Incorporating LDA With Word Embedding for Web Service Clustering

Yi Zhao, School of Computer Science, Wuhan University, Wuhan, China
Chong Wang, School of Computer Science, Wuhan University, Wuhan, China
Jian Wang, School of Computer Science, Wuhan University, Wuhan, China
Keqing He, School of Computer Science, Wuhan University, Wuhan, China

ABSTRACT

With the rapid growth of web services on the internet, web service discovery has become a hot topic in services computing. Faced with the heterogeneous and unstructured service descriptions, many service clustering approaches have been proposed to promote web service discovery, and many other approaches leveraged auxiliary features to enhance the classical LDA model to achieve better clustering performance. However, these extended LDA approaches still have limitations in processing data sparsity and noise words. This article proposes a novel web service clustering approach by incorporating LDA with word embedding, which leverages relevant words obtained based on word embedding to improve the performance of web service clustering. Especially, the semantically relevant words of service keywords by Word2vec were used to train the word embeddings and then incorporated into the LDA training process. Finally, experiments conducted on a real-world dataset published on ProgrammableWeb show that the authors’ proposed approach can achieve better clustering performance than several classical approaches.

KEYWORDS

Latent Dirichlet Allocation, Web Service Clustering, Word Embedding, Word2vec

1. INTRODUCTION

With the rapid development of SOA (Service Oriented Architecture), the number of available Web service resources on the Internet is increasing rapidly. As of 30 Dec. 2017, for example, over 18,000 Web services have been published on ProgrammableWeb1 (PW), one of the most popular service registries. These Web services follow various protocols, including SOAP (Simple Object Access Protocol) (Li, 2008), XML-RPC (Cerami, 2002) (XML Remote Procedure Call), REST (Wagh, 2012) (Representational State Transfer), and so on. Meanwhile, they have diverse service description formats, such as WSDL (Web Service Description Language), WADL (Web Application Description Language), and natural language text. Currently, almost all the Web services in PW are described in unstructured short texts. These heterogeneous and unstructured service descriptions bring many difficulties in service discovery. Therefore, how to accurately discover appropriate Web services for users becomes an important issue in services computing.

DOI: 10.4018/IJWSR.2018100102

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Web service search engines or online Web service directories are the major sources of service discovery. However, these search engines based on keyword matching usually suffer from using synonyms or variations of predefined keywords, and thus lead to returning inaccurate results (Al-Masri, 2008). Towards this problem, many semantic Web service discovery approaches (e.g. Mier, 2016) have been proposed to improve service discovery by semantically annotating attributes of Web services with domain ontologies. However, this is still a time-consuming task, which makes it difficult to apply these approaches in practice. In recent years, Web service clustering has shown its advantages in improving the performance of Web service discovery (Chen, 2011; Richi, 2007), by grouping Web service descriptions into clusters based on their functionalities. Among the existing service clustering approaches, LDA (Latent Dirichlet Allocation) (Blei, 2003) is the most widely adopted model, since it can be used to extract unobserved groups that explain why some parts of the service descriptions are similar and capture the underlying domain semantics. However, the word distributions in Web service descriptions are usually very sparse, so that the latent topics learned by LDA and other topic models are still inaccurate. As a result, this may lead to unsatisfactory clustering performance.

Actually, many efforts have been made to enhance the capability of LDA models in service clustering, such as label assisted service clustering and word embedding service clustering. These approaches leveraged auxiliary features to enhance the capability of classical LDA models and then achieve better clustering performance. However, these extended approaches still have some limitations in processing data sparsity and noise words. For example, Shi et al. (Shi, 2017) presented an augmented LDA model by leveraging word vectors to improve Web service clustering. They firstly grouped service descriptions into word clusters using K-means++ (Arthur, 2007), and then incorporated these word clusters to semi-supervise the LDA training process. More specifically, they leveraged the topic distributions of a word’s friend words (the words within the same word cluster) to refine those of this word during the training process. Since some words within the same word cluster may be irrelevant, these noise words may hinder the training process of LDA. In addition, these approaches paid more attention to the co-occurrence relations among words or the latent semantics, and neglected the explicit semantics of words. This may hinder these approaches from achieving better clustering performances. Therefore, how to further improve service clustering performance remains a hot topic in services computing.

In this paper, we propose a novel word embedding augmented LDA model. Our approach leveraged the high-quality semantically relevant words generated by Word2vec to improve the performance of Web service clustering. More specifically, the descriptions of Web services are firstly preprocessed by performing tokenization, stemming, and stop word removal, and then trained by the Word2vec algorithm. After that, semantically relevant words of a number of selected keywords are extracted from the word vectors produced by Word2vec. Each Web service description document can then be enriched by incorporating the synonyms of its keywords. Finally, the enriched Web service documents are modeled by applying the LDA algorithm. The contributions of this paper are as follows:

1. We proposed a novel word embedding augmented LDA model for service clustering. The proposed model can leverage the high-quality semantically relevant words for service keywords generated by Word2vec to improve the performance of Web service clustering;
2. We conducted a series of experiments on a real-world data set to evaluate our approach. The experimental results show that the proposed approach can achieve better clustering performance than several existing approaches.

The remainder of the paper is organized as follows. Section 2 discusses the related work. Section 3 provides the overall framework of our proposed model and the details of our approach for service clustering. Section 4 shows the experiments and the results. The conclusion and future work are summarized in Section 5.
End-to-End Web Service Recommendations by Extending Collaborative Topic Regression
Bing Bai, Yushun Fan, Wei Tan, Jia Zhang, Keman Huang and Jing Bi (2018).
International Journal of Web Services Research (pp. 89-112).
www.igi-global.com/article/end-to-end-web-service-recommendations-by-extending-collaborative-topic-regression/193863?camid=4v1a

Describing and Selecting Collections of Georeferenced Media Items in Peer-to-Peer Information Retrieval Systems
Daniel Blank and Andreas Henrich (2012). Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications (pp. 1-20).
www.igi-global.com/chapter/describing-selecting-collections-georeferenced-media/65106?camid=4v1a