An Improved Concept of Semantic Similarity Computation Algorithm based on Domain Ontology

Yongliang Jiang[1, a], Yamin Zhang [2, b]

[1]Network Center, Hainan Normal University, Haikou 571158, China

[2]Department of Computer, Luohe Medical College, Luohe 462300, China

[a]yongliangjiang@126.com, [b]zymluohe@126.com

Keywords: domain ontology, semantic similarity, concept similarity.

Abstract. To effectively solve the problem of semantic similarity between concepts, the existing concept semantic similarity computation methods were studied and an improved concept semantic similarity computation algorithm based on domain ontology was put forward. In the process of computing concept semantic similarity the algorithm not only considered the basic relationships but also the custom relationships between concepts. In addition, the algorithm also took into account the concept of properties, and the concept of instance impact on semantic similarity computation. All these measures make the effectiveness of the algorithm in computing concept semantic similarity improved. The example showed that in the aspect of concept semantic similarity computation the proposed algorithm is more effective than the existing algorithms.

Introduction

To the problem of concept semantic similarity computation the domestic and foreign scholars have carried on the related research and achieved some results. Richardson determined the weights of the concept according to the node density, depth and strength and then commutated the semantic similarity between concepts based on the weights [1]. Yuhua Li used concept depth, density and length between concepts to construct a nonlinear function, and then used this function to commutate semantic similarity between concepts[2].Lin-tao Lv presented a computing concept similar model based on context[3].Literature[4] proposed a concept semantic similarity computation method based on the number of the concept properties. According the upper and lower relationship between concepts and other relationship Jie Chen gave a new computation method[5].Mei-rong Yang proposed a concept of semantic similarity computing model based on the main similarity between concepts[6] etc. The above semantic similarity computation algorithms only considered certain aspects of the impact semantic similarity. That make the algorithms to achieve good results in certain applications. So an improved semantic similarity computation algorithm was proposed. The algorithm took into account the basic relationships between concepts, custom relationships between concepts, the properties of the concept, the instances of the concept and other factors influencing in the process of computing the semantic similarity between concepts. That effectively improved the algorithm's effectiveness and versatility.

Related definitions

Concept: A concept is an abstract description of the objective world of anything. It can be described by a four-tuple: {ID, L, P, I}.In the four-tuple, ID is a unique identifier of the concept. It can be represented by a URI. L represents the concept of language vocabulary. P is the concept of the collection of properties. I represents a collection of instances of concepts. Conceptual similarity if the two concepts C1 and C2 have certain common characteristics, we can say that these two concepts are similar. The degree of similarity between the two concepts is the similarity between them. Ontology is the concept of a formal specification. An ontology is composed of different concepts. Domain Ontology: Domain ontology is a professional ontology. It describes the relationship between the concepts of specific areas. In this paper, semantic similarity between the concepts based on domain ontology is represented by Sim (C1, C2) $\mathrm{Sim}(\mathrm{C}_1, \mathrm{C}_2) \in [0,1]$. If Sim (C1, C2) = 1, it indicates that C1 and C2 are the same concepts two concepts. If Sim (C1, C2) = 0, it indicates that C1 and C2 are two completely different concepts.

The Improved Concept Semantic Similarity Computation Algorithm based on Domain Ontology

In the domain ontology, the concept of semantic information is usually described by the relationship between the concepts. However in practice, there is not only hyponymy but also other relationships. To improve the accuracy of the computation the non-hyponymy must be considered in the process of computing the semantic similarity between concepts. The practical applications showed that the properties and instances of the concept to compute the semantic similarity between concepts were influential. Based on the analysis of the above, not only the hyponymy and custom relationship were considered but also the properties and instances of the concept were considered. The computation process of the algorithm is shown in figure 1.

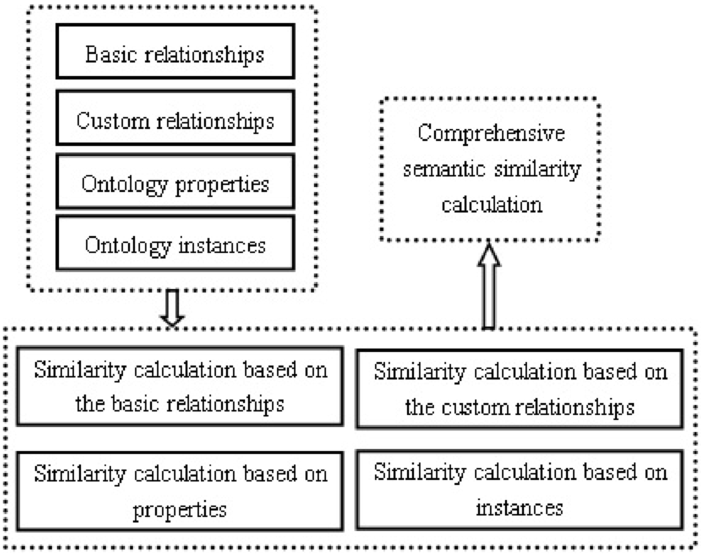
Basic relationships
Custom relationships
Ontology properties
Ontology instances
Comprehensive semantic similarity calculation
Similarity calculation based on the basic relationships
Similarity calculation based on the custom relationships
Similarity calculation based on properties
Similarity calculation based on instances

Fig. 1 Algorithm computation process

C_1, C_2 are two concepts of a domain ontology. Dep (T) represents the height of the domain ontology tree. Dis (C_1, C_2) represents the shortest path length between C_1 and C_2. P (C_1) is C_1’s parent node. P (C_2) is the parent node of C_2. CP (C_1, C_2) is the common parent node of C_1 and C_2. $Count(P(C_1) \cap P(C_2))$ Represents the number of the common parent node of C_1 and C_2. Max (Dep (C_1), Dep (C_2)) is the maximum value of Dep (C_1) and Dep (C_2). Wid (C_1) is a semantic density of C_1. Its value is the number of C_1’s siblings. Wid (T) is the maximum density of all nodes in the domain ontology tree. Based on the above assumptions, the distance impact factor can be defined as $Y_1 = \sqrt[x1]{1 - \frac{Dis(C_1,C_2)}{Dep(T)}}$. Coincidence of the impact factor can be defined as $Y_2 = \sqrt[x2]{\frac{Count(P(C_1) \cap P(C_2))}{Max(Dep(C_1), Dep(C_2))}}$. Deep impact factor can be defined as $Y_3 = \sqrt[x3]{1 - \frac{Dis(C_1, P(C_1,C_2)) + Dis(C_2, P(C_1,C_2))}{2 \times Dep(T)}}$. Density factor can be defined as $Y_4 = \sqrt[x4]{\frac{wid(C_1) + wid(C_2)}{2 \times wid(T)}}$. The semantic similarity between C1 and C2 based on the basic relations can be computed according to the equation $\mathrm{Sim}_\mathrm{B} = \mathrm{Y}_1 \times \mathrm{Y}_2 \times \mathrm{Y}_3 \times \mathrm{Y}_4$.

Semantic similarity between C_1 and C_2 based on custom relationships is expressed by Sim_C. Literature [5] described the method of computing the value of Sim_C.

If two concepts have similar properties, the two concepts are similar. If the domain and range of the two properties are similar, the two properties are also similar [7]. The properties of the concept are divided into data type properties and object type properties. Assume similarity based on datatype properties is Sim_D, similarity based on object properties is Sim_O and similarity based on properties is Sim_A, then the similarity based on properties can be computed by equation Reference [7] described the method of computing the value of Sim_A.

$$\mathrm{Sim}_\mathrm{A} = k_1 \times Sim_D + k_2 \times Sim_O .$$

If the instances are all same, then the two concepts are the same concept. If the proportion of the same instances between the two concepts is same, then the two concepts are similar. The instances of a concept is also the instances of its ancestor concepts [8]. Let $P(C_1 \cap C_2)$ be the number of instances not only belonging to C_1 but also belonging to C_2. $P(C_1 \cap \overline{C_2})$ Represents the number of instances belonging to C_1 but not belonging to C_2. $P(\overline{C_1} \cap C_2)$ Represents the number of instances belonging to C_2 but not belonging to C_1. The semantic similarity between C_1 and C_2 based on instances can be computed according to the formula:

$$Sim_E = \frac{P(C_1 \cap C_2)}{P(C_1 \cap C_2) + P(C_1 \cap \overline{C_2}) + P(\overline{C_1} \cap C_2)} .$$

According the above description the following formula will be used to computed the semantic similarity between two concepts:

$$\mathrm{Sim(C1,C2)} = \alpha \times Sim_B + \beta \times Sim_C + \gamma \times Sim_A + \eta \times Sim_E ,$$

$$0 \le \alpha \le 1, 0 \le \beta \le 1, 0 \le \gamma \le 1, 0 \le \eta \le 1, \alpha + \beta + \gamma + \eta = 1 .$$

Examples of application

To test the effectiveness of the algorithm an experimental platform was build. The platform was made up of the Eclipse, Protégé, Lucene, Java and JenaAPI. Take the data from the literature [9] as the experimental data. Respectively, used the literature [9], [10] and the proposed algorithm for same computation tasks. Semantic similarity computation results obtained by different algorithms is shown in figure 2. Figure 2 shows that the semantic similarity computed by the proposed algorithm is closer to the experience of experts than the other two algorithms.

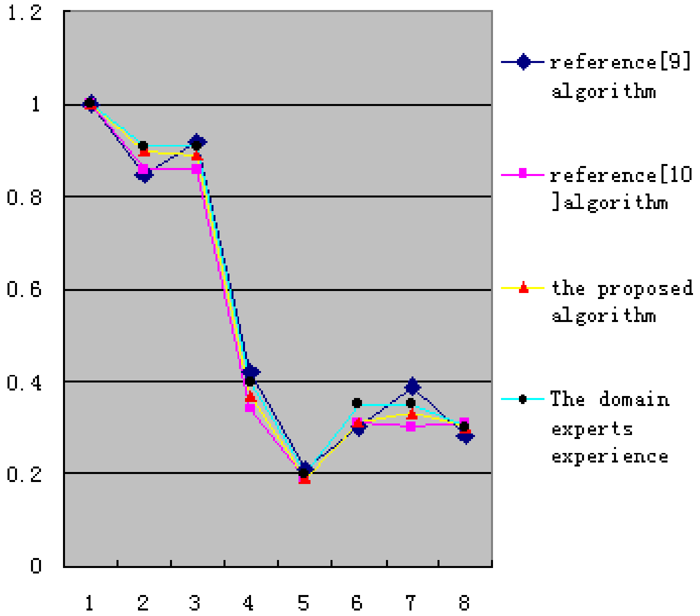

Fig. 2 Semantic similarity computation results compilation

Conclusion

After studying the existing concept semantic similarity computation algorithms the writer gave an improved semantic similarity computation algorithm based on domain ontology. The algorithm took into account of the various factors that affected the computation of semantic similarity. That improved the accuracy and validity of the computation of the semantic similarity. The example showed that the computed results of the proposed algorithm was more in line with experience of the domain experts. Compared with other algorithms the proposed algorithm has higher practical value.

Acknowledgment

I would like to express my gratitude to all those who helped me during the writing of this thesis. This work was supported by Science Foundation of Hainan Province, China (612136) and Scientific Research Foundation of the Higher Education Institution of Haina Province, China (Hjkj2012-59).

References

[1] Richardson R, Smeaton A F, Murphy J. Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words [EB/OL]. ftp://ftp. compapp.dcu.ie/pub/ w-papers/1994/ CA1294. ps.Z, 2012-03-09.

[2] Lv Lin-tao, Dong Ying. Concept Semantic Similarity Computation Model Based on Context [J]. Computer Engineering, 2010, 36(21): p.59-61.

[3] Lv Lin-tao, Dong Ying. Concept Semantic Similarity Computation Model Based on Context, J. Computer Engineering. 36(2010) p.59-61.

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

[5] Chen Jie, JIang Zu-hua. Concept Similarity Computation for Domain Ontology, J. Computer Engineering and Applications, 33(2006) p.163-166.

[6] Yang Mei-rong, Shao Hong-yu, etc. Research into Improved Model for Concept Similarity Computation in Domain Ontology, J. Information Science, 32(2014)72-77.

[7] Li Rong, Yang Dong, Liu Lei. Research of Ontology Based Conceptual Similarity Computation, J. Journal of Computer Research and Development, 48 (2011) p.312-317.

[8] Zheng Xiao-jie Zhang Lin. Modification of Similarity Computation in Ontology Mapping. Computer Science, 40 (2103) p.108-112.

[9] Liu Zi-yu, Huang Lei. Research on Concept Semantic Similarity Computation Based on Domain Ontology Model, J. Research of Ontology Based Conceptual, 33 (2011) p.52-57.

[10] Cui Qi-wen, Xie Fu. An Improved Computational Method for Conceptual Semantic similarity in Domain Ontology, J. Computer Applications and Software, 29 (2012) p.173-175.