ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 10 (2020) No. 3, pp. 86-98
DOI:10.33168/JSMS.2020.0307

A Content Mining and Network Analysis Method for the Thematic Evolution of Educational Leadership

Qingyun Li[1*], Jie Han[2*], Yue On Ko[3], Qingling Li[4] and Chui-Man Lo[2]

[1] The Hong Kong Polytechnic University

[2] The Open University of Hong Kong

[3] The Education University of Hong Kong

[4] Shenyang University of Chemical Technology

[1] qingyun.li@gmail.com, [2] chan@ouhk.edu.hk, [3] jamesko@eduhk.hk, [4] waiyu9650@163.com, [5] cmlo@ouhk.edu.hk

Abstract. This study presents a novel approach to analyze the thematic evolution in the research area -'Educational Leadership'. This approach combines the content mining techniques and Network Analysis to detect and visualize the different research thematic topics. The co-word analysis method is used to longitudinally detect the frequency of popular topics and topic words in a given time period (1996-2015). The weighted network structure is used to analyze the relationships between themes through the co-occurrence of themes, with highlighted centrality and community of the network to show the hot topics in the theme and their relationships, and the changes and relationships between themes are presented through dynamic Network Analysis among the 4 periods (1996~2000, 2000~2005, 2006~2010, 2010~2015). With highlighted network centrality & community to show the hot topics in the theme and their relationship, and dynamic Network Analysis to present the changes and trends among the 4 periods (1996~2000, 2000~2005, 2006~2010, 2010~2015). Compared with traditional methods, our method provides direct visual effects, thereby making the structure scalable and providing easy-to-understand methods for public end users. The method in the current study can be widely used in other research fields.

Keywords: Educational leadership, thematic evolution, co-word analysis, network analysis.

1. Introduction

Educational leadership is the process of enlisting and guiding the talents and energies of teachers, pupils, and parents toward achieving common educational aims (Wikipedia). Since the 1960s, different researches have been conducted to study educational leadership's concept landscape by exploring literatures. Meanwhile, some new sub-filed has been developed of the educational leadership, e.g. diverse conceptual lenses (Bates, 1980; Donmoyer, 2001), diverse methods (Briggs *et al.*, 2012). various philosophical propositions (Evers and Lakomski, 2012), and debatable epistemological issues (Oplatka, 2012). Moreover, the thematic evolution has been found in the education leadership filed, because of changes of the politics and policies.

Because of the "information explosion" in the research area in recent 20 years, the problem appears to be in part an oversupply of information in the form of so many journals and books. The demands of researchers to generally catch up the key concepts and researchers; look for patterns and changes in the above relationships; argue for the superiority of this methodology: transparency, falsifiability, objectivity, and generalizability among such big data in a short time rise recently.

With the development of content mining techniques and the research documents digitalization nowadays, the demands of the document collections' accessing, organizing and analyzing rise recently. Therefore, several relevant questions are required for comprehensive and systematic understanding of this field. What researches on educational leadership have been studied? What are the important topics of existing researches? And what are further research hotspots? These questions require researchers to find out an effective method to analysis in both research and practice.

To visualize the structure and dynamics of the target research fields, one prominent bibliometric technique will be employed in this study, co-word analysis (Cho, 2014; Wang *et al.*, 2012). Co-word analysis is a well-developed method, which is widely applied in several theoretical and empirical studies, e.g. scientometrics (Courtial, 1994), biological safety (Cambrosio *et al*., 1993), autism (Courtial and Gourdon, 1997) information retrieval (Ding *et al*.,2001) and chemical engineering (Zhang *et al.*, 2012), etc. for exploring the research thematic evolution in years. It is shown the practical value and advantages in these studies; therefore, it is rarely using network analysis together to present the knowledge. In this study, we apply the content mining techniques and network analysis to detect and visualize the different research thematic tops.

This article is organized as follows: section 2 introduces the analysis methodology, including the data processing, co-word analysis and the network analysis, findings and discussion are presented in section3; finally, conclusions and contributions are drawn in section 4. There were several phases of increase in

Khulna City. Thanks to the extension of railway line from Jessore to Khulna the primary growth came during British period in 1885 (Naznin, 2012). Forest came during the partition of India in 1947, due to the influx of the refugees. Industrialization of Khulna within the 1960s caused third phase of increase during this region. Consistent with Mondal (2012), Khulna City experienced an incredible growth of population following liberation in 1971, which was mainly thanks to the agricultural urban migration and natural increase of population, thus contributed to extend about 4.13 percent per annum. This suggests that migration from other places to Khulna had been the dominant factor of the increase. His study revealed that, among the entire population, about 60 percent belong to the age bracket of 14-44 years, 22 percent belong to the age bracket 14 years and fewer and 5.9 percent have age 60 years and over. Among these three groups, the age bracket 14-44 is that the most economically active above the economy of Khulna City. Moreover, he showed, monthly income of about 66 percent employed people of Khulna City is within Tk. 5,000 while 30 percent of them is within Tk. 2,500. Only 3.5 percent of the employed people belong to monthly income group of Tk. 15,000 and above (Mondal, 2012).

2. Methodology

In the -'Educational Leadership' filed, the 8 top journals: Educational Administration Quarterly; Educational Management Administration & Leadership; International Journal of Educational Management; International Journal of Leadership in Education; Journal of Educational Administration; Leadership and Policy in Schools; School Effectiveness and School Improvement; School Leadership & Management are selected in this study as the data source for the educational leadership research, as these journals has been consistently considered as the most prestigious journals. Our methodology is applied to all research articles from these Journals (1996~ 2015). The system follow of our research design is illustrated in Fig. 1.

1) Data Collection

To collect the text data in an efficient manner, we used the plugin-DownThemAll of Firefox to download all pdf files from the journal's websites. Secondly, only original research articles are included in our study. Thirdly, we used the software-Nitro Pro 10.2 to convert the text content from PDF format to text format. Last, all the txt files are bulky imported into SQL Server database for the co-word analysis. In the meanwhile, each article is separated into 5 fields: title, journal name, abstract, body, and publication years, etc. Following items, with abbreviations, meanings and examples, are used in our analysis.

PT: publication title, e.g. exploring the web visibility of world-class universities.

JT: Journal title. E.g. Educational Management Administration & Leadership

AB: Article's abstract,

MB: Main body of the article.
PY: publication year. E.g. 2015

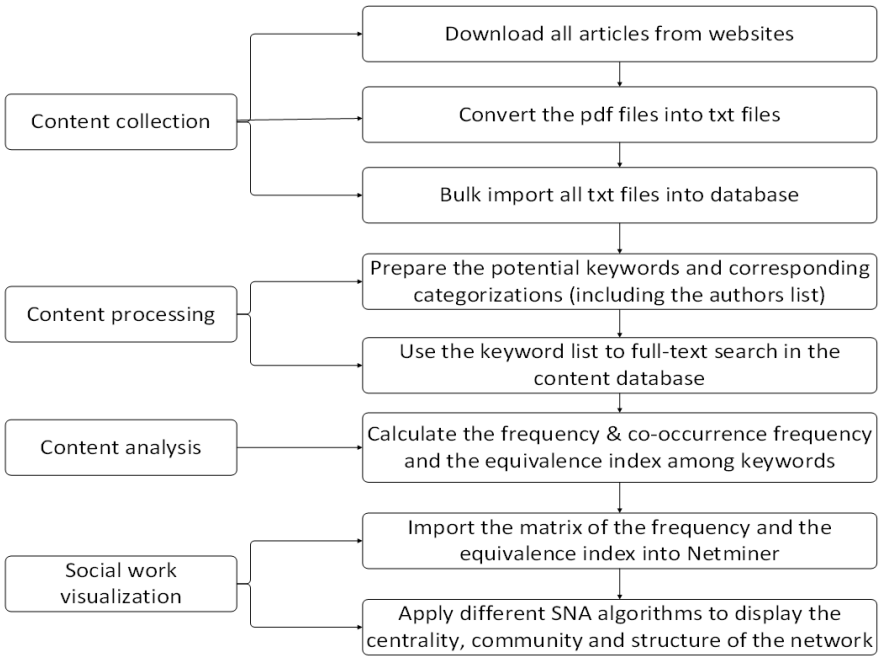


Fig. 1: Flow chart of the research design.

2) Data Processing

Keywords and their frequency are extracted from the papers of target journals. Traditional methods either use all keywords or use the top x (e.g. 20) frequency keywords for the words used to identify research themes, while this may introduce too many keyword data or ignore some important keywords. In this article, 120 keywords are extracted by the content experts in Leadership research area (Appendix 1). The Keywords covers most interested topics & research methodology are categorized into different groups to avoid this omission and this also helps research know more clearly about the relations between the keywords and groups. (8 categories: Leadership, Methods, Personal Variables, Actions, Dependent variables, Contextual Variables, Results and Researchers.)

To calculate keywords' frequency, and co-occurrence frequency, we write a program called Content-search for the full-text search in the content database. The keyword frequency is counted by the number its occurrences in the included articles, and the frequency of a pair of keywords occurring simultaneously in the same articles is calculated as the co-occurrence frequency.

The keywords occurrence frequency in articles reflects the important topics, and co-occurrence of multiple keywords means the themes' relevance to each other's.

That means the more frequent (stronger correlation) the co-occurrence of keywords, the more correlative the topics they indicate. The equivalence index is calculated to analysis the relevance to keywords.

The structure data and keywords list from the steps above are imported into SQL Server for calculating the word frequency, and the keyword frequency and co-occurrence frequency results are shown in the frequency matrix and co-occurrence matrix. The equivalence index among keywords (He, 2001) is used to calculate the keywords relationship, which is defined below:

$$e_{ij} = \frac{c_{ij}^2}{c_i \cdot c_j} \tag{1}$$

Where c_{ij} is the number of articles, in which the two keywords i and j co-occur, and c_i c_i is the number of articles in which each one occurs. The more the keywords occur together, the equivalence index e_{ij} is more closed to 1 and when the keywords are never associated, the equivalence index equals 0.

3) Network Analysis

Network Analysis is the process of investigating social structures through the use of networks and graph theory. A network is composed by a set of nodes and links. In this study, the keywords structure is constructed into a network, in which the nodes represent the keywords and the links are the equivalence index of these keywords. The node size presents the frequency of the keyword, and the edge thickness reflects the equivalence index of these keywords. We focus on the following 2 aspects of the network: closeness centrality and modularity.

a) Closeness centrality

Closeness centrality measures the centrality nodes in the network, which is defined as follows:

$$C_i = 1 \Big/ \sum_{j=1}^{n} d(i,j) \tag{2}$$

Where, d(i,j) is the distance from node i to node j.

It reflects the efficiency about how a node spreads information to other nodes in the network.

b) Modularity

Modularity is a commonly used method to measure the network structures. Nodes are divided into different modules (also called communities). The algorithm in (Aaron, 2004) is applied in our study to analyze the keywords network modularity structure.

3. Result and Discussion

1) Frequency analysis

Base on the keyworks frequency, following findings are achieved: The topics on 'Instructional Leadership', 'Teacher Leadership', 'Shared Leadership', 'Curriculum Leadership' and 'Servant Leadership' become more popular since 1996. The topics on 'Transformational Leadership', 'Transactional Leadership', 'Collaborative Leadership', and 'Strategic Leadership' increase sharply in last 5 years. However, the topics on 'Moral leadership' is decreasing (Fig.2).

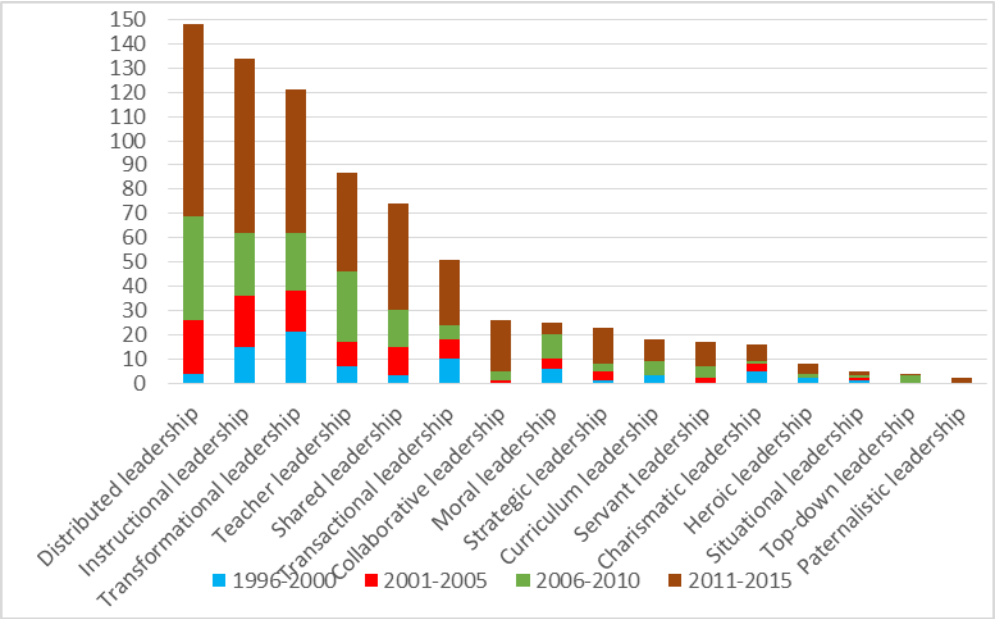


Fig.2: The keywords frequency of research topics on 'Leadership'.

Furthermore, from the citation view. The citation numbers have increased in the last 5 years, due to the increase in total volume of publication; Leithwood and Hallinger have become leading researchers since 2011; before that Harris, Spillane, Fullan and Gronn were comparatively famous in the field (Fig.3).

For 'Personal variables': 'role', 'knowledge', 'experience', 'decision', 'value', 'vision' and 'commitment' are the top 7 factors in the field (Fig.4).

For 'Research method', 'Interview', 'Case study, 'Correlation', and 'Regression' are the most used research method in the Educational Leadership (Fig.5).

'Culture', 'Community', 'Policy', 'Structure', 'Environment', 'Team', 'Reform', 'Challenge' and 'Climate' contextual variables are most used int the studies (Fig.6).

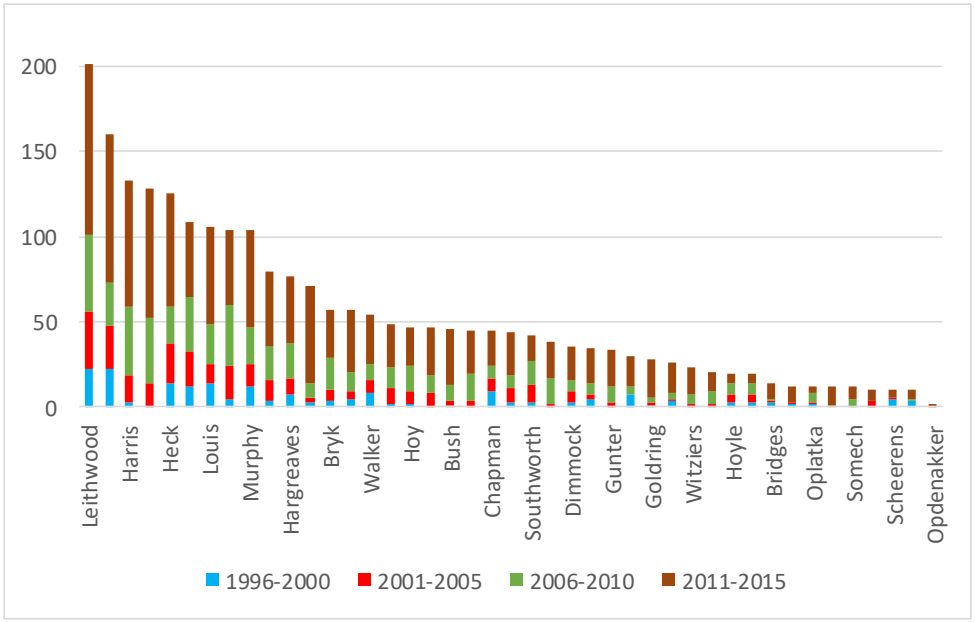


Fig.3: The keywords frequency of researcher.

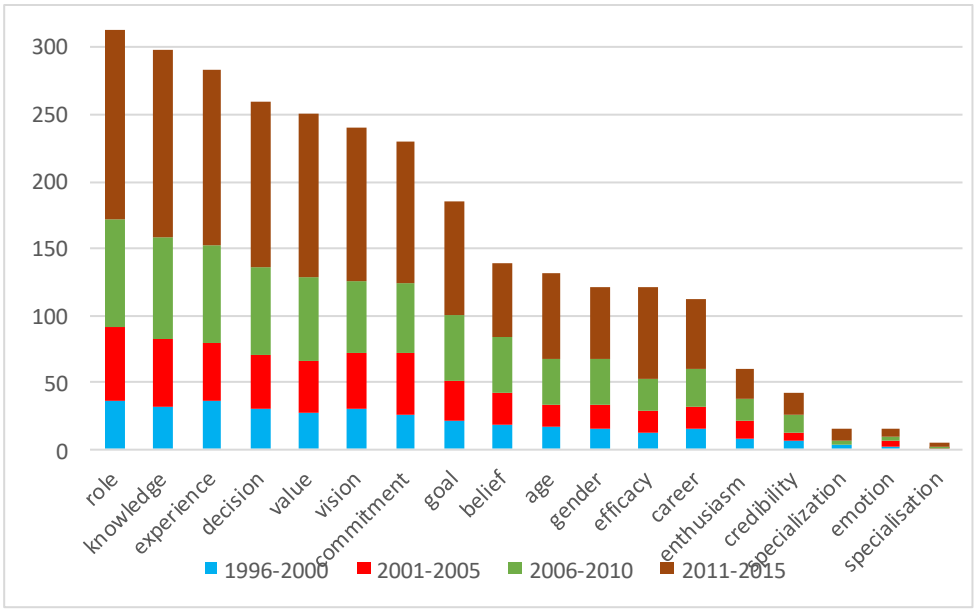


Fig.4: The keywords frequency of Personal variables.

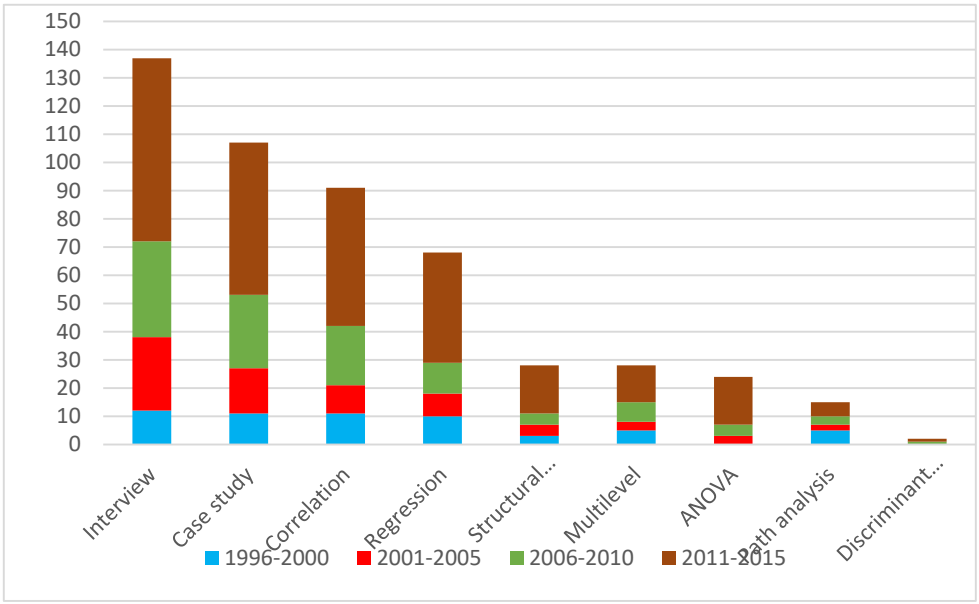


Fig.5 The keywords frequency of Research method

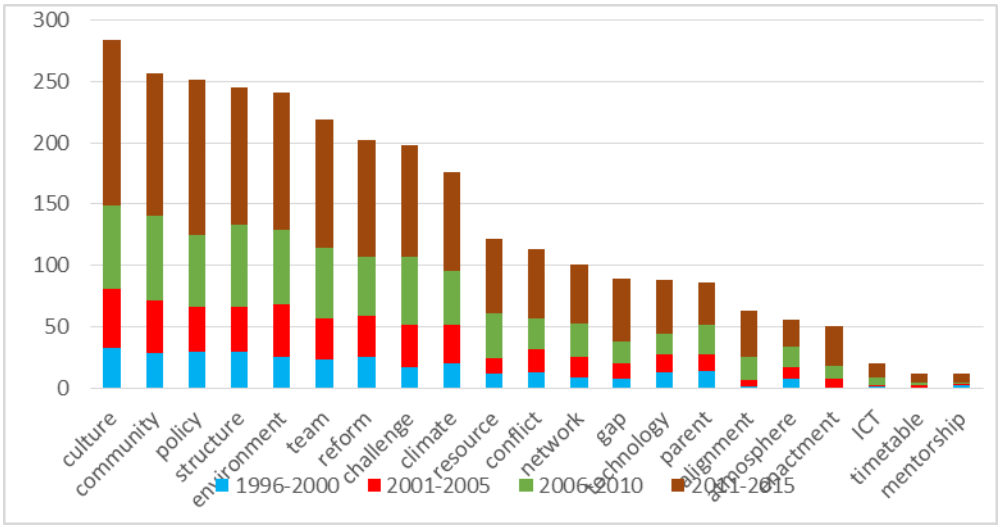


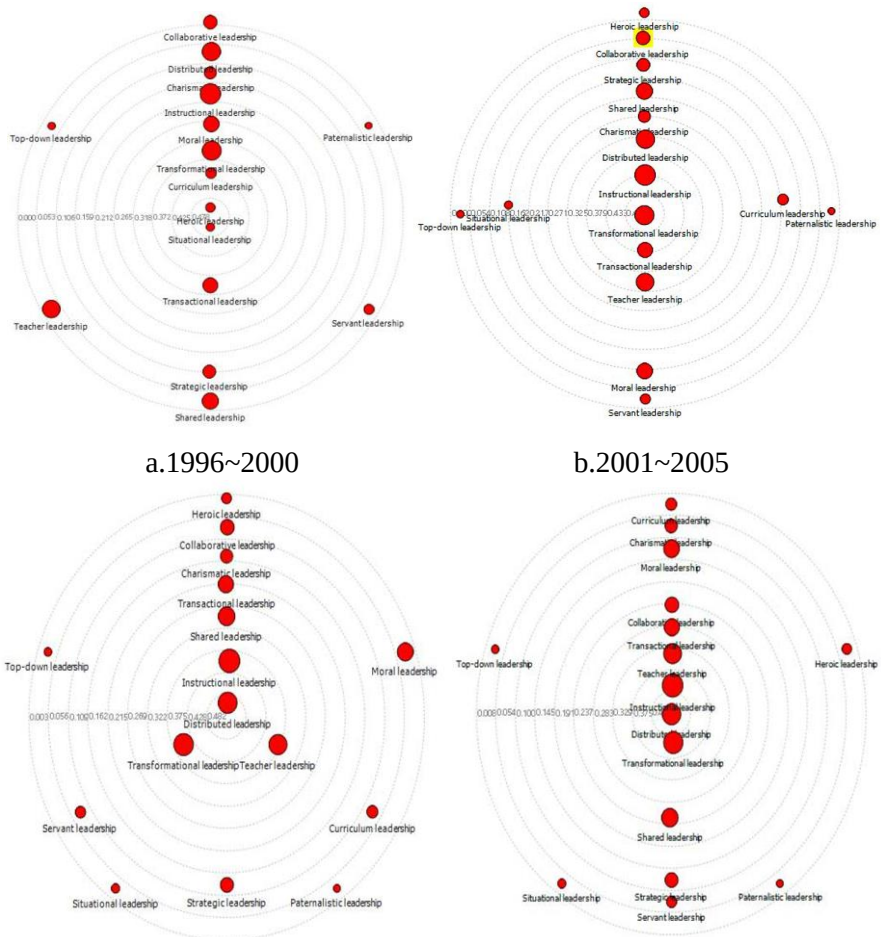
Fig.6: The keywords frequency of Contextual variables.

2) Network Analysis

In this study, the centrality result of the SNA is in Fig. 7, and modularity result is presented in Fig.7, in which the nodes represent and the links are the co-occurrence of these keywords. The size of nodes indicates the keywords frequency, and the thickness of the edges indicates the co-occurrence frequency of keywords pairs.

The normalized scores with nodes are shown in Fig. 7, and the scores are bound between 0 and 1. It is 0 if a node is an isolate, and 1 if a node is directly connected

all others. Fig. 6 shows that the keyword -'Heroic Leadership' in the period 1996~2000, the keyword-'Transformational Leadership' in the period 2001~2005, the keyword-'Distributed Leadership' in the period 2006~2011, as well 'Instructional Leadership' in the period 2012~2015, are common in topics and close to others else in the different periods. It is in an excellent/key position to monitor the information flow in the network -- it has the best visibility in the sub network.



a.1996~2000 b.2001~2005

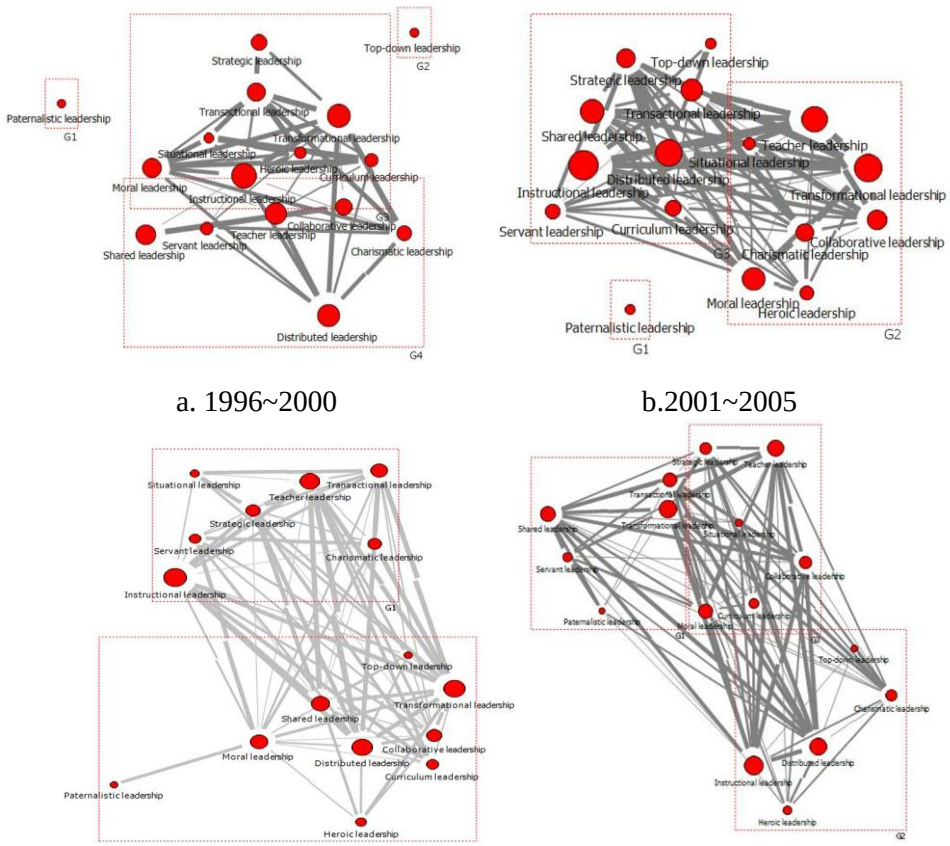
c. 2006~2011 d. 2012~2015

Fig. 7: The Closeness centrality of the keywords network in the 4 periods.

The modularity structure in Fig. 8 shows that there are 4 groups of modularity among the keywords. The size of node stands for its frequency and the link width is the equivalence index among different keywords.

By viewing the largest node and its connections with other nodes in Fig. 8, we know that 'Distributed Leadership', 'Instructional Leadership', and 'Transformational Leadership' are the most popular topics, and they are in 2

'communities' in the period 1996~2000. The topic-'Paternalistic Leadership', and 'Top-down Leadership' attract little focus, isolated from the main research area; In the next period- 2001~2005, the topics-about 'Distributed Leadership', 'Instructional Leadership', and 'Transformational Leadership' still act the core in the 2 'communities', and more studies focus on the 'Top-down Leadership'; 'Transitional Leadership' join the same 'community' with 'Instructional Leadership', means these topics are closer than before in the period 2005~2010; 'Strategic Leadership', 'Teacher Leadership', 'Situational Leadership', 'Moral Leadership', 'Curriculum Leadership', and 'Collaborative Leadership' form into a new 'community' in the period 2011~2015. The changes show the hot research interests have changed across different time period, which can be regarded as a kind of trend of research topics.



a. 1996~2000 b.2001~2005

c. 2006~2010 d. 2011~2015

Fig. 8: The Modularity of the keywords network in the 4 periods.

4. Conclusion

In this paper, we have conducted content mining and network analysis in the field of Educational Leadership research to clearly understand the development trend of research topics from 2000 to 2015. By providing some reasonable and clear results, we found the main research topics and the relationships between these research topics.

From the result, following keywords are extracted - Moral leadership, Curriculum leadership, Shared leadership, Transactional leadership, Transformational leadership, Distributed leadership, Teacher leadership Heroic leadership, Charismatic leadership, Instructional leadership, Servant leadership, Collaborative leadership, Situational leadership, collaboration, Curriculum, Instruction, Challenge, culture, policy, change and improvement.

Our method provides direct visual effects, which makes the structure expandable and easy for users to understand. Our method has hardly changed and can be widely used in other research fields.

References

Cambrosio A., Limoges C., Courtial J. P., and Laville F. (1993). Historical scientometrics? Mapping over 70 years of biological safety research with coword analysis. Scientometrics, journal article, 27 (2), 119-143, doi: 10.1007/bf02016546.

Clauset A., Newman M. E. J., and Moore C. (2004). Finding community structure in very large networks. *Physical Review E*, 70 (066111).

Briggs A., Coleman M., and Morrison M. (2012). Research methods in educational leadership & management: 3. ed. *Los Angeles.*

Evers C. and Lakomski G. (2012). Science, systems, and theoretical alternatives in educational administration: The road less travelled. *Journal of Educational Administration,* 50, 57-75, 01/27, doi: 10.1108/09578231211196069.

Labaree D. F. (2011). CONSUMING THE PUBLIC SCHOOL. *Educational Theory*, 61, (4), 381-394, doi: 10.1111/j.1741-5446.2011.00410.x.

Oplatka I. (2009). The Field of Educational Administration: A Historical Overview of Scholarly Attempts to Recognize Epistemological Identities, Meanings and Boundaries from the 1960s Onwards. *Journal of educational administration,* 47(1), 8-35, doi: 10.1108/09578230910928061.

Oplatka I. (2012). Fifty Years of Publication: Pondering the Legacies of the "Journal of Educational Administration. *Journal of educational administration,* 50(1), 34-56, doi: 10.1108/09578231211196050.

Courtial J. P. (1994). A coword analysis of scientometrics. *Scientometrics, journal article* .31(3), 251-260, doi: 10.1007/bf02016875.

Courtial J. P., and Gourdon L. (1997). A scientometric approach to autism based on translation sociology. *Scientometrics, journal article,* 40(2), 333-355, doi:10.1007/bf02457442.

Zhang J. et al. (2012). Mapping the Knowledge Structure of Research on Patient Adherence: Knowledge Domain Visualization Based Co-Word Analysis and Social Network Analysiss. *PLOS ONE,* 7(4), e34497, doi: 10.1371/journal.pone.0034497.

Cho J. (2014). Intellectual structure of the institutional repository field: A co-word analysis. *Journal of Information Science,* 40(3), 386 – 397.

Hallinger P. (2013). A conceptual framework for systematic reviews of research in educational leadership and management. *Journal of educational administration,* 51(2), 126-149, doi: 10.1108/09578231311304670.

Hallinger P. (2014). Reviewing Reviews of Research in Educational Leadership: An Empirical Assessment. *Educational Administration Quarterly,* 50(4),539-576, doi: 10.1177/0013161X13506594.

He Q. (2001). Component study of co-word analysis.

Bates R. J. (1980). Educational Administration, the Sociology of Science, and the Management of Knowledge. *Educational Administration Quarterly,* 16(2), 1-20, doi: 10.1177/0013161X8001600204.

Donmoyer R. (2001). Evers and Lakomski's search for leadership's holy grail(and the intriguing ideas they encountered along the way. *Journal of Educational Administration,* 39(6), 554-572, doi: 10.1108/EUM0000000006053.

Heck R. H., and Hallinger P. (2005). The Study of Educational Leadership and Management: Where Does the Field Stand Today? *Educational Management Administration & Leadership,* 33(2), 229-244, doi: 10.1177/1741143205051055. Wikipedia. Educational leadership."https://en.wikipedia.org/wiki/Educational_leadership#:~:text=Educational%20leadership%20is%20the%20process,toward%20achieving%20common%20e

ducational%20aims.&text=Several%20universities%20in%20the%20United,educational%20leadership%20can%20be%20overcome. (accessed.Wikipedia. Social network analysis."https://en.wikipedia.org/wiki/Social_network_analysis (accessed.

Ding Y., Chowdhury, G.G., Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, 37,817-842.

Wang Y., and Bowers A. J. (2016). Mapping the field of educational administration research: a journal citation network analysis. *Journal of educational administration*, 54(3), 242-269, doi: 10.1108/jea-02-2015-0013.

Wang Z.-Y., Li G., Li C.-Y., and Li A. (2012). Research on the semantic-based co-word analysis. *Scientometrics, journal article*, 90(3), 855-875, doi: 10.1007/s11192-011-0563-y.

Appendix. The keyword list used in the study

Leadership	Methods	Personal Variables	Actions	Dependent variables	Contextual Variables
Moral leadership	ANOVA	age	affect	access	alignment
Curriculum leadership	Correlation	career	allocate	appraisal	alignment
Shared leadership	Discriminant	career	assign	assessment	atmosphere
Transactional leadership	Multilevel	commitment	cause	belonging	challenge
Transformational leadership	Path	credibility	coach	citizenship	climate
Distributed leadership	Regression	decision	control	collaboration	community
Teacher leadership	Structural equation	efficacy	dampen	communication	conflict
Heroic leadership		emotion	enhance	confusion	culture
Paternalistic leadership		enthusiasm	hinder	cooperation	enactment
Charismatic leadership		experience	improve	creativity	enactment
Top down leadership		experience	increase	curriculum	environment
Instructional leadership		gender	influence	emphasis	gap
Servant leadership		goal	manage	expectation	ICT
		quality	mentor	grievance	mentorship
Collaborative leadership		role	nurture	instruction	network
Situational leadership		specialization	promote	involvement	parent
		values	result in	job satisfaction	policy
		vision	review	knowledge	reform
			train	management	reform
				motivation	resources
				professional development	structures
				professional learning	team
				reflection	technology
				relationship	timetable
				strategy	workload
				time	
				trust	