The Journal of Systems and Software 85 (2012) 370–381

Contents lists available at SciVerse ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Concept vector for semantic similarity and relatedness based on WordNet structure

Hongzhe Liu[a,b,*], Hong Bao[a,b], De Xu[a]

[a] Beijing Jiaotong University, Beijing, China
[b] Beijing Union University, Beijing Key Laboratory of Information Service Engineering, Beijing, China

ARTICLE INFO

Article history:
Received 16 November 2010
Received in revised form 9 August 2011
Accepted 24 August 2011
Available online 31 August 2011

Keywords:
Concept similarity
Concept relatedness
Concept vector model
Hierarchical concept tree
Hierarchical concept graph
WordNet

ABSTRACT

We define WordNet based hierarchy concept tree (HCT) and hierarchy concept graph (HCG), HCT contains hyponym/hypernym kind of relation in WordNet while HCG has more meronym/holonym kind of edges than in HCT, and present an advanced concept vector model for generalizing standard representations of concept similarity in terms of WordNet-based HCT. In this model, each concept node in the hierarchical tree has ancestor and descendent concept nodes composing its relevancy nodes, thus a concept node is represented as a concept vector according to its relevancy nodes' local density and the similarity of the two concepts is obtained by computing the cosine similarity of their vectors. In addition, the model is adjustable in terms of multiple descendent concept nodes. This paper also provides a method by which this concept vector may be applied with regard to HCG into HCT. With this model, semantic similarity and relatedness are computed based on HCT and HCG. The model contains structural information inherent to and hidden in the HCT and HCG. Our experiments showed that this model compares favorably to others and is flexible in that it can make comparisons between any two concepts in a WordNet-like structure without relying on any additional dictionary or corpus information.

1. Introduction

Semantic similarity between concepts is becoming a common problem for many applications of computational linguistics and artificial intelligence such as information retrieval and word sense disambiguation (Alexander and Hirst, 2006). The notion of similarity is to identify concepts having common "characteristics." Humans can judge relatedness between concepts even if they do not know how to define that relatedness formally. For example, even a small child can tell that "apple" and "orange" have more to do with each other than "apple" and "toothpaste". Formally, the way in which these pairs of concepts are related to each is called an "is-a" hierarchy. However, even dissimilar entities may be semantically related in some way. For example, "apple" and "orange" have some similarity, while "glass" and "water," "tree" and "shade," or "gym" and "weights" have no formal similarity but are still related in some way. Semantic similarity is a type of semantic relatedness. Usually similarity methods make use of "is a" hierarchy only, while relatedness methods consider more relations (Pedersen et al., 2004).

There are several WordNet implemented similarity measure and relatedness measures, they are generally grouped into four categories that including structure-based, information-content-based, feature-based and the hybrid approaches which are listed in the following Table 1.

In this paper, we propose a novel structure-based concept vector that can use cosine similarity to compute concept similarity in WordNet. Our approach distinguishes from previous work in the sense that we do not need a training corpus to fine-tune the algorithm, and it has good time performance as edge-based methods. Overall, the approach has very good human correlation. These are very important for building a new application.

2. Need for a new measurement

WordNet is the product of a Princeton University research project that has attempted to model the lexical knowledge of a native speaker of English (WordNet website, 2010). The system uses both online thesauri and online dictionaries to organize each part of speech (such as nouns and verbs) into taxonomies that render each node into a set of synonyms (synset). These synsets are represented as one sense. Words with more than one sense appear in multiple synsets. WordNet also defines the semantic and lexical relations between synsets and word senses, respectively, as follows:

- Semantic relations are hyponym/hypernym ("is-a") and meronym/holonym ("part of," "member-of," "substance-of") relations.

* Corresponding author at: Beijing Union University, Beijing, China. Tel.: +86 010 64900942; fax: +86 010 64900942.
E-mail addresses: xxtliuhongzhe@buu.edu.cn (H. Liu), baohong@buu.edu.cn (H. Bao), dxu@bjtu.edu.cn (D. Xu).

0164-1212/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2011.08.029

Table 1
Classification of existing methods.

Category		Methods	Description	Advantages and disadvantages
Structure based	Edge based	Rada et al. (1989), Wu and Palmer's (1994)	Based on edge counting	Simple, but not so accurate
	Structure based	CP/CV (Jong Wook and Selçuk, 2006)	Based on WordNet structure	Relatively simple, use of WordNet structure only
Information content based		Resnik (1999), Leacock and Chodorow (1998), Lin (1998), Jiang and Conrath (1997)	Mainly based on information content	Need a additional corpus
Feature based		Tversky's (1977), Banerjee and Pedersen (2003), Patwardhan's gloss vector (Patwardhan, 2003)	Based on attributes or WordNet gloss	Need a complete attribute or gloss set
Hybrid		Hybrid: Li et al. (2003), SSA: Marco and SeungJin (2007), OHIIC: Bin et al. (2009), and Peng et al. (2009), ZhongCheng (2009). Improved model: Songmei and Zhao (2010), Lei et al. (2009)	Combination of above	Depend on its component methods

- Lexical relations consist of both derived form relations and antonym relations.

The ground truth data commonly used to evaluate similarity measures between words comes from an experiment performed by Miller and Charles (1991). The authors carried out a user study in which assessors were given 30 pairs of words and asked to rate these words for similarity in meaning on a scale from 0 (dissimilar) to 4 (highly similar). Examining the similarity values of these 30 pairs of words, we found that, strictly speaking, the Miller and Charles similarity values not only covered similarity but also included relatedness inside. By relatedness, we mean the inner connection between two terms while being used in the same context. For example, the similarity between "journey" and "car" got value 1.16. As we know from common sense, "journey" and "car" have no formal similarity but are related to each other in the sense that one can use a car to go on a journey.

Rada et al. pointed out that the assessment of similarity in a semantic network can be in fact thought of as involving only taxonomic "is-a" relations (Rada et al., 1989). Most previous work mainly focuses on the "is-a" hierarchy of WordNet (Rada et al., 1989; Wu and Palmer, 1994; Jiang and Conrath, 1997; Lin, 1998; Leacock and Chodorow, 1998). But as our examination of Miller and Charles's similarity values has shown, other relations should also play role in the computation of similarity. This role is shallow, however, because the hyponym/hypernym relation accounts for nearly 80% of all link types. Nevertheless, it does affect the extent of relatedness (Yang and Powers, 2005; Hirst and St-Onge, 1998; Banerjee and Pedersen, 2003; Patwardhan, 2003). For example, in WordNet, the nearest common node of sense 1 of "jewel" is denoted as "jewelry#1" ("a precious or semiprecious stone incorporated into a piece of jewelry") together with sense 4 of "stone", denoted as "stone#4" ("a crystalline rock that can be cut and polished for jewelry"). Both of these are types of "physical entity", which is one layer below "entity". If we ignore the meronym/holonym relation, the word similarity search still takes place in the "is-a" hierarchy. By the edge-counting-based methods, the word sense pair "jewelry#1" and "stone#4" shows only a very small similarity value. But when we consider the meronym/holonym relation between them, the similarity increases, which accords with commonsense judgment. This interconnectivity of hyponym/hypernym and meronym/holonym relationship hierarchies produces benefits such as permitting the more accurate evaluation of the relatedness of word pairs such as "stone" and "jewel".

According to the analysis of paper (Marco and SeungJin, 2007; Varelas et al., 2005) and our study of the WordNet based similarity measures, we conclude that edge-counting-based methods ignore most of the structure of WordNet, so they are simple but produce some unreliable results. Information-content-based methods need an additional large text corpus to compute word frequency. In addition, it ignores part or all of the structure of the taxonomy, so it normally generates a coarse result for comparison of concepts (Jiang and Conrath, 1997). Feature-based methods rely on a complete attribute or WordNet gloss set. Structure-based method like CP/CV method by Jong Wook and Selçuk (2006) includes an iterative concept propagation process. Although combined approach improved human correlation to some extent, they did not solve the above problem in essence.

In this paper, we propose a novel structure based method which makes use of hyponym/hypernym, meronym/holonym relations and the full structure of the WordNet, and the method does not need to relay on any additional corpus or dictionary information.

3. Semantic similarity and relatedness for WordNet-based HCT and HCG

In this section, we address to the means by which concept similarity and relatedness are computed based on WordNet structure. We first define HCT and HCG in the WordNet taxonomy (Section 3.1). Then we explain in details how semantic similarities/relatedness can be calculated from HCT (Section 3.2) and HCG (Section 3.3).

3.1. WordNet-based HCT and HCG

3.1.1. Definitions of HCT and HCG

Definition 1 *(Hierarchical concept tree).* HCT is denoted as T (N, E), a rooted tree where N is the set of concept nodes in the tree and E is the set of edges between the parent/child pairs in H. The semantic coverage of the child concept nodes is the partition of the semantic coverage of their parent concept node. WordNet-based HCT contains all hyponym/hypernym relations and concepts (nouns) connected to them.

The HCT is the basis of our method, and our similarity computation is derived from cosine similarity, which is based on the orthogonality of its components, so the semantic coverage of the concept nodes should be independent. The limits of the semantic coverage of the child concept nodes are the partition (instead of covering) of the semantic coverage of their parent concept nodes. That is, the concepts subsumed by sibling concept nodes are usually non-overlapping; the relationship between two siblings is captured only through their ancestor concept nodes.