Engineering Applications of Artificial Intelligence 65 (2017) 268–281

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A graph-based semantic relatedness assessment method combining wikipedia features

Pu Li [a,*], Bao Xiao [b], Wenjun Ma [c], Yuncheng Jiang [c,*], Zhifeng Zhang [a]

[a] *Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450000, China*
[b] *School of Electronics and Information Engineering, Qinzhou University, Qinzhou 535000, China*
[c] *School of Computer Science, South China Normal University, Guangzhou 510631, China*

ARTICLE INFO

Keywords:
Semantic relatedness
Wikipedia features
Graph-based relatedness
Naive-ESA algorithm

ABSTRACT

Semantic relatedness assessment between concepts is a critical issue in many domains such as artificial intelligence, information retrieval, psychology, biology, linguistics and cognitive science. Therefore, several methods assess relatedness by exploiting knowledge bases to express the semantics of concepts. However, there are some limitations such as high-dimensional space, high-computational complexity, fitting non-dynamic domains. Considering that Wikipedia, a domain-independent encyclopedic repository, which provides very large coverage, has been exploited by many methods as a huge semantic resource. In this paper, we propose a novel graph-based relatedness assessment method using Wikipedia features to avoid some of the limitations and drawbacks mentioned above. Firstly, for each term in a word pair, the top *k* most relevant Wikipedia concepts are returned by the Naive-ESA algorithm to reduce the dimensional space of Explicit Semantic Analysis (ESA) method. Secondly, for each different candidate concept in two relevant concept sets, we collect its categories set from the Wikipedia Category Graph (WCG). Based on the categories in WCG network, the relatedness between concepts at the correspondence position of the two sorted concept sets is computed as the association coefficient. Thirdly, based on this parameter, a novel relatedness assessment metric is presented. The evaluation is performed on some datasets well-recognized as benchmarks, using several widely used metrics and a new metric designed by ourselves. The result demonstrates that our method has a better correlation with the intuitions of human judgments than other related works.

1. Introduction

Semantic relatedness between concepts is considered as an important problem for many tasks in Natural Language Processing (NLP) such as automatic detection and correction of spelling errors (Budanitsky and Hirst, 2006), word sense disambiguation (Han and Zhao, 2010; Leacock and Chodorow, 1998), semantic annotation (Sanchez et al., 2011b), information retrieval (Baziz et al., 2005; Finkelstein et al., 2002; Formica, 2008; Gurevych et al., 2007; Tapeh and Rahgozar, 2008), and knowledge acquisition (Liu et al., 2012). Research in semantic relatedness has increased the accuracy of knowledge representation especially for the massive online data. The assessment of semantic relatedness can improve the understanding of data resources and has been widely used in knowledge based applications, such as search (Fellbaum and Miller, 1998) as well as document categorization or clustering (Batet, 2011; Luo et al., 2011).

Considering the significance of semantic relatedness assessment, many research results have been proposed recently. Among these studies, Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) method created by Gabrilovich and Markovitch shows good performance on the correlation with human judgments for both words and text fragments. This well known method has been extensively studied in many applications (Cimiano et al., 2009; Egozi et al., 2011; Müller and Gurevych, 2010; Potthast et al., 2008; Sorg and Cimiano, 2008).

However, ESA requires distilling the background knowledge provided by Wikipedia (or Wiktionary) (Hovy et al., 2013; Medelyan et al., 2009) into a large matrix, where terms are associated with a weighted list of the articles in which they appear during its calculation process for semantic relatedness. It means that with the expanding scale of data in Wikipedia, this matrix for inverted index may contain millions of columns with many 0-value elements. Therefore, when computing

[*] Corresponding authors.
E-mail addresses: superlipu@163.com (P. Li), ycjiang@scnu.edu.cn (Y. Jiang).

http://dx.doi.org/10.1016/j.engappai.2017.07.027
Received 21 September 2016; Received in revised form 7 July 2017; Accepted 30 July 2017
Available online 30 August 2017
0952-1976/© 2017 Elsevier Ltd. All rights reserved.

semantic relatedness using Cosine formula (Salton and Buckley, 1988) over this huge sparse matrices, the efficiency of ESA will obviously slow down and many worthless overheads are incurred on the computing over those 0-value elements.

In this paper, we present a new graph-based relatedness assessment method in order to improve the efficiency and performances of ESA method. In other words, we approach the problem of feature based semantic relatedness between words with a novel perspective by making use of Wikipedia Category Graph (WCG). As we know, each page in Wikipedia has its own title and the corresponding article. So, in many related studies, such as Gabrilovich and Markovitch (2007), Hadj Taieb et al. (2014), Jiang et al. (2015), researchers call the titles of Wikipedia articles as ‘‘Wikipedia concepts’’. Since each Wikipedia concept has its own categories in WCG network, the method to relatedness assessment presented in this paper can compute the relatedness between two concepts using their category features in Wikipedia. In this way, for a given word, our method allows users to only focus on its most relevant concepts in Wikipedia and discard the less relevant ones. So this new graph-based relatedness assessment method can reduce the dimensional space and improve the computational efficiency to some extent. Moreover, since Wikipedia is a rich encyclopedia (or corpus, thesaurus, network structure) that covers almost all imaginable sources, the method combining Wikipedia features presented in this paper can process lots of terms and has a good performance in dynamic domains.

The remainder of this paper is organized as follows. Section 2 briefly reviews the state of the art in computing semantic relatedness using different computing strategies and different knowledge resources (e.g. WordNet (Miller, 1995) or Wikipedia). In Section 3 we design a Naive-ESA algorithm which returns the top *k* most relevant Wikipedia concepts as the concept vector for each term in word pair. Then, we use this algorithm to generate the intersection set and difference list of these two concept vectors. By combining the category features in Wikipedia, a new graph-based semantic relatedness assessment method is presented in Section 4. Section 5 details the experiments that evaluate the effectiveness of our method and reports the analysis of results. Finally, we remark our conclusion and present some perspectives for future research in Section 6.

2. Related work

Several semantic relatedness assessment methods have been proposed in the last years. In this section, we present the main studies.

2.1. Using different computing strategies

From the view of the computing strategies of semantic relatedness, the studies can be mainly divided into the following four families: (1) structure-based methods, (2) Information Content (IC) based methods; (3) feature based methods; (4) hybrid methods.

Structure-based methods base the relatedness assessment on 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). Wikirelate! (Strube and Ponzetto, 2006) and its successor, Wikitaxonomy (Ponzetto and Strube, 2011) which is no longer just an assessment method but also has been developed into an ontological and taxonomic resources, created by Strube and Ponzetto are based on category structure. They derive a taxonomy from the system of categories in Wikipedia. Considering that the categories of Wikipedia form a graph which can be taken to represent a conceptual network with unspecified semantic relations (Ponzetto and Strube, 2007; Strube and Ponzetto, 2006), Wikitaxonomy transforms a graph with unlabeled semantic relations into a semantic network where the links between categories are augmented with *isa* relations. Then, for a given word pair, Wikitaxonomy runs through the category tree to extract the categories that each term belongs to and compute relatedness based on the number of links of the paths connecting categories in the graphical network. Wikipedia Miner (Milne and Witten, 2013) derived from Wikipedia Link Vector Model (WLVM) (Milne, 2007) designed by Milne and Witten is another useful toolkit that contains summarized versions of Wikipedia’s content and structure. It extracts the value of semantic relatedness for word pairs from the Wikipedia’s link structure and takes both the count of page-links and their directions (*pageLinksIn* and *pageLinksOut*) into consideration. Another method, named as WikiWalks, is proposed by Yeh et al. (Yeh et al., 2009). With the strategy of random walks based on Personalized PageRank (Haveliwala, 2003) with the teleport vector being provided from the ESA, this method returns the relatedness assessment employing the link structure of Wikipedia.

The structure-based methods require a low computational cost, due to their simplicity, however, these methods offer a limited accuracy (Batet et al., 2011; Sanchez and Batet, 2013). To acknowledge some of the limitations of structure-based methods, IC based methods quantify the similarity between two concepts as a function of the information content in a given ontology. There are two primary branches: for the researches in Jiang and Conrath (1997), Lin (1998), Resnik (1995), IC was typically computed from concept distribution in tagged or non annotated textual corpora, while others infer IC of concepts from the topological parameters in an ontology structure (Hadj Taieb et al., 2014; Meng et al., 2012; Sanchez and Batet, 2013; Sanchez et al., 2011a; Sebti and Barfroush, 2008). However, the IC based methods only rely on ontological knowledge. This is also a drawback of these methods because they completely depend on the degree of coverage and detail of the unique input ontology (Sanchez and Batet, 2013).

To further improve the accuracy and versatility, feature based methods are addressed by considering the degree of overlapping between sets of ontological features (Sanchez et al., 2012). As a well-known research, ESA, which is mainly based on text features and links within articles, performs quite well on the correlation with human judgments and is fit for both words and text fragments. Moreover, Jiang et al. presented a Wikipedia feature-based method to evaluate the semantic similarity between concepts (Jiang et al., 2015), while some other methods for computing semantic relatedness using WordNet features (HadjTaieb et al., 2014) or Wikipedia features (Hadj Taieb et al., 2013) are studied by Hadj Taieb, Ben Aouicha, and Ben Hamadou.

Besides, several methods combine some of aforementioned strategies like (Ben Aouicha et al., 2016b) which integrate taxonomy-based IC and WordNet–Wiktionary–Wikipedia glosses together for semantic relatedness.

2.2. Using different knowledge resources

From the view of the knowledge resources, semantic relatedness assessment methods can be grouped into three types: (1) WordNet-based methods, (2) Wikipedia-based methods, (3) Domain dependent ontologies-based methods.

Benefitting from the logical hierarchy and high-quality knowledge representation of WordNet conceived and operated by experts, some researchers employ this dictionary as background knowledge resource to conduct their studies on semantic relatedness (Hadj Taieb et al., 2014), text clustering (Wei et al., 2015) and so on.

Questions arise about the need for a larger coverage with the advent of Big Data Era. As already cited, Wikipedia is a free, multilingual, wide coverage online encyclopedia that is collaboratively maintained by volunteers (Ponzetto and Strube, 2007). So, recently many researchers like Ponzetto and Strube (Ponzetto and Strube, 2011; Strube and Ponzetto, 2006), Gabrilovich and Markovitch (Gabrilovich and Markovitch, 2007, 2009), Zesch, Muller and Zesch et al. (2008), Hadj Taieb, Ben Aouicha, and Ben Hamadou (Hadj Taieb et al., 2013), Yazdani and Popescu-Belis (Yazdani and Popescu-Belis, 2013) have worked on the assessment of semantic relatedness by applying Wikipedia or Wiktionary.

To pursue higher accuracy in a specialized field, some semantic relatedness or similarity methods are designed based on the domain dependent (e.g. biomedical) ontologies, such as Gene Ontology (Couto