Dynamic Evolution Analysis of Domain Knowledge Clusterability (Postprint)

Authors: tranquility, Teng Guangqing, Bai Shuchun, Bi Qiang, Han Shangxuan

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] Investigating the clustering evolution in the development process of domain knowledge contributes to revealing the characteristics and patterns of knowledge clustering, and holds significant importance for comprehending the aggregation of related knowledge during knowledge growth and evolution. [Method/Process] Grounded in complex network theory, we construct a time-series domain knowledge network based on occurrence values of label adjacency relationships. Specifically, employing network motif theory and network clustering coefficient analysis methods, we conduct dynamic tracking and analysis of the domain knowledge network; furthermore, integrating indicators such as network density, characteristic path length, node degree values, and closed triads, we analyze the clustering evolution phenomenon in domain knowledge development from three dimensions: random factors, degree correlation, and proximity association. [Results/Conclusion] The research findings demonstrate: Domain knowledge consistently maintains high clustering throughout its development process; The clustering of domain knowledge simultaneously encompasses both random and structural (non-random) factors; The dynamic state of domain knowledge clustering oscillates between small-world and scale-free networks; The clustering state of domain knowledge exhibits certain disparities between the global network and local nodes.

Full Text

Preamble

Dynamic Evolution Analysis of Domain Knowledge Clustering

An Ning[1], Teng Guangqing[1], Bai Shuchun[2], Bi Qiang[3], Han Shangxuan[1]
[1]School of Information Science and Technology, Northeast Normal University, Changchun 130117

[2]Library of Jilin University, Changchun 130012
[3]School of Management, Jilin University, Changchun 130022

Abstract

[**Purpose/Significance**] Exploring the clustering evolution in the process of domain knowledge development helps reveal the characteristics and patterns of knowledge clustering, which is significant for understanding the aggregation of correlated knowledge during knowledge growth and evolution. [**Method/Process**] Based on complex network theory, this study constructs time-series domain knowledge networks using the occurrence values of tag adjacency relationships. Specifically, employing network motif theory and network clustering coefficient analysis methods, we dynamically track and analyze domain knowledge networks. Combined with network density, characteristic path length, node degree values, triadic closure, and other indicators, we examine the clustering evolution phenomenon in domain knowledge development from three perspectives: random factors, degree correlation, and adjacent correlation. [**Result/Conclusion**] The findings indicate: (1) Domain knowledge maintains high clustering throughout its development process; (2) The clustering of domain knowledge incorporates both randomness and structuration (non-randomness); (3) The dynamic state of domain knowledge clustering evolves with oscillation between small-world and scale-free networks; (4) The clustering status of domain knowledge exhibits certain differences between global network and local node levels.

Keywords: domain knowledge; knowledge network; knowledge clustering; clustering coefficient

2. Literature Review

The clustering phenomenon of domain knowledge has long been a focal issue explored by the library and information science community. Numerous studies demonstrate that knowledge units within any discipline are not completely isolated or free-floating, but rather exhibit certain clustering and grouping characteristics based on underlying associative relationships. These relationships continuously change with domain knowledge development, causing knowledge clusters to evolve and transform. On one hand, hot topics and core knowledge within a domain attract correlated knowledge to aggregate; on the other hand, the emergence of new knowledge continuously alleviates this aggregation state. Therefore, dynamic analysis of domain knowledge clustering from a time-series perspective to grasp and reveal the evolutionary characteristics and patterns of knowledge clustering during domain knowledge development has become an urgent issue in knowledge management research.

The earliest scholars to introduce network thinking into library and information science were E. Garfield[1] and D. J. S. Price[2], who published papers in *Science* during the 1950s-60s, constructing citation networks based on scientific paper citation relationships to study the inheritance and development of scientific knowledge from a network perspective. With the revival of network science in the late 20th century[3], the clustering and grouping issues of knowledge were reinterpreted through the lens of network science. The aggregation degree among knowledge nodes creates complex topological structures in knowledge networks, attracting scholarly attention to domain knowledge clustering based on network analysis.

N. Shibata et al.[4] conducted comparative studies on different types of citation networks using SCI and SSCI data, finding that direct citation networks have the highest clustering coefficient, indicating that papers connected through direct citations share the greatest content similarity, and core literature included in the largest network component faces the lowest risk of omission. Li Yating and Ma Feicheng[5] constructed a social tag co-occurrence network based on the Folksonomy knowledge organization mode, discovering through computational analysis that the tag knowledge network exhibits high clustering ($C = 0.816$). Hu Changping and Chen Guo[6] applied triadic closure-based clustering structures to analyze the hierarchical structure of keyword knowledge networks, revealing that such structures effectively uncover the diversity of micro-level associations in knowledge networks.

As research has deepened, knowledge network studies have gradually entered the most challenging domain of dynamic analysis. M. E. J. Newman[7] empirically studied the temporal evolution of collaboration networks in physics and biology using bibliographic data, revealing patterns of clustering and preferential attachment in growing networks through comparative analysis of successive time windows. J. Makani and L. Spiteri[8] measured three indicators in tag knowledge networks—tag growth, tag reuse, and tag discrimination—finding that the number of unique tags steadily decreased over time, reflecting enhanced domain stability of tag vocabulary representing community knowledge. W. Liu et al.[9] used publication datasets from the American Physical Society to construct annual bibliographic coupling networks (BCN), identifying clusters representing different research fields and visualizing long-term knowledge evolution in physics research as alluvial diagrams, exploring how new knowledge builds upon old knowledge. Liu Xiang et al.[10] introduced degree preferential attachment and time priority connection to detect the inheritance and renewal process of scientific knowledge, where degree preferential attachment ensures connection to important knowledge while time priority facilitates acceptance of the latest knowledge and updates.

In summary, as network analysis theories and methods mature, studying various knowledge networks through network thinking has gained widespread academic recognition. Research on knowledge network clustering has yielded rich results, with recent work evolving from static to dynamic analysis. However, domain

knowledge development is always accompanied by knowledge growth, decline, derivation, and fusion. Dynamic analysis based on cumulative data (references [13], [10-11]) focuses on the growth characteristics inherited from previous states, while analysis based on occurrence values (references [8-9], [12]) concentrates on knowledge aging and innovation during evolutionary changes. Considering the stronger timeliness of social tagging systems and current academic recognition of knowledge network construction based on such systems (references [8], [11], [13]), this study specifically examines knowledge clustering using occurrence values to better capture the impact of knowledge aging and innovation on clustering.

---

3. Knowledge Clustering Theory

Knowledge units within any discipline possess certain correlations, either direct or indirect, preventing them from being discrete and disorderly but instead forming knowledge clusters to some degree. In the dynamic process of disciplinary development, new knowledge and its associations always emerge based on existing knowledge and relationships, more prominently reflecting changes such as knowledge growth, decline, derivation, and fusion within specific periods. Therefore, this study employs occurrence values of domain knowledge as foundational data, using network analysis concepts and methods to investigate knowledge clustering during dynamic evolution.

Clustering is essentially a process of reclassifying objects within a set, which in knowledge networks manifests as the formation of knowledge cohesive subgroups. During domain knowledge evolution, knowledge nodes and their relationships continuously change over time, causing continuous differentiation and aggregation phenomena in the time series. In occurrence value-based knowledge evolution analysis, while some knowledge nodes and relationships from the previous period are implicitly inherited in the next period, some fade away, and new nodes and relationships emerge. This represents a micro-rule cyclic process and an iterative development of domain knowledge. Network analysis can quantify these clustering phenomena, enabling computation and measurement of real knowledge networks.

This study primarily investigates knowledge network clustering based on network motif theory. Network motifs, first proposed by R. Milo et al.[14] in their *Science* paper on building blocks of complex networks, are fundamental construction units of networks. Among various network motifs, closed triads are the most suitable basic building blocks for revealing clustering relationships. Based on this, we adopt M. E. J. Newman' s[15] definition of clustering coefficient: the proportion of closed paths among all length-2 paths in a network. Since we construct undirected knowledge networks, the clustering coefficient formula is:

Formula (1) calculates the clustering coefficient where TC represents the number of closed triads (clustering motifs), i.e., the number of closed paths among length-

2 paths; TP represents the number of connected triads, i.e., all length-2 paths including both closed and non-closed paths. Since each closed triad contains three connected triads, the coefficient is 3.

Figure 1 [Figure 1: see original paper] illustrates this formula. In Figure 1, (A) shows a length-2 path a→b→c, where nodes a and c share a common neighbor b. In knowledge networks, this is understood as nodes a and c having a common correlated knowledge node b. If the triad {a, b, c} is closed (with a solid connection between a and c), then a and c are also directly correlated. Thus, Newman' s clustering coefficient can be understood as the probability that two knowledge nodes directly correlated with a common knowledge node are also directly correlated with each other—the probability of three points forming a clustering motif. Equivalently, it represents the average probability that two knowledge nodes sharing a common neighbor are directly connected.

Figures 1(B), (C), and (D) show networks with equal size (N=10 nodes, L=20 edges). (C) is a regular network with no closed triads (clustering motifs), giving it a clustering coefficient C=0. (B) and (D) are regular and irregular networks respectively of the same scale and density. Both contain closed triads and thus have higher clustering coefficients, with (B)' s coefficient being constant regardless of network size. This reveals that clustering coefficient is independent of whether the knowledge network is regular; the critical factor is the presence of clustering motifs—closed triads where knowledge nodes have direct mutual relationships.

---

4. Research Method

4.1 Research Data

This study uses the social bookmarking and publication sharing website BibSonomy as its data source. Domain knowledge selection can range from broad disciplines to specific topics or even finer-grained problem domains. Given increasingly prominent interdisciplinary integration and the openness of social bookmarking, seemingly isolated knowledge becomes incorporated into relevant domains through various associations, providing broader research perspectives. Using "folksonomy"as the target tag, we developed a web crawler to collect 5,470 relevant documents spanning 2006-2015. Using natural years as time units, we divided the 2006-2015 period into 10 time windows (t_0, t_1, ⋯, t_9). Statistics on documents and corresponding tags for each window yielded the foundational data shown in Table 1 .

---

5. Analysis and Findings

The analysis reveals that during domain knowledge development, the number of closed triads and 2-hop paths around knowledge nodes changes with evolution, with both trends generally aligning. This reflects that as domain knowledge develops, new 2-hop paths (indirect associations) and numerous closed triad relationships (direct associations) continuously emerge around knowledge nodes. Overall, results from Table 5 and Table 6 show that higher ratios of closed triads (clustering motifs) to 2-hop paths (associations through length-2 paths) typically correspond to higher clustering coefficients. In time windows t_7 and t_8, the 2-hop path count is 0, indicating no non-directly correlated neighbor nodes for "ontology" at this stage, demonstrating phase-specific stability and maturity in the ontology domain' s evolution.

However, knowledge network development as a complex system has unique characteristics. While numerous connected triads (including closed triads and 2-hop paths) facilitate high hub status for nodes, a high proportion of closed triads to 2-hop paths does not always guarantee high clustering coefficients. In Table 6 , time window t_1 shows the "ontology" node with far more closed triads than 2-hop paths and a high ratio (8.914), yet its numerous 2-hop paths constrain the clustering coefficient from reaching extremely high values. Strictly speaking, relative to individual nodes' 2-hop paths (non-direct associations), closed triads (clustering motifs) are necessary but not sufficient conditions for high clustering coefficients.

---

6. Conclusions and Discussion

This study constructs time-series domain knowledge networks based on tag co-occurrence relationships in social tagging systems to explore knowledge clustering states and influencing factors during domain knowledge evolution. Through temporal tracking and analysis of random factors, degree correlation, and adjacent correlation, we reveal evolutionary patterns and underlying influences of knowledge clustering.

Based on comprehensive dynamic tracking and analysis, we draw the following conclusions:

(1) Domain knowledge maintains consistently high clustering throughout its growth and development. Table 3 shows that occurrence value-based domain knowledge networks are sparse networks. However, compared with random networks of equal density and scale, domain knowledge networks exhibit significantly higher clustering coefficients that persist throughout the evolution cycle. Although using real data differs from the simulation data used by D. J. Watts et al.[20] under laboratory conditions, our findings similarly validate differences between knowledge networks and both completely regular and random networks from a clustering perspective.

(2) Domain knowledge clustering incorporates both randomness and structuration (non-randomness). Real scientific research always builds upon previous work, with knowledge associations in subsequent time windows undergoing subtle updates and iterations equivalent to network rewiring. Combined with high clustering (Table 3) and the decline in Pearson correlation coefficient from extremely strong (0.8-1.0) to moderately strong (0.6-0.8) (Rce = 0.9868, Rck = 0.7323), structuration (non-randomness) becomes prominent. While structuration dominates most periods, randomness never completely disappears, as shown by the discrete distribution scatter points in time window t_5 in Figure 2 [Figure 2: see original paper].

(3) The dynamic state of domain knowledge clustering evolves with oscillation between small-world and scale-free networks. High clustering coefficients (Table 3) and short characteristic path lengths (Table 4) indicate that occurrence value-based domain knowledge networks are small-world networks. The concentration of high clustering coefficients in medium-low degree regions (Areas A and B in Figure 2 [Figure 2: see original paper]) further confirms small-world characteristics. However, degree correlation analysis reveals varying tail distributions across time windows (e.g., Area C in the lower right of t_0 scatter plot), demonstrating scale-free network features. Different prominence levels across scatter regions during evolution reflect oscillation between small-world and scale-free network states.

(4) Clustering status shows spatial dimensional differences (global vs. local). Temporal analysis reveals differences across time dimensions. Globally, network clustering coefficient positively correlates with the ratio of closed triads to 2-hop paths (Rcr = 0.9739). However, local node clustering coefficients, while maintaining this overall positive correlation, show individual temporal differences. In Table 6 , the high ratio in time window t_1 does not yield extremely high clustering, reflecting differences between global and local knowledge clustering—statistically, the distinction between sample means and individual variations.

This study specifically examines occurrence values to capture knowledge aging and innovation' s impact on clustering. Compared with cumulative values, occurrence values better reflect these dynamics. The discovered clustering characteristics also apply to social networks, information dissemination networks, and other real networks facing relationship termination/formation or channel obstruction/creation. Future research should combine cumulative and occurrence values for smoother yet dynamic perspectives to more comprehensively explore domain knowledge evolution patterns.

References

[1] GARFIELD E. Citation indexes for science: a new dimension in documen-

tation through association of ideas[J]. Science, 1955, 122(3159): 108-111.
[2] PRICE D J de S. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.
[3] BARABÁSI A-L. Network science[M]. Cambridge: Cambridge University Press, 2016: 20-41.
[4] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Comparative study on methods of detecting research fronts using different types of citation[J]. Journal of the association for information science and technology, 2009, 60(3): 571-580.
[5] LI Yating, MA Feicheng. Research on social network analysis based on tag co-occurrence[J]. Journal of intelligence, 2012, 31(7): 103-109.
[6] HU Changping, CHEN Guo. Research on triadic relationship patterns in conceptual knowledge networks from a hierarchical perspective[J]. Library and information service, 2014, 58(4): 11-16.
[7] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Physical review E, 2001, 64(2): 025102.
[8] MAKANI J, SPITERI L. The dynamics of collaborative tagging: an analysis of tag vocabulary application in knowledge representation, discovery and retrieval[J]. Journal of information & knowledge management, 2010, 9(2): 93-103.
[9] LIU W, NANETTI A, CHEONG S A. Knowledge evolution in physics research: an analysis of bibliographic coupling networks[J]. PLoS ONE, 2017, 12(9): e0184821.
[10] LIU Xiang, MA Feicheng, WANG Xiaoguang. Structure and process models of knowledge networks[J]. Systems engineering-theory & practice, 2013, 33(7): 1836-1844.
[11] TENG Guangqing. Dynamic evolution of tight domain knowledge communities in Folksonomy mode[J]. Journal of library science in China, 2016, 42(4): 51-63.
[12] ZHU Na, WANG Fang. Research on knowledge evolution path identification based on topic correlation: a case study of 3D printing[J]. Library and information service, 2016, 60(5): 101-109, 82.
[13] MA J. The sustainability and stabilization of tag vocabulary in CiteULike: an empirical study of collaborative tagging[J]. Online information review, 2012, 36(5): 655-674.
[14] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.
[15] NEWMAN M E J. Networks: an introduction[M]. GUO Shize, CHEN Zhe, trans. Beijing: Publishing House of Electronics Industry, 2014: 126-130, 152-173.
[16] BARABÁSI A-L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.
[17] TENG Guangqing, HE Defang, PENG Jie, et al. Research on dynamic evolution of domain knowledge based on network centrality[J]. Library and information service, 2016, 60(14): 128-134, 141.
[18] MCGLOHON M, AKOGLU L, FALOUTSOS C. Statistical properties of social networks[C]//AGGARWAL C C. Social network data analytics. New York:

Springer, 2011: 17-39.
[19] TENG Guangqing, CHANG Zhiyuan, LIU Yashu, et al. Research on dynamic evolution laws of domain knowledge in Folksonomy knowledge organization mode[J]. Library and information, 2016(4): 96-101.
[20] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440-442.
[21] LEWIS T G. Network science: theory and applications[M]. CHEN Xiangyang, JU Xiulian, et al., trans. Beijing: China Machine Press, 2011: 138-140.
[22] FRONCZAK A, HOLYST J A, JEDYNAK M, et al. Higher order clustering coefficients in Barabási-Albert networks[J]. Physica A: statistical mechanics and its applications, 2002, 316(1): 688-694.

Author Contributions:
An Ning: data analysis and paper writing;
Teng Guangqing: research design, data analysis, paper writing and revision;
Bai Shuchun: data collection and analysis;
Bi Qiang: research conceptualization and design;
Han Shangxuan: data analysis and paper revision.

Note: Figure translations are in progress. See original paper for figures.