Chapter 85
Research on Methodology of Correlation Analysis of Sci–Tech Literature Based on Deep Learning Technology in the Big Data

Wen Zeng
Institute of Scientific and Technical Information of China, China

Hongjiao Xu
Institute of Scientific and Technical Information of China, China

Hui Li
Beijing Institute of Science and Technology Information, China

Xiang Li
Torch High Technology Industry Development Center, China

ABSTRACT

In the big data era, it is a great challenge to identify high-level abstract features out of a flood of sci-tech literature to achieve in-depth analysis of data. The deep learning technology has developed rapidly and achieved applications in many fields, but has rarely been utilized in the research of sci-tech literature data. This article introduced the presentation method of vector space of terminologies in sci-tech literature based on the deep learning model. It explored and adopted a deep AE model to reduce the dimensionality of input word vector feature. Also put forward is the methodology of correlation analysis of sci-tech literature based on deep learning technology. The experimental results showed that the processing of sci-tech literature data could be simplified into the computation of vectors in the multi-dimensional vector space, and the similarity in vector space could be used to represent similarity in text semantics. The correlation analysis of subject contents between sci-tech literatures of the same or different types can be made using this method.

1. INTRODUCTION

With the rapid development of science and technology, the relationship between science and technology is getting closer and closer while their boundary is becoming more and more blurred. Relevant researches on sci-tech data, especially on sci-tech literature, are of great significance for understanding the connection between sci-tech literature, and assessing their achievements and the level of innovation. A large number of studies have shown that correlation analysis of sci-tech literature is useful for understanding sci-tech trend, measuring innovation level, etc. (Zhiyuan, Aryya, George, Michael, and Claire, 2007; Congfeng, Junming, Dongyang, Yumei, and Lifeng, 2018). The environment and the characteristics of big data make necessary and the valuable analysis on the correlations between the contents of sci-tech literature, which is also a new perspective and new application direction of the sci-tech information analysis. Taking sci-tech papers and patents for example, papers and patents are the two most important types of sci-tech Literature. Paper is an important output of the natural science and the social science research while the information contained in patents is more practical, technical and innovative than that in other sci-tech literature. The traditional academic evaluation standards based on citation methods are not in line with the needs of the big data era.

Deep learning, as a branch of machine learning and a hot field of artificial intelligence research, is intended to establish a mechanism in simulation of the neural network of human brain to analyze, learn, and interpret data such as images, sounds and texts. The concept of deep learning was put forward (Hinton and Salakhutdinov, 2006; Silver, Huang, and Maddisonc, 2016). Based on the deep belief network, the unsupervised, greedy and layer-wise training algorithm was put forward, bringing hope to solve the optimization problems related to deep structure. Later on, the deep structure of multi-layer auto-encoder was proposed. The deep machine learning methods can mainly be divided into two types: supervised learning or unsupervised learning. The learning models constructed under the frameworks of supervised and unsupervised learning are different. For examples, the convolutional neural network is a machine learning model under the framework of deep supervised learning, while the deep belief network is a machine learning model under the framework of unsupervised learning. "Deep learning" is relative to "shallow learning" methods such as support vector machine, lifting scheme and maximum entropy. The shallow learning relies on the artificial experience to extract sample characteristics, and after the model learning, it obtains the single-layer features with no hierarchical relations. While the deep learning can get hierarchical feature representation through automatic learning, using multi-layer feature transformation of input data signal to realize the change of original data sample feature representation to the new feature space, thus facilitating data classification or visualization of features. The deep network structure acquired through deep learning has the characteristics of neural networks. Thus, deep networks are actually neural networks at deep levels, namely deep neural networks (DNN). In recent years, continuous development and innovation has taken place in the deep learning field, which fuels new discoveries in its research and applications, and much progress has been achieved in model, algorithm and large-scale applications. At present, the increasing scale of sci-tech literature data poses a great challenge to the information analysis, and calls for application of the deep learning technology in the sci-tech literature data analysis.

Based on the characteristics of big data, this paper uses deep learning technology to achieve an effective dimensional reduction of the massive sci-tech data, establish a word vector representation method with contextual semantic relationship, and finally realize the content correlation analysis of the sci-tech literature. This paper will focus on the methodology of correlation analysis of sci-tech literature

Related Content

A Framework for an Artificial-Neural-Network-Based Electronic Nose
Mudassir Ismail, Ahmed Abdul Majeedand Yousif Abdullatif Albastaki (2022). *Research Anthology on Artificial Neural Network Applications (pp. 350-374).*
www.irma-international.org/chapter/a-framework-for-an-artificial-neural-network-based-electronic-nose/288965

Artificial Higher Order Neural Networks for Modeling Combinatorial Optimization Problems
Yuxin Ding (2013). *Artificial Higher Order Neural Networks for Modeling and Simulation (pp. 44-57).*
www.irma-international.org/chapter/artificial-higher-order-neural-networks/71794

Neuromorphic Computing: Implementation and Open Issues
Shivakumar Hunagund (2023). *Neuromorphic Computing Systems for Industry 4.0 (pp. 30-62).*
www.irma-international.org/chapter/neuromorphic-computing/326833

Medical Diagnosis
 (2014). *Medical Diagnosis Using Artificial Neural Networks (pp. 1-10).*
www.irma-international.org/chapter/medical-diagnosis/110995

Other Diseases Detection: Hepatitis
 (2014). *Medical Diagnosis Using Artificial Neural Networks (pp. 170-180).*
www.irma-international.org/chapter/other-diseases-detection/111008