An approach to improving the analysis of literature data in Chinese through an improved use of Citespace

Weichen Jia  
Jun Peng  
Na Cai  
City University of Macau, Macau

Recommended citation:  
An approach to improving the analysis of literature data in Chinese through an improved use of Citespace

Weichen Jia
School of Education
City University of Macau, Macau
E-mail: jwc19890114@163.com

Jun Peng*
School of Education
City University of Macau, Macau
E-mail: 4588775@163.com

Na Cai
School of Education
City University of Macau, Macau
E-mail: 1337803185@qq.com

*Corresponding author

Abstract: Citespace, a visualization-based analysis tool, has been used to analyze the literature data by visualizing the patterns and potential trends of a field. Previous studies show that when used for analyzing the literature in Chinese, Citespace could only conduct very basic analysis, different from its use in analyzing the literature data in English. To address this limitation, this study presents an approach to improving the use of Citespace for effective analysis of literature data in Chinese. The approach employs data-processing and data-analysis scripts in data collection, knowledge map generation, and interpretation steps to improve the accuracy and comprehensiveness of analysis of literature data in Chinese. An empirical evaluation has been conducted to demonstrate the effectiveness of the approach.

Keywords: Citespace; Literature analysis; Chinese social sciences citation index; China national knowledge infrastructure

Biographical notes: Weichen Jia is a PhD student of School of Education, City University of Macau. His research interests include educational technology, natural language processing.

Dr. Jun Peng is assistant professor, programme coordinator of School of Education, City university of Macau.

Na Cai is a PhD student of School of Education, City University of Macau. She has completed all of the requirements for the doctoral degree with the exception of the dissertation. Her research interest includes foreign students’ cross culture adaptation.
1. Introduction

With the rapid development of information visualization and data mining technologies, visualization-based software or tools for analyzing literature data have proliferated (Sumangali & Kumar, 2017). With the support of such tools, knowledge mapping for analyzing the structure and trends of a field has received increased attention (Lu & Hu, 2019). Among various visualization-based analysis software or tools, Citespace (http://cluster.cis.drexel.edu/~cchen/citespace/), an information visualization software developed by Dr. Chaomei Chen (Chen 2006), has been applied for analyzing the literature data in many academic fields (Hou & Hu, 2013; Chen, 2010; Van Eck & Waltman 2010; Li, 2018). It has been used to generate and interpret diverse knowledge maps based on literature data (Chen, 2006) and explore research hotspots, frontiers, and new trends in a field (Li & Chen, 2016). However, previous studies point out that Citespace, when used for analyzing the literature data in Chinese, could only conduct very basic analysis (Guo & Chen, 2019; Lin & Dai, 2018; Yu & Zhou, 2018), different from its use in analyzing the literature data in English.

2. Literature review

2.1. Citespace for literature analysis

Knowledge mapping is becoming increasingly important in educational and social studies (Chen, 2017). Consequently, a number of applications have been developed in recent years for analyzing literature data, such as Citespace (Chen, 2006), UCINET (Borgatti et al., 2002), BibExcel(Persson, Danell, & Schneider, 2009), Sci2(Sci2 Team, 2009), VOSViewer (Van Eck & Waltman, 2010), and CitNetExplorer(Van Eck & Waltman, 2014). These tools share their main functions in common with subtle differences involved in their own features and design focuses.

UCINET (Borgatti et al., 2002) is a software package for the analysis of social network data which is usually used to analyze the relationship among the authors and institutions. BibExcel (Persson et al., 2009) is designed to assist users in analyzing bibliographic data, or any data of a textual nature formatted in a similar manner. It focuses on the keyword frequency distribution and co-occurrence metrics. Sci2(Sci2 Team, 2009) is a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of academic datasets at the micro (individual), meso (local), and macro (global) levels. This software allows users to customize the database as a plug-in extension, which means this software has a stronger network constructing functionality. VOSViewer (van Eck & Waltman, 2010) is another tool for constructing and visualizing bibliometric networks. It offers text mining functionality that can be used to construct and visualize co-occurrence networks of key terms extracted from a body of scientific literature. CitNetExplorer (Van Eck & Waltman, 2014) focuses on visualizing and analyzing citation networks of scientific publications. It allows citation networks to be imported directly from the Web of Science database. Citation networks can be explored interactively, by drilling down into a network and by identifying clusters of closely related publications.

Comparing with the functionality of these tools, Citespace, VOSViewer, and Sci2 particularly emphasize on the literature data analysis, the analysis of data from citation indexes, and social network analysis, while CitNetExplorer only focuses on the analysis
of data from citation indexes. In China, Citespace is widely accepted by most users for its strong graphics display capability and large-scale data capacity.

As an information visualization application developed by Dr. Chaomei Chen from Drexel University, USA (Chen, 2006), Citespace has been used to analyze the literature of a field (Chen, 2006; Chen, Hu, Liu, & Tseng, 2012; Chen, 2017) by bibliometric analysis techniques involving author co-cited analysis (ACA) and scientific revolution structure analysis (Kuhn, 1962; White & Griffith, 1981). It provides various functions to facilitate the analysis of underlying patterns of a domain, such as identifying the fast-growth study areas, finding citation hotspots, classifying research types according to keywords, and identifying geospatial collaborations (Chen, 2006). In addition, Citespace can support both structural and unstructured analyses of a variety of networks derived from academic publications, including collaboration networks, author co-citation networks, and document co-citation networks (Chen, 2006).

Citespace has also been extensively applied in teaching and learning of many subjects, such as Big data analysis (Wang, Chen, Wang, & Yang, 2016), science education (Tho et al., 2017), foreign language learning (Xu & Nie, 2015), and education of information literacy (Zhao, Shan, Dong, & Hu, 2016). As a visual-based knowledge mapping and interpretation, Citespace could help users to predict education trends, identify research orientations, and make decisions (Chen, 2006). Besides, the author co-citation networks and document co-citation networks generated by using Citespace could reveal the relationships between authors and research topics in a visual form, which is significant for novices to grasp the status in quo of certain research fields (Chen, 2006).

2.2. Citespace for analyze the literature in English

Two representative studies on Citespace (Chen, 2017; Chen et al., 2012) summarize the typical usage of Citespace in English literature. In general, it consists of three steps: data collection, map generation, and map interpretation (Chen, 2017; Chen et al., 2012), which are briefly presented in Fig. 1.

- **Data collection.** Literature data are searched and collected from Web of Science (Wos). After that, they would be inputted into Citespace for further processing (Chen, 2017; Chen et al., 2012).

- **Map generation.** In this step, various visual-based knowledge maps, such as “concept tree map”, “time-line map” and “cluster map”, would be generated by Citespace based on the inputted data (Chen, 2017; Chen et al., 2012).

- **Map interpretation.** With the aid of diverse analysis measures provided by Citespace (e.g., “discipline analysis”, “topic analysis”, “co-citation analysis”, “typical cluster analysis”, etc.), a comprehensive interpretation involved in research hotspots, core scholars, frontiers, and trend predictions would be afforded (Hu, 2017; Chen, 2017).

2.3. Citespace for analyzing the literature in Chinese

The quality of source data has a strong correlation with the reliability and credibility of the analysis results of Citespace (Hu, 2017; Chen, 2017; Huo & Shi, 2018). However, Chinese literature data are not fully compatible with the Citespace. In practice, usually CSSCI (Chinese Social Sciences Citation Index) or CNKI (China National Knowledge Infrastructure) database would be chosen as the data source to provide literature data to
Citespace. However, the CSSCI data lacks the abstract field, while the CNKI data lacks the reference field (Chinese Social Science Research Assessment Center, 2016; Hou, 2014). Therefore, when using Citespace to analyze the Chinese literature data, the structure of data source would be incomplete seriously. Besides, many relevant studies in recent years have pointed out the insufficient generated knowledge maps and the lack of in-depth map-interpretation methods have been the main limitation of using Citespace on Chinese literature (Huo & Shi, 2018). In most cases, there are just a few knowledge maps (usually only “time-line map” and “cluster map”) could be provided to Chinese users (Guo & Chen, 2019; Lin & Dai, 2018) and they have to relied on their existing knowledge and experience to understand the literature, which is contrary to the original purpose of Citespace that “it offers a new platform for the newcomers to have an objective overview of the target areas” (Guo & Chen, 2019; Lin & Dai, 2018; Yu & Zhou, 2018; Li & Chen, 2016; Chen, Chen, Hu, & Wang, 2014).

![Fig. 1. The typical usage of Citespace in analyzing the literature in English](image)

The typical use of Citespace in Chinese literature is with similar three steps: data collection, map generation, and interpretation (Guo & Chen, 2019; Lin & Dai, 2018; Yu & Zhou, 2018; Huo & Shi, 2018). As mentioned above, the accuracy and comprehensiveness are far less than its English counterpart, as shown in Fig. 2.

- **Data collection.** Either CSSCI or CNKI database is searched by a single or multiple keyword. After that, the raw incomplete data would be input into Citespace without any further processing such as inspection and correction (Guo & Chen, 2019).
• **Map generation.** Data are only used to generate a few visual-based knowledge maps such as "timeline map" and "cluster map" (Guo & Chen, 2019; Lin & Dai, 2018; Yu & Zhou, 2018).

• **Map interpretation.** Some basic analysis measures are offered to interpret the maps generated in the previous step, which may result in the improper interpretation of knowledge maps. (Guo & Chen, 2019; Lin & Dai, 2018; Yu & Zhou, 2018; Huo & Shi, 2018).

![Diagram of Citespace process](image)

**Fig. 2.** The typical usage of Citespace in analyzing the literature in Chinese

### 3. An improved use of Citespace

To address the aforementioned problems, an improved usage (Chinese) is presented in this study. It employs data-processing and data-analysis scripts in data collection, knowledge map generation, and interpretation steps to improve the accuracy and comprehensiveness of analysis of data in Chinese.

#### 3.1. Features

##### 3.1.1. New data field

The abstract is a brief summary of a manuscript, which summarizes the purpose, methods and final conclusions of the study (Wu & Yang, 2020). Therefore, a full-text analysis of
the abstract data could be a comprehensive overview of certain subject. Thus, it is promising to put the abstract into a new data field of the improved usage.

3.1.2. New map generation and interpretation methods

Previous studies indicated that “concept tree map” would be an appropriate method to analyze the abstract data (Chen, Yao, & Yang, 2016; Gong, You, Guan, Cao, & Lai, 2018; Jelodar et al., 2019; Pavlinek & Podgorelec 2017; Shiryaev, Dorofeev, Fedorov, Gagarina, & Zaycev, 2017; Guan, Wang, & Fu, 2016). Concept tree map is a kind of knowledge map that extracts a list of semantic topics and the relationships between the topics in a visual topic map based on co-occurrence analysis of topics in different documents. It is also extensively adopted in Citespace for analyzing literature data in English. It has been used to mine research hotspot (Yang, Li, & Jin, 2012), identify research topic evolution (Li, Li, & Tan, 2014; Li, Zhang, & Yuan, 2014), and predict research trends (Huang, Zhang, Wu, & Tang, 2016; Fan & Ma, 2014). In this study, additional scripts are used to enable Citespace to generate this kind of map and to perform corresponding interpretation of literature data in Chinese.

3.2. Framework

The framework of proposed usage is presented in Fig. 3. As shown, under the support of data-processing script, the raw literature data obtained from CNKI and CSSCI would be merged and refined. Then, “concept tree map” (including a list of topics and a visual topic map) would be produced with the aid of data-analysis script. Finally, various analyses could be achieved in map interpretation step.

![Fig. 3. An improved use of Citespace in analyzing the literature in Chinese](image)
3.2.1. Data collection
First, a data-processing script is used to merge the literature data searched from CSSCI and CNKI. As such, a completed Chinese literature dataset with abstract and reference information is obtained. Then, various measures including missing value detection, setting, and removal of duplicate records would be conducted by the script to enhance the quality of the merged data.

3.2.2. Map generation and interpretation
Data-analysis script is used to assist CiteSpace to achieve “concept tree map”, whereby a list of topics and a visual topic map would be produced. Accordingly, built-in interpretation methods of Citespace would be functionated.

4. Evaluation
In this section, a primary evaluation of the proposed usage is presented, which analyzed the literature data in Chinese in the field of “teacher professional development”. The CSSCI database was chosen as the main data source, where the CNKI database was selected as the supplement to provide abstract data. The time range of the literature is from 2001 to 2018.

4.1. Process
First, a dataset of 1068 CSSCI records without abstract data field were obtained by keyword search. Then, data-processing script was used to inspect, correct and merge the raw data with corresponding abstract data field. After that, data-analysis script was used to assist Citespace to generate a list of topics and a visual topic map. At last, abstract topics interpretation and high-cited interpretation were processed by Citespace.

4.2. Result
Table 1 presents a list of six topics: Rural Teacher, Theory, University Teacher Professional Development, Physical Education Teachers, Teacher Professional Development School, and Preschool Teacher) extracted from 1068 abstracts in the selected field by using Citespace in an improve way proposed in this study. Each topic is associated with a dozen of keywords, based on which the topic can be defined semantically. The visual topic map generated from the data is presented in Fig. 4. The map also shows that the six topics are segmented into 4 regions according to the distance between topics. The inter-topic distances represent the similarity in meaning between topics. Topics 1, 2 and 3 construct the largest region in the middle of the figure, while Topics 4, 6, and 5 are in three other regions with more distance. The areas of the circles are proportional to the relative prevalence of the topics in the corpus. The largest region typically reflects the core topics of the cluster. For example, topics such as rural teacher, university teacher, and theory research are the primary interests of this cluster. The overlap of circles represents cross-topic studies.

Fig. 5 and Table 2 demonstrated the top 9 highest-cited authors and their publications provided by high-cited interpretation, which may be conducive to reveal the Chinese prevailing scholars and knowledge development path of this field over the past decade or so.
Table 1
A list of topics extracted from the abstract data in “Teacher Professional Development”

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Key words of topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>Development, Rural Teacher Professional Development, Improvement, System, Knowledge…</td>
</tr>
<tr>
<td>Topic 2</td>
<td>Realization, Reflection, Understanding, Profession, Theory, Development, Teaching, Practice…</td>
</tr>
<tr>
<td>Topic 3</td>
<td>University, Research, Atmosphere, Professional Development, Development, Promotion, Ability…</td>
</tr>
<tr>
<td>Topic 4</td>
<td>Professional Development of Physical Education Teachers, Physical Education Teachers…</td>
</tr>
<tr>
<td>Topic 5</td>
<td>Teacher Professional Development School, China, USA, Promoting Teacher Professional Development…</td>
</tr>
<tr>
<td>Topic 6</td>
<td>Preschool Teacher Professional Development, British, Planning, Decision, Degree…</td>
</tr>
</tbody>
</table>

Fig. 4. The visual topic map of “Teacher Professional Development”
Table 2
High-cited authors and their publications in “Teacher Professional Development”

<table>
<thead>
<tr>
<th>ID</th>
<th>Author</th>
<th>Literature</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L. Ye</td>
<td>Let the classroom glow with vitality</td>
<td>1998</td>
</tr>
<tr>
<td>2</td>
<td>M. Fullan</td>
<td>The three stories of education reform</td>
<td>2000</td>
</tr>
<tr>
<td>3</td>
<td>C.-T. Hsu</td>
<td>Restructuring school enable to remodel teachers' professional development</td>
<td>2004</td>
</tr>
<tr>
<td>4</td>
<td>H. Borko</td>
<td>Professional development and teacher learning: Mapping the terrain</td>
<td>2004</td>
</tr>
<tr>
<td>5</td>
<td>G. Song &amp; S. Wei</td>
<td>On teachers' professional development</td>
<td>2005</td>
</tr>
<tr>
<td>6</td>
<td>X. Zhuang</td>
<td>Pursuing excellence begins with learning: Action for teachers' professional development</td>
<td>2005</td>
</tr>
<tr>
<td>7</td>
<td>W. Yuan</td>
<td>The pedagogical content knowledge: The new perspective of teacher professional development</td>
<td>2005</td>
</tr>
<tr>
<td>8</td>
<td>A. Webster-Wright</td>
<td>Reframing professional development through understanding authentic professional learning</td>
<td>2009</td>
</tr>
<tr>
<td>9</td>
<td>T. Cao &amp; F. Li</td>
<td>Transcending the dilemma: An analysis of novice teachers' professional development under the performance-based salary system</td>
<td>2011</td>
</tr>
</tbody>
</table>

Fig. 5. High-impact authors in “Teacher Professional Development”

5. Conclusion

This study provides an improved use of Citespace for analyzing the literature in Chinese. The improvement focused on data collection, knowledge map generation, and
interpretation steps by employing data-processing and data-analysis scripts to improve the accuracy and comprehensiveness of the analysis functions of Citespace.

The empirical evaluation showed that the improved approach can collect the abstract data as a new analytic dataset, and thus generate a list of topics and a visual topic map. When analyzing the literature data in Chinese in the field of “teacher professional development”, the improved approach could figure out six topics, divide the topics into four regions, and identify prevailing scholars and knowledge development path in the field, all of which indicate the promising effects of the proposed approach in improving the accuracy and comprehensiveness of analysis of literature data in Chinese.

ORCID
Weichen Jia https://orcid.org/0000-0002-0749-2859
Jun Peng https://orcid.org/0000-0003-1802-7300
Na Cai https://orcid.org/0000-0003-3520-0072

References


Zhao, Y., Shan, C., Dong, H., & Hu, G. (2016). Evolution path and research hotspot in international information literacy education field based on citespace II. In *Proceedings of the 2nd International Conference on Communication and Information Processing* (pp. 79–82). ACM.