Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

A Prediction Method of Scientific Achievement based on Attention Characteristics

Yan Ma[1, 2], Lida Zou[1*], Yingkun Han[2], Lei Ma[2], Ke Liu[1], Jianmei Cao[2]
[1]Shandong Electric Power Research Institute, Jinan, Shandong, China
[2]State Grid Shandong Electric Power Research Institute, Jinan, Shandong, China
*Corresponding Author.

Abstract:

In order to arrange funds and increase profits, scientific management department in enterprise needs to master the situation and development of scientific achievements. Scientific achievements have the characteristic of lag. Prediction in advance is beneficial for managers to improve working efficiency and effectiveness. In the paper we propose a prediction method of scientific achievements based on attention characteristics. The prediction framework consists of four modules. The data of working status are collected by advanced monitoring devices, while the current works ignore this kind of data. The attention features are extracted and processed for the convenience of model training. Our prediction model is time-sensitive, which could find the relationship between the achievements and their working period. The experimental results demonstrate that the proposed method works as well for any kind of scientific achievements. Its estimate accuracy and topic similarity is 48% and 70.6% better than benchmark models separately.

***Keywords**: Scientific achievement, Predict, Time-sensitive, Attention characteristics, Monitoring.*

I. INTRODUCTION

The prediction of scientific achievements is important for scientific management department, which is responsible for arranging fund, promoting technology upgrades, and forecasting profit according to the future output of scientific results. Scientific achievements mean the published papers, books, intellectual property, prototypes etc. When it is predicted that there may have lots of intellectual property, funds for screening infringement and protecting the intellectual property rights need to be added. When many prototypes are predicted to be created, manpower, money and materials should be prepared in advance for large-scale production. When there would be many papers and books, budget for rewards and publicity channels should be considered.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

There are certain complexity and challenge for predicting scientific achievements. First, it is hard to obtain data for quality of work in scientific research division. Although the number of staff doing scientific and technological research is easy to get, their working status are difficult to acquire. Sometimes one scientific worker is temporarily transferred to management position. Moreover, it is almost impossible to quantify the actual working time and efficiency of scientific workers. However, data of working quality is important for prediction of scientific achievements. Second, scientific achievements are shown in the forms of paper, book, intellectual property, report, prototype etc. Current working content and working quality influence future results output in different temporal hysteresis. Report and prototype would be fed back and validated in a short time. Paper, book and intellectual property need peer reviews, which may take a long time. Therefore, achievement output of a scientific division in a future time is not easy to be predicted accurately.

Working quality of scientific and technological staff could be described based on attention characteristics. Attention is mainly viewed as the orienting and focus of certain objects in psychological activity, and is one common psychological characteristics along with the mental process such as sensation, memory, thought and imagination. There are two features: directivity and concentration. Directivity means that psychological activity tries to choose certain objects and leave the others. Concentration means the intensity of psychological activity on the chosen objects. In working time, scientific and technological personnel handle all kinds of work relying on electronic devices. Monitoring system could be used to continually obtain their attention situation, which provides the condition for getting and quantifying working quality and focusing area of scientific and technological personnel.

To address the above problems, we design a prediction approach of scientific achievements based on attention characteristics in the paper. It tries to accurately predict scientific achievements and research directions of the given scientific division in the future time. In the paper we could give the estimates for the number of different achievements and their topic keywords. Our work could help scientific and technological managers grasp up scientific and technological trends of their different divisions.

The paper is organized as follows. Section II gives the related work. The prediction framework based on attention characteristics are introduced in Section III. In Section IV, we present the concrete implementation of our proposed method. Experiments and analysis are demonstrated in Section V. Section VI concludes the paper.

II. RELATED WORK

Scientific and technological research plays an important role on enterprise development. The prediction of research hotspots is one of the newest applications in the fields of information science. In general research hotspots can be predicted by analyzing online articles [1] or academic literature [2].

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

Based on academic literature, prediction of research hotspot falls into two types according to analysis objects. One is to analyze the citation relationship between the citing reference and the cited reference [3-5]. The other is to analyze text content [6] and its concrete methods include co-word analysis [7], word frequency statistics [8], social network analysis [9] and text mining technology [10]. For text mining technique, subject model tries to find the latent topic in documents, which derives from Latent Semantic Indexing (LSI) [11]. Strictly LSI is not a subject model, but laid the groundwork for developing subject model. A method called probabilistic latent semantic indexing (pLSI) is proposed by Hofmann [12], which is a method of data analysis based on dual mode and co-occurrence. In 2003, the famous Latent Dirichilet Allocation (LDA) method [13] is presented based on pLSI. It is an almost perfect probabilistic generation model, which include words, topic and document. We use LDA algorithm in our paper.

Prediction of scientific achievements needs heterogeneous data including text information and image information. At present image processing algorithms fall into three types: objective detection [14], objective recognition [15] and OCR (Optical Character recognition) [16]. Based on the types, more sub-classifications such as semantic image segmentation, shape identification and visual tracking are forming. For text information, natural language processing (NLP) and machine learning techniques are efficient. Cognitive intelligence algorithm based on NLP consists of feature encoding, feature extraction and mapping knowledge. Feature encoding methods include ONE-HOT, Word2Vec [17], FastText [18], GLOVE [19], ELMO [20]. Feature extraction module can use RNN, LSTM [21], Transformer [22] and BERT [23] methods.

In the area of information perception, people are reluctant to obviously show their relevant information and hidden data can be used to predict their cognitive status [24, 25]. Information perception is a collective term covering people's sense and feeling. At present, the research on perception of relevant information mainly uses system log, screen recording and eye movement etc. Jiang et al. [26] use the document content and clicked data to find the relevance among documents, and following works add the attributes of staying time [27-29]. White et al. [30] indicate that the selection of texts is efficient for prediction. Now more wide data about user behavior have been used on user cognitive status [31]. At present perception of relevant information are mainly applied in information searching filed. For prediction of scientific achievements data about user behavior rarely are used.

III. THE PREDICTION FRAMEWORK BASED ON ATTENTION CHARACTERISTICS

The main idea of our proposed method is as follows. Firstly attention situations of scientific and technological personnel are automatically acquired and used to train prediction model for different kinds of achievements. Secondly data in critical time slices are chosen according to the time span of models and used to generate topic keywords.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

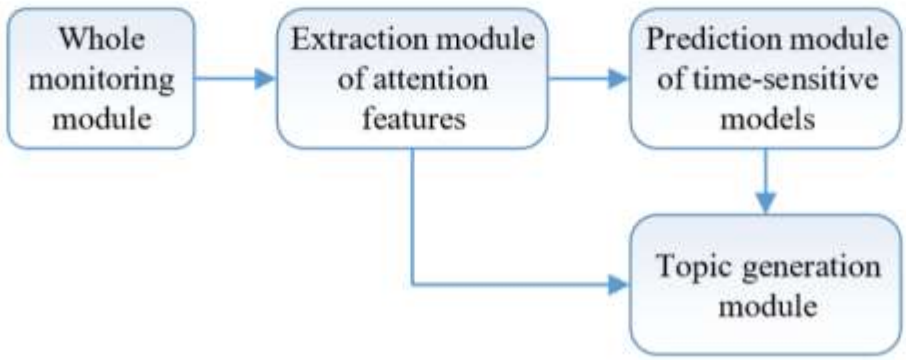

Fig 1: The prediction framework based on attention characteristics

Fig 1 gives the prediction framework of scientific achievements and research directions based on attention characteristics. The framework includes whole monitoring module, extraction module of attention features, prediction module of time-sensitive models and topic generation module. Whole monitoring module is to set different kinds of monitoring devices in working environment of scientific and technological personnel, obtain their original data of attention status and send the data to extraction module of attention features. Extraction module of attention features is to select and process the original data of attention status, and let them apply in the prediction of scientific achievements. The extracted features are sent to both prediction module of time-sensitive models and topic generation module. Prediction module of time-sensitive models is to design prediction models of scientific achievements and research directions, and used to predict the expected achievements in the appointed time. Topic generation module is to generate the corresponding topics according to time-sensitive weights and data of attention status.

IV. THE CONCRETE IMPLEMENTATION

The concrete flow chart of our proposed prediction method is shown in Fig 2. First the original data of attention status is obtained and then the original data are selected and processed to form multiple sample sets. Based on the sample sets, different models are trained. Next the optimal model is determined by double standards of both estimate accuracy and topic overlapping. Last scientific achievements and their topic words are predicted based the optimal model.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

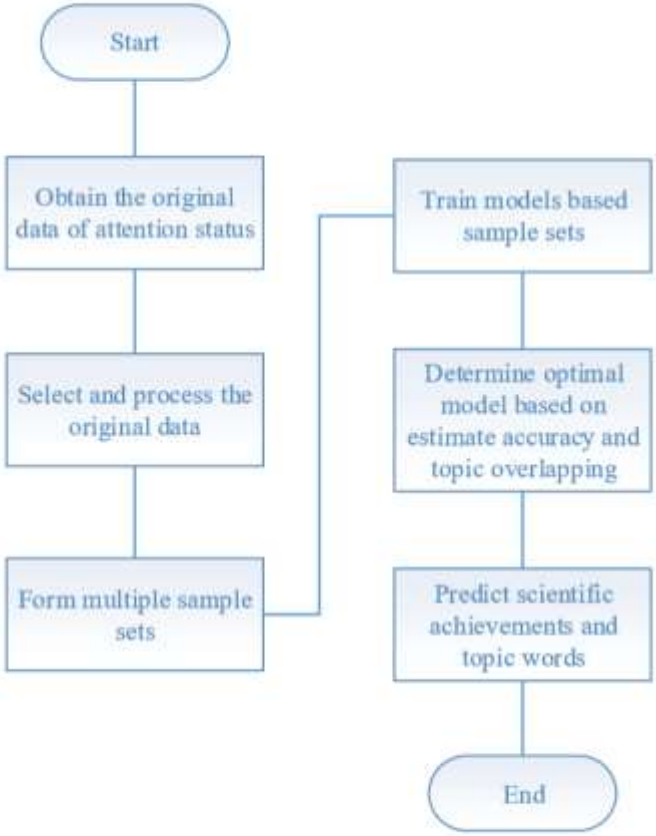

Fig 2: The flow chart of prediction method

4.1 Whole Monitoring Module

The module sets up computer monitoring unit, recording device, seat-sensing device and monitoring unit for portable smart devices. They monitor working status and working content of each person in scientific and technological division. Their concrete process is as follows.

(1) Computer monitoring unit records input information, output information and active time of working computers. Input information includes the input text content. Output information means the texts scientific workers read. Computer monitoring unit is a background supervising program installed on the working computer, which could compatibly open the common text editor or text browser such as txt, doc, pdf and etc.

i. First we give how to record the input text contents. If we could get control of current software, API call mode is used to record the input texts in real-time. If we cannot get control right, we should

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

often capture screenshots and trace the change of texts. When there exist many redundant words in screenshots, OCR technique is adopted.

ii. The way of recording the texts they read is as follows. If we could get control of current software, API call mode is used. Here, we do not get all the texts in the file but intercept part of the texts. If we use screenshots, the *n* lines of texts around current texts are recorded as the read information at the current moment. We formulate the concrete method as follows. The current moment is set as t_i, and the information during time $[t_i - a, t_i + a]$ is captured, where a defines the length of time period. OCR technique is adopted to recognize the texts. After mergence and de-duplication operations we get the final read texts at the current moment.

iii. The active time records two items: operation time and edit time. Operation time is the length of computer usage, which is the sum of reading time and writing time. Edit time is the time of editing documents, i.e., writing time.

(2) Recording device saves the voice message during working time, which is carried by scientific staff. It is open during working hours. After recognizing the unique sound of each person, the voice messages are transformed into texts.

(3) Seat-sensing device is used to record the time of setting on the seat.

(4) Monitoring unit for portable smart devices saves input and output information of the used portable smart devices such as tablet and smartphone. The monitoring unit is built into the smart device. Screenshot capture and OCR techniques are used to recognize the information they edit or read, which is similar to working computer.

4.2 Extraction Module of Attention Features

The extraction module of attention features changes the sent data from whole monitoring module into time series data in order to facilitate model training.

(1) First we set the length of a time slice, and the working status of scientific staff are recorded time interval by time interval. Let *i* denote one time slice. The length of one time slice can be one week, 10 days or one month.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

(2) Assume that p_j represents a scientific personnel, where *j* is the identifier. We let $p_{i,j}$ denote the working status of p_j during time slice *i*, i.e., $p_{i,j} = (\alpha_{i,j}^{m}, \alpha_{i,j}^{e}, \alpha_{i,j}^{k1}, \alpha_{i,j}^{k2}, \beta_{i,j}^{m}, \beta_{i,j}^{e}, \beta_{i,j}^{k1}, \beta_{i,j}^{k2}, \chi_{i,j}, \delta_{i,j}^{k1})$. Here $\alpha_{i,j}^{m}, \alpha_{i,j}^{e}, \alpha_{i,j}^{k1}, \alpha_{i,j}^{k2}$ denote the reading time, editing time, read texts and edited texts of worker p_j using computer during time slice *i*, $\beta_{i,j}^{m}, \beta_{i,j}^{e}, \beta_{i,j}^{k1}, \beta_{i,j}^{k2}$ denote the reading time, editing time, read texts and edited texts of worker p_j using portable smart devices during time slice *i*, $\chi_{i,j}$ represents the time of sitting on the working seat during time slice *i*, and $\delta_{i,j}^{k1}$ denotes the topic of worker's speech during time slice *i*. $\alpha_{i,j}^{k1}, \alpha_{i,j}^{k2}$, $\beta_{i,j}^{k1}, \beta_{i,j}^{k2}$, $\delta_{i,j}^{k1}$ are all obtained through whole monitoring module.

(3) We build a data set A, where a_i is a record in A during time slice *i* and computed as

$$a_i = (\sum \alpha_{i,j}^{m}, \sum \alpha_{i,j}^{e}, \mathrm{LDA}(\alpha_{i,j}^{k1}), \mathrm{LDA}(\alpha_{i,j}^{k2}), \sum \beta_{i,j}^{m}, \sum \beta_{i,j}^{e}, \mathrm{LDA}(\beta_{i,j}^{k1}), \mathrm{LDA}(\beta_{i,j}^{k2}), \sum \chi_{i,j}, \mathrm{LDA}(\delta_{i,j}^{k1}) \mid p_j \in P)$$

where P is the set of scientific staff, $\mathrm{LDA}()$ is topic vector extraction function. Its idea is as follows. First all the texts in brackets are summarized into a long general text and denoted as h. Second LDA technique based on natural language processing is used to analyze h and obtain topic weight vector or topic vector data.

(4) We build label data of the presented achievements B, where b_k is a record in B. The presented achievements mean the authorized intellectual property, the public report and published paper. Here b_k records the number of achievements during time slice k and the same kind of achievements are counted together. In general, the time slice of set B is longer than that of set A. One time slice of set B can be a month or three months.

(5) Next sample data of different time intervals are generated. For b_k during a given time slice k, it is hard to determine which time slice may influence the number of achievements at time k. System cannot automatically know how long it is before the working achievement could be presented. Therefore, sample data of different time slices should be prepared to train the models. As shown in Fig 3, the data of

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

achievement presentation could match with data of multiple working statuses, and sample data in different time intervals are generated as follows.

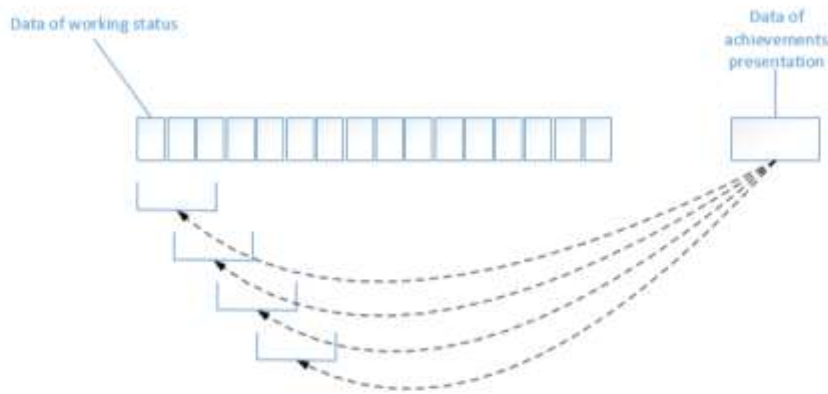

Fig 3: The generation method of sample data in different time intervals

i. Based on b_k, multiple records can be built, which is formulated as $< x_m, y_m > | m \in [0, k-t)$. Here y_m is label data, x_m is attribute data and computed as $x_m = (a_i \mid i \in [m, m+t), a_i \in A)$, where t denotes the length of time slice., m is the identifier of record and denotes the moment of the chosen working status. Label data is computed as $y_m = (b_k, \mathrm{LDA}(b_k^{k1}))$, where b_k^{k1} is the texts of achievements, $\mathrm{LDA}(b_k^{k1})$ is the topic weight vector computed in step (3).

ii. For each b_k, step i) is executed to construct data records. The record set is denoted as $\Re =< x_m, b_k > | m \in [0, k-t), b_k \in B$.

iii. Let $(k-m)$ denote the time lag between the chosen working status and achievement presentation. The data records with the same time lag are put into one sample set, i.e., based on the value of $(k-m)$, all the records in $\Re$ are partitioned into distinct sample sets. For a given sample set C_{k-m}, its records have the same time lag $(k-m)$.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

4.3 Prediction Module of Time-sensitive Models

Prediction module of time-sensitive models separately trains models based on multiple sample sets. It predicts the topic and number of achievements during the next time slice based on the optimal model. We next introduce its method.

(1) The generated sample sets from the extraction module of attention features is denoted as $C_v \mid v \in [0, n-1]$, where v is the identifier of sample set, n is the number of generated sample sets.

(2) For each C_v, we train a model. The input is part of attribute data in $C_v : \{< x_m, y_m > , \cdots\}$, i.e., partial attributes a_i are chosen from x_m in C_v . Formally, the input is $(\sum \alpha^m_{i,j}, \sum \alpha^e_{i,j}, \sum \beta^m_{i,j}, \sum \beta^e_{i,j}, \sum \chi_{i,j} \mid p_j \in P)$ chosen from $a_i = (\sum \alpha^m_{i,j}$, $\mathrm{LDA}(\alpha^{k1}_{i,j}), \mathrm{LDA}(\alpha^{k2}_{i,j}), \sum \beta^m_{i,j}, \sum \beta^e_{i,j}, \mathrm{LDA}(\beta^{k1}_{i,j}), \mathrm{LDA}(\beta^{k2}_{i,j}), \sum \chi_{i,j}, \mathrm{LDA}(\delta^{k1}_{i,j}) \mid p_j \in P)$. The output is b_k in y_m. The training model can be any regression models of machine learning. The model is denoted as E_v based on C_v. For the given tested data, let ${}^{E_v}\hat{b}_k$ represent the estimate value of E_v gives.

(3) Next the optimal model is evaluated. We apply double standards including estimate accuracy and topic overlapping. The function of $\mathrm{arcmin}_{E_v}(\lambda * \theta)$ is used to evaluate model E_v , where $\lambda = \mid {}^{E_v}\hat{b}_k - b_k \mid$ denotes the difference between estimate and real value, $\theta = \sum_{z \in \{\alpha^{k1}_{i,j}, \alpha^{k2}_{i,j}, \beta^{k1}_{i,j}, \beta^{k2}_{i,j}, \delta^{k1}_{i,j} \mid i \in [m, m+t),\ p_j \in P\}} D(\mathrm{LDA}(z), \mathrm{LDA}(b^{k1}_k))$ represents the similarity between the topic of achievements and working contents, and $D()$ is cosine distance between vectors. The function tries to select a model whose product of λ times θ is minimal.

(4) After we get the optimal model, the number of scientific achievements can be predicted.

4.4 Topic Generation Module

Topic generation module generates the topic weights according to the texts used by optimal model E_v. First the topic texts in data of working status are found according to the corresponding time lag of

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

sample set in E_v. Second the topic weights are generated based on the found topic texts. The computation method of topic weight vector is introduced in Section 4.2. Third the first n words with the largest weights are regarded as topic words and sent to scientific management department.

4.5 A Case Study

In the section we assume to predict the number of paper publication in the following half year for a scientific research institute. The prediction is based on the above four modules.

First whole monitoring module is to collect data of working status for scientific staff. Second extraction module of attention features is to process data of working status in order to form multiple sample sets with different time intervals. Third prediction module of time-sensitive models is to train multiple neural network models and choose the optimal one based on prediction accuracy and topic overlapping. Four topic generation model is to predict the topic and the number of scientific achievements. After prediction, scientific management department could prepare people and funding in advance for awards and transformation of scientific achievements.

During implementation, it is better to separately predict conference paper and journal paper, since there are different publication periods for them. Separate prediction for scientific achievements would have higher accuracy rate. We will validate the difference in Section V.

V. EXPERIMENTS AND ANALYSIS

In order to validate the performance of our proposed method, we collect the data of working status and scientific achievements for 20 scientific workers from June 2019 to June 2021. In the section we will test the estimate accuracy of scientific achievements, and the topic similarity between predicted achievements and published achievements. We will also verify publication periods for different kinds of scientific achievements and find their own relationship between time spans and estimate accuracy.

First we introduce a benchmark training model called B. Our proposed method is called A in the paper. Model B uses artificial neural network to predict scientific achievements based on data of working time. Fig 4 shows the change of estimate accuracy with the number of scientific workers. Here we use the value of λ to denote estimate accuracy as Y-axis. The number of scientific workers whose data of working status are collected is taken as X-axis. From the figure we see that the estimate accuracy becomes better when the number of scientific workers increases. It means that we should collect as much data as possible within budget in order to get high accuracy. Compared with B model, our method uses more exact and fine-grained data, and its value of λ is small relatively.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

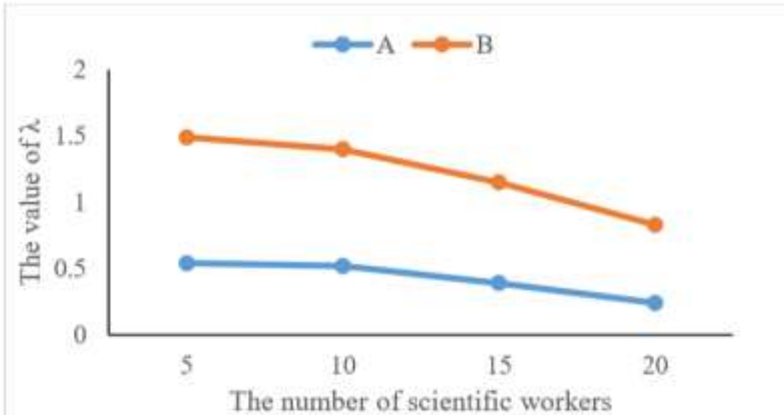

Fig 4: Estimate accuracy with the number of workers

We next test the value of λ with different time spans for various scientific achievements. The achievements are divided into paper, patent and prototype. The time spans are set as six, twelve, eighteen and twenty-four months. Fig 5 shows the comparison results, which demonstrate that our method works as well for different kinds of achievements. In the figure we conclude that each kind of achievements has its own publication period. The experimental results are coincident with the actual publication period for paper, patent and prototype. Taking patent as an example, its publication period is twenty-four months.

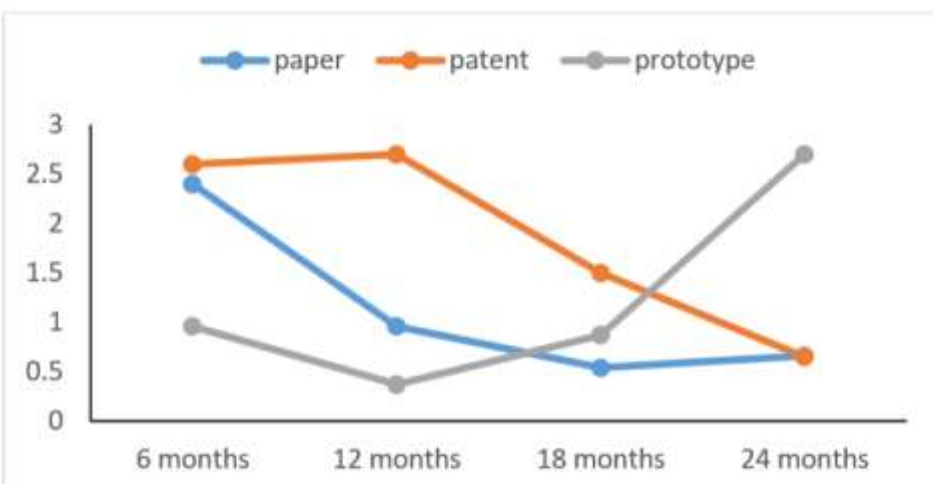

Fig 5: Estimate accuracy with different time spans for various achievements

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

Last we validate the similarity between predicted topic and actual topic of achievements. In Section 4.4, we define an indicator θ, which denotes the similarity. Fig 6 gives the comparison results on θ value for different kinds of achievements. Here we introduce a model called C. Model C tries to collect all the texts to make topic extraction generally. From the figure we see that our model A performs better than C, since A could sensitively perceive that works at what time period are related with the actual achievements. However, model C just obtain the overall topic content in general. Among three kinds of achievements, paper has the lowest θ value and prototype has the highest θ value. The lower the θ value is, the higher the topic overlapping is. Therefore, the topic of paper is clear, while the topic of prototype is relatively hard to acquire. It is better to predict the topic when there are more data of texts.

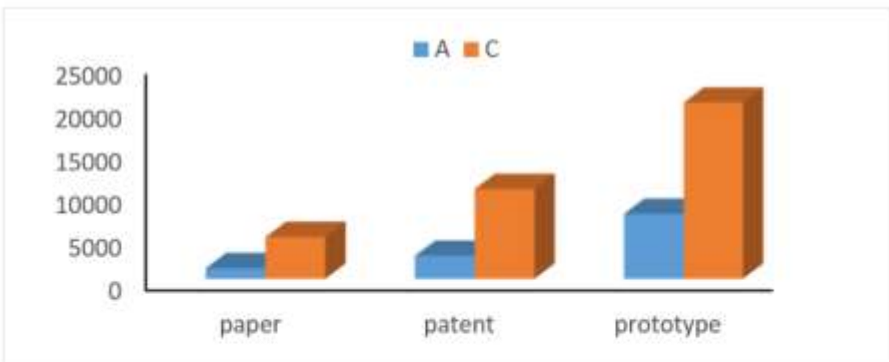

Fig 6: Topic similarity with different types of achievements

VI CONCLUSIONS

In the paper we propose a prediction method of scientific achievements based on attention characteristics of scientific personnel, which could help scientific management department estimate the topic and the number of achievements in the future. We choose multiple dimensions of input data such as read texts, edited text, voice message, working time etc. to predict the future scientific achievements, while most current works just consider simple input data. During feature extraction we use LDA method, and regression model is allowed to be adopted in model training. Two standards of both estimate accuracy and topic overlapping are introduced to validate the performance of our method. Extensive experiments demonstrate that our proposed method has high prediction accuracy no matter for any type of achievements. It performs 48% better in prediction accuracy and 70.6% better in topic overlapping than the benchmark models. The predicted publication period by our method is in consistent with the actual situations for paper, patent and prototype as well.

ACKNOWLEDGEMENTS

This research was financially supported by Scientific Projects of Shandong Electric Power Research Institute (Grant No. ZY-2021-16), Science and Technology Plan for Colleges and Universities in Shandong Province (Grant No. KJ2018BAN046).

REFERENCES

[1] Modjtaba G, Matthias D, Abdullah L, Najaf A, Abbe M, Frank ES (2021) On the relationship between PageRank and automorphisms of a graph. Information Science 579: 401-417

[2] Elizabeth G (2018) What to expect in 2018: science in the new year. Nature 553: 12-13

[3] Garfield E (2004) Historiographic mapping of knowledge domains literature. Journal of Information Science 30(2): 119-145

[4] Small H (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the Association for Information Science &Technology 24(4): 265-269

[5] Morris SA, Yen G, Wu Z, et al. (2003) Time line visualization of research fronts. Journal of the Association for Information Science & Technology 54(5): 413-422

[6] Álvaro J. Rojas L, Salvador Del BG, Juan MAP (2022) A review of three decades of academic research on brand equity: A bibliometric approach using co-word analysis and bibliographic coupling. Journal of Business Research 139: 1067-1083

[7] Ying D, Gobinda G, Chowdhury SF (2001) Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing & Management 37(6): 817-842

[8] Andrés C, Damián HZ (2021) Word frequency–rank relationship in tagged texts. Physica A: Statistical Mechanics and its Applications 574(1): 126-140

[9] Shen BJ, Guan TJ, Ma J, Yang LH, Liu YL (2021) Social network research hotspots and trends in public health: A bibliometric and visual analysis. Public Health in Practice 2(1): 100155

[10] Sunil K, Arpan KK, Vigneswara IP (2020) Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights 1: 1-14

[11] Deerwester SC, Dumais ST, Landauer TK, et al. (1990) Indexing by latent semantic analysis. Journal of the American Society for lnformation Seience 41(6):391-407

[12] Hofmann T (1999) Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International SIGIR Conference, New York, ACM Press 50-57

[13] Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3: 993-1022

[14] Felzenszwalb P, Mcallester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 1-8

[15] Papageorgiou C, Poggio T (2000) A trainable system for object detection. International journal of computer vision 38(1): 15-33

[16] Martinek J, Lenc L, Kral P (2020) Building an efficient OCR system for historical documents with little training data. Neural Computing and Applications 3: 1-12

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 304-317
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

[17] Mikolov T, Corrado G, Kai C et al. (2013) Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013) 2: 1-10

[18] Joulin A et al. (2017) Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 2

[19] Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation. Conference on Empirical Methods in Natural Language Processing 1: 221-232

[20] Peters M, Neumann M, Iyyer M et al. (2018) Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1

[21] Hochreiter et al. (1997) Long short-term memory. Neural Computation 15: 123-142

[22] Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. Advances in neural information processing systems 1: 5998-6008

[23] Devlin J et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810: 04805

[24] Back J, Oppenheim C (2001) A model of cognitive load for IR: implications for user relevance feedback interaction. Information Research 6: 342-359

[25] White RW, Ruthven I, Jose JM (2002) The use of implicit evidence for relevance feedback in web retrieval. Advances in Information Retrieval 2: 93-109

[26] Jiang S, Hu Y, Kang C, et al. (2016) Learning query and document relevance from a web-scale click graph. the 39th International ACM SIGIR Conference.ACM 1: 185-194

[27] Liu C, Liu J, Belkin NJ (2014) Predicting search task difficulty at different search stages. In Proceedings of CIKM'2014. New York: ACM: 569–578

[28] Guo Q, Agichtein E (2012) Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. International Conference on World Wide Web: 569-578

[29] Liu J, Belkin NJ (2010) Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. Proceedings of SIGIR'2010, New York: ACM: 26-33

[30] White RW, Buscher G (2012) Text selections as implicit relevance feedback. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM: 1151-1152

[31] Huang J, White R, Buscher G (2012) User see, user point: Gaze and cursor alignment in web search. Proceedings of CHI'2012. New York: ACM: 1341-1350