Information Processing and Management 54 (2018) 1002–1021

Contents lists available at ScienceDirect

Information Processing and Management

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

Computing semantic similarity based on novel models of semantic representation using Wikipedia

Rong Qu[a], Yongyi Fang[a], Wen Bai[b], Yuncheng Jiang[*,a]

[a] School of Computer Science, South China Normal University, Guangzhou 510631, China
[b] Collaborative Innovation Center of High Performance Computing, Sun Yat-Sen University, Guangzhou 510006, China

ARTICLE INFO

Keywords:
Semantic similarity
Concept similarity
Information content
Feature-based methods
Wikipedia

ABSTRACT

Computing Semantic Similarity (SS) between concepts is one of the most critical issues in many domains such as Natural Language Processing and Artificial Intelligence. Over the years, several SS measurement methods have been proposed by exploiting different knowledge resources. Wikipedia provides a large domain-independent encyclopedic repository and a semantic network for computing SS between concepts. Traditional feature-based measures rely on linear combinations of different properties with two main limitations, the insufficient information and the loss of semantic information. In this paper, we propose several hybrid SS measurement approaches by using the Information Content (IC) and features of concepts, which avoid the limitations introduced above. Considering integrating discrete properties into one component, we present two models of semantic representation, called CORM and CARM. Then, we compute SS based on these models and take the IC of categories as a supplement of SS measurement. The evaluation, based on several widely used benchmarks and a benchmark developed by ourselves, sustains the intuitions with respect to human judgments. In summary, our approaches are more efficient in determining SS between concepts and have a better human correlation than previous methods such as Word2Vec and NASARI.

1. Introduction

Semantic Similarity (SS) (Sanchez, Sol-Ribalta, Batet, & Serratosa, 2012) is a widely used technique in many areas including but not limited to Natural Language Processing (NLP) (Ahsaee, 2014; Jiang, Bai, Zhang, & Hu, 2017), Information Retrieval (Choumane, 2014), Knowledge Discovery (Mazandu, Chimusa, & Mulder, 2016), Artificial Intelligence (Liu, Bao, & Xu, 2012; Pirr, 2009) and Cognitive Science (Aouicha, Taieb, & Hamadou, 2016; Mazandu, Chimusa, Mbiyavanga, & Mulder, 2016; Meymandpour & Davis, 2016). SS can be used for weighting or ranking similar concepts based on a given ontology. It has recently received increased attention in different fields (Ahsaee, 2014; Batet, Sanchez, & Valls, 2011; Bollegala, Ishizuka, & Matsuo, 2015; Jiang et al., 2017; Jiang, Zhang, Tang, & Nie, 2015; Kang, 2012; Li, Bandar, & Mclean, 2003; Sanchiz et al., 2017; Taieb & Aouicha, 2014; Taieb, Aouicha, & Hamadou, 2013; Taieb, Aouicha, Tmar, & Hamadou, 2012) since proper assessment of concept similarity can improve the understanding of textual resources and increase the accuracy of knowledge based applications. SS would be more useful when applications need to encode hierarchical relations between concepts, such as concept expansion, question answering and concept-based retrieval.

Another related technique i.e., Semantic Relatedness (SR) considers all kinds of semantic relations between concepts.

* Corresponding author.
E-mail addresses: ycjiang@scnu.edu.cn, ycjiang21@qq.com (Y. Jiang).

https://doi.org/10.1016/j.ipm.2018.07.002
Received 24 December 2017; Received in revised form 22 May 2018; Accepted 6 July 2018

Aouicha et al. (2016) treat SR and SS as two different terms. SR is a more general term whereas SS is a special case of relatedness that relates to the resemblance among concepts. Moreover, SS measurement methods take two concepts as input and returns a numeric score that quantifies how much they are similar. This measure is usually based on *is-a* relations within the ontology in which the concepts reside. For example, *car* and *automobile* are similar while *car* and *wheel* are related but not similar.

In fact, several works about SS measures have been proposed in the past years. There are four different families listed as follows:

1. Path-based measures (Leacock & Chodorow, 1998; Rada, Mili, Bicknell, & Blettner, 1989; Wu & Palmer, 1994) assess similarities by the semantic distance of concepts in a given ontology. The measures of this type are widely used in many areas for its simple computation.
2. Information Content (IC) based measures (Jiang et al., 2017; Meng, Gu, & Zhou, 2012; Sanchez, Batet, & Isern, 2011; Taieb et al., 2013; Taieb et al., 2012) assess similarities by the IC of concepts in a given ontology. The notion of IC is based on an assumption that concrete and special entities present more IC than the general and abstract ones.
3. Feature-based measures (Jiang et al., 2015; Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006; Resnik & Philip, 1995; Rodriguez & Egenhofer, 2003; Tversky, 1977) assess similarities using the weighted sum of the common and non-common features of concepts.
4. Hybrid measures (Aouicha et al., 2016; Gao, Zhang, & Chen, 2015) combine at least two of above measures to merge the advantages, which lead to the dependency on parameters.

In recent years, extensive researches develop the use of SS/SR measures based on Knowledge Resources (KRs) such as Wikipedia, WordNet (Miller, 1995) and biomedical ontology Medical Subject Headings (MeSH) (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). As a free, online multilingual knowledge resource, Wikipedia is collaboratively maintained by volunteers and known to have a good coverage. Over the past few years, Wikipedia has persuaded many researchers to exploit the huge amounts of semi-structured knowledge available in such collaborative resources for different NLP applications (Aouicha et al., 2016; Ittoo & Bouma, 2013; Jiang et al., 2017; Kaptein & Kamps, 2013; Taieb et al., 2013; Taieb et al., 2012). Up to April 2018, the English version of Wikipedia contains 5,641,410 concepts and keeps steadily increasing. WordNet is a lexical database for English language, which divides words into the sets of synonyms called synsets, provides short and general definitions, and records the various semantic relations between these synsets. WordNet is particularly well suited for similarity measures since it organizes nouns and verbs into hierarchies of *is-a* relations (Hadj Taieb, Ben Aouicha, & Ben Hamadou, 2014). The measures based on WordNet have become mature for many years. However, the latest online version of WordNet is 3.1 and it contains 155,327 words organized in 175,979 synsets which are much fewer than Wikipedia concepts. Thus, Wikipedia draws attentions to a number of researchers (Jiang et al., 2015; Mehdi, Okoli, Mesgari, Nielsen, & Lanamki, 2017; Taieb et al., 2012; Zhao, Wu, Wang, & Li, 2016).

In Wikipedia, path-based measures and IC-based measures use category structure to assess SS. The category structure called Wikipedia Category Graph (WCG) contains about 1,500,000 concepts which are much fewer than the number of Wikipedia concepts. It is benefited from traditional properties such as synsets, glossaries and neighbors that feature-based measures are adopted universally to assess SS between concepts. However, feature-based measures rely on the available information of concepts while partial Wikipedia concepts are lack of useful features like descriptive texts. Moreover, for two concepts that are semantically similar but have few common features, or two concepts that are semantically dissimilar but have many common features, the SS between them are misjudged inevitably by traditional feature-based measures. Furthermore, the relations among various properties that reflect different aspects of a concept should be close, but they are likely to be ignored in most feature-based measures.

In order to overcome the problems that traditional feature-based measures have met, this paper gives novel models of semantic representation for concepts and proposes several SS measurement methods. We integrate the discrete properties of concepts into a single component for supplementing the missing information in properties and mining the latent semantic information among these properties. In detail, we firstly use the neighbors as an extra feature to enhance the representation of concepts. Secondly, we select some similar category pairs of two concepts and construct two subgraphs to obtain extra semantic features with the aid of IC-based measures. Thirdly, we integrate the features of both neighbors and categories into one concept representation. Finally, we propose several formulas to measure the similarity between two concepts based on the models of semantic representation. The main advantages of this work are summarized as below.

1. The missing information in a property can be completed using other related properties.
2. The latent semantic information among different properties is mined and used in SS measurement.

The remainder of this paper is organized as follows. Section 2 provides an overview on the main researches about the SS measures and briefly reviews Wikipedia. Section 3 proposes two models of representation based on two subgraphs and introduces our novel SS measures. Section 4 evaluates our models and discusses the experimental results. Finally, in Section 5, we draw our conclusion and present some perspectives for future research.

2. Related work

There are lots of SS measurement methods proposed in previous researches. In this section, we mainly focus on some IC-based and feature-based measures used in our experiments. Besides, we concern about our former works and some advanced computational models such as Word2Vec which relies on the use of word embedding.