixed position on the whole in the spatial structure and whether it has some effect on the functionality of the whole; Homeomerous refers to whether the part
is similar to the whole; Separable means part. Whether it is physically disconnected from the whole, that is, whether it can be physically separated from the whole. Some of the overall relationship categories in the WCH classification are as follows:

- **Component-Integral Object**: represents the relationship between a part and its whole. The whole has structure, and the parts in the part-whole relationship have a fixed position in the whole structure, have functionality to the whole, and can be separated from the whole.

- **Member-Collection**: represents membership in a collection. Members are part of the collection, and they do not assume a functional role in the collection. They are different from the collection, and members can be separated from the collection.

- **Portion-Mass**: represents the relationship between clique, physical dimension and its parts. Some of them are similar to the whole, can be separated from the whole, and have no functionality to the whole.

- **Stuff-Object**: describes the relationship between an object and its materials. Materials are parts and objects are whole. In the material-object relationship, the part has no functional effect on the whole, is not similar to the whole, and cannot be separated from the whole.

- **Feature-Activity**: represents the semantic relationship between various activities or processes and their phases. The part provides a certain function to the whole, which is different from the whole and cannot be separated.

- **Place-Area**: represents the relationship between an area and its specific place and location. Parts and the whole are homogeneous and have no functional effect on the whole, but cannot be separated from the whole.

Table 1 lists examples of the overall relationship of each class in the WCH classification and their values in the relationship attributes. In the table, "+' indicates that the relationship satisfies this relationship element, and "-' indicates that this relationship element is not satisfied.

| Category                | Functional | Homeomeorous | Separable | Examples                      |
|-------------------------|------------|--------------|-----------|-------------------------------|
| Component-Integral Object| +          | -            | +         | (tire, car)                   |
| Member-Collection       | -          | -            | +         | (teacher, school)             |
| Portion-Mass            | -          | +            | +         | (cm, meter)                   |
| Stuff-Object            | -          | -            | -         | (iron, car)                   |
| Feature-Activity        | +          | -            | -         | (graduation answer, graduate) |
| Place-Area              | -          | +            | -         | (Yellowstone National Park, USA) |

This paper proposes a method for calculating conceptual semantic similarity based on a tree structure. This method uses the hierarchical characteristics of the tree structure to calculate instance-based conceptual similarity according to the characteristics of the ontology concept and the relationship between concepts. Improve the accuracy of the calculation, simplify the calculation process, and verify the proposed method through experiments.

2. **Relate concepts**

We use the part-whole relationship diagram to obtain all partial whole relationship instances. The concepts in the part-to-whole relationship form the points in the graph, the part-whole relationship forms the edge in the graph, and the meaning of the edge indicates the type of part-whole relationship between the two concepts to which the edge is associated.

2.1. **Definition of M**

Part-whole relationship graph M is a 3-tuple M = (C, E, r), where
• Collection $C = \{c_1, c_2, \ldots, c_n\}$ is the set of concepts in the part-whole relationship diagram $M$
• Collection $E = \{e_1, e_2, \ldots, e_n\}$ is the set of directed edges in part-whole relationship diagram $M$, $E \subseteq C \times C$
• Mapping $r$, $E \rightarrow R$ is the mapping from $E$ to $R$. $r(e)$ represents the special part-whole relationship type. The range of $r(e)$ is collection of \{component-of, member-of, portion-of, stuff-of, ingredient-of, process-of, region-of, part-of\}.

2.2. Definition of related concepts

part concept and whole concept: If $<p, w> \in E$ in a part-whole relationship graph $M$, concept $p$ is called part concept of concept $w$ and concept $w$ is called whole concept of concept $p$. All of part concepts of concept $w$ are written as $Parts(w)$ and all of whole concepts of concept $p$ are written as $Wholes(p)$.

parent concept and children concept: If $c_1 \in C$, $c_2 \in C$, …, $c_n \in C$, $e_i \in E$, $e_2 \in E$, …, $e_n \in E$ and $e_i = <c_1, c_2>$, $e_2 = <c_2, c_3>$, $e_3 = <c_3, c_4>$, …, $e_n = <c_{n-1}, c_n>$ in a part-whole relationship graph $M$, $n$ is larger than 1. Then concepts $c_2$, …, $c_n$ are called parent concept of concept $c_1$ and concepts $c_1$, $c_2$, …, $c_{n-2}$ are called children concepts of concept $c_n$. All of parent concepts of concept $p$ are written as $GrandWholes(p)$ and All of children concepts of concept $w$ are written as $GrandParts(w)$.

commom concept: For giving concept $c$, the concepts are related to concept $c$ which are part concept or whole concept or parent concept or children concept of concept $c$. of For giving concept $c_1$ and $c_2$, the concepts being related to concept $c_1$ and related to concept $c_2$ are named by common concept for concept $c_1$ and $c_2$.

3. Method

3.1. Assumptions

We compute the conceptual similarity of concepts $c_i$ and $c_j$ based on part-whole relationship. In general, the more part concepts and children concepts $c_i$ and $c_j$ share, the higher the conceptual similarity between concepts $c_i$ and $c_j$. But some part concepts or children concepts are more important and contribute more to the calculation of similarities between their whole concept $c_i$ and $c_j$. As shown in Figure 1, when calculating the similarity between concept $w_1$ and concept $w_6$, concept $p_2$ contributes more to computing similarity than concept $p_1$.

We give two assumptions to measure the importance of the part concept or children concept.

• **Assumption 1**: The more the whole concepts of a concept $c$, the less important it is. As shown in Figure 1, concept $p_2$ is more important.

• **Assumption 2**: The longer the distance, the smaller the contribution to the similarity calculation. As shown in Figure 2, concept $p_1$ is more important than concept $p_2$ for calculating the relevance of concept $w_5$ to concept $w_6$. 

---

Journal of Physics: Conference Series 1544 (2020) 012046 IOP Publishing doi:10.1088/1742-6596/1544/1/012046
3.2. Algorithm
To define conceptual similarity, first we define the conceptual distance. The distance between concept \( c_i \) and \( c_j \) is written as \( CptsDis(c_i, c_j) \). If in a giving part-whole relationship graph \( M \), concept \( c_j \) is the whole concept or parent concept of concept \( c_i \). The Minimum number of edges among concept \( c_i \) and \( c_j \) is the value of \( CptsDis(c_i, c_j) \). Because edge is directional, \( CptsDis(c_i, c_j) = - CptsDis(c_j, c_i) \). Otherwise, we think the distance is infinite.

According to the characteristics of part-whole relationship and the assumptions proposed in 3.1, the definition of conceptual similarity is shown in Equation 1.

\[
CptsMutDeg(c_i, c_j) = \frac{\lambda \cdot \text{ShareCpts}(c_i, c_j) \sum_{c \in \text{ShareCpts}(c_i, c_j)} \frac{\text{WholesMst}(c)}{|\text{CptsMst}(c_i) \cup \text{CptsMst}(c_j)|}}{\max\{|\text{CptsDis}(c_i, c)|, |\text{CptsDis}(c_j, c)|\}}
\]

Where, \( \lambda = 1 \) when \( r(c_i, c) = r(c_j, c) \) that is, concepts \( c \) and \( c_i \), and concepts \( c \) and \( c_j \) satisfy the same part-whole relationship. \( \text{ShareCpts}(c_i, c_j) \) represents common concepts for concept \( c_i \) and \( c_j \). \( \text{WholesMst}(c) \) represents all whole concepts and parent concepts of concept \( c \). \( \cup_{c \in \text{ShareCpts}(c_i, c_j)} \text{WholesMst}(c) \) is used to measure the importance of common concepts of concept \( c_i \) and \( c_j \).

The value of \( CptsMutDeg(c_i, c_j) \) is less than or equal to 1.

4. Result
The calculation results heavily depend on size of the part-whole relation graph. More part-whole relationship examples, more accurate results. The algorithm proposed in this article is more suitable for calculating similarity of concepts that are completely different in characters. For example, based on our algorithm, the similarity between computers(in Chinese 电脑) and computers(in Chinese 计算机) is very high.

Acknowledgments
This research was supported by National Key Technology R&D Program (2016YFF0204205, 2018YFF0213901, 2017YFF0206503, 2017YFF0209004), and grands from China National Institute of Standardization (522019Y-6771, 522019Y-6781, 522019C-7044, 522019C-7045).
References

[1] Huang, G., Zhou, Z.R., Zhou, T.(2007) Research on the Domain-Ontology-Based Semantic Similarity Computing. Computer Engineering & Science, 29:112-117.

[2] Fan, H.Y., Zhang, Y.S.(2014) Computing method for semantic similarity of words based on How Net. Journal of Beijing Information Science & Technology University, 29:42-45.

[3] Cruse, D.A.(1979) On the Transitivity of the Part-Whole Relation. Journal of Linguistic, 29-38.

[4] Shanks, G., Tansley, E., Nuredini, J., Tobin, D., Weber, R.(2002) Representing Part-Whole Relationships in Conceptual Modeling: An Empirical Evaluation. In: Twenty-Third International Conference on Information Systems. Barcelona. pp.89-100.

[5] Winston, M.E., Chaffin, R., Hermann, D. (1987) A Taxonomy of Part-Whole Relations. Cognitive Science, 11: 417-444.

[6] Anne Iris, M., Litowitz, B. (1989) Problems with Part-Whole Relation. In: Martha, E.(Eds.), Relational Models of the Lexicon. Cambridge University Press, London. pp.261-288.