Journal of Visual Languages and Computing 31 (2015) 246–251

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

Semantic Similarity Assessment Using Differential Evolution Algorithm in Continuous Vector Space ☆

Wei Lu, Yuanyuan Cai *, Xiaoping Che, Kailun Shi

School of Software Engineering, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Available online 3 November 2015

Keywords:
Differential evolutionary
Semantic similarity
Continuous vector space
Vector similarity metrics
WordNet

ABSTRACT

The assessment of semantic similarity between terms is one of the challenging tasks in knowledge-based applications, such as multimedia retrieval, automatic service discovery and emotion mining. By means of similarity estimation, the comprehension of textual resources can become more feasible and accurate. Some studies have proposed the integration of various assessment methods for taking advantage of different semantic resources, but most of them simply employ average operation or regression training. In this paper, we address this problem by combining the corpus-based similarity methods with the WordNet-based methods based on a differential evolution (DE) algorithm. Specifically, this DE-based approach conducts similarity assessment in a continuous vector space. It is validated against a variety of similarity approaches on multiple benchmark datasets. Empirical results demonstrate that our approach outperforms existing works and more conforms to the human judgement of similarity. The results also prove the expressiveness of continuous vectors learned from neural network on latent lexical semantics.

1. Introduction

Textual semantic similarity assessment, which measures the degree of likeness among terms or documents, enables the semantic-based applications, such as automated service discovery [17] and semantic data mining [28]. For example, a user who queries the *bank* service may obtain the services whose profile attributes consist of *deposit* and *interests*. Throughout the years, massive data and heterogeneous resources distributed on the web have made the semantic extraction and computation more challenging. Accordingly, improving the accuracy of semantic similarity measurement is still an attractive research work.

Many approaches to measuring semantic similarity have been developed, which can be divided into corpus-based and knowledge-based approaches in terms of the semantic resources available. Corpus-based approaches primarily map a given corpus into a vector space [27]. The words close together in the vector space tend to be semantically similar or at least occur in similar contexts. However, most of the corpus-based approaches are limited by the distributional vector space model (VSM) where a corpus is modeled as “bag of words”. The bag-of-words model only scratches the surface of words but neglects sufficient semantic association of words. As a result, the distributional VSMs often suffer from data sparsity problem since the large size of corpus. Even though there are many dimension reduction techniques can address this problem, they could hardly decode implied semantics and syntactic information of context from corpus into vector spaces.

On the other hand, knowledge-based approaches take advantage of structured knowledge bases to measure semantic similarity, such as thesauri and ontology.

☆ This paper has been recommended for acceptance by Henry Duh.

* Corresponding author.
E-mail addresses: luwei@bjtu.edu.cn (W. Lu), yycai@bjtu.edu.cn (Y. Cai), xpche@bjtu.edu.cn (X. Che), shikailun@bjtu.edu.cn (K. Shi).

http://dx.doi.org/10.1016/j.jvlc.2015.10.015
1045-926X/

WordNet [14] is a general-purpose ontology commonly used for similarity calculation. WordNet-based measures can be roughly classified into path-based, information content (IC)-based, feature-based and hybrid measures. The path-based measures and the IC-based measures mainly exploit the path distance and IC difference between concepts in WordNet, while the feature-based measures [4,18] rely on estimate the vector similarity by means of representing the intrinsic attributes as conceptual vectors. As an instance, Liu et al. used local densities to construct conceptual vectors and evaluated the cosine similarity of vectors [12]. Besides, hybrid approaches [10,20] mainly combine multiple computing factors derived from different kinds of WordNet-based approaches and thus have excellent performances. However, these WordNet-based approaches are limited by the coverage of the knowledge base so that rare words are often poorly estimated. Besides, the feature-based approaches are considered more suitable for measuring semantic relatedness rather than semantic similarity.

In this work, aiming to integrate the semantics from corpus and WordNet, we find the optimal weighting sum of different vector-based methods using a differential evolutionary (DE) algorithm. In addition to the vector similarity of features supplied by WordNet, we integrate various similarities of the lexical vectors learned from corpus to capture varying degrees and aspects of semantic similarity. E.g., the cosine distance can determinate the angle distance between two vectors (directional similarity) in the vector space, whereas the Euclidean distance evaluates straight-line distance between two vectors (magnitude similarity). Specially, since the vector-based similarity approaches highly depend on the quality of vectors, our DE-based similarity assessment uses the continuous distributed word representation derived from deep learning technologies.

The rest of this paper is organized as follows: the related works are presented in Section 2. Section 3 presents the vector-based similarity metrics used in the DE algorithm. The methodology and experimental result analysis of this paper are given in Section 4. Conclusions and future work are summarized in Section 5.

2. Related works

As an alternative of one-hot vector, the continuous word representation has significantly benefited the vector-based semantic similarity measurement recently [26]. Continuous word representation, namely distributed word embedding, is a real-valued and low-dimension vector [13]. The latent semantics of terms are encoded into the vectors via unsupervised neural network training, which can better understand and express the lexical resources. On the basis of distributed VSMs, the powerful expressiveness of continuous vectors has been validated in many natural language processing tasks, such as semantic disambiguation and analogy reasoning.

In addition to the quality of vectors, similarity metric is an important part of vector-based similarity assessment. Most of the previous similarity measures are performed with a single computational metric such as cosine distance, vector overlaps or Euclidean distance [1]. Faruqui and Dyer evaluated the concept similarity based on the cosine similarity of continuous word embeddings [8]. Pennington et al. learned distributed vectors from supervised global log-bilinear regression model with matrix factorization, and took the cosine value of the vectors as concept similarity [19]. However, a single metric merely provides a certain degree of semantic similarity and fails to suit all types of input data. Besides, the cosine similarity of vectors is commonly used whereas other vector-based similarity metrics are less applied. To capture different aspects of semantic similarity between concepts, a variety of studies focus on the integration of different computational metrics. Yih and Qazvinian incorporated different vector measurements based on the heterogeneous lexical sources such as Wikipedia, web search engine, thesaurus and WordNet [25]. Alves et al. proposed a regression function where lexical similarity, syntactic similarity, semantic similarity and distributional similarity are input as independent variables [2]. Similarly, Bär et al. introduced a linear regression model integrating multiple content similarity values at the aspects of string, semantic, structure [5]. In the study of [6], WordNet-based semantic similarity measures are combined by means of a meta-heuristic algorithm. Kiela and Clark studied the effects of different computational metrics on semantic similarity estimation, as well as data source, dimensionality reduction strategy, term weighting scheme and feature granularity of vectors [11]. However, the related researches mainly focus on the average or regression of different similarity approaches.

Differing from previous studies, we integrate different vector-based similarity metrics based on the unsupervised differential evolutionary (DE) algorithm. The algorithm of DE [24] is a population-based stochastic search strategy for solving global optimization problems. It derives from evolutionary algorithm and has multiple variants according to the strategy for the generation of new candidate members, fitness computation, mutation, etc. [7,16]. These variants have been proved applicable for continuous function optimization in a large number of research domains such as heat transfer [3]. Our study aims to assess the semantic similarity between concepts in continuous vector space and optimize the measurement by combining multiple vector similarities derived from different semantic resources.

3. Semantic similarity measurement based on differential evolution algorithm

In this section, we describe the proposed DE-based approach which incorporates heterogeneous vector similarity metrics. The basic idea is to use a DE algorithm for an optimal semantic similarity assessment based on different results of similarity. Fig. 1 illustrates the DE algorithm in our work. It performs with the similarity values provided by various vector-based metrics. All the metrics evenly contribute to evaluate the degree of semantic similarity between two concepts at the beginning of the differential evolution. Then each metric is iteratively assigned a specific weight. Finally, the metric that provides the most similar results to the human judgement is offered the highest weight after