Evolutionary Characteristics of Specialized Domains Based on Knowledge Element Citation Networks

Authors: Mao Jin, Hou Bowen, Wang Yimeng, Mao Jin

Date: 2023-04-14T14:12:47+00:00

Abstract

Purpose/Significance Understanding the developmental and evolutionary processes of scientific knowledge facilitates the advancement of scientific research; tracking the structural and evolutionary characteristics of knowledge in specialized sub-fields from a micro perspective holds significant importance for knowledge evaluation and knowledge services. Methods/Process Taking knowledge elements in medical informatics as an example, this study employs semantic types to delineate treatment-related sub-domains for each disease, constructs knowledge element citation networks for 125 diseases at various time points, utilizes the Leiden algorithm to identify knowledge communities, reveals the evolutionary characteristics of individual diseases from dimensions such as community knowledge evolution and community knowledge competition status, and proposes three measurement indicators—richness, balance, and difference degree—to uncover diversity characteristics at both individual disease and holistic levels. Results/Conclusion The research demonstrates that communities within knowledge element citation networks can effectively reflect the structure and evolutionary state of disease knowledge. The diversity characteristics of disease knowledge at the overall level encompass: a continuous increase in the number of disease knowledge communities and expanding differences in scale and composition among communities; different diseases exhibit conventional, early controversial, and general evolution patterns; and diseases studied earlier generally demonstrate lower balance and higher difference degree.

Full Text

Preamble

Research on the Evolution Characteristics of Subdivision Fields Based on Knowledge Unit Citation Networks

Mao Jin[1],[2], Hou Bowen[1],[2], Wang Yimeng[2]

Abstract

[**Purpose/Significance**] Understanding the developmental evolution of scientific knowledge facilitates scientific research. Tracing the structural and evolutionary characteristics of knowledge within subdivision fields from a micro perspective holds significant importance for knowledge evaluation and knowledge services. [**Method/Process**] This study takes knowledge units in medical informatics as examples, utilizes semantic types to define treatment-related subdivision fields for each disease, constructs knowledge unit citation networks for 125 diseases at different time points, employs the Leiden algorithm to identify knowledge communities, and reveals the evolutionary characteristics of individual diseases from dimensions such as community knowledge evolution and community knowledge competition states. Three measurement indicators—richness, balance, and disparity—are proposed to reveal the diversity characteristics of both individual diseases and the overall population. [**Result/Conclusion**] The research demonstrates that communities in knowledge unit citation networks can reflect the knowledge structure and evolutionary state of diseases. The overall-level diversity characteristics of disease knowledge include: the number of disease knowledge communities continues to increase, while differences in scale and composition between communities continue to expand; different diseases exhibit conventional, early-controversial, and generalized evolutionary patterns; and earlier-studied diseases generally show lower balance and higher disparity.

Keywords: SPO triples, knowledge unit, knowledge community, evolution characteristics, knowledge diversity
Classification Number: G203

With the rapid growth of global scientific publications, mining and analyzing the evolutionary characteristics and patterns of scientific knowledge has become increasingly important, and revealing the structure and evolution of scientific knowledge has emerged as a significant research question in library and information science. Domains and subdivision fields represent aggregations of scientific knowledge at different levels. A domain often contains multiple distinct topics, while a subdivision field denotes the connotation of similar types of knowledge within a specific topic in a domain. For example, if all scientific knowledge related to treatment in the biomedical discipline is regarded as a research domain, then disease treatment knowledge concerning diabetes and asthma can each be considered a subdivision field. Current research on scientific knowledge primarily focuses on the discipline or research domain level, yet analyzing the knowledge structure of subdivision fields and revealing their characteristics will contribute to a deeper understanding of scientific knowledge development processes.

From the perspective of analytical objects, related studies predominantly em-

ploy keywords or subject terms for domain topic mining and analysis; however, keywords or subject terms merely reflect surface-level knowledge features of documents [1]. A knowledge unit is a knowledge element parsed from scientific literature content that can more granularly reflect the internal structure of scientific knowledge. An SPO triple is a knowledge unit triple consisting of Subject, Predication, and Object that represents certain semantic content and relationships. Related research has focused more on drug treatment [2], gene diagnosis [3], and other knowledge graph [4] studies based on triples, overlooking the important role of triples in knowledge evolution analysis. In fact, as an operational data model for representing knowledge units, triples can establish inheritance relationships during the literature citation process, providing a feasible approach for knowledge measurement [5]. Unlike the excessive abstraction and condensation of keywords or subject terms, the rich semantic types of triples can more precisely define the subdivision fields to which knowledge belongs. On the other hand, communities composed of triples in citation networks can more comprehensively reflect the knowledge composition of subdivision fields.

Therefore, this paper proposes an evolution analysis method for subdivision fields targeting SPO triples. Based on knowledge unit citation networks, we identify knowledge communities within subdivision fields, analyze the diversity and evolution characteristics of subdivision fields, and finally mine the diversity evolution patterns of the overall domain. This study is the first to propose constructing citation networks using knowledge unit triples. By leveraging the semantic types of triples and knowledge communities, we simultaneously consider the integrity and structurization of scientific knowledge, providing a new perspective for analyzing characteristics such as the content and paths of knowledge evolution. Meanwhile, this paper utilizes knowledge communities to measure the diversity of subdivision fields and summarizes potential scientific knowledge evolution patterns, which can provide reference for domain knowledge diversity evaluation and enrich theoretical research on knowledge units.

1.1 Fine-Grained Scientific Knowledge Representation

Fine-grained representation of scientific knowledge is a prerequisite for discovering its structure and deep characteristics. The knowledge unit is currently one of the important forms of fine-grained scientific knowledge representation, which has inseparable connections with triples and knowledge memes.

Early scholars considered knowledge units as a knowledge structure composed of guide information and knowledge [6], while subsequent research redefined the connotation of knowledge units as "logical combinations of N semantic triples" and accordingly formed the SPO triple description model for knowledge units [7]. Current studies have conducted certain knowledge discoveries by mining contradictory knowledge from massive SPO triples [8] or discovering regular semantic patterns for specific diseases [9]. Knowledge memes are stable terms that accompany topic evolution in citation or co-occurrence relationships [10], often being short text units that are inherited and replicated in citation relation-

ships [11], with some research using them to analyze the disciplinary structure of interdisciplinary fields [12].

Thus, the subject, predicate, and object in triples, as knowledge memes with actual semantics, can define domains and subdivision fields and become an important entry point for knowledge unit citation network evolution analysis.

1.2 Domain Knowledge Evolution

Research on knowledge evolution and knowledge communities draws on analogies to biological problems. Popper first viewed knowledge growth from a biological evolution perspective [13], and subsequent scholars have continuously enriched knowledge evolution theory, proposing that knowledge evolution follows patterns such as inheritance and variation mechanisms [14]. The inheritance mechanism refers to knowledge achieving continuity through inheritance and transmission, while the variation mechanism refers to the process where different knowledge fragments recombine into new knowledge (genes). Analogous to the conceptual model of biological communities, knowledge communities are defined as knowledge clusters with certain biological attributes formed based on potential internal connections between knowledge or specific objectives [15], and citation relationships between knowledge can shape the community structure of domain knowledge [16]. In library and information science, analysis of knowledge communities is similar to clique analysis methods in complex network theory and has been gradually applied to explore knowledge development contexts and patterns [17]. Inspired by this, this paper conducts community-level evolution research and diversity analysis on knowledge unit citation networks to address gaps in previous research.

Topic evolution reveals results that are often more macroscopic than knowledge evolution processes [18], with the former being more suitable for the evolution of different topics and having poor portability for knowledge evolution issues. Some studies identify evolutionary topics for specific research domains and construct evolution paths for analysis. For example, conducting word frequency statistics on SAO structures from medical patent data and building semantic networks to identify core technology topics and development stages [19], extracting keywords from titles and abstracts of NSF data in the AI field for topic mining [20], or directly using SPO triples from the biomedical field as document substitutes to build predicate-based semantic networks and identify emerging research topics [21]. More studies focus on constructing semantic networks of keywords [22], co-occurrence networks [23], or association networks of subject terms [24], combined with LDA models [25] for topic mining. Evolution path construction is based on similarity between topics in different time windows, such as using keyword co-occurrence to assign weights and construct distance matrices to determine evolution paths from co-word networks [26].

Unlike topic evolution, knowledge often evolves within a specific topic through citation relationships. Numerous scholars have analyzed knowledge evolution

around knowledge meme citation networks. Some studies mine knowledge connections between citations from three perspectives: knowledge convergence, knowledge aggregation and divergence, and topic dynamics [27], or use knowledge lifecycle theory to summarize knowledge evolution processes and construct knowledge evolution paths through direct and indirect citation methods of keyword pairs [28]. Additionally, some research uses citation relationships to construct diffusion cascade networks of knowledge memes and discovers four diffusion patterns in medical informatics [29], or analyzes inheritance and variation of knowledge between citations based on knowledge gene flow and diffusion mechanisms [30]. Some scholars have begun conducting knowledge evolution research through knowledge units, such as measuring the migration and recombination of knowledge unit sets in ESI frontier fields across different periods to reveal micro-level knowledge evolution processes and patterns [31], but these studies generally neglect describing knowledge structure, while the fine-grained triple model provides a feasible path for conducting subdivision field knowledge evolution.

2.1 Construction of Subdivision Field Knowledge Unit Citation Networks

A knowledge unit citation network refers to a knowledge network constructed by taking knowledge units in the same subdivision field as nodes and using citation relationships between the documents to which knowledge units belong as edges. The overall domain is defined by the semantics expressed by the predicate, while the subdivision field is defined by the semantic types of the subject and object in the knowledge unit (as shown in Figure 1 [Figure 1: see original paper]). Taking the triple "Rivastigmine-TREATS-Alzheimer's Disease" as an example, the predicate "TREATS" first determines that the triple belongs to the treatment domain, and then the drug "Rivastigmine" and the disease "Alzheimer's Disease" assign the knowledge unit node to the knowledge unit network within the treatment domain with Alzheimer's disease as the subdivision field. The formal representation is as follows: as of time t, the knowledge unit node set is V, and the directed edge set is E.

Figure 1 Example of Knowledge Unit Citation Network Construction

According to the definition of knowledge unit citation networks, the specific construction process is described as follows:

Step 1. Initialize the knowledge unit citation network G(V, E), select document set P and subdivision field Q, where the domain is determined by the predicates of SPO triples extracted from document set P, and the subdivision field is determined by the semantic types of subjects and objects in SPO triples.

Step 2. Extract all SPO triples belonging to subdivision field Q from document set P to obtain node set V.

Step 3. Use citation relationships between documents to which node set V

belongs as edges to obtain edge set E.

Step 4. Divide the network collection into different time periods Ti according to the publication time of document set P.

Step 5. Output the dynamic knowledge unit citation network collection for the subdivision field.

2.2 Community Identification Based on Leiden Algorithm

The Leiden algorithm is a hierarchical clustering algorithm that modifies the Louvain algorithm's problem of loosely connected communities, achieving excellent community identification effects through local node movement, partition refinement, and network aggregation [32]. In knowledge unit citation networks, knowledge communities refer to semantic collections of knowledge units that cluster similar knowledge and are connected through citation relationships. Partitioning network communities at different times can identify knowledge communities within subdivision fields and present knowledge diffusion and transfer situations. This paper applies the Leiden algorithm to identify communities from knowledge unit citation networks. Since overly small communities may not form mature knowledge communities, this paper sets the node threshold for knowledge communities at 3.

Previous studies have identified core nodes as topic labels through network measurement indicators such as degree centrality and node count [33] or other comprehensive indicators [34]. Since node types in knowledge unit citation networks are identical and have high repetition rates, after setting thresholds, this paper uses one or more knowledge units with high proportions as the knowledge label for the community. Taking a knowledge unit Ki in community c as an example, the calculation process for knowledge label pKi is:

$$p_{Ki} = \frac{m_i}{\sum m_i}$$

where mi represents the quantity of knowledge unit Ki in the community, and Σmi represents the total quantity of knowledge units in the community. Subsequently, to unify dimensions across communities at different times, we standardize pKi values for communities at each time point:

$$LM_t = \{K_i | p_{Ki} > \text{threshold}\}$$

$$p_{Ki}^{\text{std}} = \frac{p_{Ki} - p_{\min}}{p_{\max} - p_{\min}}$$

where LMt represents the set of knowledge labels with pKi greater than the threshold, and pmin and pmax represent the minimum and maximum values of pKi in the network, respectively. Finally, we obtain the standardized pKi.

2.3 Diversity Measurement of Domain Knowledge Communities

Knowledge communities share certain commonalities with biological communities. This paper draws on three fundamental dimensions of biodiversity measurement—Variety, Balance, and Disparity—to conduct measurement and analysis of domain diversity. Richness measures the variety of knowledge community types in the current subdivision field, with biology commonly using the Chao index (i.e., community count) to measure richness. Balance reflects the degree to which scientific research in a subdivision field tends toward different knowledge communities, as widely recognized knowledge communities often accumulate more knowledge units. This paper uses the grouping calculation method, employing the proportion of community quantities to calculate the Gini coefficient as the balance indicator. Disparity assesses the degree of difference in composition between different knowledge communities within a subdivision field, with changes in disparity degree demonstrating the applicability of the current knowledge set. Biology commonly uses β-diversity analysis to measure differences between communities. This paper maps community composition within a subdivision field into vectors, calculates pairwise similarity to obtain a distance matrix, and computes domain disparity degree. The calculation formulas are as follows:

$$V = n + \frac{\sum p_i^2}{\sum p_i}$$

$$B = 1 - 2\sum_{i=1}^{n} \frac{(n+1-i)u_i}{n\sum u_i}$$

$$D = 1 - \overline{S_{ij}}$$

where n is the number of communities in the domain, ui is the node count of the ith community, yi and yj are vectors representing the knowledge unit type composition of the ith and jth communities (including knowledge unit types and quantities), sij = cos(yi, yj) is the cosine similarity between communities i and j, and Sij is the distance matrix composed of sij. Similar to changes in biological community diversity, improvements in knowledge community diversity initially manifest as increased richness of knowledge communities, improved balance, and increased disparity. Real-world knowledge systems often undergo multiple simultaneous diversity changes during evolution. The following sections conduct empirical analysis of knowledge community evolution characteristics from specific domains.

3.1 Domain Selection

Constrained by data, tools, and the development degree of disciplinary informatics [35], current knowledge discovery research based on SPO primarily concentrates in the biomedical domain, which features complex semantic relationships where triples extracted from literature possess richer semantic expression capabilities than keywords. This study uses data from the SemMedDB knowledge base (version 43) developed by the U.S. National Library of Medicine. This database uses the medical knowledge extraction tool SemRep to extract titles and abstracts from the PubMed database to form knowledge triples. SemRep identifies sentence subjects and objects according to 21 types of predicate relationships, with each predicate representing a category of knowledge triple sets in medical research [37], such as DIAGNOSES representing disease diagnosis research and TREATS representing disease treatment research.

The core schema of SemMedDB describes the basic attributes of SPO triples, involving the semantic types of subjects and objects, the pmid number of the document to which the triple belongs, and simultaneously links necessary data items such as triple extraction location and document publication time [37] (as shown in Figure 2 [Figure 2: see original paper]). Since SemMedDB is built upon the PubMed database, its pmid numbers correspond to citation relationships between documents contained in PubMed, thereby enabling the acquisition of citation relationships between triples.

Figure 2 Example Data Item Description

Basic statistics on SemMedDB reveal that the TREATS domain, particularly treatment-related disease domains, is a hotspot of biomedical research, accounting for 28.7% of total SPO frequency (see Table 1). Related semantic patterns involve knowledge of therapeutic drugs, treatment measures, and treatment equipment for diseases, namely three semantic patterns: phsu-TREATS-dsyn, topp-TREATS-dsyn, and horm-TREATS-dsyn. Therefore, this paper selects disease research within the TREATS domain as the scope for knowledge community evolution analysis.

Table 1 Predicate Domains and TREATS Domain Semantic Patterns (Top 10 SPO Frequencies)

Predicate	SPO Frequency	Semantic Pattern
TREATS		topp-TREATS-podg
AFFECTS		topp-TREATS-dsyn
COEXISTS_{WITH}		topp-TREATS-neop
INTERACTS_{WITH}		phsu-TREATS-dsyn
CAUSES		phsu-TREATS-podg
ASSOCIATED_{WITH}		hlca-TREATS-humn
STIMULATES		phsu-TREATS-mamm
ADMINISTERED_{TO}		topp-TREATS-fndg

Predicate	SPO Frequency	Semantic Pattern
INHIBITS		hlca-TREATS-dsyn
AUGMENTS		topp-TREATS-popg

Note: topp: therapeutic or preventive procedure; podg: patient or disabled group; dsyn: disease or syndrome; neop: neoplastic process; phsu: pharmacologic substance; hlca: health care activity; humn: human; mamm: mammal; fndg: finding

3.2 Data Collection and Preprocessing

This study first conducts preliminary screening in SemMedDB using TREATS as the predicate retrieval term. Previous research has been limited to empirical analysis of knowledge discovery issues for specific disease types. To obtain more objective evolution characteristics of knowledge communities, this paper expands the data collection scope to multiple diseases with high research attention. Specifically, we screen semantic patterns involving dsyn in SPO triples, obtaining 125 disease names with the highest quantities, comprising a total of 1,048,577 triples.

The preprocessed data undergo the following procedures: Delete triples extracted from title sentences (TYPE attribute = ti). Remove triples whose SUBJECT_{SEMTYPE} attribute is human (humn) or other types that do not express actual information, including Physicians, Author, etc. Delete frequently occurring broad terms describing therapeutic drugs, measures, and equipment across different diseases according to the SUBJECT_{NAME} attribute, including Pharmaceutical Preparations, Therapeutic procedure, etc. Deduplicate pmid numbers from multiple SPO triples originating from the same document.

Finally, we conduct temporal classification and disease classification according to the publication time of documents to which nodes belong and disease names, respectively, to obtain node sets for different times and diseases. We construct directed edges using citation relationships between pmid numbers of triples. For any given disease, we select triples related to that disease from 1960 to 2019, constructing cumulative dynamic knowledge unit citation networks every three years.

3.3 Evolution State Analysis from Community Competition Perspective

Knowledge units within a subdivision field are composed of the same type of knowledge memes. The knowledge unit citation network constructed in the TREATS domain covers all treatment knowledge applicable to a disease. Different knowledge communities within each network maintain constant competitive

relationships, while multiple knowledge units in a community's knowledge label set exist in symbiotic relationships.

After community identification and knowledge label representation for a disease network, this study uses changes in knowledge labels across different times to characterize community evolution and employs cosine similarity of community composition as the similarity between communities at times t and t+1. The similarity between communities at times t and t+1 is defined as:

$$Sim(M_t, M_{t+1}) = \cos(\vec{M_t}, \vec{M_{t+1}})$$

where $\vec{M_t}$ and $\vec{M_{t+1}}$ represent the knowledge unit compositions of communities at times t and t+1, respectively. After obtaining similarities between knowledge communities at different times, we set thresholds to filter evolution relationships between knowledge communities and construct evolution paths. Knowledge communities below the threshold are considered to have no significant evolution relationship, and paths are finally visualized. Some scholars define community evolution in networks as six patterns: birth, growth, merge, decay, split, and death [38]. This study combines information lifecycle theory [39] to define evolution states from the knowledge community competition perspective (as shown in Table 2).

Table 2 Definition of Evolution States from Community Competition Perspective

Evolution State	Definition
Birth	Community knowledge labels appearing at t+1 are different from all those at time t and before
Growth	Community knowledge labels are identical between times t and t+1
Merge	Different knowledge communities at time t integrate into a new knowledge community at t+1
Split	A knowledge community at time t differentiates into two distinct knowledge communities at t+1
Decay	Community knowledge labels at t+1 differ from those at time t

Evolution State	Definition
Death	Community knowledge labels at time t no longer appear at t+1 and thereafter

4.1 Knowledge Community Identification Results in Disease Treatment Domain

We construct knowledge unit citation networks for 125 diseases in the TREATS domain (hereinafter referred to as TREATS) at different time points, obtaining a total of 2,032 networks across 20 timestamps, after which we identify knowledge communities in the networks and 统计相关变量. Since most diseases have formed communities since 1992, this study selects 1992, 2004, and 2016 as time slices to present the density distribution of community count, average community size, network average length, and citation relationship quantity across different periods (as shown in Figure 3 [Figure 3: see original paper]).

The number and scale of disease knowledge communities gradually increase over time, and differences between diseases also expand. The distribution of disease community counts shows that annual disease community numbers mostly concentrate between 0 and 200, and as knowledge accumulates, community counts increase toward the 100-200 range. Meanwhile, the average community size in 2016 is overall 10 nodes larger than the 10-node average in 1992, and the distribution of average community counts and sizes across different diseases becomes more uniform in 2016. Additionally, the depth and density of knowledge unit citation networks gradually increase. Every decade, the number of citation relationships for almost all diseases increases by nearly an order of magnitude compared to the previous stage. Moreover, as citation relationships accumulate and node counts increase, the distribution of network average path lengths becomes more uniform and increases significantly.

Preliminary community identification results indicate that disease knowledge unit citation networks become denser over time, with stable increases in both community counts and internal scales within networks, providing a basis for domain diversity analysis. The gradually expanding differences in community development degrees across different diseases also suggest that community diversity evolution may be influenced by other factors.

Figure 3 Distribution of Knowledge Community Identification Results

4.2 Example Subdivision Field Community Evolution Analysis

Alzheimer's disease treatment research first emerged in 1980 and has since accumulated substantial knowledge, remaining an incompletely conquered research hotspot with typical analytical value. This study uses Alzheimer's disease as an

example to characterize and analyze knowledge labels and community knowledge evolution states, finally explaining the disease's diversity evolution characteristics. This analytical process is equally applicable to other disease subdivision fields.

4.2.1 Community Knowledge Label Evolution Analysis

After representing community knowledge labels for each period, this study screens out knowledge that has persisted for over ten years in larger knowledge communities with continuously expanding scales, totaling 10 types. We 统计其在网络中全部所属知识群落中的数量并绘制河流图 (as shown in Figure 4 [Figure 4: see original paper]).

The above knowledge can be regarded as mainstream treatment options for Alzheimer's disease, which can be roughly divided into four stages. At the beginning of disease research, physostigmine gained absolute attention within limited knowledge communities. Starting in 1989, tacrine rapidly expanded and gradually replaced physostigmine; simultaneously, non-steroidal anti-inflammatory agents and selegiline also formed relatively large knowledge communities. Until 1998, with the explosive development of research, differentiated knowledge formed in different communities, and early community scales gradually stagnated. During this period, acetylcholinesterase inhibitors, donepezil, rivastigmine, and galantamine replaced previous knowledge, while immunotherapy and memantine gradually gained research attention. Between 2010 and 2019, immunotherapy and glutamate inhibitors received more recognition, while rivastigmine disappeared during knowledge label evolution. Meanwhile, new drugs or methods such as curcumin, deep brain stimulation, and acupuncture procedure, although not included in the statistics due to insufficient years, also formed relatively large knowledge communities.

Figure 4 River Diagram of Alzheimer's Disease Knowledge Labels

4.2.2 Community Knowledge Competition State Analysis

Previous evolution studies set the similarity threshold between community topics at 0.7. Considering the high repetition rate of knowledge unit types within TREATS, after sufficient experimentation, this study sets the knowledge community similarity threshold for Alzheimer's disease at 0.8 and partially presents paths containing the above mainstream knowledge labels.

This study finds that after calculating knowledge labels, the 10 mainstream knowledge units of Alzheimer's disease all have competitive relationships in their respective periods. As research develops, increasingly diverse knowledge deepens competition, while knowledge symbiotic relationships only exist in the merge and split processes of some knowledge. According to our definition of knowledge evolution states, taking knowledge labels containing donepezil and memantine from 2004 to 2016 as examples, node colors represent different years (as shown in Figure 5 [Figure 5: see original paper]). After donepezil first

underwent knowledge split in 2007, one community experienced knowledge inheritance and continued to split in 2013, while the other community gradually underwent knowledge substitution and merged with another memantine community to maintain the latter's inheritance state.

After summarizing other knowledge paths, this study finds that new knowledge basically evolves from attaching to current mainstream knowledge communities to independently generating new knowledge communities, rather than gradually replacing original knowledge communities. In the growth process of mainstream knowledge, knowledge inheritance is the most typical characteristic of single knowledge labels represented by six knowledge units including rivastigmine. Knowledge obsolescence means the disappearance of knowledge labels or stagnation of community scale, the former stemming from knowledge substitution (physostigmine) or community reorganization (rivastigmine), the latter manifesting as similarity between adjacent time communities approaching 1. This is preliminarily proven by the community similarity performance of tacrine around 2004, while memantine's similarity maintained above 0.99 from 2013 to 2019, indicating it may face knowledge limitations. This study is limited to extraction results from the SemRep tool, and some knowledge labels still have relatively similar concepts, such as the pharmacological overlap between cholinesterase inhibitors and drugs like rivastigmine. Limited by research length, this study does not deeply explore large-scale knowledge evolution mechanisms, but tracking mainstream knowledge evolution through knowledge labels provides a feasible path for characterizing knowledge lifecycles.

Figure 5 Partial Path of Knowledge Evolution

4.2.3 Community Diversity Characteristics Analysis

Community diversity evolution indicators show that Alzheimer's disease knowledge communities grew rapidly in early stages, with richness scale continuously increasing to reach 63 communities in 2019. From the perspective of community balance, early disease research had high Gini coefficient spans, with balance decreasing significantly. Combined with knowledge label evolution, this indicates that tacrine accumulated extensively within limited knowledge during this period, while after 2000 it roughly maintained a balance value around 0.8. Disease research maintained consistently high disparity, with community disparity approaching 1 in early stages due to fewer knowledge communities and more homogeneous knowledge unit types, after which community disparity fluctuated and decreased to stabilize around 0.9 (as shown in Figure 6 [Figure 6: see original paper]).

Biodiversity theory suggests that higher richness, balance, and disparity generally mean higher domain diversity, which means Alzheimer's disease research shows poor performance on the balance indicator. However, for disease treatment domains, the lack of large exclusive communities means the subdivision field has not formed dominantly recognized knowledge, indicating that research

on the disease has not yet found a consensus solution. Therefore, analysis of knowledge community diversity characteristics must be combined with specific semantics of the subdivision field. Combined with the above knowledge label evolution and competition state analysis, as disease research gradually matures, the variety of Alzheimer's disease knowledge labels gradually increases, accompanied by continuous growth in knowledge community richness and disparity.

Figure 6 Diversity Characteristics of Alzheimer's Disease

4.3 Diversity Evolution Characteristics of the Overall Research Domain

In addition to Alzheimer's disease, this study measures the diversity evolution characteristics of the selected 125 diseases and analyzes their similar or dissimilar diversity performances.

4.3.1 General Diversity Evolution Characteristics

This study first 统计了 125 种疾病的三项多样性指标的平均值与方差, analyzing the 共性 of disease diversity evolution characteristics after fitting (as shown in Figure 7 [Figure 7: see original paper]).

Richness of diseases in TREATS gradually increases, with average community scale growth approaching quadratic function growth (red curve), but exponential variance growth indicates significantly increasing differences in community scale between diseases, suggesting that even among the 125 diseases with the most SPO triples, research knowledge still shows substantial enrichment differences. Additionally, the average Gini coefficient of knowledge communities roughly presents S-shaped growth within the 统计年限, indicating that domain balance in knowledge evolution systems continuously declines with larger mid-term changes. In 2019, many diseases in TREATS exhibited a few knowledge communities occupying absolutely dominant positions, meaning that drug knowledge composing these communities received high recognition for disease treatment efficacy. The continuous decline in balance variance after 1992 proves that differences in balance between diseases gradually narrow, with some diseases' Gini coefficients gradually stabilizing.

Finally, community disparity of diseases slightly declines over time but remains generally high, with knowledge communities maintaining large composition differences. The fluctuating decline in disparity variance indicates that composition differences within diseases become more similar. Further statistics on disparity data for 125 diseases reveal a divergence into two distinct trends: substantial increase and slight decrease, with most diseases' similarity stabilizing around 0.09 in 2019. Combined with the example disease, we infer that low similarity performance may stem from diversified knowledge labels and deepened community competition states.

Figure 7 Similar Characteristics of Overall Domain Disease Diversity Evolution

4.3.2 Locational Characteristics of Diversity Evolution

The comprehensive presentation of richness, balance, and disparity indicators can reflect the locational distribution of different diseases. Using the horizontal axis to represent the balance indicator and the vertical axis for the disparity indicator, with point size indicating disease community richness scale and point color representing different years, we obtain a scatter plot of 125 diseases (as shown in Figure 8 [Figure 8: see original paper]). From the color distribution between 1962 and 2019, we can see that disease evolution roughly follows three patterns: one moves from low-disparity, high-balance small-scale disease research to high-disparity, low-balance large-scale research; the other two move from high-disparity, high-balance small-scale disease research to large-scale research with medium and low balance, respectively, with disparity slightly decreasing.

Different development patterns reflect internally differentiated knowledge evolution processes. Stable growth in disease knowledge community scale indicates continuously increasing knowledge richness in subdivision fields. However, evolution patterns with substantially increased disparity indicate that early disease research stages had relatively single treatment drugs, with community composition becoming increasingly complex as research deepens and disease treatment gradually diversifying, which can be regarded as a "conventional" evolution pattern. Conversely, high disparity in early disease research stages may stem from several controversial treatment options that are only recognized within their respective communities, while as disease research develops, a certain class of drugs or new drugs begins to occupy the main body of disease research. This study considers diseases evolving toward high-disparity, low-balance as showing an "early-controversial" pattern. Finally, medium balance performance in some diseases in 2019 indicates they have not yet formed absolutely dominant knowledge communities. To preliminarily explore the causes of this phenomenon, this paper 统计了部分中等均衡与低均衡疾病的名称及其在 2019 年的基尼系数 (as shown in Table 3). We find that concepts of medium-balance diseases are relatively abstract, mostly being general descriptions of certain body systems or organs, while low-balance characteristics more often point to specific diseases. This study considers high-disparity, medium-balance diseases as showing a "generalized" knowledge evolution pattern.

Table 3 Description Differences Between Low-Balance and Medium-Balance Diseases

Low-Balance Diseases	Medium-Balance Diseases
Type 2 Diabetes	Infectious Diseases
Hypertension Disease	Viral Diseases
Acute Myocardial Infarction	Nervous System Diseases

Low-Balance Diseases	Medium-Balance Diseases
Cerebrovascular Disease	Non-alcoholic Fatty Liver Disease Polycystic Ovary Syndrome

4.3.3 Temporal Characteristics of Diversity Evolution

Considering the factor of disease research start time, this study explores whether it relates to the development of disease knowledge community diversity. The selected diseases in this study have continued without interruption since their start time, and we roughly obtain research duration for each disease by acquiring the number of timestamps in the disease sample. We cluster diseases with different research durations and calculate average diversity indicators to obtain a bubble chart of disease research time differences (as shown in Figure 9 [Figure 9: see original paper]). The vertical axis represents increasing research time from bottom to top, the horizontal axis represents the Gini coefficient, color depth represents similarity magnitude, and point size represents community count.

Although the number of diseases with earlier start times far exceeds that of short-term research diseases, the results still indicate that earlier-started diseases have higher Gini coefficients and lower similarity. Combined with previous general diversity evolution patterns, this shows that earlier-started research forms higher disparity and lower balance. However, diseases with different start times do not show large differences in richness, indicating that knowledge community scale growth is not entirely constrained by start time factors, and some newly emerging diseases may receive more attention and research in a short time, thus rapidly accumulating community scale.

Figure 9 Bubble Chart of Disease Research Duration Differences

5 Summary and Outlook

This paper takes knowledge unit triples as the research object, proposes knowledge unit citation networks, and designs an analysis framework and research process for knowledge community evolution in subdivision fields. Drawing on three diversity measurement indicators—richness, balance, and disparity—we conduct a relatively comprehensive diversity evolution analysis in the biomedical domain. The main conclusions are as follows:

(1) The method enriches related research on knowledge communities, and verifies the feasibility of analyzing knowledge evolution paths and states through example disease community evolution. This study 梳理了阿尔兹海默症四个阶段的 10 种主流治疗方案, analyzes their knowledge evolution states such as inheritance and substitution in their lifecycles, and finally proposes that knowledge diversity evaluation should be combined with specific semantic connotations based on disease balance performance.

(2) This paper discovers that biomedical domains related to treatment follow three aspects of diversity evolution patterns. First, disease richness in the domain shows quadratic function growth while balance shows S-shaped decline. Second, based on comprehensive presentation of diversity indicators, we can roughly classify three disease evolution patterns: conventional, early-controversial, and generalized. Third, diseases with earlier research start times tend to have higher disparity and lower balance.

This paper still has certain limitations. For example, it does not consider the semantic function of knowledge units in articles; future research could incorporate functions such as methods and conclusions to attach more minable information to knowledge units. Additionally, the empirical analysis does not deeply explore associations with other domain knowledge of diseases, and the combination with exogenous disease information is not tight enough to enable deeper medical interpretations. How to more purposefully extract SPO triples and systematically construct paths will be the focus of future research.

References

[1] Sun Z, Leng F. A knowledge evolution analysis method for ESI research fronts based on knowledge unit co-occurrence[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(11): 1095-1113.

[2] ZHANG R, HRISTOVSKI D, SCHUTTE D, et al. Drug repurposing for COVID-19 via knowledge graph completion[J]. Journal of Biomedical Informatics, 2021, 115: 103696.

[3] SONG M, HAN N, KIM Y, et al. Discovering implicit entity relation with the gene-citation-gene network[J]. PloS one, 2013, 8(12): e84639.

[4] Du J, LI X. A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development[J]. JMIR medical informatics, 2020, 8(4): e18323.

[5] Du J. Progress and prospects of measuring uncertainty in medical knowledge[J]. Data Analysis and Knowledge Discovery, 2020, 4(10): 14-27.

[6] Wen Y. Knowledge organization and retrieval based on "knowledge units"[J]. Computer Engineering and Applications, 2005, (01): 55-57+91.

[7] Suo C, Gai S. Research on the connotation, structure, and description model of knowledge units[J]. Journal of Library Science in China, 2018, 44(04): 54-72.

[8] Wang X, Yang X, Li P, et al. Discovery of contradictory drug knowledge based on semantic models[J]. Journal of Intelligence, 2020, 39(07): 159-165.

[9] Cai M, Li X, Zhao J, et al. Disease knowledge discovery based on SPO semantic triples[J]. Data Analysis and Knowledge Discovery, 2022, 6(01): 134-144.

[10] KUHN T, PERC M, HELBING D. Inheritance patterns in citation networks reveal scientific memes[J]. Physical Review X, 2014, 4(4): 041036.

[11] MAO J, LIANG Z, CAO Y, et al. Quantifying cross-disciplinary knowledge flow from the perspective of content: Introducing an approach based on knowledge memes[J]. Journal of Informetrics, 2020, 14(4): 1751-1577.

[12] Cao Y, Liang Z, Mao J. Analysis of disciplinary structure in interdisciplinary research fields from the perspective of knowledge memes[J]. Library Tribune, 2019, 39(07): 84-90.

[13] K · Popper. Objective Knowledge: An Evolutionary Approach[M]. Translated by Shu W, et al. Shanghai: Shanghai Translation Publishing House, 1987: 114.

[14] Wang Y, Ding Y. Construction and simulation of an evolution model for scientific literature dissemination networks based on knowledge evolution perspective[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(09): 966-973.

[15] Teng G, Yang M, Tian Y, et al. Knowledge communities and core knowledge analysis in Folksonomy mode[J]. Library and Information Service, 2015, 59(22): 124-129.

[16] PHAM M C, KLAMMA R, JARKE M. Development of computer science disciplines: a social network analysis approach[J]. Social Network Analysis and Mining, 2011, 1(4): 321–340.

[17] Teng G, He D, Peng J, et al. Research on growth mechanisms of domain knowledge communities based on network evolution[J]. Information Studies: Theory & Application, 2016, 39(10): 16-20+15.

[18] XU J, DING Y, BU Y, et al. Interdisciplinary scholarly communication: An exploratory study for the field of joint attention[J]. Scientometrics, 2019, 119(3): 1597–1615.

[19] Ma M, Wang C, Zhou Y, et al. Research on methods for identifying and analyzing evolution trends of core technology topics based on semantic information[J]. Information Studies: Theory & Application, 2021, 44(09): 106-113.

[20] Jin J, Wang Y, Ba Z, et al. Topic mining and dynamic evolution analysis of funded project research—Taking the AI field in NSF data as an example[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(09): 967-979.

[21] HU Z, ZENG R, PENG L, et al. Discovering Emerging Research Topics Based on SPO Predications[C]//L, Uden et al. Communications in Computer and Information Science. Zamora, Spain: Springer, 2019: 110-121.

[22] Chen X, Huang L, Ni X, et al. Research on topic evolution path identification based on dynamic semantic network analysis[J]. Journal of the China Society for Scientific and Technical Information, 2021, 40(05): 500-512.

[23] Huang C, Huang S, Fu H. International comparative study on knowledge flow and research themes in AI governance from an interdisciplinary perspective[J]. Journal of Information Resources Management, 2022, 12(06): 98-110.

[24] Hu J, Yang Z, Zhu G, et al. What is the Chinese government doing—Content structure analysis of Government Work Reports based on word association[J]. Journal of Information Resources Management, 2023, 13(01): 115-128.

[25] Li W, Tan L, Zhang G, et al. Research on identifying key core technology topics in industrial chains based on multi-source information fusion—Taking the AI field as an example[J]. Journal of Information Resources Management, 2022, 12(01): 116-126.

[26] WANG X, HE J, HUANG H, et al. MatrixSim: A new method for detecting the evolution paths of research topics[J]. Journal of Informetrics, 2022, 16(4): 1751-1577.

[27] KIM E, JEONG Y K, KIM Y H, et al. Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction[J]. Journal of Informetrics, 2022, 16(1): 1751-1577.

[28] ZHANG X, XIE Q, SONG C, et al. Mining the evolutionary process of knowledge through multiple relationships between keywords[J]. Scientometrics, 2022, 127(4): 2067–2087.

[29] Liang Z, Mao J, Cao Y, et al. Analysis of domain knowledge diffusion patterns based on knowledge meme cascade networks[J]. Information Studies: Theory & Application, 2020, 43(04): 40-46+39.

[30] Bai R, Sun Y, Zhang Q. Research on constructing scientific and technological innovation paths based on knowledge gene expression[J]. Information Studies: Theory & Application, 2020, 43(04): 137-144+176.

[31] Sun Z, Leng F. A knowledge evolution analysis method for ESI research fronts based on knowledge unit migration[J]. Journal of the China Society for Scientific and Technical Information, 2021, 40(10): 1027-1042.

[32] TRAAG V A, WALTMAN L, VAN ECK N J. From Louvain to Leiden: guaranteeing well-connected communities[J]. Scientific Reports, 2019, 9(1): 5233.

[33] He N, Li D, Gan W, et al. Survey on discovering important nodes in complex networks[J]. Computer Science, 2007(12): 1-5+17.

[34] An S, Nie P, He G. Comprehensive measurement method for node importance in node-weighted networks[J]. Journal of Management Sciences in China, 2006(06): 37-42+52.

[35] Dai B, Hu Z. Recent research review on literature-based knowledge discovery[J]. Data Analysis and Knowledge Discovery, 2021, 5(04): 1-12.

[36] KILICOGLU H, ROSEMBLAT G, FISZMAN M, et al. Broad-coverage biomedical relation extraction with SemRep[J]. BMC Bioinformatics, 2020,

21(1):188.

[37] KILICOGLU H, SHIN D, FISZMAN M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications[J]. Bioinformatics, 2012, 28(23): 3158–3160.

[38] PALLA G, BARABÁSI A L, VICSEK, T. Quantifying social group evolution[J]. Nature, 2007, 446(7136): 664–667.

[39] Suo C. On the concept and research content of information lifecycle[J]. Library and Information Service, 2010, 54(13): 5-9.

Author Contributions: Mao Jin: Proposed research ideas, revised the paper; Hou Bowen: Conducted literature review, data processing and analysis, wrote and revised the paper; Wang Yimeng: Collected and processed data.

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