A Competency Mining Method Based on Latent Dirichlet Allocation (LDA) Model

Jiying Wu, Guandong Song*, Sihui Wang
School of Humanities and Law, Northeastern University, Shenyang 315211, China
*1710014@stu.neu.edu.cn
wjy_wjy606@126.com

Abstract. A text mining approach based on latent Dirichlet allocation (LDA) is proposed to analyze the competency characteristics. First, we briefly introduce the principle and hypothesis of latent Dirichlet allocation (LDA) model. Second, we elaborate the idea of using LDA topic model to extract competency. Then, we use Chinese text materials such as biographies and interviews of Chinese scientific and technological talents as data sources to conduct experiments, and obtain four competency topics including knowledge, attitude, quality and values. The research results preliminarily verify the effectiveness of latent Dirichlet allocation (LDA) model in competency research, but there are still many details to be improved in the future.

1. Introduction
With the advent of the era of knowledge economy in the 21st century, knowledge is gradually transformed into the main production factors, knowledge workers have gradually become the leading productive forces, and the importance of human resources in the comprehensive strength of various countries has been unprecedented. The comprehensive quality of individuals is an important factor affecting their work behavior and performance, which is mainly reflected in the competency. The concept of competency was first put forward by Professor McClelland of Harvard University. He published an article entitled "Testing competency rather than intelligence", pointing out that researchers should try to find a relatively fair and truly affect job performance personal condition factors and behavior characteristics[1]. Based on this, he put forward the idea of measuring competency and set off an upsurge of competency research. The so-called competency refers to "motivation, self-concept, attitude, values, knowledge and skills, etc., which can be reliably measured and can effectively distinguish between general performers and excellent performers"[2]. Since the competency was first proposed, many scholars have studied the competency of different positions, such as leader competency, manager competency, entrepreneur competency, etc. By combing the competency research literature, it is found that the most widely used competency modeling method is behavior event interview, which mainly takes the employees in the target position as the research object, and understands their key events in the process of performing their duties through in-depth interviews. The behavior event interview method is a highly professional method with high technical requirements for interviewers. Researchers generally lack corresponding practical experience, which leads to the failure to ensure the scientificality and effectiveness of information collection in the process of competency modeling. In fact, some existing text data such as job records, biographies of outstanding employees, record important turning events and behaviors in their work growth process,
which contains more abundant and in-depth growth details. However, there is a lack of research on competency modeling through the systematic mining of these text data.

Therefore, this paper introduces a data mining method: Latent Dirichlet Allocation (The following is abbreviated as LDA), which can extract topic and key information from large text data. In this paper, we propose the method of using LDA model to extract the competency. We use the biographies and interviews of academicians of Chinese Academy of Sciences and Chinese Academy of engineering as data sources to do experiments, in order to provide technical support for the future research and application of competency.

2. LDA model

LDA topic model is a more generalized text topic model proposed by Blei et al. on the basis of probabilistic latent semantic analysis (PLSA) model [3]. It is an unsupervised topic modeling method widely used in natural language processing. Because of its three-tier structure of words, topics and documents, it is also called three-tier Bayes Hierarchy Model. The emergence of this model completes the expansion of topic model in Bayesian level and has been widely used.

The core idea of the LDA model is that each document corresponds to a topic subject to the Dirichlet distribution $\theta$, and the words corresponding to each topic subject to the Dirichlet distribution $\phi$, where the document—topic distribution $\alpha$ parameters and topic—word distribution $\beta$ parameters obey the Dirichlet distribution $\alpha$ and $\beta$ [4].

For each document in the corpus, LDA defines the following generation process[5]:

1. For each topic: generate the "topic—word" distribution parameter $\phi_k \sim \text{Dir}(\beta)$.
2. For each document: generate the "document—topic" distribution parameters $\theta_m \sim \text{Dir}(\alpha)$.
3. For each location of the current document:
   (a) generate the topic of the current location: $z_{m,n} \sim \text{Cat}(\theta_m)$.
   (b) According to the topic of the current location and the "topic - word" distribution parameter, Generate the word corresponding to the current position $w_{m,n} \sim \text{Cat}(\phi_{z_{m,n}})$.

![Fig.1 Graphical representation of the LDA model](image)

The document generation process assumed in LDA can also be represented in Figure 1. The double circles in the figure represent observed variables, single circles represent latent variables, and arrows represent conditional dependencies between variables. The box represents repeated sampling, $K$ is the number of topics, $M$ is the number of documents, $\tilde{\beta}$ is the Dirichlet prior parameter of the multinomial distribution of words under each topic, and $\tilde{\alpha}$ is the Dirichlet prior parameter of the multinomial
distribution of topics under each document. $Z_{m,n}$ is the topic of the nth word in the m-th document, $W_{m,n}$ is the nth word in the m-th document, and the remaining $\theta_m$ and $\phi_k$ respectively represent the topic distribution parameters of the m-th document and the word distribution parameters of the k-th topic.

3. Competency mining based on LDA model

Competency mining based on the LDA model mainly includes three links: data collection and preprocessing, topic discovery, topic extraction and analysis. The specific process is shown in Figure 2. Firstly, the collected data are stored to form a text corpus for analysis, and then the data are preprocessed, including data cleaning, word segmentation, removing stop words and word normalization. According to the language, content and quality of the corpus, the methods needed for data preprocessing are different. For example, for Chinese and Japanese language corpora, word segmentation is required for texts, while for English corpora, word normalization such as stem extraction is required. After data preprocessing, it is necessary to select appropriate methods to determine the optimal number of topics. Generally, the most commonly used methods are empirical method, calculation of topic perplexity, and calculation of topic coherence, and so on. After determining the optimal number of topics, the topic can be extracted according to the optimal number of topics. The extracted topic keywords represent the specific competency details. According to these topic keywords, the competency categories of each topic can be determined and named. Finally, the topic model of competency features established is explained.

![Fig.2 Mining competency process based on LDA model](image)

4. Experimental process and results

4.1. Data acquisition

"Academician" is the highest academic title in science and technology established by the state. It is a lifelong honor. In China, academicians include members of the Chinese Academy of Sciences and the Chinese Academy of Engineering (hereinafter referred to as academicians). By analyzing the reasons why academicians are competent for scientific research, we can discover the competency of Chinese scientific and technological workers. Therefore, the empirical analysis of this paper takes the interviews and biographies of the newly elected academicians in 2017 and 2019 as the data sources. The manual collection method is to ensure the accuracy of the data, and not to mix a large number of irrelevant materials to affect the results of text mining. After preliminary data collection, a total of 87...
academicians were obtained, including 52 academicians of Chinese Academy of Sciences and 35 academicians of Chinese Academy of engineering, with a total of 281089 words of text materials.

4.2. Data preprocessing
The collected data come from multiple heterogeneous data sources. In order to make it suitable for text mining tasks, we preprocess the data properly to eliminate redundant and unreliable data before formally entering the analysis process[6]. The pretreatment process consists of a series of successive steps.

First of all, the original text data is cleaned to remove the low value and repetitive contents such as interviewers' comments, subject and professional interpretation, and special symbols.

Then, the cleaned text data is segmented. Unlike English and other Western languages, there is no word separation mark in Chinese writing. Therefore, word segmentation is needed. This process is completed with the help of Jieba package in Python. Jieba is a popular Chinese word segmentation Library in Python. It supports three word segmentation modes: precise mode, full mode and search engine mode. This paper selects the precise mode to segment the text.

Finally, Remove the stop words in the text data. Although some words (such as "some", "so far", "since" and so on) appear frequently in the text, they have no practical significance for the interpretation of the text analysis, and even bring interference. Before the text mining, the stop words list is used to eliminate the stop words. In order to make the stop list more targeted, a series of stop words which have nothing to do with the interpretation of the results were added to the general stop words list of Harbin Institute of technology, such as "research", "remember", "patient", etc. It forms a special stop words list for the research.

After data preprocessing, the original text data is reduced to 187171 words, and the personnel structure in the data has not changed.

4.3. Determine the number of topics
At present, most studies use the index of perplexity to select the optimal number of topics in LDA model. However, Michael R and Both A et al. have proved that the coherence index is the most consistent measure of human interpretability[7]. Therefore, the topic coherence score is selected as the quantitative index of the optimal number of topics in the model. Generally speaking, the higher the topic coherence score, the better the quality of the model. It can be seen from the results in Fig. 3 that when the number of topics is 4, the topic coherence score is the highest, which is 0.469. From the perspective of coherence score, the optimal number of topics in LDA model is 4. In order to prevent too few topics from making the information in the topic too abstract and difficult to interpret, we tested several topics with higher coherence score, and after many experiments, when the number is 4, the model is the best, so the number of topics in the model is finally set to 4.

Fig.3 Topic coherence score on LDA model

4.4. Topic extraction and analysis
By solving the topic model, we get the topic of the competency of scientific and technological workers and the keywords that can represent the competency of scientific and technological workers. In order
to increase the interpretability of the topic, after obtaining the keywords, we first return to the original text, and assign meanings to the keywords according to the context. On this basis, we construct the sub topics and further aggregate them into topic names. The results are shown in TABLE I. Among them, the topic of competency is the category of the competency of scientific and technological workers, and the sub topic is the competency summarized according to the key words of the subject. The topic words reflect the details of the competency of science and technology workers in the text.

**TABLE I. TOPIC EXTRACTION RESULTS**

| Number | Topic      | Sub topic                                           | Topic Key Words                                                                 |
|--------|------------|-----------------------------------------------------|---------------------------------------------------------------------------------|
| 1      | Values     | patriotism, integrating theory with practice, career orientation | nation innovation social laboratory field China technology effort career time goal foundation experiment motherland |
|        | topic      | mathematical basis, professional knowledge, professional interests | Mathematics field knowledge major interest discipline physics direction foundation likes solve reading grades chemistry thinking |
| 2      | Knowledge  | responsibility, enterprise                          | do well excellent rigorous promote effort difficulty problem career foundation theory responsibility night technology knowledge likes |
|        | topic      | hardworking, diligent                                | Office night colleagues habit doctor laboratory explore honour foresight |
| 3      | Attitude   |                                                     | from beginning to end consciousness interest nationwide effort career |
|        | topic      |                                                     |                                                                                  |
| 4      | Personality|                                                     |                                                                                  |

Topic 1 is named as values topic. From TABLEI, we can see that the values of scientific and technological talents are mainly manifested in patriotism, integrating theory with practice and career orientation. From the list of key words, we can see that in the eyes of scientific and technological talents, national innovation, social needs, field development and career development are their primary objectives for scientific and technological research.

Topic 2 is named as knowledge topic, and the competency is manifested in the aspects of mathematical foundation, professional knowledge and professional interest. The general science and technology talents will show their ability in mathematics and physics at the beginning of their studies, and consolidate their professional basic knowledge under the guidance of their professional interests.

Topic 3 is named as attitude topic, and sense of responsibility and enterprise are important manifestations of attitude topic. Returning the key words to the original text, we find that the scientific and technological talents are conscientious and responsible for their work, aggressive, brave in challenging difficult problems, and have an objective and rigorous scientific attitude.

Topic 4 is named quality topic, including hardworking and diligence. Many scientific and technological talents mentioned in the interview that diligence and hardworking are the most important reasons for their success in their work. Their colleagues also mentioned that working all year round and forgetting to eat and sleep is their normal work. They even did not stop working when they were hospitalized due to illness.

In order to make the topic distribution clearer, pyLDAvis is used to draw the topic distribution map of the competency characteristics of scientific and technological workers, as shown in Figure 4. On the left side of Figure 4, circles represent different topics. The number of topics in the circles corresponds to Table II. The larger the circle size, the more common the topic is. On the right side of the figure is a list of 30 key words with the highest frequency. It can be seen from Figure 4 that topic 1 is the most common topic, which indicates that there are more scientific and technological talents who mention the theme of values in the text, followed by the competency of Topic 2 on knowledge dimension. Returning to the original text, it is found that this may be due to the fact that academicians are generally willing to talk about their educational experience in interviews or biographies, followed by topic 3 and topic 4, which are less common, attitude and quality competency features are mostly presented by the behaviors and events recorded in the text, which is more difficult to mine than topic 1 and topic 2.
5. Conclusion

In this paper, we propose a method of using LDA model to mine competency, and conduct experiments with Chinese texts related to the academic growth of academicians of Chinese Academy of Sciences and Chinese Academy of engineering. The results show that the idea of using LDA model to mine competency is feasible. Compared with the traditional behavioral event interview, questionnaire survey and other methods, LDA model makes use of the existing literature, which can achieve more real-time and efficient. However, it needs to be acknowledged that more efforts should be made in data cleaning and preprocessing, such as building a special word segmentation dictionary, in order to increase the interpretability of the results of topic extraction. In addition, this paper only uses Chinese text as the data source, which has certain reference value for related research, and provides ideas for competency research. The formation of the system method depends on more comprehensive data support.

Acknowledgments

This work was financially supported by the National Academy of Innovation Strategy project (2020020200015).

References

[1] McClelland, David C. Testing for competence rather than for "intelligence." [J]. American Psychologist, 1973, 28(1): 1-14.
[2] Spencer L M, McClelland D C, et al. Competency assessment methods: History and state of the art [M]. Hay/McBer Research Press, 1994.
[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of Machine Learning Research 3 Jan (2003): 993-1022.
[4] Luo L. Network Text Sentiment Analysis Method Combining LDA Text Representation and GRU-CNN[J]. Personal and Ubiquitous Computing, 2019, 23(3): 405-412.
[5] Zong Chengqing, Xia Rui, et al. Text data mining [M]. Beijing, China: Tsinghua University Press, 2019 (in Chinese).
[6] Garcia, Salvador, Julián Luengo, and Francisco Herrera. Data preprocessing in data mining. Cham, Switzerland: Springer International Publishing, 2015.
[7] Röder, Michael, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures." Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015.