Application of Deep Learning in Text Mining

Haoriqin Wang
College of Mathematics
Inner Mongolia University for the Nationalities
Tongliao, China

Mingyang Jiang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China
e-mail: jiang_ming_yang@163.com

Jianhong Qi
Function Department
City Hospital
Tongliao, China

Xinhong Zhang
Department of Neurology
The Affiliated Hospital of Inner Mongolia University for Nationalities
Tongliao, China

Qinghu Wang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Yuxin Zhou
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Mingyu Bai
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Lisha Liu
Automation Workstation People's Laborationa
Army 65367 Troops
Tonghua, China

Zhili Pei*
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China
e-mail: zhilipei@sina.com

Abstract—With the advancement of science and technology development and the development of social , more and more current information technology and computer technology has been used in the various aspects of life and work, the current rapid development of network technology has penetrated into all aspects of life. With the development of network technology and the text messaging of network gradually strengthened, wide and messy network information, in the vast network of information, how to efficiently access the information that we need quickly, which is related to the text mining technologies. The deep learning is a new learning method of the machine learning, it simulates the human brain and analysis the neural network through the imitation of the human brain and the interpretation of the relevant data. In the text mining, the application of the deep learning can be a very good text clustering and text classification, it is easy to find the desired text information, so the application of the deep learning plays an important role in the deep learning. Research for this paper analyzes the deep learning in text mining and the related content of knowledge**.**

Keywords- deep learning; text mining; characteristics and applications; network technology; machine learning

I. INTRODUCTION

With the development of information technology, the current information appears more and more diverse in the network, the speed of information mining that people need become faster and faster, in order to enhance the speed of text mining, it needs to research new technologies and the new text mining methods. Deep learning is a new learning method in network text mining, it can improve the mining speed of the text information in order to access the desired text information quickly. The deep learning has an important significance in the text mining.

II. 1、BASIC KNOWLEDGE OF THE DEEP LEARNING

Deep learning is a new form of learning, it build a simulation network of brain through the analysis of the human brain. Deep learning reveals the brain networks that they understand through images, sounds, text, etc. This learning method is an unsupervised learning. Deep learning is a multi-level learning, which is mainly produced from artificial neural networks, it can simulate the network of the brain to study the multi-level learning and the deep learning .It can optimize the information of knowledge and improve the utilization of information.

The depth of the deep learning can be represented by a flow chart, it can be represented by figure of calculation, each node will have a corresponding point calculation in the flow chart, but also produce a corresponding calculations. That is depth which is a special property, it has the flow characteristics and properties of this graph, and the length of the longest path [1] in the entire circuit from input to output is the depth.

The deep learning can be applied to solve the problem of text mining encounter. There are some features in deep learning application process, such as, It can solve the problem of inadequate depth, it build a network of simulated human brain has deep structure, layer by layer depth learning, and gradually abstraction. If you use the deep learning, you need to look the structure learning content and text information as a network, and then build the idea of an unsupervised learning, learning step by step. [2][3]

The classical algorithm of deep learning build a relatively generic model applicable on the basis of theory and experiment and optimize the choice about the relevant data for the mode , then select the best data set from it. Build the neural network model to analyze the deep learning from the data and algorithms in order to achieve deep learning network structure transformation, it can reduce the difficulty of the network structure and the time complexity[4-6]. According to the analysis of the model, establish the deep learning models neurons accordingly to form a topology model structure, so the algorithm of the deep learning will be updated. Deep learning algorithm flowchart below:

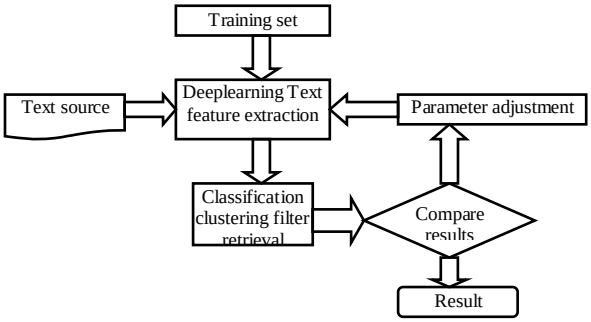

Figure 1. Algorithm Model of Deep Learning

The subject model of the deep learning mainly used for the input document collections, let the documents and other topics lexical item enter the main model while the context probability model is introduced into the body through the algorithm of deep learning to optimize in order to achieve the estimation of parameter, the word under each topic by a probability distribution and document theme probability distribution. After complete the whole subject model training, infer new samples of the model of the subject matter, the subject model the documents in the transition to a new theme in the space, forming a document theme probability distribution[7-9]. Deep learning of the theme algorithm flowchart below:

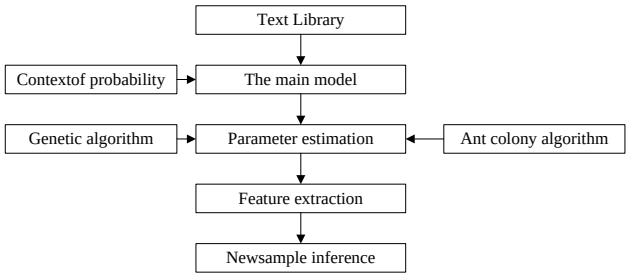

Figure 2. Theme of Deep Learning Model

III. DEEP LEARNING APPLICATION IN TEXT MINING

With the process of IT and the network technology evolving, a variety of information appears in people's vision in order to ensure you can get the exact information that they need, we can obtain the text information that we need with the text mining technology. Text mining is a newly formed data mining theme, text mining is actually classified information retrieval and information, and to find information. Information retrieval and text mining information that can be extracted as a pretreatment process of information, we can use the text mining after the completion of pre-treatment information.

Analysis from the process of text mining, text mining's main features are:

Firstly, text mining is a large-scale centralized documents, it is not a small amount of text messages that carried out in a text.

Second, the reason for text mining is concentrated from a large number of documents to obtain the hidden knowledge which is new, it is not previously discovered.

Thirdly, excavation of information and the knowledge has a certain value in the text information mining, the excavation of information and the knowledge can be directly used, it may be interest of a particular user, perhaps is the answer to a problem and so on.

Fourth, analysis from the characteristics of text information, text information mining algorithm has certain complexity . its complexity is established on the space and time.

Fifth, text information mining is in order to get the new information of text, while the scope for text mining involves multiple disciplines, such as data mining, statistics, language processing, database technology and so on.

Analyzed about the text mining, it can be divided into two types, one is a single document data mining, and the other is a document set. And one single document is only on the information in a single document classification; it has nothing to do with the other document information. There are many text information in a single document, such as text, text information, phrases and phrases and so on. With the progress and development of science and technology, information technology and network technology widely used in people's live, and the produce a wide variety of text information, in order to obtain the required information of text in the multifarious document information, so text mining is required.

Text mining is a variety of document information preprocessing, document information will be standardized extract text features and structures in the form of

information, and then set for the text features for exclusion of information, draw text data formalized, extraction of knowledge from which to get text messages and text knowledge required, this process is called text mining. With the development of text mining, it formed the new methods and technology in order to use text mining in the study of people's constantly , and it can improve the speed and quality. Deep learning is a model of network text mining algorithms newly formed, text mining process will be simplified and optimized to improve the speed of text mining quality.

The current application and development of network technology has been involved in all aspects, and then with the development of network technology, it is difficult for people to get the desired text and date mining from all kinds of text information on the network. Text mining also proposed a new direction for research and development because of the increase of the difficulty of text mining. Deep learning has been widespread concern in recent years, it has been widely used in a variety of pattern classification now ,it produced a great influence in the application of image processing, In the application of text research, combine the text information up and down with structural information effectively to form a new method of text information.

The neural network create based on deep learning shown in Fig .3, the picture is not all links of neural network structure, its each of neurons only accept themselves bestow domain transmission signal, as shown in Fig .4, each of the connection matrix are the same in the formation of neurons [10][11].

The first k characterized by plane in the neural network is called hk, a layer of connection weight matrix is called Wk , and a deviation is called bk , so we can draw a function, the feature mapping is:

$$(f * g)(x) = \int f(u)g(x-u)du \qquad (1)$$

$$(f * g)(n) = \sum_m f(m)g(n-m) \qquad (2)$$

The first k characterized by plane in the neural network is called hk, a layer of connection weight matrix is called Wk , and a deviation is called bk , so we can draw a function, the feature mapping is:

$$h^k_{ij} = \tanh(W^j * x)_{ij} + b^k \qquad (3)$$

i, j of formula is the position of neurons in the neural network where the feature plane appear.

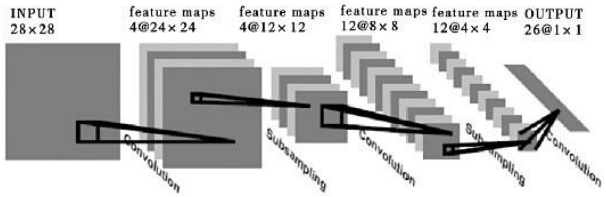

Figure 3. Establish Neural Network Structure

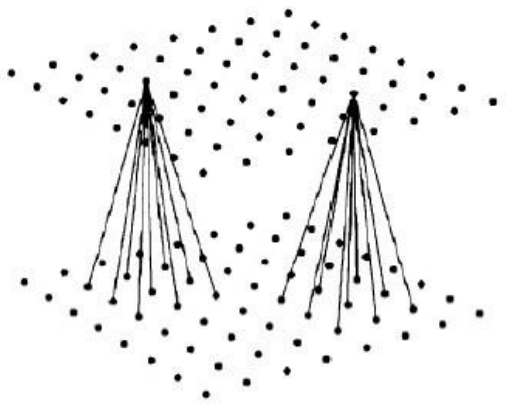

Figure 4. Structure of Neural Network Domain

Deep learning applications in text mining, including text clustering and text analysis, it can improve the text mining speed, accuracy, quality .a simulated network structure is constructed based on human neural network structure ,it formed learning model .we can extract different characteristics of the observed object characteristics and structure extraction through building these analog network learning model, These information is used as input to the next layer of research of information retrieval. Learning the hierarchical structure of deep learning , gradually deepening abstraction layer by layer, in order to form a deep learning model.

Analyze the text representation of the text mining , text representation model has space model and probabilistic models, the neural network model based on deep learning is built on the basis of the structure of space and time on the formation, So we analyze space model as an example:

In the model feature of the text : D= (t_1, t_2, t_3, t_4, ...t_k, ...t_n) Which feature in the first k t_k is model characteristics, thus as a dimension of the model.

there are n documents in weight algorithm of feature model ,W_{M*N}= (W_{ij})

The representation of a model text d: V (d) = (w_1, w_2, w_3, w_4, ...w_k, ...w_n)

The space model of the vector is expressed as:

$$V(d) = (t_1, w_1(d); \ldots; t_i, w_i(d); \ldots; t_n, w_n(d)) \qquad (4)$$

We can calculate t, d from the vector space model and form a word frequency matrix result:

TABLE I. PPLICATION FORM OF THE WORD FREQUENCY MATRIX

	d1	d2	d3	d4	d5	d6
t1	322	85	35	69	15	320
t2	361	90	76	57	13	370
t3	25	33	160	48	221	26
t4	30	140	70	201	16	35

In text mining, build a neural network model stratified by the application of deep learning , each layer is carried deep unsupervised learning, output the item before a layer to obtain the characteristics is used to retrieve information on the next layer, forming a equilibrium, and layer by layer for text mining, deep learning in text mining applications

forming a network model. Text mining process will be simplified and optimized to produce an optimal network model from analyzing the process of the model in order to achieve text mining quickly.

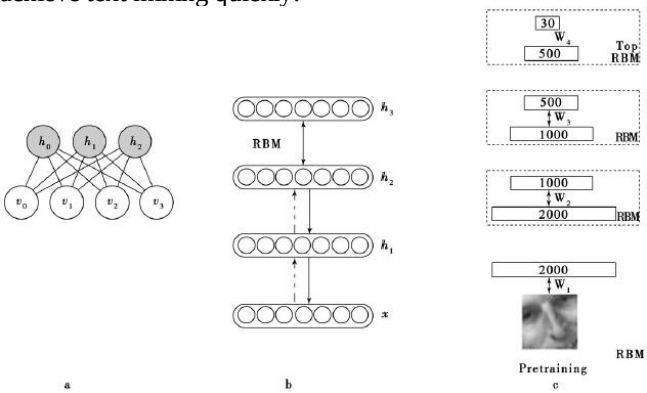

Figure 5. Deep Learning Model

A is RBM model, b is a typical structure of DBN, c is an example of using the image of the text structure DBN

The h and x of DBN typical structure form a probability distribution, the relationship of h and x is:

$$P(x,h^2,...,h^l)=(\prod_{k=1}^{l-2}P(h^k\left|h^{k+2}\right.))P(h^{l-2},h^l) \qquad (5)$$

For example, train each layer of sample in the model of deep learning . Assuming that the unknown for the joint probability distribution f（x1,.....,x k） , and extract n samples, sampling probability:

$$p(x_j\left|x_1,x_2,...x_{j-1},x_{j+1},...x_k\right.)= \frac{p(x_1,...x_k)}{p(x_1,x_2,...x_{j-1},x_{j+1},...x_k)}\propto p(x_1,x_2,...,x_k) \qquad (6)$$

Extract each layer of the sample In deep learning mode, assuming additional information and variables of the text mining are known, complete sample extraction on the basis of the probability distribution until drawn to the required samples.

IV. SUMMARY

The neural network model of deep learning is built on the basis of the human brain networks, centering on the network model for data mining. Clustering analysis method can improve the accuracy of the text in the text mining, but the speed of text mining needs to improve .Applications of deep learning in text mining can not only increase the speed of text mining , but also can improve quality and accuracy of text mining . With the development of network technology and information technology, applications of deep learning will continue to increase in the field of text mining.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61163034, 61373067), the Grassland Excellent Talents Project of Inner Mongolia Autonomous Region(2013), the supported By Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region(NJYT-14-A09), the Inner Mongolia Natural Science Foundation(2013MS0911), the 321 Talents Project the two level of Inner Mongolia Autonomous Region(2010), the Inner Mongolia talent development fund(2011), the First Trade Leading Talent of Tongliao, and the Science Research Project of Inner Mongolia University(NMD1229, NMD1322). The Corresponding author is Zhili Pei, and the Corresponding email is zhilipei@sina.com.

REFERENCES

[1] D. Erhan,Y. Bengio,A. Courville,P.-A. Manzagol,P. Vincent,S. Bengio.Why does unsupervised pre-training help deep learning?. Journal of Machine Learning Research.2010

[2] Summary of Zhijun Sun, Lei Xue, Yang Xu, Wang. Deep learning research [J] Computer Applications Research.2012(08):2806-2810.

[3] G.E. Hinton,S. Osindero,Y. W. Teh.A fast learning algorithm for deep belief nets. Neural Computation.2006

[4] Key Technology Research [D] Shozhong Tang. Text mining. Beijing Forestry University 2013.

[5] Computer and telecommunications Xiaoyan Zhang, Ying Hua. Text mining methods and applied research [J]. 2011(12):68-69

[6] Salton G,Wong A,Yang CS.A vector space model for automatic indexing. Communications of the ACM . 1975

[7] Fang Li.Text mining of certain key technologies [D]. Beijing University of Chemical Technology .2010.

[8] Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B. Rule extraction from support vector machines: an overview of issues and application in credit scoring. Studies in Computational Intelligence, 2008, 80: 33-63.

[9] Carter P H. The Rocchio classifier and second generation wavelets. In: Proceedings of the International Society for Optical Engineering. Orlando, USA: SPIE, 2007: 1-11.

[10] Esuli A, Fagni T, Sebastiani F. Boosting multi-lable hierarchical text categorization. Information Retrieval, 2008, 11(4):287-313.

[11] Wang J, Yao Y, Liu Z J. Web page automatic categorization based on non-linear SVM decision tree. Journal of Computationl Information Systems, 2008, 4(2): 449-454.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.