Computing Concept Relatedness Based on Ontology

Yanping Lu, Xingwei Hao and Shaocun Tian

Abstract Concept relatedness is widely used in information retrieval, text classification, semantic extension, and other fields. So measuring the concept relatedness efficiently is an important task. Previous studies rarely distinguish between relatedness and similarity; they usually use a common formula. We suggest that concept relatedness consists of similarity and relevance, which should be computed differently. In this paper, we first give a similarity measure based on path length, taxonomy depth, and different relations between concepts. Then we propose a method to measure the specific association relation besides basic relations. Finally, incorporating both similarity and specific relevance, we get an overall formula of computing concept relatedness. Compared to existing methods, our measure of concept relatedness is more consistent with human judgment.

Keywords Domain ontology · Relevance · Similarity · Concept relatedness

1 Introduction

Relations between words are very complex in natural language. One is related to another is only a simplified summary. Actually semantic relatedness represents the degree of how words are related; it can be quantified by some general measure.

Semantic relatedness is widely used in information retrieval, text classification, semantic extension, and other fields. In particular, finding the semantic relatedness between two words has been one central problem in information retrieval for many

Y. Lu (✉) · X. Hao · S. Tian
School of Computer Science and Technology, Shandong University, Jinan, China
e-mail: luyanping1994@163.com

Z. Wen and T. Li (eds.), *Knowledge Engineering and Management*,
Advances in Intelligent Systems and Computing 278,
DOI: 10.1007/978-3-642-54930-4_12, © Springer-Verlag Berlin Heidelberg 2014

years. By extending the query with closely related words, performance of information retrieval system can be significantly improved [1]. The computation of concept relatedness is very important for many NLP applications. Most of the previous studies calculate similarity as the relatedness between the two concepts, while Resnik [2] has given an example to explain the difference between them. He points out that cars are dependent on the gasoline to move, while cars and bicycles are both vehicles and have some same components, they also share the common attribute of transportation. If we compute the relatedness by models considering only similarity, the relatedness of cars and bicycles is certainly greater than that of cars and gasoline, but from our knowledge of the real word, we know that cars and gasoline are more closely related. Therefore, Resnik [2] points out similarity was a special kind of relatedness; similar concepts are related to each other. In addition to similarity, there are other kinds of relations between words. We consider those special relations as semantic relevance. Then semantic relatedness includes both similarity and relevance.

In this paper, we propose a method of computing the relatedness between two concepts. Our method is based on measurement of similarity and relevance. For the part of similarity, a number of factors such as different semantic relations, shortest path length, etc., are considered. For the relevance part, we propose a computing method based on distance.

2 Related Works

There are two kinds of model for computing semantic relatedness. One is based on word co-occurrence of real corpus. It requires large-scale data for statistical analysis to get convergent results [3, 4]. The other is based on linguistic knowledge and taxonomy system. Usually, a common formula is used to calculate the semantic relatedness ignoring the difference between similarity and relatedness. When the relation is *is-a*, we get a measure of similarity, otherwise we get a measure of relatedness [5].

According to the corpus-based model, more times two words co-occur, more closely they are related. To some extent, this method reflects the degree of relatedness, but it can't further explain the particular semantic relations between words, and semantic relatedness is more about concepts than words, which makes it a less satisfying method to measure semantic relatedness.

As we have already mentioned above, relatedness and similarity computation share same calculation formulas in many previous models. There have been a number of algorithms proposed. For example, Liu et al. [6] proposed an algorithm based on HowNet, considering the distance of concepts. The simplest algorithm [7] only utilizes the shortest path among the possible paths between concepts. Short distance means high similarity. In spite of its simplicity, it has been applied to multiple constraints medical semantic web [8] and gives a rather good result. Leacock and Chodorow [9] extend the idea by scaling the path. Their method shows some improvements. But all methods above have the common flaw that same distance results in same relatedness, whatever their depths are. Wu and Palmer [10] not only consider the distance between concepts, but also take common parent nodes of two concepts into consideration, as is shown in the following formula:

$$\mathrm{Sim}(X,Y)=\frac{2*\mathrm{depth}(\mathrm{msc}(X,Y))}{\mathrm{len}(X,Y)+2*\mathrm{depth}(\mathrm{msc}(X,Y))} \tag{1}$$

msc(X, Y) denotes the parent concept of concept X and Y. Their algorithm has better results compared to the previous two methods. Lin [11] defines similarity in term of information content besides the factors of length and depth. More common parent nodes two concepts share, they are more related, and otherwise less related.

Duan [12] proposes a new method which has better results than previous algorithms. The method is a nonlinear combination of path length, concept intersection, the union set of concepts, and the depth level. The formula is as follows:

$$\mathrm{Sim}(X,Y)=\begin{cases}1 & X=Y\\ \dfrac{\alpha\times\beta\times|N\mathrm{Set}(X)\cap N\mathrm{Set}(Y)|}{(\mathrm{Dist}(X,Y)+\alpha)\times|N\mathrm{Set}(X)\cup N\mathrm{Set}(Y)|\times(\gamma|d(X)-d(Y)|+1)} & X\neq Y\end{cases} \tag{2}$$

Although the above models have considered many factors, they have their own scope of application. For example, the target application of Liu' algorithm is machine translation. It considers the structure and the interpretation of the word, but does not consider cases where words have low similarity but high relevance. For example, by the algorithm, the similarity between “孔子” (Confucius) and “孟子” (Mencius) is 1,while the similarity between “孔子” (Confucius) and “论语” (The Analects) is 0.130233.

3 Concept Relatedness

3.1 Ontology and Conceptual Relation

Domain Ontology [13] is an abstraction of domain knowledge, including concepts of the discipline, attributes of concept and relations between concepts and attributes. The relatedness is the quantification of the relationship of the ontology. In the domain ontology, relation between concepts contains the basic relation and the associated relation. The basic relation contains *is-a, part-of*, *attribute-of*, *made-of* [8, 14]. The associated relation is defined by experts in particular field who are familiar with domain knowledge. This particular relation determines the relevance between concepts. A simple ontology graph of virus knowledge is given in Fig. 1. In the graph, solid lines represent the basic relation, while dotted lines represent the associated relation.

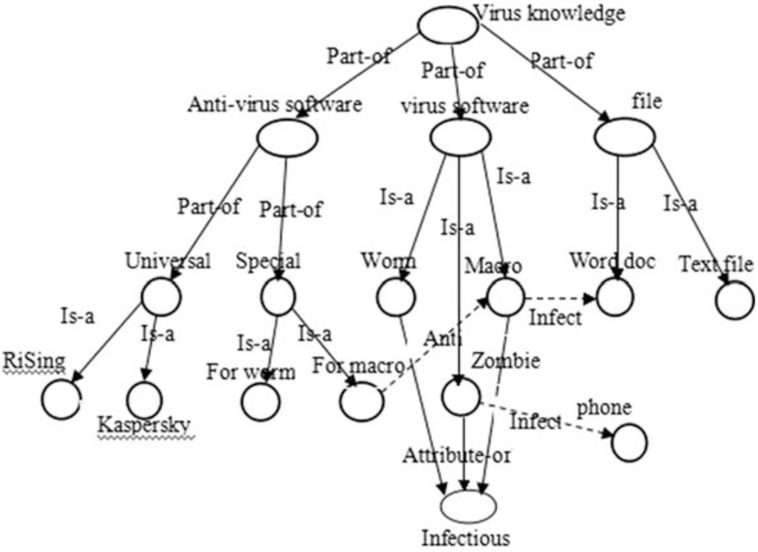

Fig. 1 A fragment of virus ontology

Though several basic relationships are marked in Fig. 1, it is not complete. Concepts usually have many attributes and many other components which the figure doesn't show.

3.2 Concept Similarity

In general, given an ontology graph, factors that affect the similarity between concepts are as follows: the shortest path length of the concept, the hierarchy depth of the concept, the density of concept, and the maximum common ancestor set [15]. In this paper, we give the following definition.

Definition 3.1 The length of relational edge, it refers to the weight of different relations between two concepts in the ontology. Because different relations have different contributions to the similarity, so we assign different weights to different relational edges. We define that d(*Is-a*) = a_1, d(*Part-of*) = a_2, d(*Is made of*) = a_3, d (*an attribute of*) = a_4. If it is required to define new basic relation, the length of which is max{a_1, a_2, a_3, a_4}. For all i, we have $a_i \geq 1$.

Definition 3.2 The shortest path distance, it refers to the weighted sum of edge length in the shortest path between two concepts X and Y. We denote it as dist(X, Y). When the two nodes are not connected, dist(X, Y) = ∞.

Definition 3.3 The depth. In the ontology, the depth of root node is defined to be 1. The depth of any concept X except root node is calculated as:

depth(X) = depth(parent(X)) + 1

Definition 3.4 The sum of depth, it refers to the recursive sum of the depth of node X and its parent nodes. Here we use the symbol Sumdepth(X), then by definition, Sumdepth(X) = $\sum_{i=1}^{\text{depth}(X)} i$.

Definition 3.5 The upper set of concepts, the set of nodes in the shortest path from concept X to the Root node. It is denoted as US(X).

It is clear that the contribution to similarity of node from different levels is different. Deeper level represents finer concept granularity, accordingly, hence the contribution to similarity is larger. On the contrary, the contribution will be smaller. Similarity calculation is divided into two parts. The first part is determined by the upper set of concepts, the depth and the sum of depth. The second part is calculated based on the shortest distance. Then values of these two parts are combined, as:

$$\mathrm{Sim}(X,Y) = \begin{cases} 1 & X = Y \\ \alpha \dfrac{\sum\limits_{Z \in \{US(X) \cap US(Y)\}} \mathrm{depth}(Z)}{\max\{\mathrm{Sumdepth}(X), \mathrm{Sumdepth}(Y)\}} + \beta \dfrac{\lambda}{\lambda + \mathrm{disc}(X,Y)} & X \neq Y \end{cases} \tag{3}$$

α and β are parameters that act as weights of the two factors (the upper set of concepts and the shortest distance) in the integrated semantic similarity. The only constraint is 0 ≤ α, β ≤ 1 and α + β = 1, but the specific values depend on specific application. The interval of the similarity is [0, 1]. Equation 3 clearly shows that nodes in different level have different weights. For a parent node Z, the depth of Z is depth(Z), the weight of Z is $\frac{\mathrm{depth}(Z)}{\max\{\mathrm{Sumdepth}(X), \mathrm{Sumdepth}(Y)\}}$. We can know that for the upper set of concepts, the deeper the node is, the greater the weight is.

And Eq. 3 satisfies the following conditions:

1. If the distance of the two concepts is 0, the similarity of them is 1;
2. The value of the similarity ranges from 0 to 1.
3. The greater the distance of two concepts is, the smaller the similarity is. The smaller the distance is, the greater the similarity is.
4. If the distance is infinite, the similarity is 0;
5. The more nodes the intersection of two concepts' upper sets has, the greater the similarity is.

3.3 *Concept Relevance*

The associated relation is defined by experts of specific area. These relations determine the relevance between concepts. For example, personalization is a relatively new word in the field of computer science. With the development of user-centered Web2.0, personalized search has become an important concept.

While generally personalization is not similar to search, in the field of computer science, we see a strong relatedness between these two concepts. Then an expert may define a special associated relation between them, which in turn will facilitate our calculation of relatedness.

The relevance is based on distance. We define that two concepts X and Y are relevant if only there is path between them that contains edges of associated relation. Since associated relation is less transitive than basic relations, every appearance of it will cause significant decrease in relevance.

$$\mathrm{Re}\,l(X, Y) = \frac{\gamma}{\gamma + \underset{d \in \mathrm{Allpath}}{\Pi} d} \tag{4}$$

where, Allpath is an aggregation of all the edges from concept X to Y, d is the length of the edge. The product in the denominator guarantees that the relevance will greatly decrease as the path becomes long. γ is a parameter controlling the maximum value of relevance.

3.4 Concept Relatedness

The semantic relatedness of concepts X and Y is the integration of similarity and relevance of concepts X and Y. It is calculated as:

$$\mathrm{Sim\_Re}\,l(X, Y) = \mathrm{Sim}(X, Y) + \mathrm{Re}\,l(X, Y) - \mathrm{Sim}(X, Y) * \mathrm{Re}\,l(X, Y) \tag{5}$$

The upper bound of relatedness is 1.

4 Experiments and Result

In this section, we give the experiment result of our method based on Fig. 1. We choose different pairs of words to show the influence of depth and other factors. The calculation follows the description in Sect. 3. In this paper, we set parameters as follows:

$$a_1 = 1.5,\ a_2 = a_3 = a_4 = 3,\ a_5 = 2;$$
$$\alpha = 0.5, \beta = 0.5, \lambda = 4, \gamma = 6;$$

In addition, we give a detailed comparison with other classic methods. They are introduced respectively by Wu and Palmer [10] (Eq. 1) and Duan [12] (Eq. 2). Results are summarized in Table 1. Column $R2$ shows the result of Wu and Palmer. Column $R2$ shows the result of Duan, we set parameters as the original paper $\alpha = 5$, $\beta = 1$, $\gamma = 0.2$. The last column is the result of our method.

Table 1 Results of the relatedness computation

Concept pairs		$R1$	$R2$	Col
Virus knowledge	Virus software	0.66667	0.34722	0.45238
Virus knowledge	Universal	0.50000	0.17007	0.28333
Virus knowledge	Worm	0.5	0.17007	0.31863
Anti-virus	Universal	0.80000	0.46296	0.53571
Anti-virus	For worm	0.66667	0.25510	0.38529
Universal	Rising	0.85714	0.52083	0.66364
Virus software	Zombie	0.80000	0.46296	0.61364
Zombie	Phone	0.85714	0.52083	0.75
Virus software	Phone	0.66667	0.25510	0.66667
Macro	Word doc	0.66667	0.52083	0.80929
For Macro	Word doc	0.5	0.34014	0.670
Special	Word doc	0.40000	0.19531	0.61063
Warm	Infectious	0.85714	0.52083	0.58571
Rising	Kaspersky	0.75	0.42857	0.58571
Warm	Macro	0.66667	0.35714	0.53571

From Table 1, it is clearly seen that Wu and Palmer [10] only considers the semantic distance and common nodes, it does not consider the concept granularity, which usually has impacts on relatedness. Duan's method [12] considers more comprehensive factors, thus the result is more reasonable, but it doesn't distinguish the different semantic relations. For example, the relatedness of Virus knowledge and Universal is equal to that of Virus and Worm.

In our method, edges of different relations of instance, part, properties, and composition have different length. In addition, relevance decreases rapidly as the number of associated edges included in the path between two concepts increases. Just as Table 1 shows, the relatedness of Macro and Word doc is much larger than that of For Macro and Word doc. Besides, for paths that contain different relationship between concepts, the results are different. The relatedness of Virus software and phone is 0.66667. The path between them contains two edges: one is the relationship "is-a", another is associated relationship. While the relatedness of For Macro and Word doc is 0.670. The path between these two concepts also contains two edges, but they are all the associated relationship. The result is more consistent with people's intuition.

5 Conclusions and Future Work

In this paper we proposed a novel algorithm of measuring the semantic relatedness between concepts. The model is based on a weighted graph for some domain ontology. Different relations have different weights. According to the experiment result, our algorithm is better than the previous approaches. And it can be further applied in data mining and information retrieval, etc.

The method in this paper is based on domain ontology, which largely depends on experts to define the semantic relations and semantic distance. The relationship between concepts must be well defined to achieve a better result. In future work, we will focus more on this challenging problem.

References

1. Li Y, Bandar ZA, McLean D (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15(4):871–882
2. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007
3. Zazo ÁF, Figuerola CG, Berrocal JLA, Rodriguez E (2005) Reformulation of queries using similarity thesauri. Inf Process Manage 41(5):1163–1173
4. Dagan I, Lee L, Pereira FC (1999) Similarity-based models of word cooccurrence probabilities. Mach Learn 34(1–3):43–69
5. Liu H, Xu D (2012) Ontology based semantic similarity and relatedness measure review. Comput Sci 39(2):8–13
6. Liu Q, Li S (2002) Based on the "Text" lexical semantic similarity computation. Chin Comput Linguist 7(2):59–76
7. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 19(1):17–30
8. Kamps J, Marx M, Mokken RJ, De Rijke M (2004) Using WordNet to measure semantic orientations of adjectives
9. Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database. 49(2):265–283
10. Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics, pp 133–138. Association for Computational Linguistics
11. Lin D (1998) An information-theoretic definition of similarity. In: ICML, vol 98, pp 296–304
12. Duan J, Yang Z, Gan J (2009) The comprehensive quantification research of semantic similarity and relevance based on domain ontology. Comput Sci Technol 11:011
13. Ruotsalo T, Hyvönen E (2007) A method for determining ontology-based semantic relevance. In: Proceedings of database and expert systems applications, pp 680–688. Springer, Berlin, Heidelberg
14. Alvarez MA, Lim S (2007) A graph modelling of semantic similarity between words. In: International conference on semantic computing 2007, ICSC 2007, pp 355–362. IEEE
15. Song L, Ma J, Lian L, Chen Z (2006) Fuzzy similarity from conceptual relations. In: IEEE Asia-Pacific conference on services computing 2006, APSCC'06, pp 3–10. IEEE