LDA Meets Word2Vec: A Novel Model for measuring technology similarity

INTRODUCTION

Technology similarity is an important basis for identifying potential technology competition and partners among subjects. Technical similarity needs to be identified through the similarity of R&D results. Patent information is the most important embodiment of global R&D achievements (Cantwell, et al.,1999). Therefore, how to use patent information to analyze the technical similarity between patentee is particularly critical. In previous studies, scholars have applied text mining for research, mainly include keyword analysis, Co-word analysis method(Bergmann, et al.,2003), LDA topic model(Rus, et al.,2013), etc.

With the rapid development of technology, traditional topic classifications are too general, short of timeliness and scientific. As a result, relevant frontier areas and technical topics cannot be meticulously represented. Further, the accuracy of technical similarity measurement is affected. Word2Vec is a wordembedding model to predict a target word from its surrounding contextual words(Wang, et al.,2016). It makes up for the shortcomings of LDA topic model.Therefore, this paper propose a new method by using both Word2vec and LDA to extract the technology topics of patentee in a semantic space and combines with analytical method of multilayer complex networks. To investigate its performance, we compared with LDA methods.

METHODOLOGY

We design the research framework including three steps and show it in **Figure1**:

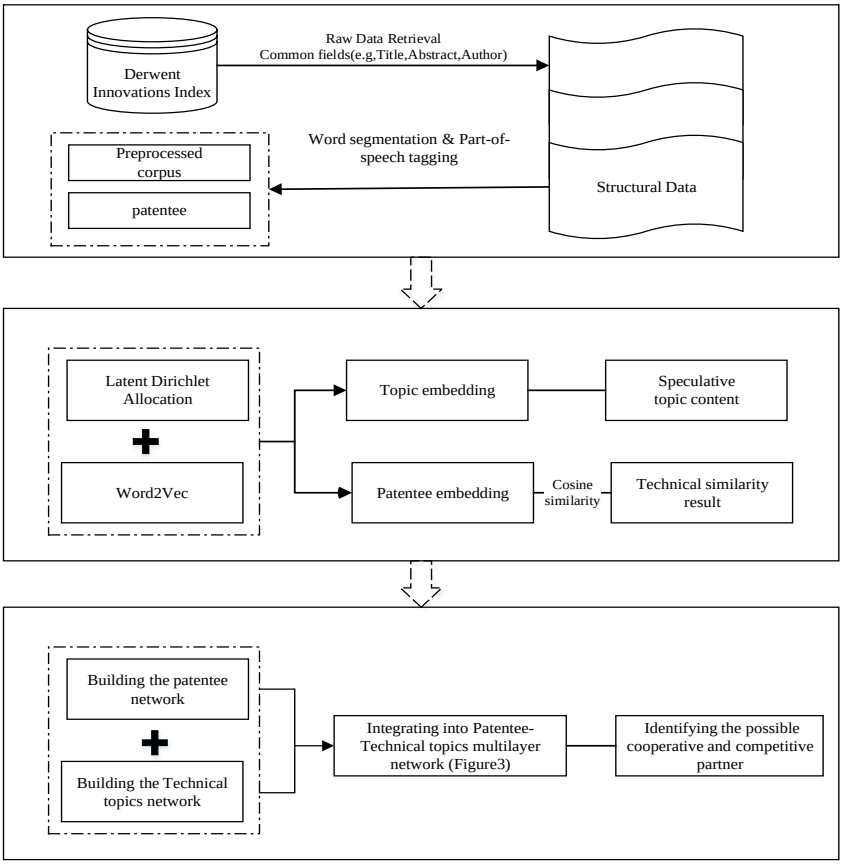

Fig.1 The research framework

1)Step1: The patent data was retrieved and downloaded through the DII database, and text preprocessing mainly including the cleaning of Assignee/Applicant and the Abstract-DWPI and Title fields.

2)Step2: Our new method as shown in **Fig.2** projects topic, documents and patentee in a high-dimension semantic space.

A document vector is considered as a single vector, which is the centroid of all words in the document as what Word2Vec does in the projection layer. In addition, each document has its individual length, thus its vector is divided by the number of words in the document to guarantee the measurements with same scale. Considering the mapping relationship between the patentee and the document, A patentee vector is considered to be the sum of the document vectors to which it belongs, divided by the number of documents. we measure Cosine distances from each patentee to identify potential competitors and partners.

We construct topic vectors in a similar way, but it is a little more complicated. A subset of high-probability words in each topic is employed to represent the topic, and then their probabilities are rescaled as the weights of words. Hence different words have different contributions to the topic. We measure Cosine distances from each patentee to topics to represent technology layout of the patentee.

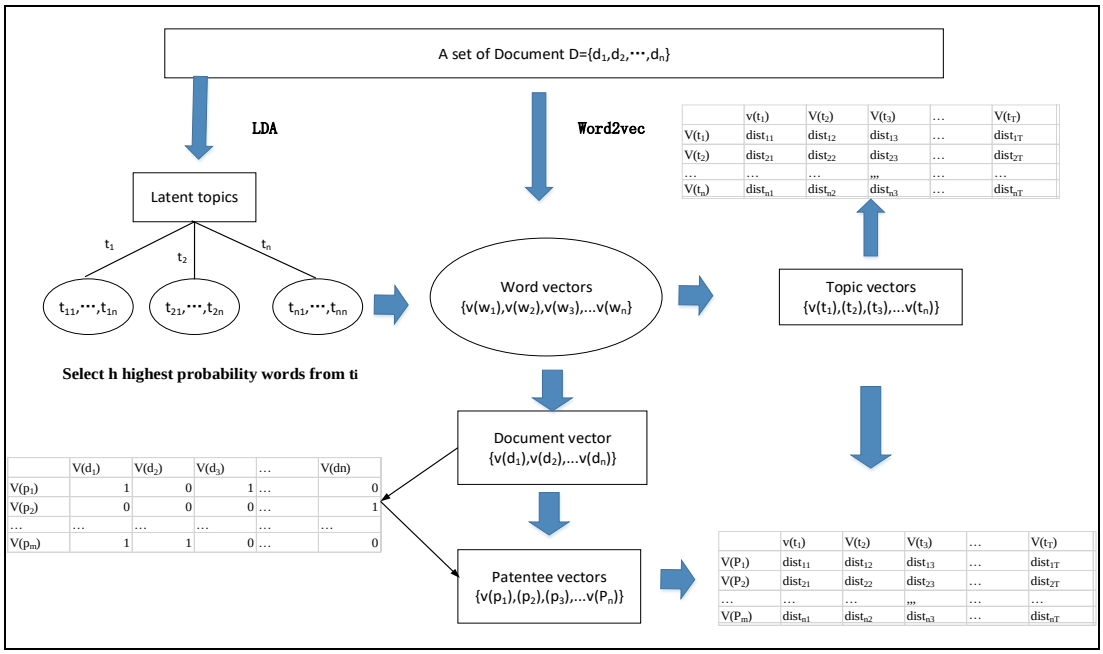

Fig.2 processes and results of our hybrid method

3)Step3: We apply the approaches of multilayer complex networks to present the situation of Patentee macroscopically. The model of Patentee - Technical topics multilayer network is shown in **Figure3**.

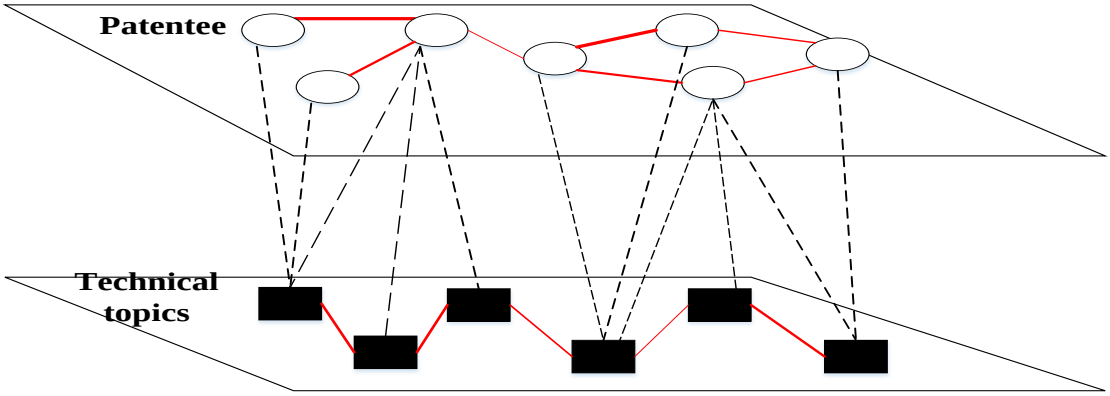

Fig.3 The model of Patentee - Technical topics multilayer network

EXPERIMENT RESULTS

NEDD has shown great potential in the treatment of cancer and cerebral-nerve diseases, which has attracted wide attention from the biomedical field. Meanwhile, it is a key field concerned by the “Made in China 2025”. Thus, we use the NEDD as a case study. The estimated results are that: 1) Our new hybrid method performs a better job of measuring the technical similarity of the patentee than LDA topic model. In addtion ,the research of this paper hopes to provide reference for the research methods of technical similarity.2) The research results not only showcase the technical layout of the field, but also provide a reference for enterprises to identifying the company's competitors and partners and university partners.

REFERENCE

Bergmann I., Butzke D., Walter L., et al.(2008). Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis: the Case of DNA Chips. R&D Management, 5: 550-562.

Cantwell John., Fai Felicia.(1999). Firms as the source of innovation and growth: the evolution of technological competence. Journal of Evolutionary Economics, 9(3): 331-366.

Rus V., Niraula N., Banjade R.(2013). Similarity measures based on latent dirichlet allocation.International Conference on Computational Linguistics & Intelligent Text Processing.

Wang Z., Ma L., Zhang Y.(2016). A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec.2016 IEEE First International Conference on Data Science in Cyberspace (DSC). IEEE Computer Society.