Information Processing and Management 000 (2016) 1–18

Contents lists available at ScienceDirect

Information Processing and Management

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

Wikipedia-based information content and semantic similarity computation

Yuncheng Jiang*, Wen Bai, Xiaopei Zhang, Jiaojiao Hu

School of Computer Science, South China Normal University, Guangzhou 510631, China

ARTICLE INFO

Article history:
Received 26 July 2015
Revised 4 September 2016
Accepted 6 September 2016
Available online xxx

Keywords:
Information content
Semantic similarity
Concept similarity
Wikipedia
Category structure

ABSTRACT

The Information Content (IC) of a concept is a fundamental dimension in computational linguistics. It enables a better understanding of concept's semantics. In the past, several approaches to compute IC of a concept have been proposed. However, there are some limitations such as the facts of relying on corpora availability, manual tagging, or predefined ontologies and fitting non-dynamic domains in the existing methods. Wikipedia provides a very large domain-independent encyclopedic repository and semantic network for computing IC of concepts with more coverage than usual ontologies. In this paper, we propose some novel methods to IC computation of a concept to solve the shortcomings of existing approaches. The presented methods focus on the IC computation of a concept (i.e., Wikipedia category) drawn from the Wikipedia category structure. We propose several new IC-based measures to compute the semantic similarity between concepts. The evaluation, based on several widely used benchmarks and a benchmark developed in ourselves, sustains the intuitions with respect to human judgments. Overall, some methods proposed in this paper have a good human correlation and constitute some effective ways of determining IC values for concepts and semantic similarity between concepts.

1. Introduction

The Information Content (IC) of a concept is a fundamental dimension in computational linguistics. It states the amount of information provided by the concept when appearing in a context. In this manner, the basic idea is that general and abstract entities present less IC when found in a discourse than more concrete and specialized ones. A proper quantification of the IC of concepts improves text understanding by enabling assessing the degree of semantic generality or concreteness of words referring to these concepts (Sanchez, Batet, & Isern, 2011). Informally, IC is defined as a measure of the informativeness of concepts and computed by counting the occurrence of words in large corpora (Pirro, 2009). That is, IC measures the amount of information provided by a given term based on its probability of appearance in a corpus. Up to the present, research on IC has been very active and many results have been achieved in the theoretical and application aspects. In particular, IC has been applied to the computation of semantic similarity, which acts as a fundamental principle by which humans organize and classify objects (Batet, Sanchez, & Valls, 2011; Buggenhout & Ceusters, 2005; Formica, 2008; Jiang & Conrath, 1997; Lin, 1998; Pirro, 2009; Resnik, 1999; Resnik, 1995; Sanchez & Batet, 2013; Sanchez & Batet, 2011; Sanchez, Batet, & Isern, 2011).

* Corresponding author. Fax. +0862085215418.
E-mail addresses: ycjiang@scnu.edu.cn, ycjiang21@qq.com (Y. Jiang).

http://dx.doi.org/10.1016/j.ipm.2016.09.001
0306-4573/

Making judgments about the semantic similarity of different concepts is a routine yet deceptively complex task. To perform it, people need to draw on an immense amount of background knowledge about the concepts. As a result, any attempt to compute semantic similarity automatically must also consult external sources of knowledge. Usually, these resources can be search engines (Bollegala, Matsuo, & Ishizuka, 2007; Cilibrasi & Vitanyi, 2007; Martinez-Gil & Aldana-Montes, 2013), topical directories such as Open Directory Project (Maguitman, Menczer, Erdinc, Roinestad, & Vespignani, 2006), well-defined semantic networks such as WordNet (Ahsaee, Naghibzadeh, & Naeini, 2014; Liu, Bao, & Xu, 2012), or more domain-dependent ontologies such as Gene Ontology (Couto, Silva, & Coutinho, 2007; Mathur & Dinakarpandian, 2012) and biomedical ontologies MeSH or SNOMED CT (Batet, Sanchez, Valls, & Gibert, 2013; Pedersen, Pakhomov, Patwardhan, & Chute, 2007; Sanchez & Batet, 2011). In fact, several works about semantic similarity measures with external sources of knowledge have been developed in the past years. According to the concrete knowledge sources exploited and the way in which they are used, different families of methods can be identified (Sanchez & Batet, 2013; Sanchez et al., 2011). These families are (Martinez-Gil, 2014; Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006): (1) edge counting measures: which consist of taking into account the length of the path linking the concepts (or terms) and the position of the concepts (or terms) in a given dictionary (or taxonomy, ontology) (Li, Bandar, & McLean, 2003); (2) feature based measures: which consist of measuring the similarity between concepts (or terms) as a function of their properties or based on their relationships to other similar concepts (or terms) (Petrakis et al., 2006; Rodriguez & Egenhofer, 2003; Sanchez, Batet, Isern, & Valls, 2012); (3) information content measures: which consist of measuring the difference of the information content of the two concepts (or terms) as a function of their probability of occurrence in a text corpus (or an ontology) (Buggenhout & Ceusters, 2005; Lin, 1998; Resnik, 1999; Resnik, 1995; Sanchez & Batet, 2013; Sanchez, Batet, Valls, & Gibert, 2010); (4) hybrid measures: which consist of combining all of the above (Batet et al., 2013; Pirro, 2009).

Information theoretic approaches (i.e., IC-based approaches) assess the similarity between two concepts as a function of the IC that both concepts have in common in a given ontology. In the past, IC was typically computed from concept distribution in tagged textual corpora (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995). However, this introduces a dependency on corpora availability and manual tagging that hampered their accuracy and applicability due to data sparseness (Sanchez et al., 2010). To overcome this problem, in recent years several researchers have proposed various ways to infer IC of concepts in an intrinsic manner from the knowledge structure modeled in an ontology (Sanchez & Batet, 2013; Sanchez & Batet, 2011; Sanchez et al., 2011). From a domain-independent point of view, these approaches provide accurate results when relying on large and general-purpose knowledge sources such as biomedical ontologies MeSH or SNOMED CT (Batet et al., 2013; Pedersen et al., 2007; Sanchez & Batet, 2011) and tagged corpora such as SemCor (Fellbaum, 1998a, b) or Brown Corpus (Francis & Kucera, 1982). However, there are still some limitations in these methods of ontology based IC computation. The fact that intrinsic IC-based measures only rely on ontological knowledge is a drawback because they completely depend on the degree of coverage and detail of the unique input ontology (Sanchez & Batet, 2013). Especially, with the emergence of social networks or instant messaging systems (Martinez-Gil & Aldana-Montes, 2013; Retzer, Yoong, & Hooper, 2012), many (sets of) concepts or terms (proper nouns, brands, acronyms, new words, conversational words, technical terms and so on) are not included in domain ontologies. Therefore, IC computation that is based on these kinds of knowledge resources (i.e., domain ontologies) cannot be used in these tasks. On the other hand, the prerequisite of ontology based IC computation is the existence of a (or several) predefined domain ontology (or ontologies). Such ontologies are established by panel experts in the given domains. Clearly, the construction process of these domain ontologies is time-consuming and error-prone and maintaining these ontologies also requires a lot of effort from experts. Thus, the methods of ontology based IC computation are also limited in scope and scalability. These limitations are the motivation behind the new techniques presented in this paper which compute IC of a concept from a kind of new source of information, i.e., a wide coverage online encyclopedia, namely Wikipedia (Hovy, Navigli, & Ponzetto, 2013; Medelyan, Milne, Legg, & Witten, 2009). As everyone knows, Wikipedia was launched in 2001 with the goal of building free encyclopedia in all languages. Today it is the largest, most widely used, and fastest growing encyclopedia in existence (Medelyan et al., 2009).

The purpose of this paper is to present several new methods to IC computation of a concept and similarity computation between two concepts to solve the shortcomings of existing approaches for IC computation and semantic similarity computation. The presented paper focuses on the IC computation of a concept and similarity computation between two concepts drawn from the Wikipedia category structure. In other words, we approach the problems of IC computation and similarity computation of concepts from a novel perspective by making use of Wikipedia. Thus, the terminologies of Wikipedia categories, categories, concepts, and words can be used interchangeably. In this paper, we utilize the Wikipedia category structure to act as a knowledge source. It is well known that the Wikipedia category structure is a very complex network. Comparing with traditional taxonomy structures, the Wikipedia category structure is a graph structure. Faced with such a complex structure, how do we compute the semantic similarity between concepts (categories)? Since the Wikipedia category structure is a graph, naturally, we may assess the semantic similarity between concepts by extending traditional information theoretic approaches (i.e., IC based approaches). The first thing we need to do is to compute the IC value of a concept (category) in a graph. Because the Wikipedia category structure is too complex, we are not sure which computation method for IC of a concept is most appropriate. Therefore, according to the characteristics of the Wikipedia category structure we have to present several IC computation approaches by extending traditional methods. Based on these IC computation approaches, we need to propose an approach for semantic similarity computation. How to give the method of similarity computation? Clearly, we may generalize existing approaches to the similarity measure for Wikipedia categories. It is noted that in traditional IC based methods, the key issue is to find the LCS (Least Common Subsumer) of two concepts. However, the Wikipedia

Please cite this article as: Y. Jiang et al., Wikipedia-based information content and semantic similarity computation, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.09.001