Text mining based theme logic structure identification: application in library journals

Qing Zhu and Yiqiong Wu
Institute of Cross-Process Perception and Control, Shaanxi Normal University, Xi'an, China
Yuze Li
Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada
Jing Han
Institute of Cross-Process Perception and Control, Shaanxi Normal University, Xi'an, China, and
Xiaoyang Zhou
Institute of Cross-Process Perception and Control, Shaanxi Normal University, Xi'an, China and
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Received 22 October 2017
Revised 16 December 2017
17 December 2017
31 December 2017
Accepted 31 December 2017

Abstract

Purpose – Library intelligence institutions, which are a kind of traditional knowledge management organization, are at the frontline of the big data revolution, in which the use of unstructured data has become a modern knowledge management resource. The paper aims to discuss this issue.
Design/methodology/approach – This research combined theme logic structure (TLS), artificial neural network (ANN), and ensemble empirical mode decomposition (EEMD) to transform unstructured data into a signal-wave to examine the research characteristics.
Findings – Research characteristics have a vital effect on knowledge management activities and management behavior through concentration and relaxation, and ultimately form a quasi-periodic evolution. Knowledge management should actively control the evolution of the research characteristics because the natural development of six to nine years was found to be difficult to plot.
Originality/value – Periodic evaluation using TLS-ANN-EEMD gives insights into journal evolution and allows journal managers and contributors to follow the intrinsic mode functions and predict the journal research characteristics tendencies.

Keywords Big data, Knowledge management, Machine learning, Text mining, ANN, EEMD
Paper type Research paper

1. Introduction

The rapid developments in mobile communications technology and the commensurate rise in big data have resulted in a large-scale technological revolution in information interactions and applications (Chen *et al.*, 2014; Hashem *et al.*, 2016; Issam *et al.*, 2014;

The authors thanks those that have given constructive comments and feedback to help improve this paper. Supported was provided by the National Natural Science Foundation of China (71401093, 71350007, 91646113).

Library Hi Tech
Vol. 36 No. 3, 2018
pp. 411-425
Emerald Publishing Limited
0737-8831
DOI 10.1108/LHT-10-2017-0211

Su *et al.*, 2016; Deng and Liu, 2017), which has had a profound impact on information acquisition and access to media and objects, and revolutionized information utilization (Hashem *et al.*, 2016; Zhang *et al.*, 2016; Wu *et al.*, 2013). Big data has a wide range of applications, such as artificial intelligence, data mining, machine learning, and information aggregation (Chen *et al.*, 2014; Bello-Orgaz *et al.*, 2016; Xu *et al.*, 2016). The big data index change at http://trends.google.com has tracked the use of big data from its beginnings in around 2012, and found that it has been receiving significantly more attention from science, business, and governments (Gandomi and Haider, 2015). As traditional knowledge management organizations, library intelligence institutions are at the frontline of the big data revolution, but also face great challenges in the big data era (Ko *et al.*, 2016; Lu *et al.*, 2016).

The expansion of knowledge applications can be seen in the improvements made to automated unstructured data technology in terms of data and text integration and the enabling of more intelligent data analysis processing (Baars and Kemper, 2008; Kai *et al.*, 2008). Al-Daihani and Abrahams (2016) analyzed tweets from ten academic libraries using a text mining approach, and found that applying text mining to social media data could assist managers make better customer service decisions. Compared to Bayesian statistics, humans can more easily conduct objective measurements of significant subjective tendencies, thus enabling future trends to be accurately predicted for more effective decision making (Tsui *et al.*, 2014).

Hu *et al.* (2013) applied co-word analysis to explore the research advances in the Library and Information System (LIS) in China, and used keyword clustering to identify the current status and trends in the LIS research topics. Figuerola *et al.* (2017) applied a Latent Dirichlet allocation to identify the key topics in LIS academic productions and grouped them into four main areas.

These studies have harnessed the power in the LIS. Therefore, in this paper theme logic structure (TLS) and the trends in each journal are examined as these can directly affect a manager's knowledge management behavior. However, as it is difficult to transform unstructured topic representations into structured data to conduct further mathematical research, subjective choices based on unstructured data must be made, which could lead to inconsistent standards, making the results unsuitable for reference.

The purpose of this study, therefore, is to use the text mining process to examine the possibility of using the text mining process for Library Intelligence Knowledge Management to identify the research characteristics, influence, and future development trends in library information and knowledge management journals. The main contributions of this paper are summarized as follows: first, A novel text mining based approach is proposed for prediction and transforming unstructured data into structured data with less human intervention. Second, the TLS building process of this study can be used to develop a novel retrieval method, which would result in more satisfactory results. Third, by using TLS-artificial neural network-ensemble empirical mode decomposition (TLS-ANN-EEMD), the overall research characteristic tendencies of journal and knowledge preferences of the editors, reviewers, and journal contributors can be identified. And all these results not only can assist the managers determine and/or change the scope of their journal, but can assist the authors in choosing the most suitable journal for their work.

The remainder of this paper is organized as follows. In Section 2, the research methodology is briefly introduced, after which in Section 3, the research design and data collection methods are outlined. The results are discussed in Section 4, and Section 5 presents some concluding remarks.

2. Related work

2.1 Text mining

Text mining technology, which can effectively extract useful information from unstructured text data, was developed from data mining technology (Zeng *et al.*, 2012). Tan (1999) found that about 80 percent of the information in an organization was hidden in the text files related

to the organization. Text mining technology now has a mature technical foundation in text classification, text clustering, sentiment analysis, automatic summarization, and association analysis (Thijs, 2017; Pang *et al.*, 2008). Therefore, this paper selected text mining to extract topic information from journals.

2.2 The EEMD

The Hilbert-Huang transformation (Huang *et al.*, 1998) involves a core algorithm called empirical mood decomposition (EMD) to deal with nonlinear and non-stationary signals. Based on the extreme value distribution of the signals, the EMD method decomposes a signal with poor performance into a set of intrinsic mode functions (IMFs) and a residual term called the trend term (Lei *et al.*, 2013). IMFs must meet the following two conditions: in the data set, the number of extrema and the number of zero-crossings must either be equal or differ at most by one; and at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. With these IMFs constraints, the EMD method requires several steps to decompose a signal, as follows:

(1) Identify all the maxima and minima in the original signal $y(t)$ ($t = 1, 2, \ldots, n$).

(2) Use the spline interpolation method to connect all the maximum points to form the maximum envelope $e_{max}(t)$. At the same time, connect all the minimum points to form the minimum envelope $e_{min}(t)$.

(3) Calculate the mean of maximum and minimum envelop $p(t)$: $p(t) = (e_{max}(t) + e_{min}(t))/2$.

(4) Calculate the difference between $y(t)$ and $p(t)$, denoted by $h(t)$: $h(t) = y(t) - p(t)$.

(5) Repeat steps 1-4 until $h_k(t)$ satisfies the conditions of become an IMF. The following equation is used to decide whether $h_k(t)$ can become an IMF:

$$C_k = \frac{\sum_{t=1}^{n} \left| h_{(k-1)}(t) - h_k(t) \right|^2}{\sum_{t=1}^{n} \left| h_{(k-1)}(t) \right|^2} \tag{1}$$

If C_k is less than a predetermined value, $h_k(t)$ can be considered as an IMF.

(6) Once the first IMF $P_1(t)$ is obtained, a residual term $r(t)$ can be separated from $P_1(t)$, the formula for which is: $r(t) = y(t) - P_1(t)$, where $r(t)$ is regarded as a new primitive signal $y(t)$. Then repeat steps 1-5 until the second IMF is obtained.

(7) When $r(t)$ is a monotonic function or a function that has an extreme value and an IMF cannot be extracted, the process terminates (Huang *et al.*, 2003), after which the original signal $y(t)$ can be expressed as the sum of the IMFs and the residual term:

$$y(t) = \sum_{i=1}^{m} P_i(t) + r_m(t) \tag{2}$$

where m is the number of IMFs, and $r_m(t)$ represents the final residual term.

However, EMD has a mode mixing problem. To overcome the EMD disadvantages, Wu and Huang (2011) proposed an efficient decomposition nonlinear, non-stationary signal noise assisted data analysis method, the EEMD, which was able to distribute the additional white noise evenly across the time-frequency space, which allowed for the separation of the signal regions of the different scales while at the same time resolving the EMD Mode Mixing problem.

The steps used in the EEMD are as follows:

(1) Add white Gaussian noises $g(t)$ into the original signal $y(t)$, and get the signal $X(t)$:

$$X(t) = y(t) + g(t). \tag{3}$$

(2) Set the ratio of the standard deviation of the added noises to 0.2-0.3.

(3) Using the EMD method, decompose $X(t)$ into several IMFs.

(4) Repeat steps 1-3 to obtain the corresponding IMFs for the different white noise sequences.

(5) Calculate the mean value of all IMFs and the mean of the residual terms.

2.3 ANN

ANNs are information processing systems that simulate the behavior of the human brain and can estimate the output of the problem based on a given training set. In recent years, ANN has been widely used in complex system prediction and decision making (Panapakidis and Dagoumas, 2016; Kuo *et al.*, 2010; Ahn *et al.*, 2017; Wang *et al.*, 2006; Zeng *et al.*, 2017). A number of studies have shown that ANN can automatically approximate the function forms that best represent the data characteristics (Keles *et al.*, 2008; Wang *et al.*, 2015). And the performance of neural networks can also be optimized by some other algorithms, such as the fruit fly optimization algorithm and its variants (Wang *et al.*, 2016).

The ANN model is a forward-feeding neural network that can estimate the input-output relationships without any prior knowledge (Hosseinpour *et al.*, 2016). Typically, ANN consists of an input layer, multiple hidden layers, and an output layer. Because the hidden layers have a significant impact on the efficiency of neural networks (Asfaram *et al.*, 2016), to determine the optimal number of hidden layers for each neural network, this study adopts a random sampling method to extract 10 percent of a specific journal as the training set, and changes the number of hidden layers multiple times to train the neural network. All ANN classification accuracies in this study are maintained at a 98.5-99.0 percent level. Figure 1 shows a schematic diagram for the ANN.

The input vector of ANN's input layer is expressed as $I = (I_1, I_2, ..., I_n)^T$. The output of X neurons in the hidden layer is expressed as $H = (H_1, H_2, ..., H_x)^T$, and the output vector of the

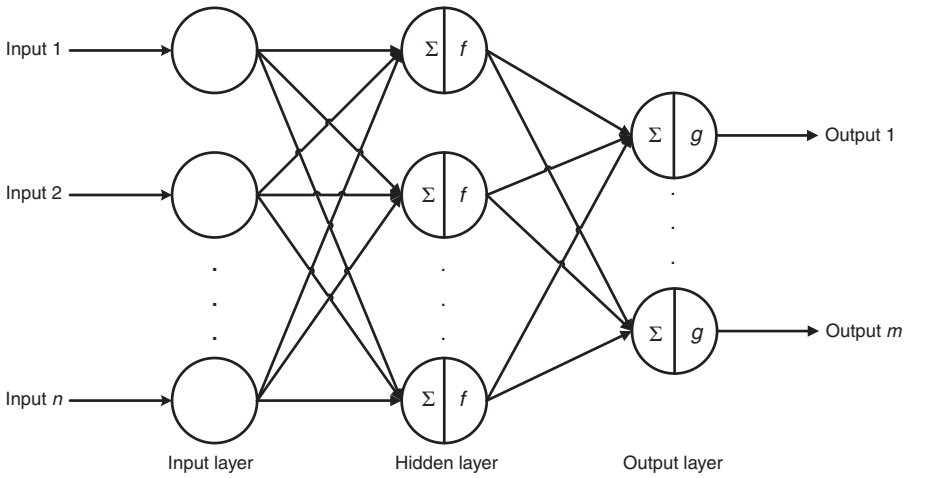

Figure 1.
ANN structure

output layer is represented as $O=(O_1, O_2, ..., O_m)^T$. Suppose the weight between the input layer and the hidden layer is w_{ij}, the weight between the hidden layer and the output layer is w_{jk}, and the input I_i is known, then, the output of hidden layer and the output layer can be calculated by the following equation:

$$H_j = f\left(\sum_{i=1}^{n} w_{ij} I_i - \theta_j\right) \tag{4}$$

$$O_k = g\left(\sum_{i=1}^{q} w_{jk} H_i - \theta_k\right) \tag{5}$$

where $f(\cdot)$ and $g(\cdot)$ in Equations (4) and (5) represent the activation functions. In ANN, all neurons except for the input layer neurons need activation functions to map the neuron inputs and outputs.

3. Research design and data collection

3.1 Modeling and processing

In this study, four consecutive processing processes were arranged in order as follows. In the first stage, all the original data from the abstracts and the body of the paper in the Web of Science database were extracted, after which the TM package in *R* was used to preprocess the text, and the main ingredients extracted from each journal and the ultra-high component set as output. In the second stage, the ultra-high component was processed and the χ^2 distance criterion used to solve the component association matrix. The relationship between the components was then transformed into vectors, with vectors with associations lower than a certain threshold being eliminated. Then, combination and text disambiguation were conducted (Zhong and Enke, 2017). In the third stage, the weights of the relationships were merged, and the merged components input into the ANN network input layer. The weights of the input layer units were then adjusted to train the ANN network, and all samples input into the ANN network for annotation. In the final stage, the ANN annotations were transformed into a time series, and the residual term set obtained using EEMD, which was a non-parametric pseudo decomposition method or a trend set (Lei *et al.*, 2013). Figure 2 shows how the processing processes act on a single journal. The details of the final three processing stages are shown in the following list.

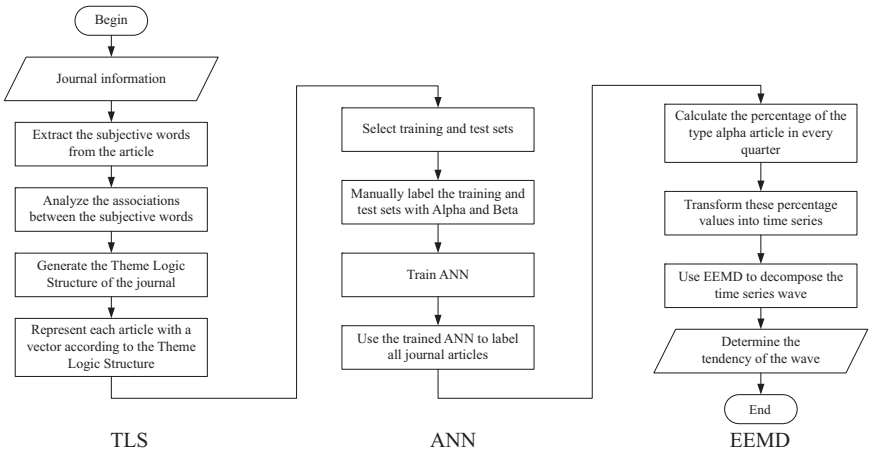

Figure 2.
Processing processes of single journal

Details of TLS-ANN-EEMD:

1: Journal $\leftarrow \{x_i\}_{i=1}^{n}$, where n is the number of article in this journal.
2: $x_i \leftarrow \{x_{ik}\}_{k=1}^{t}$, where x_{ik} is the weight of the corresponding subjective word, t is the number of subjective words of this journal.
3: Initialize x_{ik} with one.
4: $L_i \leftarrow$ the label of the number i article.
5: **FOR** i in 1:n **do**
6: **FOR** k in 1:t **do**
7: **IF** the corresponding subjective word of x_{ik} associates with other words
8: Update the value of x_{ik}, and $x_{ik} \in [3,5]$.
9: **ENDIF**
10: **ENDFOR**
11: **ENDFOR**
12: train set $\leftarrow$ sample(n,10%n)
13: test set $\leftarrow$ sample(n,10%n)
14: Manually label the L_i of the training set and test set with Alpha and Beta.
15: ***Train ANN:***
16: ANNInput $\leftarrow x_{ik}$
17: ANNOutput $\leftarrow L_i$
18: **WHILE** prediction accuracy $\notin$ [98.5%,99.0%]
19: Change the number of units in the hidden layer.
20: **ENDWHILE**
21: Use the trained ANN to label all articles in this journal.
22: Calculate the percentage of the article labeled Alpha in every quarter.
23: Transform the results from step 22 into a time series.
24: Use EEMD to decompose the time series wave to determine the overall tendency of the wave.

3.2 Data collection

Based on the modeling and processing process, this study examined the data collection conditions. Due to the non-structural data processing characteristics of text mining, the data processing objects were collected more broadly, and all data except graphs and tables were collected and processed (Tseng *et al.*, 2007). Then, using the Journal Citation Reports (JCR) partition scientific principle (Bensman and Leydesdorff, 2009), an equal amount of data was collected from the four sub-regions in the JCR partitions. In total, 20 *Library Information Science* journals were identified, which accounted for 23 percent of all relevant journals in the entire region, as shown in Table I.

From the 2016 JCR partitions, to fully distinguish between *Library Information Science* and *Information Systems* journals, all journals from information systems, information management, and information were eliminated. The total size of the original data was 2.76×10^9 bits (2.76 billion bits), or 15,317 papers. It was found that generally, most journals were published quarterly. Even though Qian (2007) found that high-frequency fluctuations caused few disturbances in the long run, high frequency and seasonal items were ignored in the EEMD decomposition modeling process, and monthly issues for the JCR partitions were supplemented with quarterly issues and quarterly interval processing performed on the data. After testing, no statistically dominant red shift was observed. While the Science Citation Index (SCI) dynamically retrieved all samples, some parts were not completely retrieved; therefore, all data collected by the SCI from 2000 were used as the standard to normalize the data start time. After the preliminary data processing, no data continuity defects were found.

JCR-Q1	*Information and Management (IM)*
	International Journal of Geographical Information Science (IJGIS)
	Journal of the Association for Information Science and Technology (JAIST)
	Research Evaluation (RE)
	Scientometrics (S)
JCR-Q2	*Information Technology & People (ITP)*
	Journal of the Medical Library Association (JMLA)
	Learned Publishing (LP)
	Qualitative Health Research (QHR)
	Telecommunications Policy (TP)
JCR-Q3	*Journal of Librarianship and Information Science (JLIS)*
	Library Quarterly (LQ)
	Library Hi Tech (LHT)
	Program-electronic Library and Information Systems (PLIS)
	Revista Espanola de Documentation Cientifica (REDC)
JCR-Q4	*Journal of Scholarly Publishing (JSP)*
	Library Trends (LT)
	Libri (L)
	Restaurator-international Journal for the Preservation of Library and Archival Material (RJPLAM)
	Serials Review (SR)

Table I. Journals used in this paper

4. Results and discussion

The question that needed to be resolved was whether the decline in the information entropy as a result of the unstructured data time series signal derivation process would skew the results and affect the final conclusions. There were three obvious information entropy folds during data processing. In the first text mining stage, the entire information entropy was folded into a set of mapping relationships between the component set and the component with the highest dimension, and dimensionality reduction and disambiguation was carried out. Using ANN, the set mapping relationships in the third stage were transformed into classifications after input, and then the second information entropy folding took place. Finally, in the fourth stage, the EEMD model transformed the posterior class probabilities into high, medium and long frequency waveforms in the time series, and a third information entropy took place.

To prove the modeling and processing appropriateness, the process rationality problem was transformed into the following target to ensure the reverse end to start process that guaranteed the existence of a solution set, and proved the relationship between the forward and reverse derivations. When the inverse operational relationships approach symmetry, a determination was made as to whether the information entropy loss was acceptable. The details of the different stages are discussed in the results.

4.1 Detecting the TLSs

Preprocessing was conducted using the TM package in *R* to remove all stop words in the sample, to generate word frequencies based on English grammar rules, and to run the *findAssocs*(·) function. Associations were found between the 22 highest-frequency words in each journal, as shown in Figure 3.

In the comparative study, three different models were used to determine the compositional research characteristics in each journal. The TOPIC model found that the average number of topics in a single journal was 73.22, the TLS generated by the model was fixed at 22 dimensions, and the average number of dimensions generated by the keyword system was 6.09. Three sets of components were input into the support vector machine (SVM) model, which was then instructed to locate the correct file number, with the files being arranged from highest to lowest according to the prediction probability.

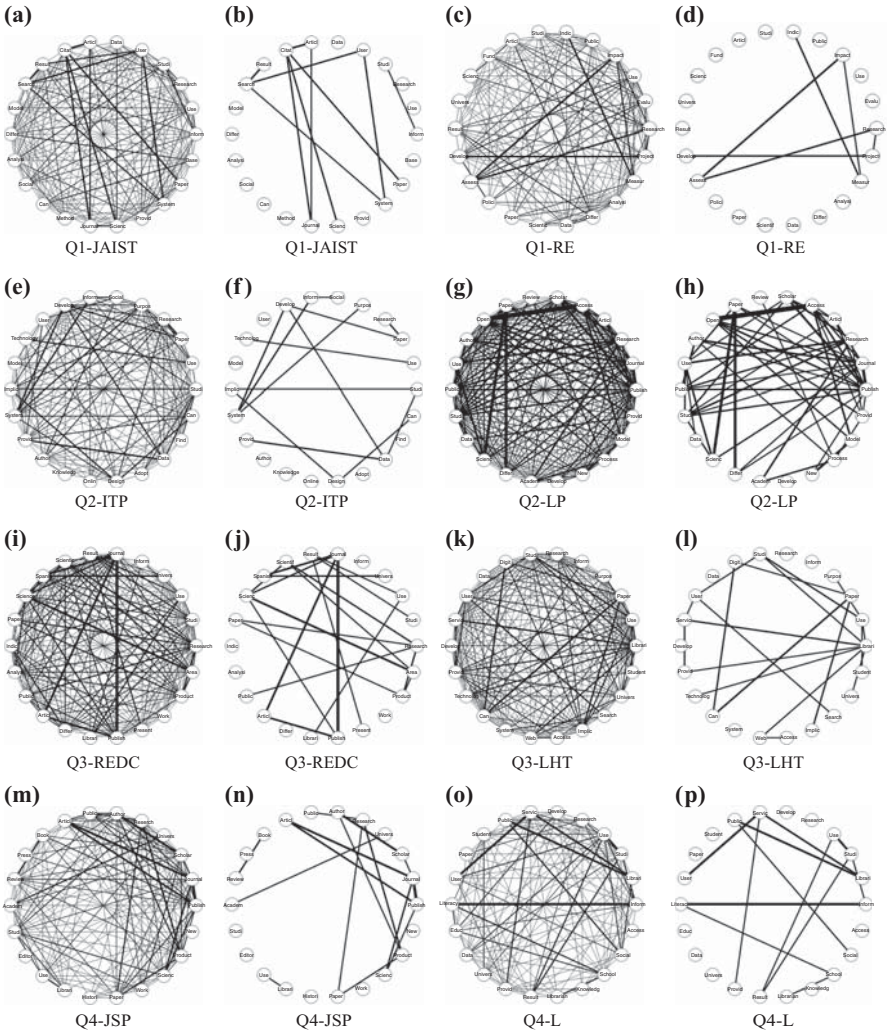

Figure 3.
Association between the 22 highest-frequency words in each journal

When the target number of the first five highest probabilities in the solution set taken by SVM was taken into account, it was concluded that the SVM could successfully backtrack based on composition. When 55 files were randomly selected from the journals, the accuracies when using TOPIC input, TLS input, and the SVM keyword system input were, respectively, 71.80, 89.01 and 55.60 percent (additional experiments using adaptive fuzzy-neural network achieved an accuracy of 99.3 percent. However, as ANN was used, to avoid self-auto-demonstration, ANN and its alternative forms were not used for backtracking). As it was understood that information entropy was inevitable for unstructured data regardless of the processing method adopted, there may have been a quantitative comparative advantage value in the component dimension.

The fundamental purpose of traditional keyword systems is to facilitate library information searches; therefore, to maximize information collection, the dimensions needed

to be reduced and the class widths increased. As the design principles and application field have little relationship with the research characteristics, the backtracking effectiveness was the worst. The TOPIC model used a maximum dimension design using a greedy algorithm and retained the high dimension component set; however, it ignored the relationships between all components, which may have incorrectly folded or covered the necessary component relationships. Busch and Ferretti-Gallon (2017) found that serious endogeneity in high-dimensional structures could lead to a decrease in prediction and location accuracy. The results of the backtracking experiment showed that the entropy failure of the TLS was acceptable, and was also a better choice.

The purpose of the TLS is a dynamic development based on a similar publication process for journal submissions, peer reviews, and publications, and is able to clearly show the relationships between the composition proportion and the correlation strengths. Through editing management, the journal editors and reviewers jointly promote the formation of and changes in the journal research characteristics. Unfortunately, even though the dynamic process was understood, during the calculation process, if no journals accorded with the statistical conditions in a single published period (more than 55 articles) of seasonal continuous data, the structural changes were unable to be fully measured on the time scale. As all data can only be compressed on the cross section to obtain a static overlapping shape, if the TLS changes slowly on the time axis, then the static structure is very similar to the present state. This study found that in the later part, the changes in the structure of the subject logic were quite slow, and six to nine years were required to make a structural change in the overall journal features.

As shown in Figure 3, the logical structure of the subject was simplified, and the JCR partitions did not impose any constraints on the centrality of the intended logical structure. However, the JCR Q1 region journals were found to have one of the most distinctive features and the least component associations in the study. Articles published in the journal shared the most narrowly studied subjects and research topics. In contrast to the other four partitions, a continuous spectrum relationship was not found in the JCR partition, possibly because of the following: the number of *Library and Information Science* journals in the SCI catalog was insufficient to find any continuous change states in the cross-sectional data; and each journal's research setting was decentralized, so there was no continuous state distribution between the journals.

The TLS was confirmed to be a better choice when seeking to resolve the research characteristics composition of a journal. Therefore, the TLS building process was applied to detect the Keyword Logic Structure in each article, and a novel retrieval method developed. Users tend to input multiple keywords to retrieve documents, and usually there is an implicit assumption that there is a logical relationship between these keywords. However, it was found that traditional retrieval methods generate results that only contain these keywords without considering the logical relationships between them. This proposed retrieval method, however, could allow users to search articles using keywords and their logical relationships, which would result in more satisfactory results.

4.2 Transforming the TLSs into signal waves

The TLS purpose for using a general approach for the simplified weight conversion process is because the relationship between the components is non-directional, with the two components sharing an edge. The two connected points were assigned equal weights as the input for the ANN. In the ANN model for each journal, all 22-k-1 structures were tested to determine the number of K neurons until the neural network achieved the highest accuracy rate in the randomly selected training sets (10 percent of all samples). The ANN accuracy for all modified structures was found to be between 98.5 and 99.0 percent. Then, the TLS was used as the basis for the mathematical decisions for each article, and all articles in a journal

were marked with Boolean mathematics. If the Bayesian probability was shown to approximate the TLS, then it was labeled Alpha, otherwise it was labeled Beta. At the same time, the journal publishing time series ruler was introduced to mark a proportion of the Alpha class on the time scales. All text information was transformed into a complex mixing time series, as shown in Figure 4.

From a comparison of the ANN input and output dimensions, 22 dimensions were folded into a single Boolean mathematical value. Therefore, it was necessary to re-examine the inverse operation. By inputting the article category, the solution was a distribution similar to the ANN annotation. After selecting a random full article from the journal, *X*-Means clustering was conducted on the two classes labeled by ANN (Alpha and Beta Class) with the assumption that the error value was the number of articles in the smallest class; the calculation results for the Alpha and Beta and the error value class were 4.2, 7.9, and 12.1 percent, respectively. The experiments showed that due to the large number of reduced component dimensions, the ANN information entropy folding process did not cause any serious distortions; therefore, the results could be used as the input values to derive a roughly accurate distribution.

4.3 Decomposition of the signal waves by EEMD

The transformed time series was decomposed by EEMD, and the results are shown in Figure 5. The research of Lei *et al.* (2013) on EEMD found that the volatility of the time series could be determined from the combination of the IMFs and the trend terms, in which the residual term was a long-term trend. This study found that the journal research characteristics presented a quasi-periodic fluctuation over the long-term trends. There were four journals in the Q1 area that were beginning to diverge from a high concentration and were beginning to focus on topics that had not appeared in the research features. The most direct reason for this was probably because the theme was too concentrated, meaning that the published articles were very similar, they shared common study objects and methods, and had obtained similar results. As knowledge management journal editors and reviewers tend to prefer more diversified research, through specific review and selection procedures, they gradually modify the research characteristics. The journals in Q3 and Q4 were found to have an enhanced concentration state, which presented as a narrowing of the research questions, objects, and methods. If research is too broad, the characteristics of the dominant journal may not account for all areas, particularly if the editors and reviewers highlight only papers which have a common research object. Therefore, with a focus on long-term adjustments and revisions, the characteristics can include a sum of the performances over time.

In practice, describing knowledge management activities is more complex as both previously published articles and new articles are referenced by the editors and reviewers. At the same time, many journals have diversity when faced with special or non-characteristic research. Nonetheless, it is speculated that under certain conditions, the long-term effect of the centralized characteristics (concentration and relaxation) on the knowledge management activities could result in a more balanced direction and become consistent. Under a long period of relaxed research publication in which more articles are accepted and published, the similarities between the articles could be less evident. Therefore, from a logical perspective, when relaxing the journal article submission boundaries and accepting a wider range than the traditional research content, areas such as Big Data, Cloud Computing and other emerging scientific issues would tend to appear more frequently in the journal relaxation phase.

The quasi-periodic variations in the characteristics were very slow. As observed from the time scale in Figure 5, it took six to nine years to adjust the rise or decline of the cycle. Although the contributors, editors, and reviewers had a preference for knowledge management activities, they appeared to be very random in their daily activities. The interactions between the journal subjects with the journal as an interface reached

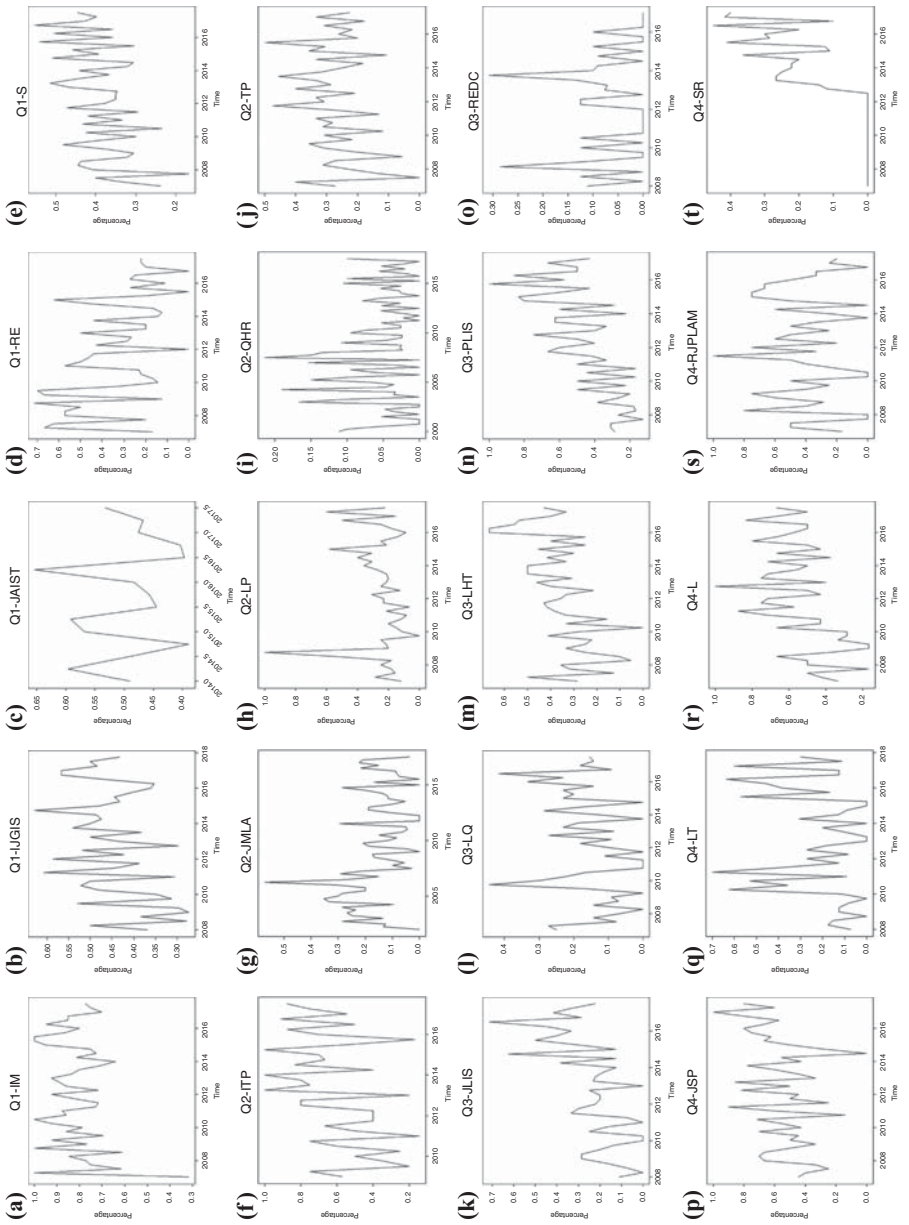

Figure 4.
Seasonal percentage time series for every journal

Figure 5.
IMFs and residual terms obtained by EEMD

consensus very slowly as high volumes were necessary to make any meaningful judgment; that is, the interactions in ten or more volumes in terms of concentration and relaxation, random daily activities, and preference levels can ultimately affect the judgment. There was also a lack of strong data to confirm that the quasi-periodic law for the research characteristics was similar in all disciplines; therefore, it was only possible to hypothesize that the quasi-periodic law had different rates of change in the different subject journals. Because of the EEMD non-parametric pseudo decomposition method, in addition to studying the quasi-periodic variations in the characteristics, the lack of factual observation and evidence affected the understanding of the high-frequency and intermediate frequency IMFs. Therefore, it was not possible to confirm that the high-frequency IMF was a result of the daily activities of the reviewers, and not possible to identify if the medium frequency IMF was part of the knowledge management activities. Therefore, determining all IMF frequencies and knowledge management activities was highlighted as a key issue for future research.

By combining the TLS of a single journal with ANN and EEMD, the overall research characteristic tendencies and knowledge preferences of the editors, reviewers, and journal contributors were identified, which could assist authors in choosing the most suitable journal for their work, and give journal managers a scientific tool to identify the research characteristics of their journals. Further, by combining text mining technology and big data thinking, this method reduces the need for intensive human intervention, which in turn reduces the time necessary to identify journal research characteristics.

The identification of journal management tendencies can also assist in knowledge management. Journal managers can ensure that they screen contributions more strictly if they have access to the research characteristics results. Conversely, if managers are wishing to decentralize their journal's research characteristics, they may choose to publish a special issue or expand the journal's topic range. Therefore, the research in this paper can assist managers determine and/or change the scope of their journal.

5. Conclusion

To provide information and structural ideas for knowledge management activities, this study presented text mining data processing technology that involved information transformation, processing, and prediction. It was proven that the information entropy folding was acceptable during the unstructured data to signal process in the forward derivation and inverse operations required for trend predictions. Research characteristics influence knowledge management activities and, depending on the degree of concentration and relaxation, can affect management behavior and lead to quasi-periodic evolution. Because of knowledge management behavioral inertia, the research characteristics were found to follow a monotonic long-term trend. It was found that when the concentration or relaxation was saturated, the research characteristics were pushed in the opposite direction, and underwent a periodic change.

There were some limitations in this research. First, there were limitations in the observation and cognition of the knowledge management activities, and consequently it was not possible to successfully determine the knowledge management activities using the IMFs. Second, when integrating a large number of algorithms, while the appropriate process was selected, the mathematical optimization of the process was not confirmed. Because of the scale of big data, the mathematical and physical conditions were limited and lacked powerful tools. Finally, this study was limited to a specialized library information disciplines, and no comparison of the knowledge management activities in different disciplines was conducted, which made it impossible to fully determine the quasi-periodic differences in the research characteristics. These limitations will be addressed in future research.

References

Ahn, J., Cho, S. and Chung, D.H. (2017), "Analysis of energy and control efficiencies of fuzzy logic and artificial neural network technologies in the heating energy supply system responding to the changes of user demands", *Applied Energy*, Vol. 190, pp. 222-231.

Al-Daihani, S.M. and Abrahams, A. (2016), "A text mining analysis of academic libraries' tweets", *Journal of Academic Librarianship*, Vol. 42 No. 2, pp. 135-143.

Asfaram, A., Ghaedi, M., Azqhandi, M.H.A., Goudarzi, A. and Dastkhoon, M. (2016), "Statistical experimental design, least square-support vector machine (LS-SVM) and artificial neural network (ANN) methods for modeling of facilitated adsorption of methylene blue dye", *RSC Advances*, Vol. 6 No. 46, pp. 40502-40516.

Baars, H. and Kemper, H.G. (2008), "Management support with structured and unstructured data – an integrated business intelligence framework", *Information Systems Management*, Vol. 25 No. 2, pp. 132-148.

Bello-Orgaz, G., Jung, J.J. and Camacho, D. (2016), "Social big data: recent achievements and new challenges", *Information Fusion*, Vol. 28, pp. 45-59.

Bensman, S.J. and Leydesdorff, L. (2009), "Definition and identification of journals as bibliographic and subject entities: librarianship versus ISI Journal citation reports, methods and their effect on citation measures", *Journal of the American Society for Information Science & Technology*, Vol. 60 No. 6, pp. 1097-1117.

Busch, J. and Ferretti-Gallon, K. (2017), "What drives deforestation and what stops it? A meta-analysis", *Review of Environmental Economics & Policy*, Vol. 11 No. 1, pp. 3-23.

Chen, M., Mao, S. and Liu, Y. (2014), "Big data: a survey", *Mobile Networks & Applications*, Vol. 19 No. 2, pp. 171-209.

Deng, Z. and Liu, S. (2017), "Understanding consumer health information-seeking behavior from the perspective of the risk perception attitude framework and social support in mobile social media websites", *International Journal of Medical Informatics*, Vol. 105, pp. 98-109.

Figuerola, C.G., Marco, F.J.G. and Pinto, M. (2017), "Mapping the evolution of library and information science (1978-2014) using topic modeling on LISA", *Scientometrics*, Vol. 112 No. 3, pp. 1507-1535.

Gandomi, A. and Haider, M. (2015), "Beyond the hype: big data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H. (2016), "The role of big data in smart city", *International Journal of Information Management*, Vol. 36 No. 5, pp. 748-758.

Hosseinpour, S., Aghbashlo, M., Tabatabaei, M. and Khalife, E. (2016), "Exact estimation of biodiesel cetane number (CN) from its fatty acid methyl esters (FAMEs) profile using partial least square (PLS) adapted by artificial neural network (ANN)", *Energy Conversion & Management*, Vol. 124, pp. 389-398.

Hu, C.P., Hu, J.M., Deng, S.L. and Liu, Y. (2013), "A co-word analysis of library and information science in China", *Scientometrics*, Vol. 97 No. 2, pp. 369-382.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Chi, C.T. and Liu, H.H. (1998), "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", *Proceedings Mathematical Physical & Engineering Sciences*, Vol. 454 No. 1971, pp. 903-995.

Huang, N.E., Wu, M.L.C., Long, S.R., Shen, S.S.P., Qu, W., Gloersen, P. and Fan, K.L. (2003), "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis", *Proceedings Mathematical Physical & Engineering Sciences*, Vol. 459 No. 2037, pp. 2317-2345.

Issam, B., mohamed, R.L., Richard, H. and Mathew, M. (2014), "Semantic ontologies for multimedia indexing (SOMI): application in the e-library domain", *Library Hi Tech*, Vol. 32 No. 2, pp. 206-218.

Kai, R.L., Monarchi, D.E., Hovorka, D.S. and Bailey, C.N. (2008), "Analyzing unstructured text data: using latent categorization to identify intellectual communities in information systems", *Decision Support Systems*, Vol. 45 No. 4, pp. 884-896.

Keles, A., Kolcak, M. and Keles, A. (2008), "The adaptive neuro-fuzzy model for forecasting the domestic debt", *Knowledge-Based Systems*, Vol. 21 No. 8, pp. 951-957.

Ko, Y.M., Song, M.S. and Lee, S.J. (2016), "Construction of the structural definition based terminology ontology system and semantic search evaluation", *Library Hi Tech*, Vol. 34 No. 4, pp. 705-732.

Kuo, R.J., Wang, Y.C. and Tien, F.C. (2010), "Integration of artificial neural network and MADA methods for green supplier selection", *Journal of Cleaner Production*, Vol. 18 No. 12, pp. 1161-1170.

Lei, Y., Lin, J., He, Z. and Zuo, M.J. (2013), "A review on empirical mode decomposition in fault diagnosis of rotating machinery", *Mechanical Systems & Signal Processing*, Vol. 35 Nos 1/2, pp. 108-126.

Lu, J., Zhou, J., Ruan, H. and Luo, G. (2016), "Establishing a university library-based health information literacy service model in the age of big data", *Journal of Medical Imaging & Health Informatics*, Vol. 6 No. 1, pp. 260-263.

Panapakidis, I.P. and Dagoumas, A.S. (2016), "Day-ahead electricity price forecasting via the application of artificial neural network based models", *Applied Energy*, Vol. 172, pp. 132-151.

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations & Trends in Information Retrieval*, Vol. 2 No. 1, pp. 459-526.

Qian, K. (2007), "Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations", *Optics & Lasers in Engineering*, Vol. 45 No. 2, pp. 304-317.

Su, Z., Xu, Q. and Qi, Q. (2016), "Big data in mobile social networks: a QoE-oriented framework", *IEEE Network*, Vol. 30 No. 1, pp. 52-57.

Tan, A.H. (1999), "Text mining: promises and challenges", *Proceedings South East Asia Research Computer Confederation in Singapore City, SEARCC, Singapore*, pp. 15-21.

Thijs, B. (2017), *Using Hybrid Methods And 'Core Documents' for the Representation of Clusters and Topics: The Astronomy Dataset*, Springer-Verlag, New York, NY.

Tseng, Y.H., Lin, C.J. and Lin, Y.I. (2007), "Text mining techniques for patent analysis", *Information Processing & Management*, Vol. 43 No. 5, pp. 1216-1247.

Tsui, E., Wang, W.M., Cai, L., Cheung, C.F. and Lee, W.B. (2014), "Knowledge-based extraction of intellectual capital-related information from unstructured data", *Expert Systems with Applications An International Journal*, Vol. 41 No. 4, pp. 1315-1325.

Wang, L., Liu, R. and Liu, S. (2016), "An effective and efficient fruit fly optimization algorithm with level probability policy and its applications", *Knowledge-Based Systems*, Vol. 97, pp. 158-174.

Wang, L., Zeng, Y. and Chen, T. (2015), "Back propagation neural network with adaptive differential evolution algorithm for time series forecasting", *Expert Systems with Applications*, Vol. 42 No. 2, pp. 855-863.

Wang, L., Zeng, Y.R., Zhang, J.L., Huang, W. and Bao, Y.K. (2006), "The criticality of spare parts evaluating model using artificial neural network approach", *Lecture Notes in Computer Science*, Vol. 3991, pp. 728-735.

Wu, X., Zhu, X., Wu, G.Q. and Wei, D. (2013), "Data mining with big data", *IEEE Transactions on Knowledge & Data Engineering*, Vol. 26 No. 1, pp. 97-107.

Wu, Z.H. and Huang, N.E. (2011), "Ensemble empirical mode decomposition: a noise-assisted data analysis method", *Advances in Adaptive Data Analysis*, Vol. 1 No. 1, pp. 1-41.

Xu, Z., Mei, L., Hu, C. and Liu, Y. (2016), "The big data analytics and applications of the surveillance system using video structured description technology", *Cluster Computing*, Vol. 19 No. 3, pp. 1283-1292.

Zeng, L., Li, L., Duan, L., Lu, K., Shi, Z., Wang, M., Wu, M., Wu, W. and Luo, P. (2012), "Distributed data mining: a survey", *Information Technology & Management*, Vol. 13 No. 4, pp. 403-409.

Zeng, Y.R., Zeng, Y., Choi, B. and Wang, L. (2017), "Multifactor-influenced energy consumption forecasting using enhanced back-propagation neural network", *Energy*, Vol. 127, pp. 381-396.

Zhang, Y., Ren, S., Liu, Y. and Si, S. (2016), "A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products", *Journal of Cleaner Production*, Vol. 142 No. 2, pp. 626-641.

Zhong, X. and Enke, D. (2017), "A comprehensive cluster and classification mining procedure for daily stock market return forecasting", *Neurocomputing*, Vol. 267, pp. 152-168.

Further reading

Kiyoumarsi, F. (2015), "Evaluation of automatic text summarizations based on human summaries", *Procedia-Social and Behavioral Sciences*, Vol. 192, pp. 83-91.

Corresponding author
Xiaoyang Zhou can be contacted at: x.y.zhou@foxmail.com