A Prediction and Push Method of Scientific Researching Issue Based on RNN

Yan Ma¹*, Dali Qi¹, Naihai Yu² and Lida Zou³*

¹State Grid Shandong Electric Power Research Institute, Jinan 250002, China
²Shandong Electric Power Industrial Boiler Pressure Vessel Inspection Center, Jinan 250002, China
³Shandong University of Finance and Economics, Jinan 250014, China

*Corresponding author e-mail: zoulida@163.com, yanpony@126.com

Abstract. Exploring and predicting the hot scientific researching issues have been a focus among recent sci-tech information research. In recent years, scholars and academic literature are increasing steadily in number. It is hard to artificially track and deal with developing trends of scientific researching hotspot. The most commonly used methods in the past are simple statistical ways for keywords and high term frequency. Most of these methods spend a lot of time and manpower, and ignore the relevance between different words. In the paper we use RNN method to design our prediction and push system, which could perceive the potential researching hotspot for some time to come. The hot issues are generated based on the relevance among sci-tech words. It finds the relevant technical application to hotspots and recommends them to academic researchers. The extensive experiments demonstrate that our proposed approach has higher accuracy rate and lower false positive ratio. It can do better forward looking forecasts than word frequency statistics as well.

1. Introduction
Sci-tech information plays an important role on the establishment and implement of strategies and plans for nation, society and enterprise [1] [2]. The prediction of scientific researching hotspot is a new application in sci-tech information area. Science researchers and scientific project managers must have forward looking ideas on topic choice [3]. They have to consider the situations of current scientific research and social development, and make a judgment about the future novel theory or potential applicable technology.

At present the prediction method of scientific researching hotspot relies heavily on senior professional officers consulting literatures and conducting market research [4]. In addition, when a new theory or technology emerges, its relative applications need lots of work to explore.

Therefore, it's imperative to design a prediction and push system of scientific researching issues. It estimates the hot research topics in a near future and push them to scientific users, which could help the work of science researchers and scientific project managers.
2. The Prediction and Push Framework

In order to solve the problem in section 1, we propose a prediction and push method of scientific researching issue based on RNN (Recurrent Neural Network). Recurrent Neural Network is a variety of ANN (Artificial Neural Network) [5]. It could simulates the internal dependence relationship among sequential data and is widely used in the area of natural language processing, Image annotation and machine translation.

Figure 1 shows its framework. The system could timely obtains the hot topics for some time ahead. It consists of five models, which are data crawler module, feature representation module, feature extraction module, model training module, and prediction and recommendation module.

Next we give the whole generating flow of scientific research hotspots, as shown in Figure 2. Our algorithm is divided into two parts: model training, hotspot prediction and recommendation. When processing features and training models, it first crawls the sci-tech articles periodically, abstract TF-IDF [6] vectors for articles and then do the feature extraction for the vectors in a period based depth boltzmann machine. It then do the training for the hotspot prediction based RNN method and get the prediction model. When predicting and pushing hot issues, it first crawls the sci-tech articles periodically, abstract TF-IDF vectors for articles and then do the hotspot prediction based RNN using the prediction model. Through the model, we get the keywords of scientific research hotspot. Combining the relevance of clustered research words in sci-tech corpus, a series of hot scientific research issues are generated and recommended to users.

3. The Functional Design of Modules

In the section, we concretely introduce each module. Each module has its own function, we will introduce its implementation process in details.

Data Crawler Module. Data crawler module uses the web crawler technology to crawl scientific and research articles from sci-tech websites and bibliographic databases. The crawled articles are usually textualized. The set of crawled articles during an interval of time is noted as $\mathcal{T}$, where $t$ is the sequence number of the period.

Feature Representation Module. Feature representation module provides an approach to represent the textual features of sci-tech information and literatures, which is the input of feature extraction module.

The keyword vectors of $\mathcal{T}$ are obtained based on weighted TF-IDF algorithm, which is denoted as $\overline{a_i} = \{b_i | i \in Q\}$, where $Q$ is the number of words in sci-tech lexicon, $b_i$ is the weighted TF-IDF value of the corresponding word component. $b_i$ is computed as follows:

1) Assuming $t_i$ is a text in $\mathcal{T}$, TF-IDF value of the $i$-th word in $t_i$, denoted as $b_{i}$, is obtained based on standard TF-IDF algorithm, where $a_{i'}$ denotes a keyword in $t_i$ and $i$ is the identifier of word component which $a_{i'}$ corresponds to in lexicon.
2) Assuming the downloading or reading quantity of $t_j$ is denoted $n_j$, and the citation quantity is $m_j$, we get $b_t = \sum_{i=1}^{T} (b_i' + \alpha b)n_j m_j/\bar{m}$, where $\beta$ is the weight of keyword, $\bar{n}$ and $\bar{m}$ are the average value of all the articles $n_j$ and $m_j$, respectively.

*Feature extraction module.* Feature extraction module uses depth Boltzmann machine [7] to extract textual features in a period, which provides the input for model training, prediction and recommendation modules.

The structure and parameters of depth Boltzmann machine are set as follows:
1) Depth Boltzmann machine use three layers, as shown in Figure 3.

![Figure 3. The structure of depth Boltzmann machine](image)

2) The first layer is visible cell layer. The visible layer is a $Q \times B$ binary matrix, where $B = \max\{b_t b_t \in \bar{\alpha}, t \in T\} \in \eta$, $T$ is a set of all the period, $\eta$ is a larger positive integer, $\lfloor \cdot \rfloor$ denotes integer conversion. $B$ means the result that the maximal word component of $\bar{\alpha}_t$ in all the periods multiplies a coefficient.
3) Assume $v_{i,j}$ is a cell in visible layer. In period $t$, for $b_t b_t \in \bar{\alpha}_t$, if $j = b_t \times \eta$ holds, we get $v_{i,j} = 1$.
4) The second layer is the hidden layer. The second layer is $Q \times B$ binary matrix, where $\beta_0$ and $\beta_1$ are coefficients larger than one. The third layer is the hidden layer, which is $2 \times 2$ binary matrix.
5) Depth Boltzmann machine uses restricted Boltzmann machine [8] to do layer-wise trainings.
6) For period $t$, assume that the output of depth boltzmann machine is $X_t$.

*Model training module.* Model training module is to predict the potential research hotspots some time in the future, which provides the prediction model for prediction and recommendation module. Our hotspot prediction model is improved based on RNN approach. Its structure and training method is as follows.

1) The structure of hotspot prediction model is shown in Figure 4, which consists of $t$ cyclic layers and three Back-Propagation (BP) neural network. Assume $X_t$ denotes the output of depth boltzmann machine in cyclic $t$.

![Figure 4. The structure of hotspot prediction model based RNN](image)

In Figure 4 input weight is $U$, $W$ is cyclic weight and $V'$ is the corresponding weight of BP neural network $l$. Vector $net_{i}^{l}$ denotes the weighted input of cyclic neuron at time $t$, which is computed as $net_{i}^{l} = U_{i} + W_{i}S_{i-1}$, where $S_{i-1}$ is the value at time $t-1$ of cyclic layer, $S_{i-1} = g(net_{i-1})$, and $g(\cdot)$ is the activation function.
Vector net\(_t\) denotes the weighted input of BP neural network at time \(t\), which is computed as net\(_t\) = \(V^t g(\text{net}^{t-1})\). \(o_t\) is the final output, which is computed as \(o_t = g(\text{net}^{t})\).

2) In Figure 4 the blue arrow means forward computing direction and is used to compute \(o_t\). The computation is iteratively computed as \(o_t = g(\text{net}^{t})\), \(\text{net}^{t} = V^t g(\text{net}^{t-1})\), \(\text{net}^{0} = Ux_t + WS_{t-1}\), \(S_{t-1} = g(\text{net}^{0})\).

3) We use \(E\) to denote model error and the error function is \(E = \frac{1}{N} \sum_{n=1}^{N} y_{(n)} \log o_{(n)}\), where \(N\) is the sample quantity, \(y_{(n)}\) is the actual value of sample \(n\) and \(o_{(n)}\) is the output of sample \(n\). \(y_{(n)}\) and \(o_{(n)}\) are word vectors. Assume \(y_{(n)} = [c_1, c_2, \ldots, c_i, \ldots, c_m]\) and \(m\) is the total number of words. If the \(i\)-th word is popular keyword, \(c_i\) is labeled as 1 and the others are labeled as 0. There can be many hot words. Assuming \(o_{(n)} = [d_1, d_2, \ldots, d_j, \ldots, d_m]\), \(y_{(n)} \log o_{(n)}\) is computed as \(y_{(n)} \log o_{(n)} = c_1 \log d_1 + c_2 \log d_2 + \ldots + c_i \log d_i + \cdots + c_m \log d_m\).

4) The red arrow denotes the direction of error back propagation. Let \(\delta\) denote the error item of each layer of each cycle time. The error item is computed as \(\delta = \frac{\partial E}{\partial W} = \sum_{t=1}^{t} \nabla_{w_t} E\).

i) Compute the gradient for weight matrix \(W_t\) of error function \(E\) for any cycle \(k\) as follows.

\[
\nabla_{w_t} E = \begin{bmatrix}
\delta_{t, x_{1,1}} & \delta_{t, x_{1,2}} & \cdots & \delta_{t, x_{1,m}} \\
\delta_{t, x_{2,1}} & \delta_{t, x_{2,2}} & \cdots & \delta_{t, x_{2,m}} \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{t, x_{k,1}} & \delta_{t, x_{k,2}} & \cdots & \delta_{t, x_{k,m}} 
\end{bmatrix}
\]

Where \(\delta_{t,j}\) denotes the \(i\)-th component of the error item \(\delta_t\) and \(s_{t,j}\) denotes the output value for the \(j\)-th nerve cell of \(s_{t-1}\). The gradient of cyclic weight matrix \(W\) is the sum of all the gradients at different times. It is computed as \(\frac{\partial E}{\partial W} = \nabla_{W} E = \sum_{t=1}^{t} \nabla_{w_t} E\).

ii) Similarly, the gradient of weight matrix \(U\) is computed as

\[
\nabla_{u_t} E = \begin{bmatrix}
\delta_{t, y_{1,1}} & \delta_{t, y_{1,2}} & \cdots & \delta_{t, y_{1,m}} \\
\delta_{t, y_{2,1}} & \delta_{t, y_{2,2}} & \cdots & \delta_{t, y_{2,m}} \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{t, y_{k,1}} & \delta_{t, y_{k,2}} & \cdots & \delta_{t, y_{k,m}} 
\end{bmatrix}
\]

Where \(x_{k,j}\) denotes the \(j\)-th component of \(X_t\) and \(\frac{\partial E}{\partial U} = \nabla_{U} E = \sum_{t=1}^{t} \nabla_{u_t} E\).

iii) Let \(H_t\) denote the value for \(l\) layer BP neural network at time \(t\), then we get \(H_t = S_t\) and \(H_t = g(\text{net}^{l-1})\). The gradient of weight matrix \(V\) is computed as

\[
\frac{\partial E}{\partial V_t} = \nabla_{V_t} E = \begin{bmatrix}
\delta_{t, h_{1,1}} & \delta_{t, h_{1,2}} & \cdots & \delta_{t, h_{1,m}} \\
\delta_{t, h_{2,1}} & \delta_{t, h_{2,2}} & \cdots & \delta_{t, h_{2,m}} \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{t, h_{k,1}} & \delta_{t, h_{k,2}} & \cdots & \delta_{t, h_{k,m}} 
\end{bmatrix}
\]

Where \(h_{t,j}\) denotes the output value for \(j\)-th nerve cell of \(H_{t-1}\).

Prediction and recommendation module. Prediction and recommendation module is to estimate and push the research hotspot in the future cycle. Its steps are as follows.
1) Gather sci-tech articles in cycles, do the feature representation, feature extraction and model prediction and get the keywords of scientific research hotspots.

2) Cluster the scientific research words in the natural language library and generate the relevance between scientific research keywords during the clustering process.

3) The scientific research hotspots are generated according to the predicted hotspot keywords and word relevance, and then pushed to users.

4. Experiments and Analysis

In the section, we choose two assessment indicators: prediction accuracy rate and false positive rate, to evaluate system performance. We use two servers to deploy our system. One is for crawling data, and the other is to deploy the system and run the prediction and recommendation algorithm. We use the labelled data from Jan. 2016 to Dec. 2017 to train our model and use the data from Jan. 2018 to Jun. 2018 to predict hot issues. Through our experimental operations for six months, we select 150 hot sci-tech issues and 150 non-hot sci-tech issues.

According to the prediction steps, data training period is an important parameter. We set the crawling and training period as 3 days, 7 days, 14 days and 30 days. Figure 5 shows the accuracy rate and false positive rate with different data periods. As shown in the figure, the larger the period is, the higher the accuracy rate is and the smaller the false positive rate is. The false positive rate is a little high, which is because our loss function did not consider the false positive ratio. We would focus it in the future work.

![Figure 5. The comparison with different time periods](image)

![Figure 6. The comparison in different prediction periods](image)

After we get the training model, we do the prediction. Considering the accuracy rate and training cost, we use 7 days as a period to evaluate the optimal time span of prediction. We assume that the time span of prediction ranges from one period to five periods. Figure 6 gives the comparison between accuracy rate and false positive ratio in different prediction periods. For the five prediction periods from the first to fifth period, the accuracy rate becomes lower and false positive ratio ascends. Therefore, for the 7-days crawling period, predicting hotspots within two periods has acceptable accuracy rate and false positive ratio.

Last we see whether our system can effectively and early find the hotspots. We set two weeks as a prediction cycle. The first two weeks is prediction cycle one, the second two weeks is prediction cycle...
two, and we choose 15 prediction cycles. We choose prediction probability in our algorithm to compare the common method named word frequency statistics. The prediction probability is the $d_i$ value corresponding to the $i$-th keyword in the output $q_i$. Figure 7 shows the comparison between prediction probability and word frequency statistics with the prediction cycle for a certain keyword. From the figure, we can see that in the prediction cycle 9 our method can predict its popularity in a high probability. It is largely better than the simple word frequency statistics in both accurate and forward looking forecasts.

5. Conclusion
In the paper we propose a method of predicting the sci-tech research hotspots some time in the future, which provides research direction and research idea for academic researchers. The relevance based on scientific research words could generate a series of hot scientific research issues. The method achieves the simultaneous delivery of relevant and potentially hot technologies or applications. Our experimental evaluations also prove that our RNN based prediction approach can predict much more accurate and timely than the common method.

Acknowledgements
This work was supported by Science and Technology Plan for Colleges and Universities in Shandong Province (KJ2018BAN046). The authors are grateful for the anonymous reviewers who made constructive comments.

References
[1] Hua bolin, Application of Big Data About Science and Technology Information in Intelligence Studies [J], library and information, 2017, 61 (16): 150-156.
[2] Bao Changhuo, Ma Dehui, Li Yan, Research on Chinese Intelligence Studies, Journal of Intelligence [J], 2015, 34 (12): 1-7.
[3] Xu Yan, Zeng Jianxun, The Policy and Mechanism of Institutional Repository Construction for Scientific Research Management [J], library and information, 2015, 59 (6): 22-27.
[4] Huang Jiaoping, Based on Microblogging Early Forecast and Analyze Negative Hot News [D], Dissertation, South China University of Technology, 2013.5.
[5] Wei Ziqing, China geodetic Coordinate System 2000 and its comparison with WGS84 [J]. Journal of Geodesy & Geodynamics, 2008, 28 (5): 1-5.
[6] Salton G. The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs, New Jersey : Prentice Hall Inc, 1971.
[7] Salakhutdinov R, Hinton G E, Deep Boltzmann Machines [C], Proceedings of International Conference on Artificial Intelligence and Statistics 2009. Brookline, MA, USA: Microtome Publishing 2009: 448-255.
[8] Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks [J]. Science 2006, 313 (5768): 504-507.