Revealing the Country-level Preference on Research Methods in the Field of Digital Humanities: From the Perspective of Library and Information Science *

Chengxi Yan [1,*], Zhichao Fang[2]

[1] School of Information Resource Management, Renmin University of China, Beijing, China, 100872
[2] Digital Humanities Research Center, Renmin University of China, Beijing, China, 100872

Abstract

Research method is a very important element for both individual scientific research and country technological development, especially for those interdisciplinary fields like digital humanities (DH) that is close to library and information science (LIS). Considering the scarcity of relevant training data, this study proposes a multi-stage recognition algorithm combining large language model and iterative learning strategy to automatically extract method mentions from DH scientific documents. According to the taxonomy of RMs in existing LIS research, we used dictionary-based mapping technology to transform these entities into RMs and their types. To clarify the differences in RM preferences across different countries, we identified the countries and established the relationship between them with the RMs. A clustering model was utilized to detect country-level RM preference. The experiments showed that quantitative research has played an increasingly central role in the international DH field, especially the experimental methods. Also, there is a distinctive distribution for RM preference among different countries.

Keywords

Bibliometric analysis, research methods, entity recognition, digital humanities

1. Introduction

For the majority of scientific researchers, identifying and understanding the research methods (RMs) in different scientific fields is not only a necessary academic basic skill, but also a significant reference for deeply getting the whole picture of its development or solving domain problems [1]. As the Stanford Encyclopedia of Philosophy defined RM as "the means of how the aims and products of science are achieved, which should be distinguished from meta-methodology and the detailed and contextual practices" [2].

The distinct characteristics of scientific approaches, technical standards and application norms can be reflected on the different use of RMs across various countries. Therefore, the comparative analysis of preference variation on RMs between countries will be conducive to a more systematic and efficient evaluation of national scientific strength and innovative ability. Moreover, it promotes the country-level awareness of the strengths and weaknesses in both international academic collaboration and competition. With the rapid development of entitymetrics-based approaches [3], the identification and measurement of RMs has become one of the hot research issues, especially for some interdisciplinary fields that integrate a large number of different technologies and methods such as rule-based and deep neural network-based methods. However, it remains highly challenging for accurately identifying all different types of RMs, due to the limitation of training corpus annotated by RM-related entities for supervised models and low prediction performance in the unsupervised way.

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2024) and the 4th AI + Informetrics (AII2024)
* Corresponding author.
20218113@ruc.edu.cn (C. Yan); fzc0225@163.com (Z. Fang)
0000-0003-1128-550X (C. Yan); 0000-0002-3802-2227 (Z. Fang);
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In addition, most previous studies analyzed the usage frequency of RMs in the field of library and information science (LIS), which ignored hot interdisciplinary fields related to LIS like digital humanities (DH) and the difficulty of their RM classification. As a research area that is inherently methodological and heavily indebted to LIS [4], DH is often viewed as a "big tent" [5] including different disciplines with an extensive range of RMs. Considering the interdisciplinary nature of DH and the close relationship between it and LIS, this study adopted DH as the analytical object. According to it, three research questions (RQs) are proposed as the following: ***RQ1***: From a global perspective, does DH research tend to be qualitative, quantitative, or mixed and which countries are the typical representation for these three method types? ***RQ2***: What are the differences in the preference of RMs among different countries? ***RQ3***: Is there a certain pattern for the country-level preference of RMs?

2. Related Work

The approaches of automatic recognition for RM entities can be divided into two main stages, namely rule-based [6-8] and machine learning-based technology. For example, Zha adopted the abbreviation patterns and regular expressions to extract candidate algorithmic entities [6]. Considering the weakness of recognition performance, more researchers turned to the approaches of machine learning [9-13]. In Zhang et al.' study [9], software entities from the *PLOS ONE* full texts were identified and reorganized into five different groups using a clustering algorithm. Wang et al. constructed a term function identification model based on the deep learning (DL) [10].

The classification of RMs can be traced back to the early study based on the content analysis in the LIS field, such as Jarvelin et al.'s systematic categorization [14]. Hider [15], Kumpulainen [16], and other LIS scientists who adopted and optimized Jarvelin's classification theory of RMs further reported on the use of RMs in the long-term evolution of the LIS field. One of the most influential studies was done by Chu and Ke [17], in which three representative LIS journals were coded computed and analyzed, yielding 16 RMs. This classification scheme has promoted a variety of development for RMs, such as the influence analysis of algorithmic entities [7], the exploration of dynamic evolution of RMs in the Chinese LIS field [18], and the survey of RMs in the practice projects [19].

3. Research Design

To answer the RQs, we proposed a research analytical framework, as shown in Figure 1, including three main steps. The latter two steps are the most crucial components.

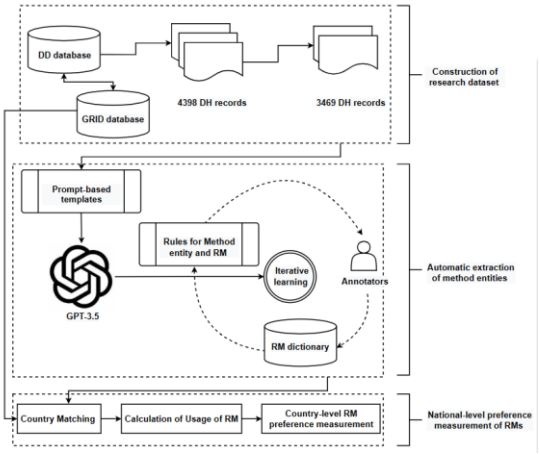

Figure 1: The entire research framework.

3.1. Construction of research dataset

To obtain the original scientific DH papers, similar to previous studies [20, 21], we used the subject term-based query strategy (including titles, abstracts and keywords) as ("digital humanit*" OR "humanit* comput*" OR "ehumanit*" OR "electronic* humanit*" OR "e-humanit*") in three well-known databases (Web of Science Database, Crossref Database and Dimensions Database, DD) to search as many relevant documents as possible. The publication timespan is set between 1900 and 2021. According to the comparative results, we found that DD almost covered all the records from the other two datasets mentioned above (mainly journal articles), and more importantly had a wide range of source types such as books, proceeding or preprint papers, and monographs. Thus, the DD database was selected as the source for the acquisition of dataset. There was a total of 4398 articles in the initial dataset. Next, we deduplicated and deleted irrelevant document records from it, finally resulting in 3469 papers.

To identify the country names in each paper, we utilized a huge global database called "GRID (Global Research Identification Database)", which is one of the most popular open repositories of authoritative research institutions. We used GRID as an institutional dictionary to link the institution entities where the authors (in DD records) are located in to its corresponding countries. The processed dataset consists of 1915 papers.

3.2. Automatic extraction of method entities

Given to the linguistic complexity (e.g. contextual features of method entities) in DH documents with dual humanities and technological aspects, the identification of RMs may confront higher technical difficulties. We proposed a three-stage method for automatic entity extraction. Firstly, we constructed prompt-based templates using a large language model (GPT-3.5) to complete zero-shot learning, which generated a coarse-grained annotation results of method entities. Secondly, a vocabulary containing normal method terminologies and their variations (e.g. abbreviation, synonyms) were built through manual collection and multiple rounds of expert evaluation. Inspired by Gupta and Manning's work [8], we next designed an iterative learning process to identify and correct method entities on the above resulting dataset in the human-in-the-loop way, where the rule-based transformation and classification for RMs were performed. During the process of RM conversion, each method entity was automatically "translated" to a regular RM in the Chu and Ke's taxonomy [17] if a rule is matched.

Given the pattern of RMs to be centrally observed and induced in the field of DH, we divided them into three categories in the wider scope in the light of Jarvine et al.' research [14], namely qualitative research, quantitative research, and mixed research. For instance, qualitative research includes "content analysis", "ethnography and field study", "historical method", "interview" and etc., while representative quantitative research are "experiment", "think aloud protocol" and "transaction log analysis", "bibliometrics". A study was judged to be "mixed" only when both types of RMs (at least one) are used.

3.3. National preference of RMs

For each DH record, we used the regular expression to match all authors and their institutions. A simple program was then designed to map them to the relevant countries based on the organizational names in the GRID database. Considering the issue of multi-path relationship between records and countries, we calculated it according to [22], in which a country used a method once when relevant RMs were mentioned in a paper regardless the occurring frequency of countries that correspond to its authors [22]. The cumulative counts of RM usage for a country were ultimately defined as its preference of RMs. Moreover, we used an agglomerative hierarchical clustering algorithm [23] to distinguish different country-level preference patterns.

4. Result Analysis

For the proportion of three types of RM, quantitative approach is observed as the most mainstream approach, which takes 82.29% records in the dataset. Compared to the qualitative approach, mixed approaches (i.e. 9.93%) turn to be slightly more common in DH research. Considering the increasingly growth of DH papers [21], it is believed that quantitative analysis is becoming a more and more important research means.

Specifically, the dominant position of the Western countries for DH studies is indisputable (seen Table 1). The United States, which has the most frequent use of RMs, has the most significant superiority compared to other countries. The United Kingdom and Germany are in the second place, especially Germany, which has a clear position of leadership in the qualitative type. The third-rank group are comprised of China, the Netherlands, Canada, Spain, Australia, and Spain. It is worth noting that as one of the few Asian countries on the list (except for Singapore and Israel), China's outstanding performance in mixed and quantitative research is quite impressive, possibly due to its diversified use of RMs in the field of DH.

Table 1
The ranking of total number of RM types used by the top 5 countries. Note: US, UK, NL, GER and SGP is for short of United States, United Kingdom, Netherlands and Germany and Singapore, respectively.

Rank	mixed	qualitative	quantitative
1	US (24)	US (29)	US (166)
2	UK (10)	GER (7)	GER (106)
3	China (9)	UK (7)	UK (86)
4	NL (7)	Australia (5)	China (48)
5	Canada (5)	SGP (3)	Italy (41)

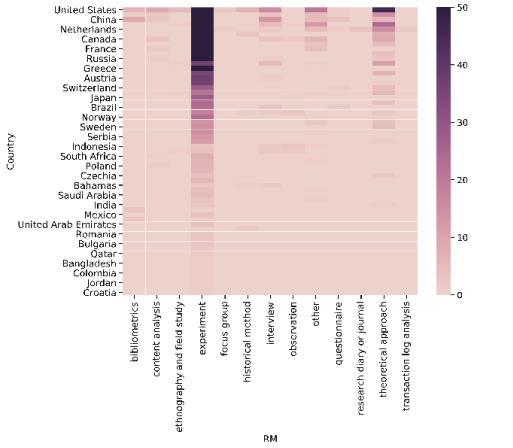

Figure 2: The statistics of national preference of RMs.

As a sign of quantitative approach, "experiment"-based approaches are the most frequently utilized (seen in Figure 2). This can be inferred according to the visual analysis on the RM usage, because the relative rates of RM usage can reach 76.98%-94.94%. Even if the samples are expanded to those countries with a usage frequency greater than 10, it also exceeds 60%. Thus, we temporarily exclude the RM of "experiment".

According to Figure 3, theoretical approaches, as the most important qualitative methods, stand out from the remaining RMs. They are frequently used by most developed countries in Europe and America, such as the United States and the United Kingdom. Chinese DH scholars seem to show more keen interest in bibliometric methods, while the Americans prefer content analysis. Both of these two countries show great attention to "interview". "observation", "transaction log analysis", "research diary or journal", and "focus group" are not highly valued by the mentioned countries. One or Two RMs are used in other lower-ranked countries, especially Finland, which is the only country with the most intensive preference for theoretical approaches. Relatively speaking, the choice of RMs is more evenly distributed for Canada, indicating that Canadian attitude towards qualitative methods may be more tolerant.

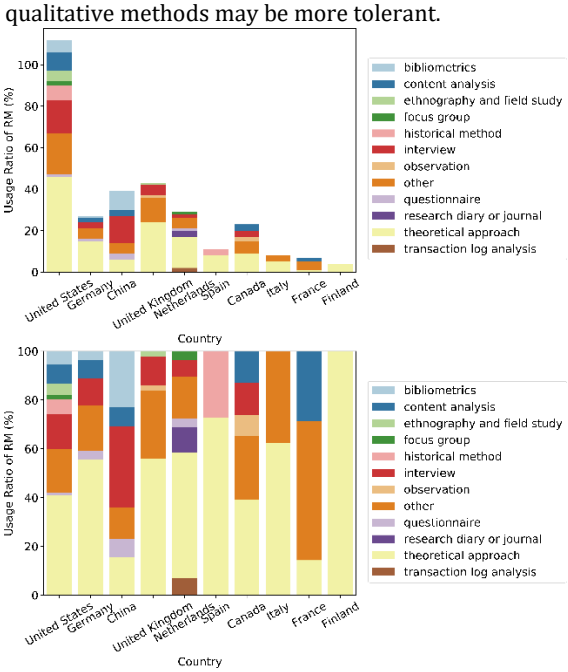

Figure 3: Preference choices of RMs measured by usage frequency (up) or usage ratio (down) in representative countries.

The clustering results of the RM preference is shown in Figure 4. There are three clusters in the entire field of DH, namely #1 (United States), #2 (Germany, China, and United Kingdom), and #3 (Other Countries).

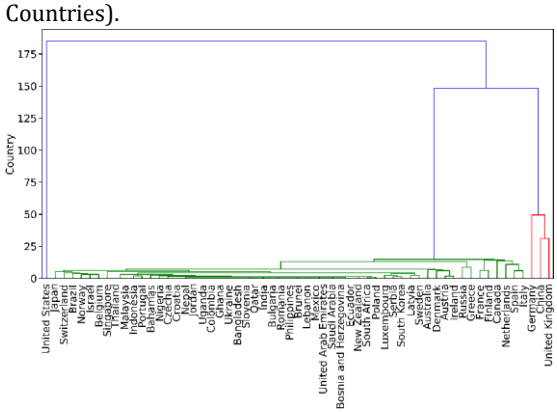

Figure 4: Different clusters of countries based on RM-related preference.

For the four RMs including "**E**xperiment", "**T**heoretical approach", "**O**thers", and "**I**nterviews" (i.e, **ETOI** approaches), the above group division based on the machine learning macroscopically and clearly provide informative results for different national preference level of RMs. #1 is the group with the strong preference of ETOI approaches. #2 and #3 are the medium-level and weak-level preference groups, respectively. Furthermore, there are some difference features among the three groups. Content analysis is heavily weighted in #1. #2 are more likely to use bibliometric analysis in DH scholarship. By contrast, #3 focuses on observational methods. The difference is not only related to the comprehensive performance of each cluster, but is also greatly influenced by the unique members in it whose preference polarity are quite overpowering, such as China's preference (in #2) of bibliometric methods.

5. Conclusion

In this paper, we proposed an optimized iterative learning for RM extraction combing large language models and rule-based transformation to extract and classify RMs from a constructed DH dataset. We compared the differences in the preference use of RMs of different countries, which revealed the distinctive country-level preference patterns of DH. As a preliminary study, our findings can provide certain guidance and assistance for further improving the level of DH development.

Acknowledgements

This project is supported by the grant from National Natural Science Foundation of China (NO. 72204258).

References

[1] Zhang, H., Zhang, C. (2021). Using Full-text Content of Academic Articles to Build a Methodology Taxonomy of Information Science in China. Knowledge Organization, 48, 2: 126-139.

[2] Hepburn, B., & Andersen, H. (2021). Scientific Method. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2021). Metaphysics Research Lab, Stanford University.

[3] Ding, Y., Song, M., Han, J., et al. (2013). Entitymetrics: Measuring the impact of entities. PloS One, 8(8): e71416.

[4] Poole, A. H. (2017). The conceptual ecology of digital humanities. Journal of Documentation, 73(1), 91-122.

[5] Jockers, M. and Worthey, G. (2011), Introduction: welcome to the big tent, Proceeding of Digital Humanities 2011 Conference, 6-7.

[6] Zha, H., Chen, W., Li, K., & Yan, X. (2019). Mining Algorithm Roadmap in Scientific Publications, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1083-1092.

[7] Wang, Y., Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. Journal of informetrics, 14(4), 1-21.

[8] Gupta, S., Manning, C. D. (2011). Analyzing the dynamics of research by extracting key aspects of scientific papers. Proceedings of 5th international joint conference on natural language processing, 1-9.

[9] Zhang, H., Ma, S., and Zhang, C. (2019). Using Full-text of Academic Articles to Find Software Clusters. Proceedings of ISSI, 2776-2777.

[10] Wang, J., Cheng, Q., Lu, W., et al. (2023). A term function–aware keyword citation network method for science mapping analysis. Information Processing & Management, 60(4), 103405.

[11] Zhang, H., Zhang, C., and Wang, Y. (2024). Revealing the technology development of natural language processing: A Scientific entity-centric perspective. Information Processing & Management, 61(1), 103574.

[12] Zhang, C., Tian, L., and Chu, H. (2023). Usage frequency and application variety of research methods in library and information science: Continuous investigation from 1991 to 2021. Information Processing & Management, 60(6), 103507.

[13] Boland, K., and Krüger, F. (2019). Distant supervision for silver label generation of software mentions in social scientific publications. Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, 15-27.

[14] Jarvelin, K., Vakkari, P. (1990). Content analysis of research articles in library and information science. Library & Information Science Research, 12, 395–421.

[15] Hider, P., & Pymm, B. (2008). Empirical research methods reported in high-profile LIS journal literature. Library & Information Science Research, 30, 108–114.

[16] Kumpulainen, K. (1991). Library and information science research in 1975. Libri, 41(1), 59–76.

[17] Chu, H., Ke, Q. (2017). Research methods: What's in the name?. Library & Information Science Research, 39(4), 284–294.

[18] Lou, W., Su, Z., He, J. et al. (2021). A temporally dynamic examination of research method usage in the Chinese library and information science community. Information Processing & Management, 58(5), 102686.

[19] Lund, B. D., Wang, T. (2021). An analysis of research methods utilized in five top, practitioner-oriented LIS journals from 1980 to 2019. Journal of Documentation, 77(5), 1196-1208.

[20] Tang, M. C., Cheng, Y. J. and Chen, K. H. (2017). A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses. Scientometrics, 113(2), 985-1008.

[21] Su, F. and Zhang, Y. (2021). Research output, intellectual structures and contributors of digital humanities research: a longitudinal analysis 2005–2020. Journal of Documentation, 78(3), 673-695.

[22] Sidone, O. J. G., Haddad, E. A. and Mena-Chalco, J. P. (2017). Scholarly publication and collaboration in Brazil: The role of geography. Journal of the Association for Information Science and Technology, 68(1), 243-258.

[23] Sasirekha, K. and Baby, P. (2013). Agglomerative hierarchical clustering algorithm-a. International Journal of Scientific and Research Publications, 83(3), 83.