Characterizing Emerging Technologies of Global Digital Humanities Using Scientific Method Entities⋆

Shaojian Li [1,2], Chengxi Yan [1,2,*]

[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

Emerging technologies support the evolvement of disciplines. Scientific method entities, as proxies of emerging technologies, provide a framework for the development of emerging technologies. Therefore, identifying and extracting scientific method entities is an important link in the study of emerging technologies. The field of digital humanities is inherently interdisciplinary, combining traditional humanities disciplines with digital tools and technologies. Thus, it is particularly important for scholars in digital humanities to stay up-to-date with emerging -technologies, as they have the potential to transform the way that we approach research and scholarship. However, there are still some problems for extracting and evaluating the emerging technologies, peculiarly in the field of digital humanities. To address these issues, this paper proposes an AI-based method to automatically extract scientific method entity, and also deeply analyzed the specific situation of emerging technologies in the field of digital humanities.

Keywords

Emerging technologies, Scientific method entity, Digital humanities, Entity extraction and evaluation

1. Introduction

The emerging technology (ET) refers to those technical innovations which represent progressive developments within a field for competitive advantage [1], they are science-based innovations that have the potential to create a new industry or transform an existing one [2]. On this basis, we believe that ETs are dominant innovative technologies or methods emerging in a specific field in a specific period. Therefore, studying ETs can quickly understand the development of a certain research field, especially some emerging interdisciplinary fields, such as digital humanities (DH). The scientific method entity (SME) is an extensively researched object in various research fields. As good proxies for ETs, they provide an objective and systematic approach to characterizing ETs of global DH.

However, there are two main issues in existing research. First, traditional methods of extracting SMEs mostly use topic models and some manual methods [3]. But the topic terms are usually general, they may not have a specific meaning and not enough to be explained. Also, manual methods are relatively labor-intensive, especially in the age of information explosion, this is not a sustainable way. Second, the previous studies about DH only focus on the landscape of knowledge topics and structures in DH, while the detailed features of ETs (e.g. knowledge distribution, temporal evolution, etc.) are still unknown. Specifically, we utilized specific SMEs extracted from DH documents to represent ETs instead of topics, and give a deep analysis of ETs based on their bibliometric relationships. The contribution of our research are two-fold: One is a newly-designed approach based on AI-enchanced algorithm to automatically extract SMEs in DH domain; The other is a feature analysis on knowledge patterns related to ETs in DH in both static and dynamic ways.

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online

*Corresponding author.
EMAIL: lishj27@ruc.edu.cn (Shaojian Li);
20218113@ruc.edu.cn (Chengxi Yan)

Therefore, we explore two research questions:

RQ1: How to identify and extract ETs.

RQ2: How is the distribution and evolution of ETs in the domain of DH?

2. Related work

Currently, the existing work on extraction and evaluation of knowledge entities has received widespread attention [4]. There are four frequently utilized methods in method entity extraction: manual annotation, rule-based extraction, traditional machine learning, and deep learning [5]. Each method has its own prons and cons. Manual annotation is precise but inefficient. Rule-based extraction has high data processing ability but lack fexibility. Statistical machine learning is more fexible but relies on feature engineering. Deep learning method has strong versatility, but needs training corpus.

To reveal the intellectual structure of DH, previous relevant studies focused on the task of topic extraction, in which bibliometric analysis approaches are frequently used. For example, Tang used the TF-IDF algorithm to identify those research topics with higher discriminative value based on author assigned keywords [6]. In Wang' research [7], the keywords co-occurrence network can macroscopically present the distribution of hot topics in DH, where each one is regarded as a group of interrelated descriptors. Similar to Wang, Su et al. have further expanded the sources of topic candidates that are not limited to keywords but rather representative terms from titles and abs

tracts, making the results of domain topic analysis more comprehensive. It is clearly observed that the recognized topic terms turn out to be a collection of general hot concepts for DH [8]. The specific pattern of ETs in the DH field is still abstract and unclear, especially due to the extracted high-frequency words that do not have detailed meaning such as "digital humanities", "cooperation", "communication" and "data".

3. Methodology

3.1. Main Framework

Based on the above analysis, we propose a solution, which includes two parts, namely the ET extraction and the ET analysis. The ET extraction part includes extracting SMEs candidate words through AI-based algorithm, and then perform lemmatization, stem extraction and filtering on them, and finally obtain ET units. The ET analysis part is mainly about the SMEs dictionary obtained after ET units are mapped to the DH collection, and the SMEs' distribution, the clustering based on ETs co-occurrence and the evolution based on word frequency.

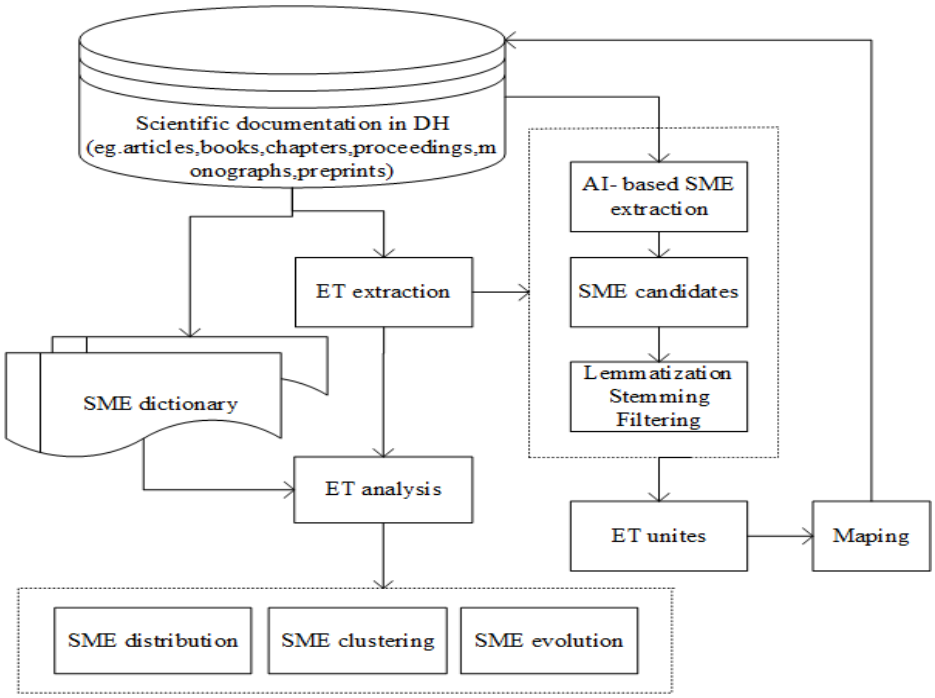

Figure 1: Framework of our work

3.2. AI-based Extraction of SMEs

To overcome the shortcomings of feature engineering-based topic extraction and rule-based recognition of scientific method knowledge, we propose an AI-empowered semi-automatic extraction method (ASAEM). This optimized approach is essentially a two-stage pipeline procedure. In the first stage, instead of manual judgement, a state-of-the-art super language model "ChatGPT"[1] is used to process documents and derive method-related entity candidates which are the fundamental units to build a dictionary of method entities. Since the key of ASAEM is to determine the most suitable prompt, an efficient automatic detection mechanism of the optimal template is designed in the algorithm. In the second stage, we leverage the above dictionary to match and extract all the method entities from DH collections whilst excluding a small amount of general high-frequency terms, which can help to significantly improve the algorithmic recall. Figure 2 presents the pseudo-code of ASAEM.

Algorithm 1 ASAEM

Input A DH document set $D = \{D_1, D_2, \cdots D_i, \cdots, D_N\}$ and N denotes the total number of the set.
Initialize a dictionary of method entities X.
for each $D_i \epsilon [1, N]$ **do**
    Constructing a collection of instruction-based templates as $T = \{t_j | 1 \ll j \ll M\}$.
    Generate the prompt input $P_i(T)$ based on T for D_i by combing its abstract and titles.
    for each $p_i^j \epsilon P_i(T)$ **do**
        Entering the prompt p_i^j using "text avinci-003" model and set the relevant parameters including max_tokens=2000, temperature=0, and top_ p=1.
        Obtaining output result O_i^j and predictive score $S(p_i^j)$.
    end for
    Computing the best result $O_i^{j^*}$ which satisfies $j^* = \underset{p_i^j \epsilon P_i(T), j \epsilon [1,M]}{\operatorname{argmax}} S(p_i^j)$.
    for each $w \epsilon O_i^{j^*}$ **do**
        Inserting a term w into X if w not in X.
    end for
endfor
Using X to match and extract method entities me_i^k in D where me_i^k denote the k-th method entity in D_i.
Output: All the method entities me_i^k.

Figure 2: The algorithm of ASAEM

3.3. Detection and Clustering of ETs in DH

We extracted a total of 24,306 SME candidates from DH collections, then screened them with the rule that the word length is less than 40 and the word frequency is greater than 3, and finally got 846 SMEs. Then we use the WordNetLemmatizer toolkit in NLTK to restore the lemmatization of these SMEs in order to remove the problem that the same word but different morphological forms of words are generated due to the number, tense, voice, etc.

We then proceeded to use NLTK to stem these re-morphed words for better mapping to the original text. In the next step, we matched some words based on pattern matching rules, such as abbreviations, synonyms, etc., and filtered some

[1] https://openai.com/blog/chatgpt

meaningless words then we got many ET unites and we use them to map the content of each DH collection and get the SME dictionary. After doing so, we get the distribution of ETs in the DH from 1903 to 2022. We mainly performed ETs co-occurrence clustering analysis and word frequency-based evolution over time on SMEs. We take relative strength index Equivalence coefficient as weights for edges and word frequency for nodes, importing them into gephi and use modular clustering to get clustering results.

4. Empirical Studies

4.1.Dataset and Implementation Details

Considering the interdisciplinary characteristics of DH, we first conducted a preliminary exploration of the data source for DH documents. Our goal is to obtain as many relevant documents as possible, which means it must not only require the largest quantity but also abundant document types. After comparing the number of retrieved papers from three well-known databases (i.e. Web of Science Database, Crossref Database and Dimensions Database), the Dimensions Database is selected to collect data. The query (digital humanit* OR humanit* comput* OR ehumanit* OR e-humanit*) is adopted to search the field of title, keywords and abstract, which yields 4398 documents. Through the removal of duplicate and unrelated documents, we finally obtain 3469 ones as the initial "target set". The descriptive statistical for the set can be seen in Figure 3.

We can see that on the one hand, compared to the number of documents, the number of DH-related terms showing an upward trend (before 2020) seems to be at a much larger order of magnitude, which implies possible significant errors of direct term-based extraction of ET. On the other hand, the types of documents in our chosen dataset are relatively rich, and can almost cover various records in the DH field.

4.2 Result Analysis

4.2.1. Distribution and Clustering of ETs

Many systems of scientific interest can be represented as networks, sets of nodes or vertices joined in pairs by lines or edges. Many networks of interest in the sciences are found to divide naturally into communities or modules. The problem of detecting and characterizing this community structure is one of the outstanding issues in the study of networked systems [9]. We performed a co-ET cluster analysis on the ET distribution results. If the word pair (W_i,W_j) does not co-occur in the document collection, the direct correlation strength $E_{i,j}$ is counted as 0. If there is a co-occurrence relationship, this paper uses the relative strength index Equivalence coefficient [10] to evaluate the word pair frequency. Inclusive processing, the multi-valued matrix is converted into a correlation matrix form with element values between [0,1], as shown in the following Formula. 1 :

$$E_{i,j} = \frac{G^2_{i,j}}{G_i * G_j} \quad (1)$$

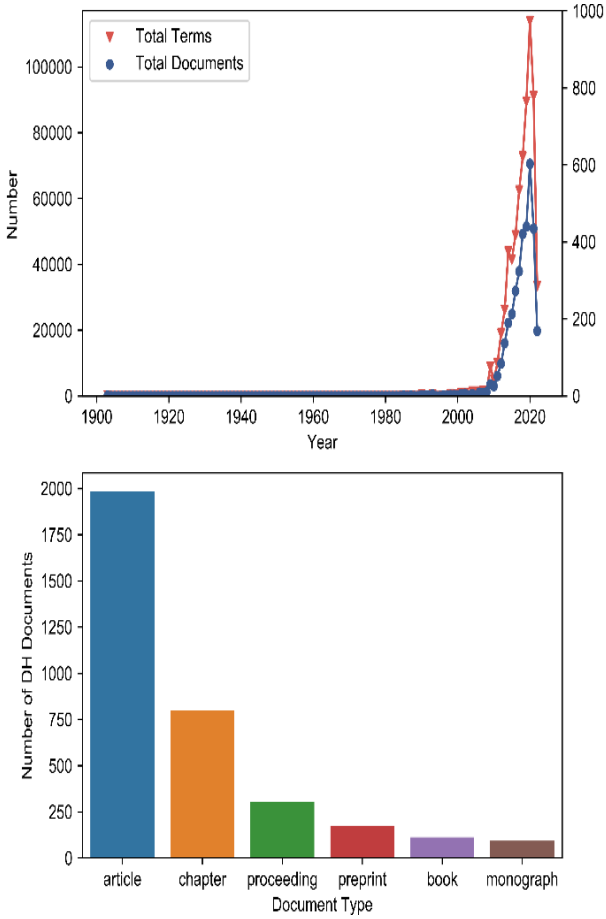

Figure 3: The basic statistics of DH dataset

Among them, $G_{i,j}$ represents the co-occurrence times of keywords W_i and W_j in the same document, and G_i, G_j represent the total frequency of keywords W_i and W_j respectively. The calculation results are used for modular clustering of method entity distributions in the DH literature, acting as weights for edges established between ETs' network.

From the clustering results of ETs, we can see that DH is a typical interdisciplinary, which contains very diverse ETs, including all aspects of data processing. (i.e. data collection, data extraction, data coding, data classification, data analysis, data mining and data visualization, etc.) In particular, some of the core SMEs in ET clustering, such as data presentation, data classification, pattern recognition, etc., are clearly the preference methods in the entire DH global community. What's more, some DH research methods come from other disciplines, such as distant reading and near reading in the field of literature, gender studies and feminism, and archival studies, this is a powerful testament to the fact that DH is an interdisciplinary field. In addition, there are also some applications of ETs

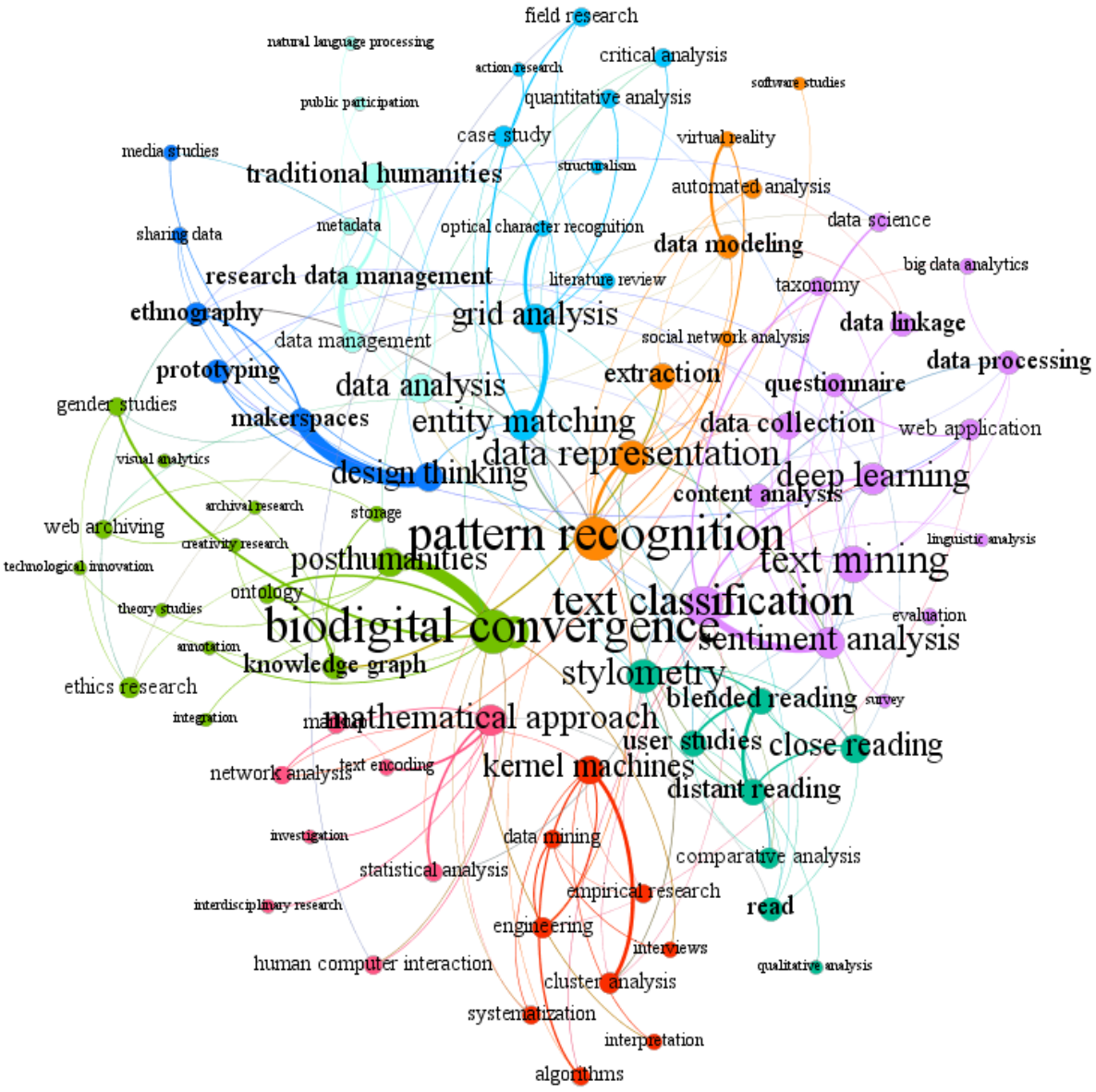

Figure 4: DH Field ET Module

Table 1
ET clustering

ETs	SMEs(degree)
#1	pattern recognition (13) , data representation (9), extraction (6) , data modeling (5) , automated analysis (3)
#2	biodigital convergence (13) , post-humanities (7) , knowledge graph (5) ,ethics research (4) , gender studies (3)
#3	mathematical approach (8) , mark up (3) , network analysis (3), human computer interaction (3) , statistical analysis (3)
#4	kernel machines (7) , cluster analysis (4) , engineering (4) , data mining (3) , systematization (3) , algorithms (3)
#5	Stylometry (9) , close reading (7) , distant reading (6) , user studies (6) , blended reading (6) , read (5)
#6	entity matching (8) , grid analysis (7) , case study (4) , critical analysis (3) , quantitative analysis (3)
#7	data analysis (7) , traditional humanities (6) , data management(4) , metadata (2) , natural language processing (1)
#8	design thinking (7) , makerspaces (5) , ethnography (5) , prototyping (5) , media studies (2)
#9	text classification (11) , text mining (10) , data collection (6) , data processing (5) , data linkage (5)

of artificial intelligence such as machine learning, which reflects the great role of AI in promoting the development of DH domain.

4.2.2. Top10 SMEs in different periods

As shown in Figure 5, most of the TOP 10 SMEs in the three stages are the same, such as geographic information systems, humanities research, artificial intelligence, history studies, retrieval, textual analysis, technological innovation, etc. are all in TOP 10 in the three stages, it shows that some methods have been used throughout the development of DH and have not been changed. These are the most widely used and mature methods in DH and represents that ETs of DH have some basic methods. In addition, there are also some method entities have a temporal phase. For example, social science only appeared in the stage 1 and 2, but it is no longer top 10 in the stage 3, while information science and cultural analysis are not significant in stage 1 and 2, but in stage 3, it has entered top10, which shows the transformation of DH's ETs. What's more, geographic information systems are almost in a dominant position in any period, which shows that DH's ETs has always attached importance to the integration with geographic methods.

4.2.3. The evolution of Top 10 SMEs

From Figure 6, we can see that the top 10 SMEs in the DH roughly experienced a general trend of rising first and then falling. From a macro perspective, the evolution of SMEs in the DH field can be roughly divided into three stages. First is the initial stage, which spans from 1903 to 2000. SMEs at this stage basically has been almost no growth, mainly due to the lacking of papers at this time. Second is the vigorous expansion stage, from 2000 to 2020. During this period many things happened to promote the development of DH, such as the release of the Digital Humanities Manifesto, the inauguration of the inauguration of the Digital Humanities Quarterly, the Global Digital Humanities Annual Conference, the development of Digital Humanities Education, etc. are proofs of its rapid development. But starting in 2020, they all showed a downward trend, which may be due to the impact of the COVID-19 epidemic and the international situation. From a microscopic point of view, geographic information systems (GIS),

digital analysis, and humanities research have always had a high frequency. Especially GIS has always maintained the highest frequency, reflecting the importance that DH scholars attach to this method. What is more interesting is that AI is also valued by digital humanities, which shows that AI technology is indeed an important help for DH research. Although the top ten SMEs like AI are all quantitative analysis methods, while traditional humanities research methods, such as history studies, cultural analysis, which are mainly based on qualitative research is also the main focus.

5. Discussion and Conclusion

DH's ETs demonstrate that DH is an interdisciplinary field of study involving multiple disciplines. ETs in the field of DH are mainly related to data processing, such as data collection and extraction, data mining and analysis, data visualization and presentation, etc. While many ETs involve the intersection with humanities and social sciences, such as history, literature, sociology, art, etc. The characteristic of interdisciplinary ETs is the guarantee of the vigorous vitality of DH. DH's ETs have the characteristics of persistence. Whether looking at the TOP 10 SEMs in different periods or looking at the overall TOP 10 SMEs, they all have a high degree of convergence. Some methods have been used all the time and have not changed over time. But at the same time, the ETs of DH also have the characteristics of partial shift. For example, in the early stage, social science was more important, and this emphasis turned to information science and cultural analysis in the later stage. DH's ETs have always maintained the characteristics of a combination of digital and humanities, and there is no bias towards one side that leads to imbalance.

Observing the evolution trend, we can find that the evolvement paths of digital technologies and humanities are completely synchronized, which shows that humanities have always been valued in the field of DH, while It's not that DH, as some scholars say, emphasizes technology over humanities. That is to say, there is no saying that DH should return to humanities in the future, because they have always been the focus. This is a strong proof of the healthy development of this discipline. In addition, a large majority of ETs are related to artificial intelligence which is an important embodiment of the word digital in DH.

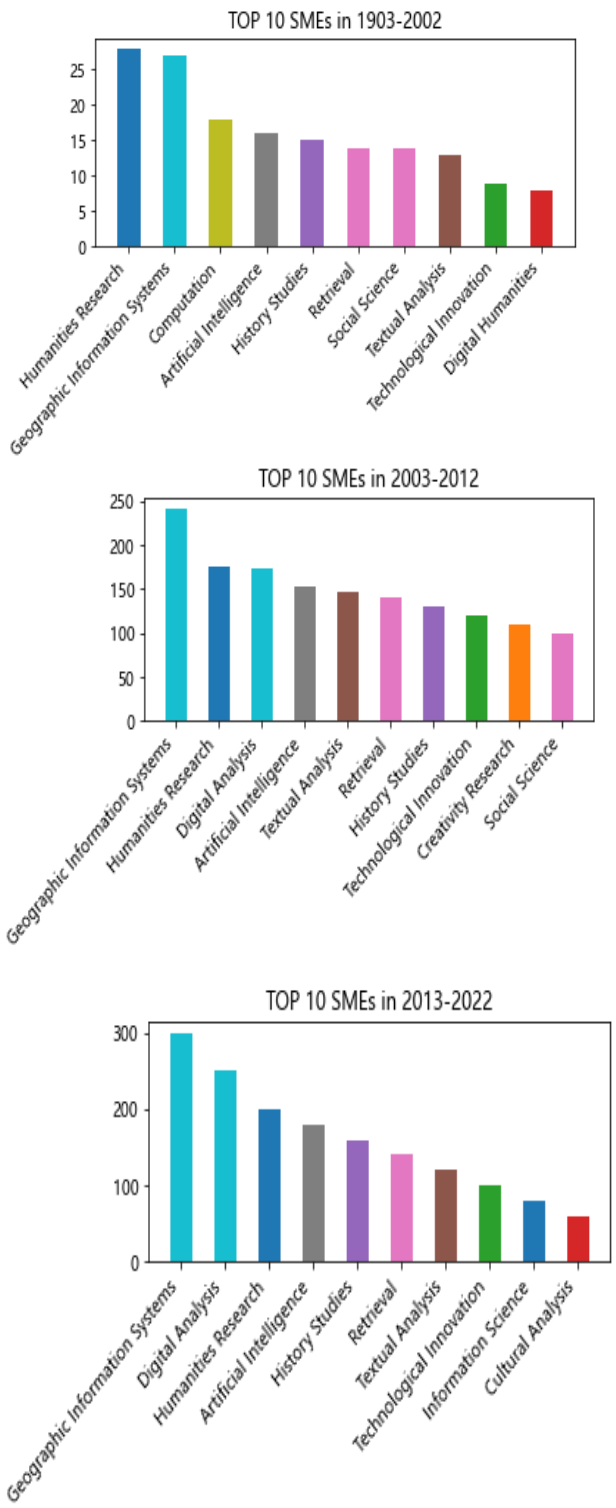

Figure 5: Top10 SMEs in different periods in DH

While many ETs involve the intersection with humanities and social sciences, such as history, literature, sociology, art, etc., which is an important embodiment of the word humanities. In addition, the application of AI has greatly promoted the development of DH.

6. Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China. (Grant No.23XNH150).

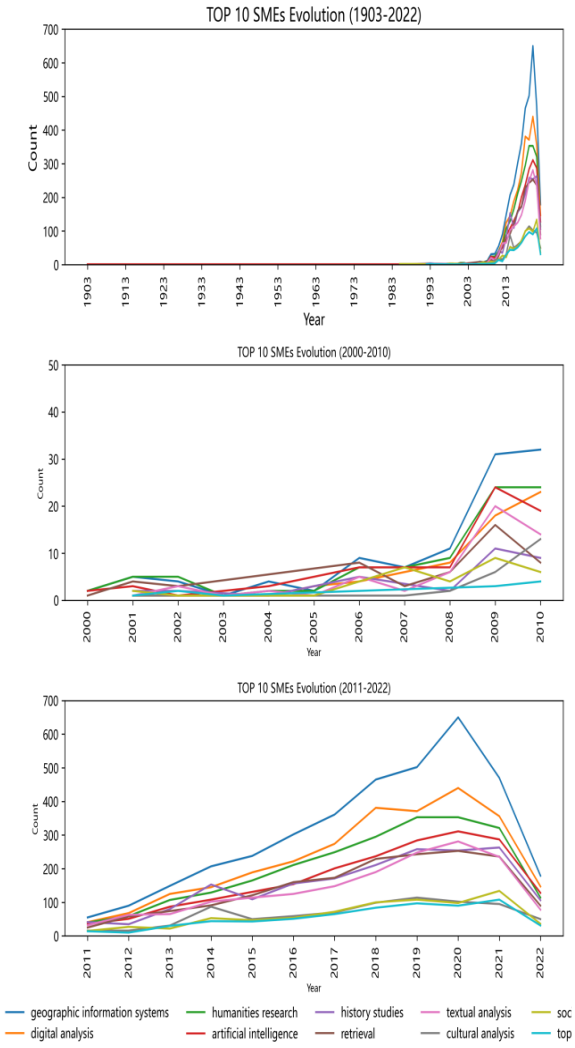

Figure 6: TOP 10 SMEs Evolution in DH(1903-2022)

7. References

[1] Innovation and technology—strategies and policies[M]. Springer Science & Business Media, 1997.

[2] Day, G.S., and P.J.H. Schoemaker. 2000. A different game. InWharton on managing emerging technologies, ed. G.S. Dayand P.J.H. Schoemaker, 1–23.New York: John Wiley

[3-4] Wang Y, Zhang C. Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing[J]. Journal of informetrics, 2020, 14(4): 101091.

[5] Wang, Y., Zhang, C. & Li, K. A review on method entities in the academic literature: extraction, evaluation, and application. Scientometrics 127, 2479–2520 (2022).

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

[7] Wang, Q. (2018). Distribution features and intellectual structures of digital humanities: A bibliometric analysis. Journal of Documentation.

[8] Su, F., Zhang, Y., & Immel, Z. (2020). Digital humanities research: interdisciplinary collaborations, themes and implications to library and information science. Journal of Documentation.

[9] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the national academy of sciences, 2006, 103(23): 8577-8582.

[10] Callon M, Courtial J P, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry[J]. Scientometrics, 1991, 22: 155-205.