Research on the Changing Trend of Employment-Relevant Terms Based on Internet Big Data Analysis

YANG Wei [1,a]

[1]Wuhan Qingchuan University, Donghu New Technology Development Zone, Wuhan, Hubei, China;

Abstract: With publicly-available data collected from mainstream information platforms, this study used the term frequency inverse document frequency (TF-IDF) algorithm to detect 74 popular terms and phrases about employment, analyzed the changes in the ranking of these terms and phrases, and visualized the changing trend in the attention to employment skills from 2017 to 2019. The research result will facilitate application of big data technology to teaching administration in colleges, and provide a guide for college students to plan their study of vocational skills.

1 Introduction

College students are a form of valuable human capital, full of vigor and creativity. Employment of college students is a matter concerned with a nation's political stability and economic sustainability. As reported by the National Bureau of Statistics , China's GDP witnessed an annual growth rate of 7.04%-11.47% from 2015 to 2019, but the number of college graduates in China grew by merely 0.69%-4.49% per year. The increase in the number of talents output by colleges has failed to catch up with the economic growth rate. Besides, as shown in the annual "Chinese College Student Employment Report" released by MyCOS , a third-party investigation agency, the employment rate of college students half a year after graduation stayed between 91% and 92% in five years, but presenting a slight decline, as shown in Table 1.

These numbers showcase two problems facing China's college education: first, the speed of talent training has been behind the economic growth rate; and second, the employment rate of college students has been declining year by year. It is not hard to tell the imbalance between the supply and demand of college students. This study attempts to visualize the changes in the attention to employment skills and capacities based on big data analysis. It is expected that this study could promote application of big data technology in teaching administration of colleges, improve supply-demand balance of college talents, and lead the Chinese economy to sustainable, healthy, and stable development.

Table 1 Comparison between college student employment and economic growth

	2015	2016	2017	2018	2019
Number of graduates form colleges and junior colleges (10,000)	680.9	704.2	735.8	753.3	758.5
Year-on-year growth rate of graduates from colleges and junior colleges	*3.26%*	*3.42%*	*4.49%*	*2.38%*	*0.69%*
Gross Domestic Product (GDP) (100 billion yuan)	688.8	746.3	832.0	919.2	986.5
Year-on-year growth rate of GDP	*7.04%*	*8.35%*	*11.47%*	*10.49%*	*7.31%*
Employment rate of college graduates half a year post graduation	92.0%	91.8%	91.9%	91.5%	91.1%

2 Materials and Methods

2.1 Materials

Tencent WeChat, Sina Weibo, and Toutiao are typical "we-media" Internet information platforms in China. These we-media platforms are, on one hand, content receivers, and on the other, distributors of secondary contents. Thus, trending topics spread fast on these platforms, and the data collected from these platforms can well reflect social trends. Therefore, document data collected and sorted by professional Internet opinion analysis agencies from these three platforms were analyzed in this study.

[a]yw0525@qq.com

2.2 Methods

In this study, the term frequency-inverse document frequency (TF-IDF) algorithm was used to process document terms. In term frequency, a higher frequency of a term in a single document indicates more prominence of the topic that the term represents. As for the inverse frequency, a term with a higher frequency in a single document and a low frequency in other documents indicates a better classification capacity. The calculation method is as follows:

$$TF_{xy} = \frac{The\ number\ of\ term\ (x) in\ a\ document(y)}{The\ number\ of\ total\ terms\ in\ a\ document(y)} \quad (1)$$

$$DF_x = \frac{The\ mumber\ of\ documents\ containing\ term\ (x)}{Total\ number\ of\ documents} \quad (2)$$

$$IDF_x = \log(\frac{1}{DF_x}) \quad (3)$$

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF} \quad (4)$$

3 Research implementation and results

3.1 Research implementation

In this study, 79,867 documents were collected containing three keywords “recruitment”, “employment”, and “job-hunting” publicly available on Sina Weibo, WeChat and Toutiao from January 1st, 2017 to December 31st, 2019, including Weibo blogs, subscription account articles, news and review texts. Using the TF-IDF algorithm, we removed terms and phrases irrelevant to vocational skills, and 74 terms and phrases related to vocational skills and capacities were selected. Terms of similar meanings are categorized into the same item. For instance, “word”, “excel”, “PowerPoint” and “ppt” are categorized into the item of “office”; and “collaboration”, “team” and “coordination” are categorized into the item “cooperation”. Finally, we obtained the top 20 terms and phrases each year, as shown in Table 2.

Table 2. Top 20 terms and phrases obtained by the TF-IDF algorithm

Ranking	2017	2018	2019
1	office	cooperation	office
2	cooperation	office	cooperation
3	communication skills	English proficiency	English proficiency
4	English proficiency	stress resilience	stress resilience
5	stress resilience	Sales experience	learning ability
6	Sales experience	communication skills	communication skills
7	Photoshop	copywriting skills	writing skills
8	learning ability	expression	internship experience
9	copywriting skills	writing skills	expression
10	business trips	learning ability	copywriting skills
11	writing skills	internship experience	Sales experience
12	programming skills	logical thinking	psychology
13	pressure	dancing	python
14	expression	formal writing	new media operation
15	office	python	pressure
16	SPSS	marketing	logical thinking
17	database	law	business trips
18	CAD	survey	student leadership experiences
19	logical thinking	database	office
20	internship experience	short video production	data modeling

3.2 Research results

As Table 1 shows, five terms, i.e., “office”, “cooperation”, “English proficiency”, “stress resilience” and “communication skills”, remain top in the ranking, and also among the top-ranking terms are “sales experience”, “copywriting skills”, “writing skills”, and “learning abilities”. It is not hard to see that office software, English proficiency and writing skills are vocational skills valued by modern enterprises. The great value that enterprises attach to cooperation, communication skills, learning abilities and stress resilience reflect the complicated labor division, fast updating, and high efficiency of the modern job market.

There are some other terms that occur in the table, such as “business trips”, “programming”, “SPSS”, “database”, “CAD”, “dancing”, “Python”, “short-video production”, “new media operation”, and “data modeling”, the changes of which in the ranking have reflected the changes in people’s recognition in these skills across the years. These changes are worth attention from both college teachers and students.

4 Limitations

The limitations of this study are as follows. First, the documents studied are from “we-media” Internet information platforms, and it is very likely that the same documents are released multiple times by different users; second, due to the limits in resources, only data of three years from 2017 to 2019 are analyzed, and the analysis result fails to show obvious trends in the changes of hot social topics; third, the research results reflect the social hot topics, and if we want to apply these results to

curriculum design and students' career planning, the real-world conditions should be considered.

5 Conclusions

This study has probed into the term frequency of terms about employment skills on "we-media" Internet information platforms in China from 2017 to 2019, analyzed the changes in the ranking of these terms, and explored the changes in the hot topics about employment skills on the Internet. Through description of the research process, the data processing methods are presented. It is expected that researchers in the future can, based on this study, collect a larger data set, process and visualize the data in a more precise manner, and in this way, the changes in the terms about employment on Internet media can be demonstrated to help colleges and students better understand the needs of the job market. With more research efforts, we can improve the balance between the supply of talents from colleges and the demand of talents in the job market, better leverage the demographic dividends and boost China's economy.

Acknowledgment

Funding: This research is funded by the Scientific Research Plan of Department of Education of Hubei Province (Grant No.: B2018398)

References

1. ZHOU, Y.J DENG, D.P CHI, J.H. (2021) A Short Text Classification Algorithm Based on Semantic Extension. Chinese Journal of Electronics, 30:153-159.
2. XIN, L. TANG, F.C. LI, M.Y ZHOU, W.X (2020) From School to Work: Improving Graduates'Career Decision-Making Self-Efficacy. Sustainability, 12:804.
3. LU, G.S. SONG, Y.P. PAN, B.C. (2021) How University Entrepreneurship Support Affects College Students' Entrepreneurial Intentions: An Empirical Analysis from China. Sustainability, 13:3224.
4. QI, L. AN, X.J. ZHANG, S. WANG, X. (2020) Research on Knowledge Gap Identification Method in Innovative Organizations under the "Internet+" Environment. Information 11:572.
5. Park, T.J, Kim, C.Y. (2020) Predicting the Variables That Determine University (Re-)Entrance as a Career Development Using Support Vector Machines with Recursive Feature Elimination: The Case of South Korea. J. Sustainability, 12: 7365.
6. Kim, H.Y. Han, Y.S. Song, J.Y, Song T.M. (2019) Application of Social Big Data to Identify Trends of School Bullying Forms in South Korea. International Journal of Environmental Research and Public Health, 16: 2596.
7. Sumayh S. Aljameel, Dina A. Alabbad, Norah A. Alzahrani, Shouq M. Alqarni, Fatimah A. Alamoudi, Lana M. Babili, Somiah K. Aljaafary and Fatima M. Alshamrani. (2021) A Sentiment Analysis Approach to Predict an Individual's Awareness of the Precautionary Procedures to Prevent COVID-19 Outbreaks in Saudi Arabia. International Journal of Environmental Research and Public Health, 18: 218.
8. Viera Maslej-Krešnáková, Martin Sarnovský, Peter Butka, Kristína Machová.(2020) Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification. Applied Sciences,10:8631
9. Aldo Mascareño, Pablo A. Henríquez, Marco Billi, Gonzalo A. Ruz.(2020) A Twitter-Lived Red Tide Crisis on Chiloé Island, Chile: What Can Be Obtained for Social-Ecological Research through Social Media Analysis. Sustainability, 12(20): 8506
10. Nizar Ahmed, Fatih Dilmaç, Adil Alpkocak.(2020) Classification of Biomedical Texts for Cardiovascular Diseases with Deep Neural Network Using a Weighted Feature Representation Method. Healthcare, 8(4): 392
11. Omar Sharif, Mohammed Moshiul Hoque, A. S. M. Kayes, Raza Nowrozy, Iqbal H. Sarker.(2020) Detecting Suspicious Texts Using Machine Learning Techniques. Applied Sciences, 10(18):6527
12. Tiancheng Tang, Tianyi Yuan, Xinhuai Tang, Delai Chen.(2020) Incorporating External Knowledge into Unsupervised Graph Model for Document Summarization. Electronics , 9(9): 1520
13. Konstantinos Demertzis, Konstantinos Tsiknas, Dimitrios Takezis, Charalabos Skianis, Lazaros Iliadis.(2021) Darknet Traffic Big-Data Analysis and Network Management for Real-Time Automating of the Malicious Intent Detection Process by a Weight Agnostic Neural Networks Framework. Electronics, 10(7):781;
14. Zhenjie Zhu, Bingjun Liu, Hailong Wang, Maochuan Hu.(2021) Analysis of the Spatiotemporal Changes in Watershed Landscape Pattern and Its Influencing Factors in Rapidly Urbanizing Areas Using Satellite Data. Remote Sens,13(6): 1168
15. Abdul Karim, Azhari Azhari, Samir Brahim Belhaouri, Ali Adil Qureshi, Maqsood Ahmad. (2020) Methodology for Analyzing the Traditional Algorithms Performance of User Reviews Using Machine Learning Techniques. Algorithms, 13(8): 202
16. Myung-Bae Park, Ju Mee Wang, Bernard E. Bulwer. (2021) Global Dieting Trends and Seasonality: Social Big-Data Analysis May Be a Useful Tool. Nutrients, 13(4):1069
17. Hyun-Jin Kim, Ji-Won Baek, Kyungyong Chung. (2020) Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score. Applied Sciences, 10(13): 4590

18. Isabella Gagliardi, Maria Teresa Artese.(2020) Semantic Unsupervised Automatic Keyphrases Extraction by Integrating Word Embedding with Clustering Methods. Multimodal Technol. Interact, 4(2): 30
19. Viju Raghupathi, Jie Ren, Wullianallur Raghupathi. (2020) Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets. International Journal of Environmental Research and Public Health, 17(10):3464
20. Tedo Vrbanec, Ana Meštrović. (2020) Corpus-Based Paraphrase Detection Experiments and Review. Information, 11(5): 241
21. Sonali Rajesh Shah, Abhishek Kaushik, Shubham Sharma, Janice Shah.(2020) Opinion-Mining on Marglish and Devanagari Comments of YouTube Cookery Channels Using Parametric and Non-Parametric Learning Models. Big Data Cogn. Comput, 4(1): 3
22. Binod Kumar Adhikari, Wanli Zuo, Ramesh Maharjan, Xuming Han, Shining Liang.(2020) Detection of Sensitive Data to Counter Global Terrorism. Applied Sciences, 10(1):182
23. Ahmad Hawalah.(2019) Semantic Ontology-Based Approach to Enhance Arabic Text Classification. Big Data Cogn. Comput, 3(4): 53
24. Celestine Iwendi, Suresh Ponnan, Revathi Munirathinam, Kathiravan Srinivasan, Chuan-Yu Chang. (2019) An Efficient and Unique TF/IDF Algorithmic Model-Based Data Analysis for Handling Applications with Big Data Streaming. Electronics, 8(11): 1331;
25. Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, Miguel Camacho-Collados. (2019) Detecting and Monitoring Hate Speech in Twitter. Sensors 19(21): 4654
26. Lixia Xie, Ziying Wang, Yue Wang, Hongyu Yang, Jiyong Zhang. (2018) New Multi-Keyword Ciphertext Search Method for Sensor Network Cloud Platforms. Sensors, 18(9): 3047