Technological Forecasting Based on Spectral Clustering for Word Frequency Time Series

Han Huang[1], Xiaoguang Wang[2,3] and Hongyu Wang[1,*]

[1] School of Management, Wuhan University of Technology, Wuhan, China, 430070
[2] School of Information Management, Wuhan University, Wuhan, China, 430072
[3] Institute of Big Data, Wuhan University, Wuhan, China, 430072

Abstract

As an essential strategy for identifying technologies that should be given priority for future development, the investigation into methods of technological forecasting holds considerable importance. This study introduces a novel method for technological forecasting, the Time Trend Clustering Model (TTCM) based on spectral clustering, and engages in an analysis and discussion utilizing word frequency time series. To verify the efficacy of the model, this study initially applies the TTCM model to analyze standard time series datasets. The experimental findings indicate the model's effectiveness in distinguishing time series data with identical trends of variation. Further, taking the Library and Information Science (LIS) discipline as an example, this study employs the TTCM model to cluster the trends of word frequency time series, identifying emerging words with burst trends, label words with high-frequency fluctuation trends, hotspot words with increasing trends, and fading words with decreasing trends. By integrating the term function, the effectiveness of the TTCM model in the discovery of domain knowledge and technological forecasting is demonstrated.

Keywords

Technological forecasting, time series, temporal trend clustering, spectral clustering, term frequency analysis

1. Introduction

In the current era, the development of the socio-economic landscape relies more heavily on the capability and efficacy of scientific and technological innovation than at any time before[1]. Nations, regions, organizations, and corporations alike are dedicating efforts towards the strategic planning and foresight of science and technology, evaluating the potential directions of technological revolutions, selecting key frontier areas of science and technology, and establishing innovation systems that align with their own realities in an attempt to secure a proactive and advantageous position in future competition[2–4]. In this context, the significance of technological forecasting has become increasingly prominent.

From the perspective of knowledge management, technological forecasting is a process that involves the continuous refinement, filtering, discovery, and creation of knowledge based on the mining of a vast amount of data information (explicit knowledge) and expert experience (tacit knowledge), which then systematically selects research areas and general technologies of strategic significance[5]. In an environment where the indices of scientific literature and patents are growing exponentially, and the hardware and software levels of technologies such as big data and artificial intelligence are continuously improving[6,7], leveraging big data analytics to mine

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (ALL2024), April 22-26, 2024, Changchun, Jilin, China and Online
* Corresponding author.
huanghan@whu.edu.cn (H. Huang); wxguang@whu.edu.cn (X. Wang); hongyuwang@whut.edu.cn (H. Wang)
0000-0002-1517-9731 (H. Huang); 0000-0003-1284-7164 (X. Wang); 0000-0002-5063-9166 (H. Wang)

scientific texts and identify different patterns of technological development, then supplemented by expert judgement to evaluate the future trends of technology constitutes a crucial implementation path for technological forecasting[8,9]. Among these, the automated determination of technological evolution stages is an initial problem that needs to be addressed.

Word frequency serves as a fundamental indicator reflecting the popularity and activity level of scientific and technological fields[10,11], with its temporal trends effectively revealing the dynamics of scientific and technological development[12,13]. Some studies utilize word frequency analysis to understand the hotspots, frontiers, and their changes within specific disciplines or technological areas by analyzing high-frequency words, new word retention rates, and time series trends[10,14], often relying on the intervention of expert knowledge for manual interpretation of these temporal trends. While some researchers have employed statistical tests like the Man-Kendall test[15], as well as curve clustering methods such as the nearest-neighbor propagation algorithm[16], to analyze the time trends of word frequency sequences in a (semi-)automated manner, these studies typically use small datasets and identify relatively simple evolutionary patterns. Indeed, the variation of word frequency within a specific time window can be considered a typical time series[17,18], allowing for the analysis of changing patterns using time series trend clustering models. By detecting word frequency trends such as bursts, growth, sudden drops, and declines, it is possible to reflect the evolutionary stages of technological points. Further integrating the different growth patterns of various technological points within a tech field, combined with expert knowledge, facilitates the foresight of key, common, and emerging technologies in the technological domain.

To this end, this study introduces TTCM and employs this model to analyze word frequency time series for technological forecasting. TTCM integrates the Dynamic Time Warping (DTW) algorithm with spectral clustering, enabling the automatic clustering of time series with similar evolution trends. To verify the model's effectiveness, this study first applied the TTCM model to cluster standard time series datasets from the UCI repository, demonstrating TTCM's capability to effectively differentiate time series data with similar evolution trends. Furthermore, taking the LIS discipline as an example, this study used the TTCM model to analyze the trends in word frequency time series, identifying four types of word frequency temporal trends: burst, increasing, decreasing, and high- frequency fluctuation. Based on these findings, the study analyzed the future research trends in the LIS discipline, further validating the scientific relevance and applicability of the TTCM model in technological forecasting.

2. Literature review

2.1. The methods of technological forecasting & foresight

Technology foresight has evolved from large-scale technological prediction activities, specifically the Delphi survey[5]. With the rapid development of science and technology, the continuous changes in the economic and social environment, and the ongoing accumulation of diverse and heterogeneous scientific and technological data, the methods and tools for technology foresight have gradually diversified[3]. The methods of technology foresight can primarily be categorized into two types: one is driven by expert experience and wisdom, primarily qualitative in nature; the other is driven by data and technology, primarily quantitative in nature.

In qualitative-oriented technology foresight studies, the Delphi method is the most widely used research approach[19]. Countries such as Japan, Germany, South Korea, and China have all conducted national-level technology foresight activities based on the Delphi survey[20,21], which has been extensively applied in various technological fields including agriculture, environment, healthcare, and ICT[22]. Besides the Delphi method, commonly used approaches also include technology road mapping, scenario analysis, brainstorming, morphological analysis, and the Analytic Hierarchy Process (AHP)[23–26]. The advantage of these methods lies in their ability to fully leverage expert experience. However, due to their strong subjective nature and the high requirements for the number of experts, their fields of expertise, and their experience, as well as the significant amount of time and expense involved, these methods are increasingly questioned and gradually becoming unsuitable in the information age, characterized by an explosive growth in data volume.

The quantitative methods of data and technology-driven technology foresight primarily involve extracting valuable information from vast datasets to construct systematic foresight models[3]. These methods identify effective information for technology foresight through the mining and visualization of scientific literature, patents, technical reports, news, etc., covering aspects such as theme identification, current state assessment, gap analysis, and trend prediction. Key techniques include growth curves[27], bibliometrics[28], patent analysis[29], social network analysis[30], data envelopment analysis[31], and data mining methods such as clustering, classification, and regression[32–34]. By leveraging the mining of objective data such as literature and patents, these methods reduce reliance on experts to some extent. However, they may also lead to decreased applicability and effectiveness in decision support due to the lack of expert experience and dependency on technological pathways.

2.2. Time series clustering analysis

Time series analysis aims at mining useful information and knowledge from a large number of complex time series data, among which cluster analysis is one of the important methods of time series data mining[35]. Time series clustering analysis method has been applied to the analysis and mining of stock data[36], social media data[37], landsat time series data[38], smart grid data[39], health detection data[40], etc.

The main process of time series clustering is similarity measurement and clustering[41]. Among similarity measurement methods, shape-based approaches are the most commonly used[42]. One of the simpler approaches to implement is the Euclidean distance, and although it has some applications in distance measurement of time series[43], it is difficult to effectively take into account the phase distortion between time series[44]. At the same time, the difference in Euclidean distance between subseries at similar locations and waveforms can also be large due to the difference in their amplitudes[45]. In contrast, the Dynamic Time Warping (DTW) distance[46], improves the process of calculating the Euclidean distance. It realizes one-to-many matching of data point in time series through the dynamic warping so that it has good robustness to the phase deviation and amplitude deformation of time series, and performs well in time series clustering task[41,47–49]. Clustering algorithms for time series can be roughly divided into hierarchical clustering, model-based clustering, partition-based clustering and density-based clustering. Partition-based clustering is the most commonly used method, such as K-Means[50], K-Medoids[51], etc. However, K-means and related methods are not fully applicable to uneven sample distribution or non-convex sample data. Spectral clustering methods applicable to various shape samples may be an effective alternative to such cases. At present, some researchers have applied spectral clustering to time series data clustering[52].

2.3. Spectral clustering algorithm

Spectral clustering is an unsupervised learning algorithm based on graph partitioning, capable of transforming the clustering problem into a graph segmentation issue on an undirected weighted graph constructed from the data to be clustered[53,54]. Unlike algorithms such as K-means that work well only for convex sample data, the spectral clustering algorithm is applicable to sample spaces of any shape and converges to a global optimal solution, and it is also applicable to high-dimensional data[55,56].

Currently, spectral clustering has been widely used in image segmentation[57], face recognition[58], earth science[59] and other related researches. Due to the good data applicability and clustering effect, scholars have also applied spectral clustering to research in LIS discipline such as scientometrics and information retrieval: Chifu et al. [60] proposed a word sense discrimination method based on spectral clustering for ranking matching documents in information retrieval, thus improving the efficiency of information retrieval, and similarly; similarly, Singh et al. [61] also used spectral clustering algorithm to improve the strategy of user ranking in community Q&A sites; Colavizza and Franceschet [62] used spectral clustering algorithm to cluster literature citations in physical reviews to find similar documents. Chen et al. [63] also used spectral clustering method in multi-perspective analysis of co-citation networks; Feng et al. [64] used spectral clustering to verify the impact of different feature combinations such as JIF, 5-Year JIF, and CiteScore on the journal classification.

In addition, some researchers have also proposed optimization schemes to address the problems of high computational complexity of spectral clustering and difficulties in data representation: for example, Wang et al. [65] proposed a linear spatial embedding clustering method to optimize the similarity matrix and clustering results of spectral clustering by adaptive neighbors; Sapkota et al. [66] optimized the initial clustering center to improve the stability of the algorithm; Some researchers have also implemented spectral clustering algorithms based on the Spark big data computing framework, Julia language, etc., which improved the algorithm running efficiency by parallel computing [67,68].

3. Methodology

3.1. Model definition

In this study, a temporal trend clustering model called TTCM based on spectral clustering is proposed to analyze the word frequency time series, which is implemented using the Spark framework. The algorithm flow is shown in Figure 1. Following the retrieval requirements, the collection and preprocessing of keywords in a given academic field are completed, and the frequency of keywords over different periods is tallied to obtain the time series data for subsequent clustering analysis. The spectral clustering algorithm encompasses three core steps: graph construction, graph partitioning, and classical clustering. As the TTCM model is based on spectral clustering, steps 1-3 in Figure 1 are also the critical steps for clustering the trends of word frequency time series in the TTCM model. These three steps are introduced in detail below.

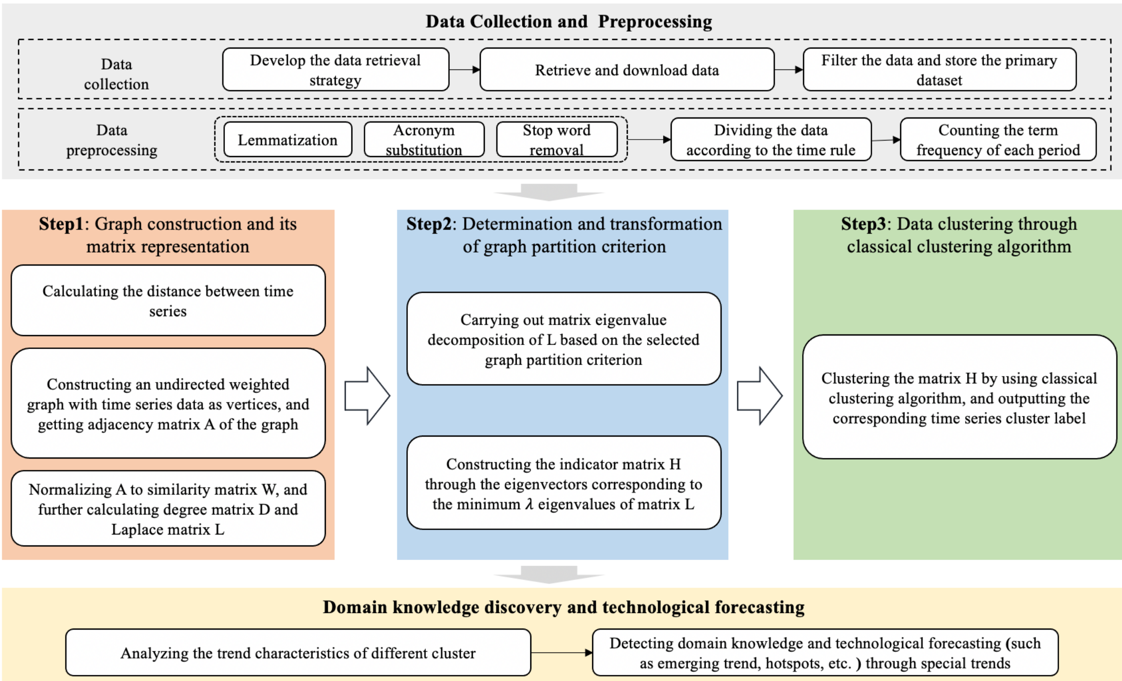

Figure 1: The algorithm flow of TTCM

3.1.1. Graph construction and its matrix representation

In this study, the time series is regarded as vertices and the DTW distance between time series is used as edge weight to construct an adjacency matrix A.

The basic idea of DTW is to find the optimal correspondence between two sequences and obtain the best match between two sequences to calculate the similarity. Its matching principle is shown in Figure 2.

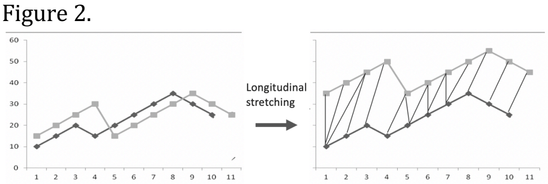

Figure 2: The matching principle of DTW algorithm

For the time series $X = \{x_1, x_2, x_3, \dots, x_m\}$ and $Y = \{y_1, y_2, y_3, \dots, y_n\}$, the DTW distance between them is calculated as formula (1).

$$D(X,Y) = \underset{W=w_1,\dots,w_k\dots,w_K}{argmin} \sqrt{\sum_{k=1,w_k=(i,j)}^{K} (x_i - y_j)^2}, i \in (1,m), j \in (1,n) \quad (1)$$

Where, $w_k = (i,j)$ represents that the ith data point of X and the jth data point of Y in the path k are corresponding points, and W is the optimal path, which can minimize the value of $D(X,Y)$.

Further, in order to reduce the dimensional difference between distances, this study uses the local scale Gaussian kernel function to normalize the DTW distances between time series to obtain the similar matrix $W = \{w_{11}, \dots, w_{1n}, \dots, w_{mn}\}$, whose calculation process is shown in Formula (2) [69].

$$w_{xy} = e^{-\frac{d_{xy}^2}{\sigma_x \sigma_y}}, \sigma_x = d_{xK}, \sigma_y = d_{yK} \quad (2)$$

Where d_{xy} is the distance between time series X and Y, σ_x is the local parameter of X, and is the distance between X and its Kth neighbor, the value of K is usually set as 7[69].

Then, the similarity matrix is transformed into Laplacian matrix. In order to prevent the analysis error caused by the non-uniform dimension between data, the symmetric normalized Laplacian matrix is used to represent the graph, and its definition is shown in Formula (3).

$$L_{sym} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \tag{3}$$

Where, I is the identity matrix and D is the degree matrix, that is, each column element of the similar matrix W is added and placed on the diagonal matrix formed by the corresponding row of the current column.

3.1.2. The determination and transformation of graph partition criterion.

The key of spectral clustering is to cut the undirected weighted graph reasonably to maximize the sum of the weights between the samples in the subgraph, that is, to minimize the sum of weights of the cut edges. According to the graph representation of the symmetric normalized Laplacian matrix determined above, this study adopt N-Cut partition criterion, whose objective function is shown in formula (4).

$$NCut(A_1, \dots, A_k) = \sum_{i=1}^{k} \frac{\frac{1}{2}\sum_{i=1}^{k} W(A_i, \overline{A_i})}{vol(A_i)} \tag{4}$$

k represents the total number of subsets, A_i represents the i'th subset, $\overline{A_i}$ is the complementary set of A_i, $W(A_i, \overline{A_i})$ represents the sum of the weights of the edges of points in subset A_i and points outside of subset A_i, $vol(A_i)$ is the sum of the weights of all edges in subset A_i. According to the mathematical derivation, the solution of the objective function can be transformed into solving the minimum eigenvalue of the Laplace matrix and its corresponding eigenvector. In this study, the eigenvectors (also known as indicator vector) corresponding to the minimum λ eigenvalues of L_{sym} should be solved, and the eigenmatrix H (also known as indicator matrix) composed of these indicator vectors is the approximate optimal solution to the graph partition problem. H is a matrix with dimension $N * \lambda$, and N is the number of time series data.

3.1.3. Data clustering through classical clustering algorithm.

After the graph is divided, the classical clustering algorithm can be used to cluster H. Based on the K-means algorithm, this study regards the row data h_n of H as a vector in the current space, and conducts cluster analysis on it, and obtain the category of h_n is the category of the n'th time series.

3.2. Algorithm parameter determination

In the implementation of TTCM model, λ, the number of feature vectors, is a parameter that needs to be set in advance. In practical processes, λ is often set as k, the final expected number of spectral clustering. However, the clustering number k is usually determined according to the change of error sum of squares or contour coefficient of K-Means model at the last stage. In order to determine λ in advance, a novel dimension determination method of indicator matrix is designed based on the meaning and properties of the Fiedler vector of the Laplace matrix.

The Fiedler vector is the eigenvector corresponding to the minimum non-zero eigenvalue (also known as the second smallest eigenvalue) of the Laplace matrix of the graph [70]. In this study, the Fidler vector of L_{sym} is first taken as the indicator matrix H, and then k-means clustering is carried out on H, and the evolution trend between the number of clustering and the error sum of squares is observed to determine the optimal number of clustering K. Then, λ is set as k, k-1 and k-2 to conduct the subsequent analysis. Finally, on the basis of ensuring that the difference between clusters, a small λ value is chosen to reduce the time and space cost of the subsequent calculation process, and avoid overfitting. In this study, the above method of λ selection is called the principle of low.

4. Experiments and result analysis

4.1. Model validation through time series standard dataset

To verify the efficiency of TTCM in time series clustering, the time series standard dataset[71] in the Knowledge Discovery Archive [72] of University of California, Irvine (UCI) is used to test the model. And the Power Iteration Clustering (PIC) model[73] and the Affinity Propagation (AP) clustering model[74], which are also based on graph theory, are selected as the baseline to compare the model recognition effects.

There are 600 pieces of data in the time series dataset, and every 100 pieces represent a trend type, which are marked as normal, cyclic, increasing trend, decreasing trend, upward shift, and downward shift. Figure 3 shows the sample data of these six trends.

According to the trends of the test dataset, the clustering number of TTCM model, PIC model and AP model is set to 6, and the maximum number of iterations is set as 30. Meanwhile, in the TTCM model, λ is set to 4,5, and 6. In addition, the three model all use the similarity matrix W calculated by formula (2).

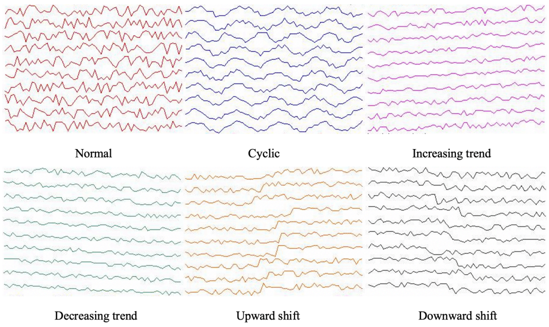

Figure 3: Sample data of the six trends in the time series dataset

After the clustering is completed, the identified cluster labels are matched to the actual labels according to the data distribution in various clusters, that is, if the identified cluster 1 contains the most increasing trend data, the cluster 1 will be marked as increasing trend. Then, the number of increasing trend and other types of data in cluster 1 is compared with the actual increasing trend data number (i.e. 100). After calculating the values of precision (P), recall rate (R) and F1 respectively, the average values of P, R and F1 in six categories are used as the evaluation value of the effect of models, and the results are shown in Table 1.

Table 1
Experimental results of TTCM/PIC/AP model on test dataset

Model	TTCM ($\lambda = 4$)	TTCM ($\lambda = 5$)	TTCM ($\lambda = 6$)	PIC	AP
Number	406	**578**	492	200	419
P	64.02%	**96.63%**	81.14%	24.44%	70.48%
R	67.67%	**96.33%**	82.00%	33.33%	69.83%
F1	64.18%	**96.33%**	81.44%	27.50%	67.11%

It can be seen that, when λ= 5, the TTCM model has a good recognition effect, it can accurately identify 578 time series data trends, F1 value up to 96.33%, much higher than PIC model, AP model and TTCM model with other values of λ. Specific to each category, the recognition results of TTCM when λ= 5 are shown in Table 2.

Table 2
Confusion matrix of the six-classification problem corresponding to TTCM model (λ= 5)

Model / Actual	Normal	Cyclic	Increasing trend	Decreasing trend	Upward shift	Downward shift	Recall
Normal	98	0	2	0	0	0	98.00%
Cyclic	0	100	0	0	0	0	100.00%
Increasing trend	0	0	100	0	0	0	100.00%
Decreasing trend	0	0	0	99	0	1	99.00%
Upward shift	1	0	11	0	88	0	88.00%
Downward shift	0	0	0	7	0	93	93.00%
Precision	98.99%	100.00%	88.50%	93.40%	100.00%	98.94%	**F_1 = 0.9633**

By observing Table 2, it can be further found that TTCM can effectively distinguish the six types of trends in the test data set, and only errors appear in the recognition of a small number of increasing and upward shifts and decreasing and downward shifts. Overall, the TTCM model proposed in this paper can effectively distinguish the evolution trends of time series and cluster time series with similar trends.

4.2. Temporal trend clustering through word frequency time series

4.2.1. Data collection and preprocessing

In order to further verify the effectiveness of TTCM in detecting trends within word frequency time series, combined with the disciplinary background of the team members, this study selected the LIS discipline for case analysis. This study adopts the same data collection principles as our previous study[13] and collects the scientific papers published in the journals included in the Social Sciences Citation Index (SSCI) in the field of LIS from 2011 to 2020. The document types of papers are limited to *research article* and *review*, and the language is limited to English. Finally, the case dataset containing 38932 scientific papers is obtained. Then the keywords of these papers are carried out the preprocessing process including denoising, morphology reduction, abbreviation conversion. After preprocessing, the number and frequency of keywords in each year are statistically analyzed as shown in Table 3.

Table 3
Confusion matrix of the six-classification problem corresponding to TTCM model (λ= 5)

Year	Number of papers	Number of keywords	Frequency of keywords	Average word frequency
2011	3298	6324	10706	1.69
2012	3414	7125	12230	1.72
2013	3618	8087	13876	1.72
2014	3732	8267	14089	1.7
2015	3822	8806	15546	1.77
2016	4155	10082	17922	1.78
2017	4139	10156	17273	1.7
2018	4079	10632	17894	1.68
2019	4205	11034	18917	1.71
2020	4470	12270	21215	1.73

4.2.2. Results

In the analysis of word frequency time series, consideration was given to the possibility that keywords with a total frequency count too low might not exhibit significant trends in time series changes (i.e., the frequency time series of such keywords could be classified as having a uniform trend). Therefore, adhering to common practice, this study filtered keywords from a pool of 57,025 distinct keywords spanning the entire study period, selecting those with a total frequency count exceeding the length of the time span. The filtration yielded 1,952 author keywords that were mentioned in more than ten articles from 2011 to 2020.

Utilizing the TTCM model, this study conducted trend identification on the time series of these 1,952 keywords. Following the principle of low introduced in Section 3.2, the study set λ to 3 and the number of clusters k to 5 for the TTCM. Subsequently, plots of the word frequency time series within each trend category were generated to facilitate a visual observation and summary of the changing characteristics of the frequency time series trends within each cluster.

In the clustering results of temporal trend of term frequency, the first kind of trend can be summarized as the burst trend. The obvious characteristic of this kind of trend is that the term frequency of keywords is low in the early and middle period of the whole-time span, but the term frequency shows a trend of rapid rise in the middle and later periods. Figure 4 shows the term frequency change curve of some keywords with a burst trend in the term frequency series, and a total of 30 keywords are clustered as such trend.

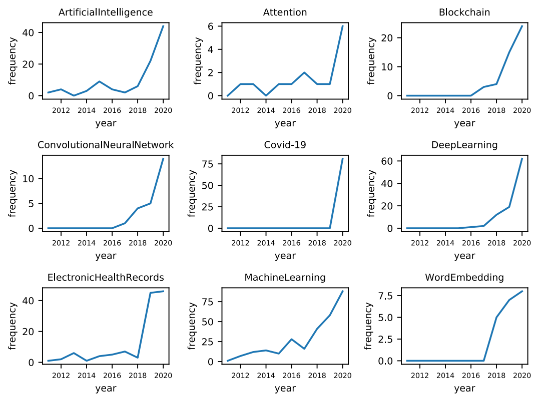

Figure 4: Part of words with burst trend in word frequency series (emerging words)

In the term frequency series of keywords, the second trend can be classified as the increasing trend. The term frequency series of this kind of trend shows a general trend of fluctuation increase in the whole-time span, but the term frequency remains at the low level in the whole-time span. Figure 5 shows part of keywords with an increasing trend in the term frequency series. There are 177 keywords with the increasing trends.

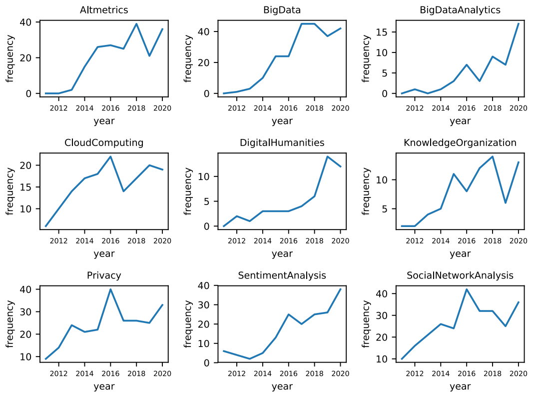

Figure 5: Part of words with increasing trend in word frequency series (hotspot words)

In the clustering results of temporal trend of term frequency, the third kind of trend can be summarized as the high-frequency fluctuation trend. The obvious characteristic of this kind of trend is that the term frequency of keywords remains at a high level in the whole-time span, and the term frequency fluctuates slightly with the passage of time. Figure 6 shows the term frequency change curves of some keywords with high-frequency fluctuation trend in term frequency series, and a total of 30 keywords are clustered as such trend.

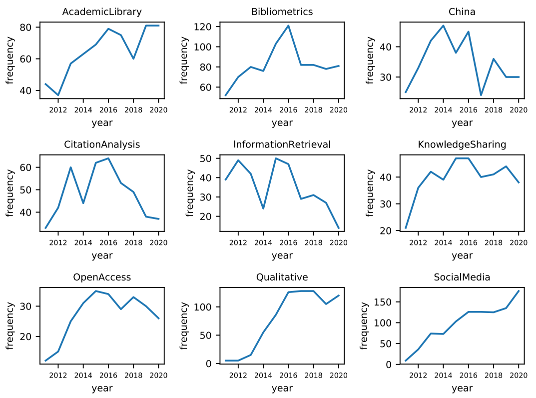

Figure 6: Part of words with high-frequency fluctuation trend in word frequency series (label words)

In the term frequency series of keywords, the fourth trend can be classified as the decreasing trend. The term frequency series of this kind of trend shows a general trend of fluctuation decrease in the whole-time span, and the term frequency remains at the low level in the whole-time span. Figure 7 shows part of keywords with a decreasing trend in the term frequency series. There are 69 keywords with the decreasing trends.

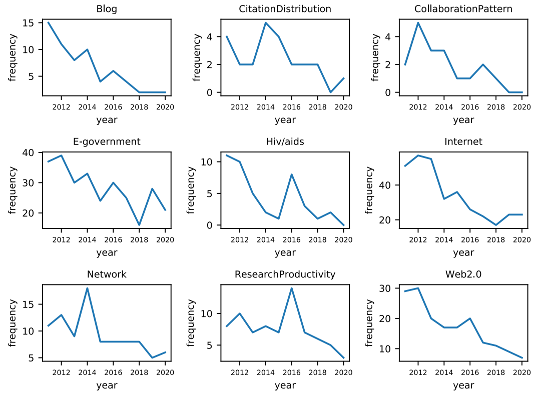

Figure 7: Part of words with decreasing trend in word frequency series (fading words)

The fifth type of trend identified by TTCM for term frequency series contains a total of 1646 keywords, and the observation of its trend curves failed to find obvious characteristics. Therefore, this paper speculates that the trend of this kind of term frequency series should be the normal trend without obvious regular fluctuation.

It can be seen from the clustering results that the TTCM model is highly effective in identifying emerging words across various disciplines that have suddenly burst onto the scene, successfully capturing the rising trend of keyword frequencies towards the end of the time span. Within the set of keywords exhibiting an upward trend, the model accurately identified research hotspots that are gradually gaining widespread attention among LIS scholars. For the keywords identified by the model as having high-frequency fluctuations, their frequency levels consistently remained high, often signifying core research sub-fields or themes within the domain. Conversely, the model effectively reflected keywords in decline, indicating words that are gradually fading from the focal interest of scholars in the discipline.

4.3. Technical foresight based on temporal trend of word frequency

Further, this study, in accordance with the term function[75], divides the keywords with significant temporal trend into two categories: research questions/objects and research methods/technologies. The count of different functional keywords showing varying trends, along with examples, is displayed as shown in Table 4.

Table 4
Keyword statistics based on different trends of term function

Trend	Research questions/objects	Research methods/technologies
Burst	**17** Electronic Health Records Covid-19 Digital Transformation Coronavirus Journalism	**13** Machine Learning Artificial Intelligence Deep Learning Blockchain Neural Network
Increasing	**117** Privacy Research Evaluation Gender Scholarly Communication Higher Education	**60** Social Network Analysis Big Data Classification Altmetrics Sentiment Analysis
Decreasing	**57** Internet E-government Digital Library Web2.0 Blog	**12** Focus Group Semistructured Interviews Nanotechnology Citation Distribution Microsimulation
High-frequency fluctuation	**18** Social Media Academic Library Information Literacy China Collaboration	**12** Bibliometrics Qualitative Citation Analysis Case Study Information Retrieval
Total	**209**	**97**

In general, the time series of word frequencies identified in this study predominantly comprise words related to research questions/objects, accounting for nearly 70% of the total.

There exists a considerable number of keywords exhibiting an increasing trend, with both functional types of words showing a relatively balanced distribution. However, due to the phenomenon of technological literature inflation, although these words continuously attract the attention of scholars in the field dynamically, the share of related research may not have expanded across the entire disciplinary spectrum in actuality. As some researchers delve into new studies, there are concurrent instances of existing scholars gradually losing focus. Should there be no emergence of new method or technology innovations or the continuation of integrating novel research objects corresponding to these characteristic keywords, the frequencies of these words will gradually transition into a decreasing trend.

Quantitatively, the number of faded keywords exhibiting a decreasing trend is less than half of the quantity of keywords showing an upward trend, a circumstance possibly attributable to the literature inflation caused by technological explosions. Within the rapid accumulation of technological literature, the conservative tendencies of some researchers and/or the attribute of knowledge application contained within certain words might prevent these fading-out keywords from becoming low-frequency words filtered out during the input phase of the TTCM model. However, these fading words, if not subject to knowledge innovation, are highly likely to decline gradually. Simultaneously, among the words demonstrating decreasing trends, words related to research questions/objects are notably higher in proportion compared to those concerning methods/technologies. This discrepancy may be attributed to the stronger applicability of methods/technologies, where researchers, even amidst shifts in research subjects or questions, tend to employ classical and established technical methodologies. The proportion of emerging words displaying burst trends is relatively low. Although there is a higher absolute number of words related to research questions/objects, the relative proportion of words related to methods is higher, indicating that, with changes in social and research environments, researchers have begun to pay attention to some new research objects, such as coronaviruses, open science, and mobile payments, while introducing more emerging technologies such as artificial intelligence, machine learning, and neural networks.

Specifically, regarding methods/technologies, technologies such as focus groups, semi-structured interviews, and microsimulation, primarily targeting small-scale data samples, exhibit a decreasing trend, while big data analysis techniques such as artificial intelligence, machine learning, deep learning, social network analysis, and sentiment analysis demonstrate a burst or increasing trend. This reflects the progress and evolution of research methods and technologies in LIS, with an increasing number of researchers adopting emerging technologies to process and analyze information to gain deeper and broader insights. Simultaneously, the application of emerging technologies also reflects the transformation of research content in the LIS field towards quantitative analysis, large-scale data processing, and deep data mining. The explosive growth trend of technologies such as artificial intelligence also indicates that future research in the LIS field may increasingly focus on leveraging advanced computing technologies to address issues related to information management, information retrieval, and user behavior analysis.

5. Discussion

Technological advancements and transformations are not only complex interplays driven by societal, economic, and political well-being but also their outcomes. Predicting and understanding the process of technological change pose challenges for decision-makers in governments and businesses[76]. Appropriately implemented and effective technological forecasting is of significant guiding value to organizations such as governments and businesses[3]. The research paradigm of technological forecasting is still evolving, promoting the effective complementary integration of qualitative and quantitative research methods. Seeking novel research methodologies to enhance research quality is currently a hotspot and focus in the field of technological forecasting. Therefore, conducting research on technological forecasting methods under this backdrop holds certain theoretical significance and practical value.

This study proposes the TTCM model based on spectral clustering. Model validation results from Tables 1 and 2 demonstrate that the TTCM model can effectively distinguish the evolution trends of time series and automatically cluster time series with similar trends. Applying the TTCM model to the analysis of word frequency time series reveals its successful identification of sudden emerging words, high-frequency fluctuating words, steadily increasing hotspot words, and gradually decreasing fading words, providing significant reference and guidance value for anticipatory analysis in disciplinary fields. Furthermore, combined with term functions, anticipatory analysis of subsequent research development and technological shifts in the field helps research institutions and relevant practitioners adjust research directions in a timely manner, grasp popular scientific research trends and frontier opportunities, and also aids governments and industrial institutions in identifying focal points and trends in the field, providing decision-making support for the formulation and planning of science and technology policies and strategies.

Essentially, the TTCM model is a clustering model whose clustering objects are time series, and the clustering basis is the evolution trend of time series, i.e., clustering time series with similar evolution trends into the same category. Therefore, the application of the TTCM model is not limited to word frequency time series. For scientific literature, analysis of research hotspots and frontier trends can be conducted using time series data such as publication volume, citation volume, and author quantity. For patent literature, technological forecasting can be conducted using time series data such as patent application volume, citation volume, and patent conversion quantity. Additionally, comprehensive analysis can be performed by combining time series data from other sources such as online news, social media, and stock securities. This aims to provide reference and guidance for organizational decision-making in governments, industries, and businesses.

Furthermore, after conducting extensive identification experiments on time series data using the TTCM model, the identified types of evolution trends can be solidified into pattern features. This can be further combined with traditional machine learning models such as Support Vector Machines, K-nearest neighbors, Conditional Random Fields, or deep learning models such as Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory Networks, to achieve rapid identification of large-scale time series evolution trends. This automation enables automated prediction of emerging research trends in the field or potential technological growth points.

6. Conclusion

The present study introduces a novel time series trend clustering model, named TTCM, and employs it to analyze word frequency time series for technological forecasting. TTCM integrates dynamic time warping algorithm and spectral clustering algorithm to automatically cluster time series exhibiting similar evolution trends. To validate the effectiveness of the model, this research initially applies TTCM to cluster standard time series datasets from the UCI repository, demonstrating its capability to effectively differentiate time series data with similar evolution trends. Furthermore, using the LIS discipline as a case study, this research utilizes TTCM to cluster the evolution trends of word frequency time series, identifying emerging words with burst trends, label words with high-frequently fluctuation trends, hotspot words with increasing trends, and decreasing fading words. The integration of term function confirms the efficacy of TTCM in domain knowledge discovery and technological forecasting.

Nevertheless, this study has certain limitations. Firstly, due to computational constraints, only ten years of data were selected for analysis, potentially overlooking evolution trends that manifest over longer time series. Secondly, the case study is limited to the LIS domain, warranting further verification of the analysis effectiveness of the TTCM model in word frequency time series from other disciplines and fields. Additionally, the analysis in this study is limited to keyword perspectives, without considering interrelations among keywords in the thematic dimension.

In the future research, in addition to addressing the shortcomings mentioned above, this study will incorporate other data sources such as patent data to achieve technology foresight with multi-source data. Moreover, after extensive experimentation to determine evolution trends in different types of time series, this study will consider solidifying these trends into pattern features and further integrating them with classification models to achieve intelligent and automated prediction of emerging research trends or potential technological growth points in large-scale datasets.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.1874129), Innovative Research Group Project of the National Natural Science Foundation of China (No.71921002).

References

[1] F. Dotsika, A. Watkins, Identifying potentially disruptive trends by means of keyword network analysis, Technological forecasting and social change 119 (2017) 114–127. doi:10.1016/j.techfore.2017.03.020.

[2] R.N. Kostoff, R.R. Scaller, Science and technology roadmaps, IEEE transactions on engineering management 48 (2001) 132–143. doi:10.1109/17.922473.

[3] C. Lee, A review of data analytics in technological forecasting, Technological forecasting and social change 166 (2021). doi:10.1016/j.techfore.2021.120646.

[4] E. Amanatidou, Beyond the veil - The real value of Foresight, Technological forecasting and social change 87 (2014) 274–291. doi:10.1016/j.techfore.2013.12.030.

[5] B.R. Martin, Foresight in science and technology, Technology Analysis & strategic management 7 (1995) 139-168.

[6] L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, Journal of the Association for Information Science and Technology 66 (2015) 2215–2222. doi:10.1002/asi.23329.

[7] C. Balili, U. Lee, A. Segev, J. Kim, M. Ko, TermBall: Tracking and predicting evolution types of research topics by using knowledge structures in scholarly big data, IEEE Access 8 (2020) 108514–108529. doi:10.1109/ACCESS.2020.3000948.

[8] A.C. Adamuthe, G.T. Thampi, Technology forecasting: A case study of computational technologies, Technological forecasting and social change 143 (2019) 181–189. doi:10.1016/j.techfore.2019.03.002.

[9] H. Lee, S. Lee, B. Yoon, Technology clustering based on evolutionary patterns: The case of information and communications technologies, Technological forecasting and social change 78 (2011) 953–967. doi:10.1016/j.techfore.2011.02.002.

[10] W. Lu, S. Huang, J. Yang, Y. Bu, Q. Cheng, Y. Huang, Detecting research topic trends by author-defined keyword frequency, Information processing and management 58 (2021) 102594. doi:10.1016/j.ipm.2021.102594.

[11] Y.H. Hu, C.T. Tai, K.E. Liu, C.F. Cai, Identification of highly-cited papers using

topic-model-based and bibliometric features: The consideration of keyword popularity, Journal of Informetrics 14 (2020) 101004. doi:10.1016/j.joi.2019.101004.
[12] T.Y. Huang, B. Zhao, Measuring popularity of ecological topics in a temporal dynamical knowledge network, PLoS ONE 14 (2019) e0208370. doi:10.1371/journal.pone.0208370.
[13] X. Wang, H. Wang, H. Huang, Evolutionary exploration and comparative analysis of the research topic networks in information disciplines, Scientometrics 126 (2021) 4991–5017. doi:10.1007/s11192-021-03963-6.
[14] M. Petrova, P. Sutcliffe, K.W.M. Fulford, J. Dale, Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: A word frequency analysis study, Journal of the American medical informatics association 19 (2012) 479–488. doi:10.1136/amiajnl-2011-000243.
[15] M. Färber, C. Nishioka, A. Jatowt, ScholarSight: Visualizing temporal trends of scientific concepts, 2019 ACM/IEEE Joint Conference on Digital Libraries, 2019, pp. 438–439. doi:10.1109/JCDL.2019.00108.
[16] M. Trevisani, A. Tuzzi, Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories, Knowledge-based systems 146 (2018) 129–141. doi:10.1016/j.knosys.2018.01.035.
[17] C. Boothby, S. Milojević, An exploratory full-text analysis of Science Careers in a changing academic job market, Scientometrics 126 (2021) 4055–4071. doi:10.1007/s11192-021-03905-2.
[18] E.S. Atlam, M. Okada, M. Shishibori, J. ichi Aoe, An evaluation method of words tendency depending on time-series variation and its improvements, Information processing and management 38 (2002) 157–171. doi:10.1016/S0306-4573(01)00028-0.
[19] J. Landeta, Current validity of the Delphi method in social sciences, Technological forecasting and social change 73 (2006) 467–482. doi:10.1016/j.techfore.2005.09.002.
[20] T. Shin, Using Delphi for a long-range technology forecasting, and assessing directions of future R&D activities - The Korean exercise, Technological forecasting and social change 58 (1998) 125–154. doi:10.1016/S0040-1625(97)00053-X.
[21] M. Rongping, R. Zhongbao, Y. Sida, Q. Yan, `Technology foresight towards 2020 in China': the practice and its impacts, Technology analysis & strategic management 20 (2008) 287–307. doi:10.1080/09537320801999587.
[22] A. Suominen, A. Hajikhani, A. Ahola, Y. Kurogi, K. Urashima, A quantitative and qualitative approach on the evaluation of technological pathways: A comparative national-scale Delphi study, Futures 140 (2022). doi:10.1016/j.futures.2022.102967.
[23] T. Heger, R. Rohrbeck, Strategic foresight for collaborative exploration of new business fields, Technological forecasting and social change 79 (2012) 819–831. doi:10.1016/j.techfore.2011.11.003.
[24] Y. Tang, H. Sun, Q. Yao, Y. Wang, The selection of key technologies by the silicon photovoltaic industry based on the Delphi method and AHP (analytic hierarchy process): Case study of China, Energy 75 (2014) 474–482. doi:10.1016/j.energy.2014.08.003.
[25] C. Flick, E.D. Zamani, B.C. Stahl, A. Brem, The future of ICT for health and ageing: Unveiling ethical and social issues through horizon scanning foresight, Technological forecasting and social change 155 (2020). doi:10.1016/j.techfore.2020.119995.
[26] M. Hussain, E. Tapinos, L. Knight, Scenario-driven roadmapping for technology foresight, Technological forecasting and social change 124 (2017) 160–177. doi:10.1016/j.techfore.2017.05.005.
[27] Y. Jeong, I. Park, B. Yoon, Forecasting technology substitution based on hazard function, Technological forecasting and social change 104 (2016) 259–272. doi:10.1016/j.techfore.2016.01.014.
[28] W. Yeo, S. Kim, H. Park, J. Kang, A bibliometric method for measuring the degree of technological innovation, Technological forecasting and social change 95 (2015) 152–162. doi:10.1016/j.techfore.2015.01.018.
[29] C. Lee, Y. Cho, H. Seol, Y. Park, A stochastic patent citation analysis approach to assessing future technological impacts, Technological forecasting and social change 79 (2012) 16–29. doi:10.1016/j.techfore.2011.06.009.

[30] M. Coccia, L. Wang, Path-breaking directions of nanotechnology-based chemotherapy and molecular cancer therapy, Technological forecasting and social change 94 (2015) 155–169. doi:10.1016/j.techfore.2014.09.007.

[31] D.-J. Lim, T.R. Anderson, O.L. Inman, Choosing effective dates from multiple optima in Technology Forecasting using Data Envelopment Analysis (TFDEA), Technological forecasting and social change 88 (2014) 91–97. doi:10.1016/j.techfore.2014.06.003.

[32] S. Jun, S.S. Park, D.S. Jang, Technology forecasting using matrix map and patent clustering, Industrial management & data systemS 112 (2012) 786–807. doi:10.1108/02635571211232352.

[33] S. Jun, A Forecasting Model for Technological Trend Using Unsupervised Learning, in: T.H. Kim, H. Adeli, A. Cuzzocrea, T. Arslan, Y.C. Zhang, J.H. Ma, K.I. Chung, S. Mariyam, X.F. Song, Database theory application, bio-science bio-technology, Springer-Verlag Berlin, Berlin, Germany, 2011: pp. 51–60.

[34] N. Gozuacik, C.O. Sakar, S. Ozcan, Technological forecasting based on estimation of word embedding matrix using LSTM networks, Technological forecasting and social change 191 (2023) 122520. doi:10.1016/J.TECHFORE.2023.122520.

[35] P. Esling, C. Agon, Time-Series Data Mining, ACM computing surveys 45 (2012). doi:10.1145/2379776.2379788.

[36] C. Guo, H. Jia, N. Zhang, Time Series Clustering Based on ICA for Stock Data Analysis, in: 4th international conference on wireless communications, networking and mobile computing, VOLS 1-31, IEEE, New York, USA, 2008: pp. 10903+.

[37] H. Zhu, Y. Mei, J. Wei, C. Shen, Prediction of online topics' popularity patterns, Journal of information science 48 (2022) 141–151. doi:10.1177/0165551520961026.

[38] Y. Zhao, L. Lin, W. Lu, Y. Meng, Landsat time series clustering under modified dynamic time warping, in: Q. Weng, P. Gamba, G. Xian, J.M. Chen, S. Liang, 4rth international workshop on earth observation and remote sensing applications, IEEE, New York, USA, 2016.

[39] H. Son, Y. Kim, S. Kim, Time series clustering of electricity demand for industrial areas on smart grid, Energies 13 (2020). doi:10.3390/en13092377.

[40] C.H. Sudre, K.A. Lee, M.N. Lochlainn, T. Varsavsky, B. Murray, M.S. Graham, C. Menni, M. Modat, R.C.E. Bowyer, L.H. Nguyen, D.A. Drew, A.D. Joshi, W. Ma, C.-G. Guo, C.-H. Lo, S. Ganesh, A. Buwe, J.C. Pujol, J.L. du Cadet, A. Visconti, M.B. Freidin, J.S.E.-S. Moustafa, M. Falchi, R. Davies, M.F. Gomez, T. Fall, M.J. Cardoso, J. Wolf, P.W. Franks, A.T. Chan, T.D. Spector, C.J. Steves, S. Ourselin, Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID symptom study app, Science advances 7 (2021). doi:10.1126/sciadv.abd4177.

[41] T. Li, X. Wu, J. Zhang, Time series clustering model based on DTW for classifying car parks, Algorithms 13 (2020). doi:10.3390/a13030057.

[42] S. Zolhavarieh, S. Aghabozorgi, Y.W. Teh, A Review of Subsequence Time Series Clustering, Scientific world journal (2014). doi:10.1155/2014/312521.

[43] X. Guo, Y. Pang, G. Yan, T. Qiao, Time series forecasting based on deep extreme learning machine, in: 29th Chinese control and decision conference, CCDC 2017, 2017: pp. 6151–6156. doi:10.1109/CCDC.2017.7978277.

[44] E.J. Keogh, M.J. Pazzani, Relevance feedback retrieval of time series data, in: 22nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 1999, 1999: pp.183–190. doi:10.1145/312624.312676.

[45] X.L. Dong, C.K. Gu, Z.O. Wang, Research on shape-based time series similarity measure, in: 2006 international conference on machine learning and cybernetics, 2006: pp.1253–1258. doi:10.1109/ICMLC.2006.258648.

[46] E. Keogh, C.A. Ratanamahatana, Exact indexing of dynamic time warping, Knowledge and information systems 7 (2005) 358–386. doi:10.1007/s10115-004-0154-9.

[47] B. Cai, G. Huang, N. Samadiani, G. Li, C.H. Chi, Efficient Time Series Clustering by Minimizing Dynamic Time Warping Utilization, IEEE access 9 (2021) 46589-46599,. doi:10.1109/ACCESS.2021.3067833.

[48] W. Wang, G. Lyu, Y. Shi, X. Liang, Time Series Clustering Based on Dynamic Time Warping, in: IEEE 9th International Conference on

Software Engineering and Service Science, Beijing, China, 2018. doi:10.1109/ICSESS.2018.8663857.
[49] V.T. Huy, D.T. Anh, An efficient implementation of anytime K-medoids clustering for time series under dynamic time warping, in: 7th symposium on information and communication technology, 2016: pp. 22–29. doi:10.1145/3011077.3011128.
[50] X. Huang, Y. Ye, L. Xiong, R.Y.K. Lau, N. Jiang, S. Wang, Time series k-means: A new k-means type smooth subspace clustering for time series data, Information sciences 367 (2016) 1–13. doi:10.1016/j.ins.2016.05.040.
[51] Y. Chen, X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, F. Pei, Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method, Landscape and urban planning 160 (2017) 48–60. doi:10.1016/j.landurbplan.2016.12.001.
[52] H. Abbasimehr, A. Bahrini, An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation, Expert systems with applications 192 (2022). doi:10.1016/j.eswa.2021.116373.
[53] M. Alshammari, M. Takatsuka, Approximate spectral clustering with eigenvector selection and self-tuned k, Pattern recognition letters 122 (2019) 31–37. doi:10.1016/j.patrec.2019.02.006.
[54] P.K. Srijith, M. Hepple, K. Bontcheva, D. Preotiuc-Pietro, Sub-story detection in Twitter with hierarchical Dirichlet processes, Information processing and management 53 (2017) 989–1003. doi:10.1016/j.ipm.2016.10.004.
[55] T. Semertzidis, D. Rafailidis, M.G. Strintzis, P. Daras, Large-scale spectral clustering based on pairwise constraints, Information processing and management 51 (2015) 616–624. doi:10.1016/j.ipm.2015.05.007.
[56] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: 15th Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 2002: pp. 849–856.
[57] K. Xia, X. Gu, Y. Zhang, Oriented grouping-constrained spectral clustering for medical imaging segmentation, Multimedia systems 26 (2020) 27–36. doi:10.1007/s00530-019-00626-8.
[58] D. Xu, C. Li, T. Chen, F. Lang, A novel low rank spectral clustering method for face identification, Recent patents on engineering 13 (2019) 387–394. doi:10.2174/187221211266618082812421 1.
[59] H. Talebi, L.J.M. Peeters, U. Mueller, R. Tolosana-Delgado, K.G. van den Boogaart, Towards geostatistical learning for the geosciences: A case study in improving the spatial awareness of spectral clustering, Mathematical geosciences 52 (2020) 1035–1048. doi:10.1007/s11004-020-09867-0.
[60] A.G. Chifu, F. Hristea, J. Mothe, M. Popescu, Word sense discrimination in information retrieval: A spectral clustering-based approach, Information processing and management 51 (2015) 16–31. doi:10.1016/j.ipm.2014.10.007.
[61] A.K. Singh, N.K. Nagwani, S. Pandey, A user ranking algorithm for efficient information management of community sites using spectral clustering and folksonomy, Journal of information science 45 (2019) 592–606. doi:10.1177/0165551518808198.
[62] G. Colavizza, M. Franceschet, Clustering citation histories in the physical review, Journal of informetrics 10 (2016) 1037–1051. doi:10.1016/j.joi.2016.07.009.
[63] C. Chen, F. Ibekwe-SanJuan, J. Hou, The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis, Journal of the American society for information science and technology 61 (2010) 1386–1409. doi:10.1002/asi.21309.
[64] L. Feng, J. Zhou, S.L. Liu, N. Cai, J. Yang, Analysis of journal evaluation indicators: an experimental study based on unsupervised Laplacian score, Scientometrics 124 (2020) 233–254. doi:10.1007/s11192-020-03422-8.
[65] Q. Wang, Z. Qin, F. Nie, X. Li, Spectral embedded adaptive neighbors clustering, IEEE transactions on neural networks and learning systems 30 (2019) 1265–1271. doi:10.1109/TNNLS.2018.2861209.
[66] N. Sapkota, A. Alsadoon, P.W.C. Prasad, A. Elchouemi, A.K. Singh, Data summarization using clustering and classification: Spectral clustering combined with k-means using NFPH, in: the international conference on machine learning, big data, cloud and

parallel computing: trends, perspectives and prospects, Faridabad, India, 2019: pp.146–151. doi:10.1109/COMITCon.2019.8862218.
[67] Z. Huo, G. Mei, G. Casolla, F. Giampaolo, Designing an efficient parallel spectral clustering algorithm on multi-core processors in Julia, Journal of parallel and distributed computing 138 (2020) 211–221. doi:10.1016/j.jpdc.2020.01.003.
[68] Z. Xing, G. Li, Intelligent classification method of remote sensing image based on big data in Spark environment, International journal of wireless information networks 26 (2019) 183–192. doi:10.1007/s10776-019-00440-z.
[69] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: 17th international conference on neural information processing systems, 2004: pp. 1601–1608.
[70] H. Qiu, E.R. Hancock, Graph matching and clustering using spectral partitions, Pattern recognition 39 (2006) 22–34. doi:10.1016/j.patcog.2005.06.014.
[71] D.T. Pham, A.B. Chan, Control chart pattern recognition using a new type of self-organizing neural network, Proceedings of the Institution of Mechanical Engineers. Part I: Journal of systems and control engineering 212 (1998) 115–127. doi:10.1243/0959651981539343.
[72] S. Hettich, S.D. Bay, The UCI KDD Archive, Irvine, CA: University of California, department of information and computer Science (1999).
[73] F. Lin, W.W. Cohen, Power iteration clustering, in: 27th international conference on machine learning (ICML-10), Haifa, Israel., 2010: pp. 655–662.
[74] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976. doi:10.1126/science.1136800.
[75] J. Wang, Q. Cheng, W. Lu, Y. Dou, P. Li, A term function-aware keyword citation network method for science mapping analysis, Information processing & management 60 (2023). doi:10.1016/j.ipm.2023.103405.
[76] V. Coates, M. Farooque, R. Klavans, K. Lapid, H.A. Linstone, C. Pistorius, A.L. Porter, On the future of technological forecasting, Technological forecasting and social change 67 (2001) 1–17. doi:10.1016/S0040-1625(00)00122-0.