A Knowledge Representation Model for Studying Knowledge Creation, Usage, and Evolution

Zhentao Liang[1], Fei Liu[1], Jin Mao[1,2], and Kun Lu[3](✉)

[1] School of Information Management, Wuhan University, Wuhan, Hubei, China
[2] Center for Studies of Information Resources, Wuhan University, Wuhan, Hubei, China
[3] School of Library and Information Studies, University of Oklahoma, Norman, OK 73019, USA
kunlu@ou.edu

Abstract. A knowledge representation model is proposed to facilitate studies on knowledge creation, usage, and evolution. The model uses a three-layer network structure to capture citation relationships among papers, the internal concept structure within individual papers, and the knowledge landscape in a domain. The resulting model can not only reveal the path and direction of knowledge diffusion, but also detail the content of knowledge transferred between papers, new knowledge added, and changing knowledge landscape in a domain. A pilot experiment is carried out using the PMC-OA dataset in the biomedical field. A case study on one knowledge evolution chain of Alzheimer's Disease demonstrates the use of the model in revealing knowledge creation, usage, and evolution. Initial findings confirm the feasibility of the model for its purpose. Limitations of the study are discussed. Future work will try to address the recognized limitations and apply the model to large scale automated analysis to understand the knowledge production process.

Keywords: Knowledge representation model · Knowledge evolution · Full-text citation analysis · Alzheimer's Disease

1 Introduction

Scientific knowledge growth has become an interesting research topic in science. In recent decades, the number of scientific publications has been soaring exponentially due to the blooming of research activities and the advance of information technology [1]. The large scale of scientific literature forms a treasurable knowledge vault, which records the trajectory of knowledge creation, usage, and evolution. The availability of academic resources has been dramatically improved as the development of scholarly databases, such as Web of Science, Scopus, Google Scholar, etc. Many researchers have leveraged these digital resources to investigate the usage and evolution of knowledge within and across research domains.

K. Toeppe et al. (Eds.): iConference 2021, LNCS 12645, pp. 97–111, 2021.
https://doi.org/10.1007/978-3-030-71292-1_9

To study this problem, many researchers attempt to identify explicit and implicit associations among the carriers of knowledge. Citations have long been recognized as a way of symbolizing knowledge transfer, based on which citation networks can be modeled to track the development of science [2]. A few methods and visualization tools have been proposed to analyze the evolution of research topics, e.g., HistCite [3] and CiteNetExplorer [4]. Citation relations do not solely demonstrate the content of knowledge. Human interpretation of citation analysis results actually relies on the context information of citations, such as the titles of citing and cited papers, although only referred by experts. This methodology is well applied in many related studies. However, it is imperative to integrate the context of citations formally into analytic methods for reducing experts' effort and subjective interference.

Alternatively, some content-based methods probe into knowledge evolution by exploring content connections among the articles in different time periods. The representations of content are different, which depend on the research purpose of specific study. Well adopted are terms and topics in most current studies. The advantage of content-based methods is that the knowledge representation is easy to interpret and the change of knowledge could be measured at a given aggregation level (e.g. by term or by topic). However, the connections among different terms/topics are not explicitly modelled or manifested by exploiting observable evidence, e.g., citations. The details of knowledge usage and evolution are often not disclosed by such methods.

A few recent studies have attempted to combine citation relations and knowledge content that citations carry to investigate knowledge diffusion [5–7]. Their underlying motivation is to identify what knowledge is spread along a citation by matching the terms in the citing and cited articles. It then enables tracing the diffusion paths of knowledge units (i.e. terms). However, the context of citations they consider is of far distance, rather than the surrounding sentences where the citations occur. It is necessary to further improve the methodology for investigating not only knowledge diffusion, but also the emergence of knowledge and the evolutionary relationship among different knowledge units.

In this study, we propose a knowledge representation model to facilitate formal studies on knowledge creation, usage, and evolution. The model captures the citation relationship among papers, internal concept structure of papers, and domain knowledge context. A multi-layer network model is used to integrate different relationships in one knowledge representation model. Citation contexts are analyzed to ascertain the knowledge usage between citing and cited papers. The model allows systematic analysis on knowledge creation, usage, and evolution using principled network models and mathematical algorithms. This study contributes to the formal methodology of studying the knowledge production process from a temporal perspective.

2 Literature Review

2.1 Knowledge Representation

LIS Perspectives. The field of Library and Information Science (LIS) has a long history in studying knowledge representation in the subfield of knowledge organization. The primary focus is on document representation since the field is specialized in managing recorded information [8]. A primary goal for knowledge representation in the LIS field has been to support information retrieval. Therefore, the representation is generally focused on the document content [9]. Rules and standards to describe and represent documents include classification systems, subject headings, and other forms of metadata [10]. The process is traditionally called cataloging in the context of library materials, and more recently resource description in the broader context of the information world. The results are bibliographic records or metadata records that contain essential features of the original documents and serve as surrogates for information retrieval purposes. Folksonomies and ontologies are additions to the traditional knowledge representation tools in LIS [11].

Another related area of knowledge representation in LIS is informetric studies that produce knowledge maps representing knowledge structures [12]. Knowledge maps at different granularity levels have been created to illustrate knowledge structures, including words, papers, journals, and disciplines. Common relationships used to create knowledge maps include word co-occurrences, semantic similarity, citation relationships (including co-citation and bibliographic coupling), and collaboration [13–17]. This subfield of studies on knowledge maps aims to reveal the structures in scientific domains rather than represent knowledge for information retrieval as in knowledge organization. Nevertheless, the maps are knowledge representations of domains. Effort has also been made to study the evolution of knowledge from a longitudinal perspective [18].

Other Perspectives. Besides LIS, cognitive science has also delved into knowledge representation, but more from the perspective of mental representation that focuses on cognitive abilities [19]. The artificial intelligence community has also discussed knowledge representation in the realm of logics, reasoning, and inferences [20].

2.2 Citation Theory

Citation relationship is widely used to reveal relationships among research work. Citations are also the foundation of many evaluative metrics for scholarly impact. Information scientists have had lengthy discussions on what citations mean and represent [21]. An influential dichotomy is the normative view versus the social constructivist view of citations [22], while the former emphasizes the intellectual functions of citations and the latter emphasizes the social factors. The two views have important implications for the use of citations because if the functions of citations are intellectual, then they can be reliably used to measure the intellectual relationship among papers; while if citations are socially constructed, the reliability of them reflecting intellectual connections becomes questionable. Empirical studies have been carried out to test the two views [23, 24]. Recent development and discussion seem to acknowledge the various factors influencing citing behaviors, but also confirm the intellectual functions of citations [25].

2.3 Citation Networks for Knowledge Evolution

The use of citation networks to describe the development of science is not new. Garfield, Sher and Torpie [26] demonstrated the feasibility of using citation data for historical analysis of science, in the case of DNA discoveries. The idea was later developed into the HistCite software that can produce an interconnected historiograph of highly cited publications for a particular topic [3]. Hummon and Doreian [27] proposed a main path analysis method to identify the mainstream of research in citation networks. The method was used by Lucio-Arias and Leydesdorff [28] on HistCite output to highlight significant paths in science development. As reviewed in [29], main path analysis and citation networks have been widely used to map technological trajectories, exploring scientific knowledge flows, and conducting literature reviews.

Despite its success in revealing the path of knowledge diffusion, citation data does not shed light on the development of knowledge content and structure. It is not immediately clear what knowledge is added to a field by a node in a citation chain/network and what knowledge is transferred from a node to the next. The model proposed in this study integrates both citation relationship and knowledge content. The use of citations is based on the intellectual functions of citations to represent knowledge usage between citing and cited papers. The model also captures the knowledge content in individual papers and the development of knowledge in a broader context of a domain.

3 Model Description

The foundation of this model is built on citation theory [21]. A paper cites a previous work to acknowledge its influence on the current paper. In the meanwhile, a paper generally covers several related concepts/entities and studies their relationships, or a paper may introduce new concepts, theories, methods, or techniques, etc. A local concept level captures the internal structure within the scope of a paper. The citation relationship between papers can be further elaborated by concepts/entities in citation contexts [30]. In addition, a global context is represented by the domain concept level that aggregates the knowledge pieces in each paper and provides an overview of the knowledge landscape.

Most of the previous studies on knowledge creation, usage, and evolution are based on single-layer networks, such as citation, collaboration, co-citation, coupling networks, or the integration of multiple networks into a single-layer composite network [31–34]. However, some researchers argue that a single-layer network is a crude approximation of reality, which ignores considerable important information existing in the corresponding multi-layer network. Furthermore, numerous phenomena and dynamic behaviors only emerge in multi-layer networks, but not in single-layer networks [35, 36]. Therefore, a multi-layer network model is used in this study to represent relationships at different levels as well as the cross connections between layers (Fig. 1).

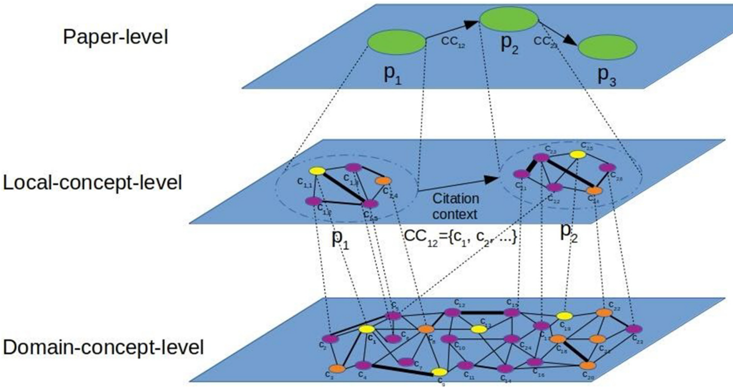

Fig. 1. A knowledge representation model (Different colored nodes in Local-concept-level and Domain-concept-level represent different entity types. The boundary of a domain is left for the users of the model to define.) (Color figure online)

4 Model Definition

According to the model description above, we construct a knowledge representation model characterized by a three-layer scientific network to study the creation, usage, and evolution of knowledge. This network is a pair $M = (L, C)$ where $L = \{L_\alpha, \alpha \in \{$ PL (Paper level), LCL (Local concept level), DCL (Domain concept level)$\}\}$ is a set of layers, and C is the set of cross connections between different layers, which can be formularized as:

$$\mathrm{C} = \{E_{\alpha\beta} \subseteq L_\alpha \times L_\beta;\ \alpha, \beta \in \{PL,\ LCL,\ DCL\}\} \tag{1}$$

In this paper, L_{PL} represents the citation network, where the vertices in L_{PL} are scientific publications, and the edge between two publications indicates the citing relations. Therefore, L_{PL} is a directed network without weights. $L_{\mathrm{PL}} = (V_{\mathrm{PL}}, E_{\mathrm{PL}})$, where $V_{\mathrm{PL}} = \{v_1^{\mathrm{PL}},\ \cdots, v_{N_{\mathrm{PL}}}^{\mathrm{PL}}\}$ is the vertex set of L_{PL}, and N_{PL} is the number of vertices in L_{PL}. E_{PL} is the edge set of L_{PL}, which can be described by the corresponding directed adjacency matrix $A^{\mathrm{PL}} = (a_{ij}^{\mathrm{PL}}) \in \mathrm{R}^{N_{\mathrm{PL}} \times N_{\mathrm{PL}}}$, where

$$a_{ij}^{\mathrm{PL}} = \begin{cases} 1 & \text{if}\left(v_i^{\mathrm{PL}}, v_j^{\mathrm{PL}}\right) \in E_{\mathrm{PL}},\quad \textit{note}\ \text{that}\left(v_i^{\mathrm{PL}}, v_j^{\mathrm{PL}}\right) \neq \left(v_j^{\mathrm{PL}}, v_i^{\mathrm{PL}}\right) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

L_{LCL} captures the internal concept structure within a publication and also specifies the concepts in citation contexts. In order to reveal the relationship among the concepts in a publication and that in different publications more clearly, we introduce the hypergraph theory [37] and establish a hypernetwork to depict L_{LCL}. The hypernetwork $L_{\mathrm{LCL}} = (V_{\mathrm{LCL}}, H_{\mathrm{LCL}}, E_{\mathrm{LCL}})$, where $V_{\mathrm{LCL}} = \{v_1^{\mathrm{LCL}},\ \cdots, v_{N_{\mathrm{LCL}}}^{\mathrm{LCL}}\}$ is the concept set of L_{LCL}, and N_{LCL} is the number of unique concepts contained in all the publications. $H_{\mathrm{LCL}} = \{H_1^{\mathrm{LCL}},\ \cdots, H_{N_{\mathrm{PL}}}^{\mathrm{LCL}}\}$ is a family of non-empty subset of V_{LCL}. Each element in H_{LCL}, $H_\gamma^{\mathrm{LCL}}(\gamma \in \{1,\ \cdots, N_{\mathrm{PL}}\})$, can be characterized by a single-layer network, namely,$H_{\gamma}^{\mathrm{LCL}} = (V_{\gamma}^{\mathrm{LCL}}, E_{\gamma}^{\mathrm{LCL}})$, where $V_{\gamma}^{\mathrm{LCL}} \subseteq V_{\mathrm{LCL}}$ represents the internal concepts of a publication γ. $E_{\gamma}^{\mathrm{LCL}}$ represents the relation between two concepts discussed in a paper, which can be operationalized by concept co-occurrences in a paragraph. The paragraph-level co-occurrence is preferred over the traditional paper-level co-occurrence because it reflects a more granular relationship of the concepts by utilizing the full-text information [38]. For instance, a shorter distance between entities in an article suggests a higher similarity between them [39]. $E_{\gamma}^{\mathrm{LCL}}$ can be formularized by a weighted adjacency matrix $A_{\gamma}^{\mathrm{LCL}} = (a_{ij}^{\gamma}) \in \mathrm{R}^{N_{\gamma} \times N_{\gamma}}$, where

$$a_{ij}^{\gamma} = \begin{cases} count\left(v_i^{\gamma}, v_j^{\gamma}\right) \text{ if } \left(v_i^{\gamma}, v_j^{\gamma}\right) \in E_{\gamma}, \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

The connection between two hyperlinks (E_{LCL}) indicates that there is a conceptual citation relationship between two papers, namely, one of the hyperlinks cites another hyperlink because of a certain concept. Therefore, there are two necessary conditions for this connection. The first condition is the citation relationship between two publications corresponding to the two hyperlinks. The second condition is that these two hyperlinks share some common concepts. We define this connection as $E_{\mathrm{LCL}}(\mathrm{co})$. To this extent $E_{\mathrm{LCL}}(\mathrm{co})$ can be expressed by a weighted adjacency matrix $A_{H_{\mathrm{LCL}}}^{\mathrm{LCL}} = (a_{ij}^{H_{\mathrm{LCL}}}) \in \mathrm{R}^{N_{H_{\mathrm{LCL}}} \times N_{H_{\mathrm{LCL}}}}$, where

$$a_{ij}^{H_{\mathrm{LCL}}} = \begin{cases} a_{ij}^{\mathrm{PL}} \times card\left(H_i^{\mathrm{LCL}} \cap H_j^{\mathrm{LCL}}\right) \text{if } H_i^{\mathrm{LCL}} \cap H_j^{\mathrm{LCL}} \neq \phi, \\ 0 \quad\quad \text{otherwise.} \end{cases} \tag{4}$$

L_{DCL} is the third layer of the knowledge representation model that can reveal the global relations among all concepts across papers. L_{DCL} can be described by a weighted network, $L_{\mathrm{DCL}} = (V_{\mathrm{DCL}}, E_{\mathrm{DCL}})$, where $V_{\mathrm{DCL}} = \{v_1^{\mathrm{DCL}}, \cdots, v_{N_{\mathrm{DCL}}}^{\mathrm{DCL}}\}$ is the collection of concepts in a domain, and N_{DCL} is the number of concepts. E_{DCL} is the weighted edge set that represents the co-occurrence frequency of two concepts from the entire domain, which can be described by a weighted adjacency matrix $A^{\mathrm{DCL}} = (a_{ij}^{DCL}) \in \mathrm{R}^{N_{DCL} \times N_{DCL}}$, where

$$a_{ij}^{DCL} = \begin{cases} count(v_i^{DCL}, v_j^{DCL}) \quad \text{if } (v_i^{DCL}, v_j^{DCL}) \in E_{DCL}, \\ 0 \text{ ohterwise.} \end{cases} \tag{5}$$

Conceptually, cross connections mean the relationship between different layers of a multi-layer network model. There are two types of cross connections between the three layers in the proposed model. The first is the connections between the nodes in the paper-level and the hyperlinks in the local-concept-level networks. It denotes the correspondence between papers in the PL layer and the hyperlinks in H_{LCL}. The second is the connections between the nodes in the local-concept-level network and the domain-concept-level network. It denotes the correspondence between the concepts in LCL layer and those in DCL layer.

5 Pilot Experiment

The operational definition of a knowledge unit in this study is a concept or a concept relationship. The proposed model is expected to support studies on knowledge creation, usage, and evolution because it incorporates both citation relationships between papers and concept structures within individual papers and in a domain. In this study, we performed a pilot experiment to investigate the creation, usage, and evolution process of the knowledge about a well-known disease, Alzheimer's Disease (AD). The same method should apply to other concepts or concept relationships. Figure 2 shows the procedure of the pilot experiment, which is divided into four primary steps.

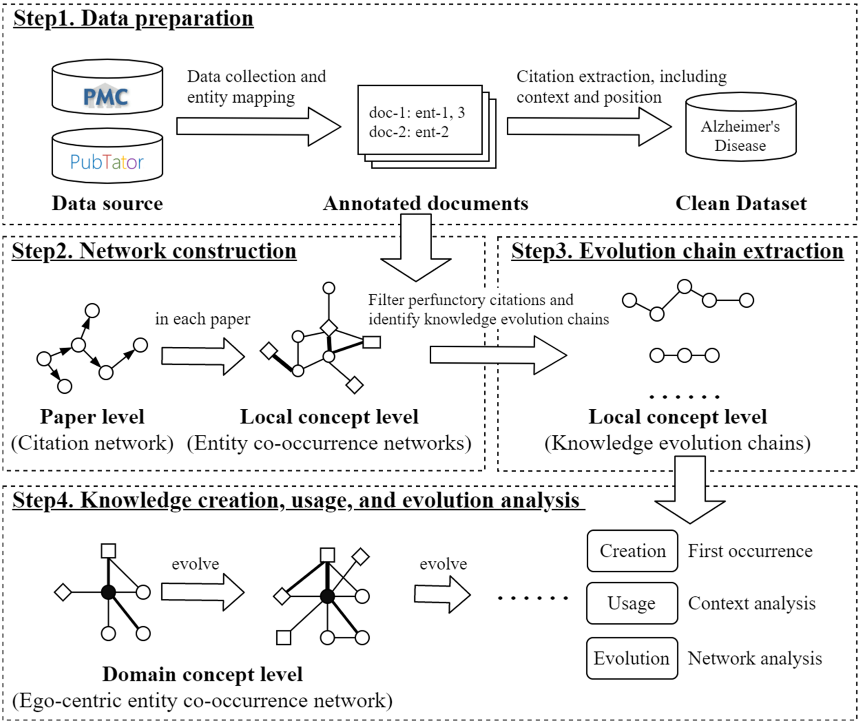
Step1. Data preparation
PMC
PubTator
Data collection and entity mapping
doc-1: ent-1, 3
doc-2: ent-2
Citation extraction, including context and position
Alzheimer's Disease
Data source
Annotated documents
Clean Dataset
Step2. Network construction
in each paper
Paper level
(Citation network)
Local concept level
(Entity co-occurrence networks)
Filter perfunctory citations and identify knowledge evolution chains
Step3. Evolution chain extraction
Local concept level
(Knowledge evolution chains)
Step4. Knowledge creation, usage, and evolution analysis
evolve
evolve
Domain concept level
(Ego-centric entity co-occurrence network)
Creation
First occurrence
Usage
Context analysis
Evolution
Network analysis

Fig. 2. An overview of the experiment procedure

5.1 Data Preparation

PubMed Central was chosen as the source of our data since it provides full-text articles in XML format, which is essential for extracting citations as well as their contextual and positional information. We collected articles related to Alzheimer's Disease from PubMed Central Open Access Subset (PMC-OA) using keyword matching in titles and abstracts (keyword = "Alzheimer's Disease"). This resulted in 22,363 articles spanning from 1995 to 2020. The dataset was further augmented by adding the cited and citing papers of the initial articles if available in PMC-OA, resulting in 119,093 articles. Biomedical entities and their relationships were considered the basic knowledge units in this study. We performed a sentence-level mapping to the Pubtator Central system [40] to identify the entities that belong to the category of disease, gene, and chemical in each paper. Finally, a clean dataset was built by extracting the citation relationships among papers, as well as the citation context and positional information.

5.2 Network Construction

Paper-level citation network was constructed by parsing the in-text citations and reference list of each paper. Then, PMIDs (PubMed IDs) were extracted and used to identify the citing and cited papers in our dataset. Isolated papers that could not be connected to any cited or citing papers in the collection were removed from the network. Since we are interested in the knowledge relationship among papers, it is more convenient to have the edges in our citation network go from the cited papers to the citing ones, the same as the direction of knowledge diffusion. For each paper, an entity co-occurrence network was established within the paper by tallying the entity co-occurrences in the same paragraphs. The paragraph-level co-occurrence offers a more accurate measure of entity relationship than a paper-level co-occurrence does. The entity co-occurrence network is also called the knowledge network for each paper as it represents the knowledge structure within the paper. To focus on the evolution of a specific knowledge unit (in our case AD), ego-centric networks were analyzed.

5.3 Extraction of Knowledge Evolution Chains

Based on the paper-level citation network and the knowledge network in each paper, we extracted the evolution chains for knowledge about AD. A knowledge evolution chain is defined as an acyclic and unbranched path in the citation network, whose nodes are papers containing specific knowledge as their major research objects and edges are citations that also include the same knowledge in their context. In this study, the top ten entities with the highest weights in the knowledge network of a paper and those that appear in the title or abstract were considered its core concepts. Perfunctory citations that do not contain any core concepts of the cited paper in their context were filtered. To ensure knowledge evolution chains concentrate on AD, we retained only citations that include the entity Alzheimer's Disease in their citation context. Knowledge evolution chains were then identified by traversing from all zero-in-degree nodes to zero-out-degree nodes, that is, from the papers not citing any other papers (source of knowledge evolution) to those not cited by any other papers in the collection (end of knowledge evolution). Each chain depicts a distinct pathway of knowledge creation, usage, and evolution process. In addition, the knowledge network of each paper in the chains can be aggregated at the domain concept level, forming a unified entity co-occurrence network that evolves with the accumulation of articles over time. In this pilot experiment, the ego-centric networks of AD in chains were used to create the domain-concept-level network. In the next section, the creation, usage, and evolution of knowledge are investigated quantitatively and qualitatively by network and context analysis.

6 Results

6.1 Descriptive Statistics

Overall, the papers in our dataset mention biomedical entities for 29,957,744 times in their full texts, with 30.8%, 45.9%, and 23.3% for disease, gene, and chemical, respectively. Among them, there are 7,569, 79,774 and 28,298 distinct diseases, genes, and chemicals. The entities with the highest frequencies of the three categories are presented in Table 1.

Table 1. High-frequency diseases, genes, and chemicals

Disease	Gene	Chemical
Alzheimer Disease	Insulin (INS)	Lipids
Neoplasms	Superoxide dismutase 1 (SOD1)	Glucose
Dementia	Apolipoprotein E (APOE)	Water
Parkinson Disease	Toll-like receptor 4 (Tlr4)	Reactive Oxygen Species
Diabetes Mellitus	Membrane associated ring-CH-type finger 8 (MARCHF8)	Sodium Chloride

Table 2 shows some descriptive statistics of the Paper-Level (PL) citation network and the Local-Concept-Level (LCL) entity co-occurrence networks. The averaged statistics of the LCL networks are presented since each paper has its own knowledge network. It is shown that the PL network is extremely sparse with a low density and the average degree of its nodes is 4.66. This is likely due to the restriction that both citing and cited articles should be in the PMC-OA set. Within the scope of a single paper, about 50 knowledge entities (i.e. diseases, genes, or chemicals) are mentioned in the full text. Each entity has co-occurrence relationships with about 10 others on average.

Table 2. Statistics of the citation network and entity co-occurrence networks

	Nodes	Edges	Density	Average Degree
PL network	118,504	552,700	3.94×10^{-5}	4.66
LCL network	49.61 (avg.)	345.18 (avg.)	0.26 (avg.)	10.49 (avg.)

On average, there are 1.89 entities in each citation context, represented as the sentence of a citing paper where the citation locates, with 90% of the citation contexts containing 0 to 4 entities. Regarding the type of entities in a single citation context, 43.0% and 42.7% are diseases and genes, while the chemicals only account for 14.3%. This is different from the distribution of entity types in the full text. By considering the entities in citation contexts and the core concepts of the cited papers, we extracted 67,427 knowledge evolution chains from the dataset, with 4,615 distinct source papers that do not cite any other papers in our dataset. The ego-centric entity co-occurrence network, whose ego node is the knowledge entity of interest (AD in this case), can be obtained from the LCL knowledge network of each paper in the chain.

The descriptive statistics of knowledge evolution chains are presented in Table 3. The average length of an evolution chain is 3.61, with about 143 entities in the Domain-Concept-Level (DCL) ego-centric entity co-occurrence network it constructs and an average time span of 7.98 years. It is also shown that the standard deviation is large for these metrics, indicating the heterogeneity among evolution chains.

Table 3. Statistics of the knowledge evolution chains | N = 67,427

	Mean	SD	Median	Q1–Q3
Length	3.61	1.75	3.00	2.00–5.00
Time span (year)	7.98	3.33	8.00	6.00–10.00
Total distinct entities on a chain	142.97	79.12	135.00	81.00–193.00

To quantitatively understand the network evolving process, we proposed a metric to capture the contribution made by each paper to the DCL network, namely Knowledge Cumulation Speed (KCS). The KCS measures how many new entities or relationships between entities are added to the DCL network constructed at three different levels, including one single chain where the paper locates (single chain), all chains that contain the paper (chain set), and all preceding papers from the domain (domain), by each paper. This shows the contribution by each paper to the cumulative knowledge of a single chain, all chains passing through it, and all preceding papers in the dataset.

As Table 4 shows, on average, a paper contributes 10.74 new entities and 50.1 new relationships to the DCL network of a single chain. By excluding 2,669 review articles, this number falls to 9.53 and 43.85 respectively. The standard deviation also slightly decreases. This shows that review articles contribute a higher than average number of new entities and relationships to a chain. This is likely the result of synthesizing entities and relationships across articles from multiple chains, which is referred to as the integrator effect in [29]. However, a paper contributes fewer new entities and relationships to the DCL network of the chain set, with 5.92 entities and 26.26 relationships. Regarding contributions to the domain DCL network, this number decreases substantially to only 0.73 entities and 11.64 relationships, respectively. This means that an article is less novel from a macro perspective (i.e. chain set and domain), compared with one single chain where it locates. More intriguingly, the average contributions to the DCL network of the chain set and domain increase after removing review articles. While review articles are the knowledge integrator of a single evolution chain, they generally contribute fewer novel entities to the domain. Instead, reviews may focus on organizing existing knowledge so the average KCS (domain) - relationship is slightly higher (11.64 vs. 11.53) if reviews are included.

Table 4 Knowledge cumulation speed (KCS) at different levels

	Mean	SD	Median	Q1–Q3
Including reviews				
KCS (single chain) - entity	10.74	8.57	9.00	5.21–13.95
KCS (chain set) - entity	5.92	6.27	4.00	1.00–8.00
KCS (domain) - entity	0.73	1.69	0.00	0.00–1.00
KCS (single chain) - relationship	50.10	53.09	40.00	15.00–65.10
KCS (chain set) - relationship	26.26	46.90	12.00	3.00–31.00
KCS (domain) - relationship	11.64	30.68	2.00	0.00–11.00
Excluding reviews				
KCS (single chain) - entity	9.53	7.04	8.14	5.00–12.16
KCS (chain set) - entity	6.32	6.53	5.00	2.00–9.00
KCS (domain) - entity	0.79	1.78	0.00	0.00–1.00
KCS (single chain) - relationship	43.85	47.00	35.60	13.00–57.17
KCS (chain set) - relationship	27.06	47.73	12.00	3.00–32.00
KCS (domain) - relationship	11.53	30.29	2.00	0.00–11.00

6.2 Case Study

To investigate the knowledge creation, usage, and evolution process in depth, we showed a case study on one knowledge evolution chain of AD. Figure 3 demonstrates the 3-layer representation model applied to this chain. Paper level shows the citation relationships between papers and the entities transferred through citations, whereas local concept level and domain concept level present the AD knowledge structure within a paper and the accumulated AD knowledge structure of this chain. For better readability and visualization, the knowledge networks were filtered by edge weight (greater than 5).

The chain begins with a paper (PMC161361) focusing on the importance of oxidative stress in the pathogenesis of AD. The paper pointed out that iron and copper are likely to be the source of oxidative stress. The LCL network of the first paper is on the link between AD and iron. The second paper PMC4132486 also focused on AD pathology, but with a different approach. They performed a differential network analysis on four region-specific gene co-expression networks. With this novel method, they also reached the conclusion that oxidative stress is a highlighted process in early AD. This paper adds new entities hippocampus (HIP) and posterior cingulate cortex (PCC) to DCL, which are the brain regions affected by AD. Similarly, the third paper PMC4718516 employed network topology analysis to identify genes related to AD, adding related genes CD4, DCN, CXCL8, PSEN1 and BACE1 to the DCL of the knowledge chain. The addition of the third paper enriches the connections between AD and related genes in this chain. Based on the related genes, paper PMC5508523 further analyzed the relationship between NRF2 (NFE2L2 officially) gene deficiency and increased oxidative stress, which may lead to AD eventually. They conducted the experiments on a house

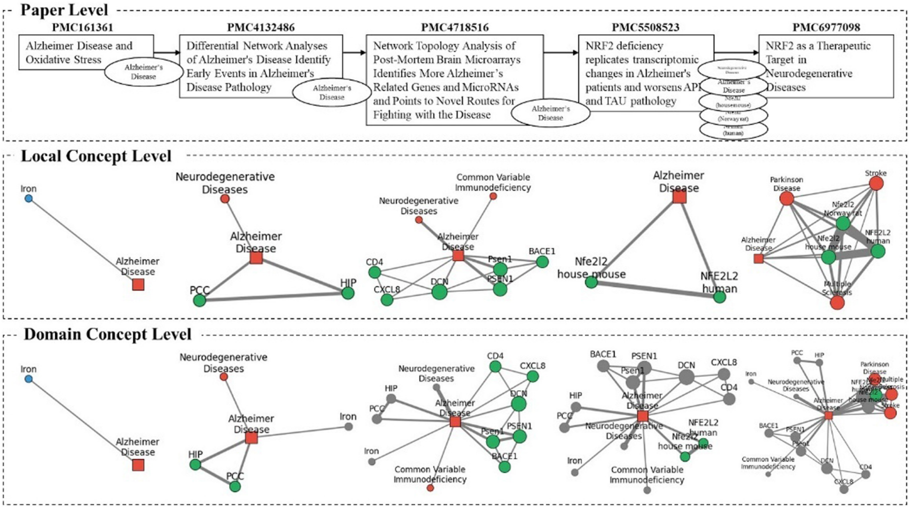
Paper Level
PMC161361
Alzheimer Disease and Oxidative Stress
Alzheimer's Disease
PMC4132486
Differential Network Analyses of Alzheimer's Disease Identify Early Events in Alzheimer's Disease Pathology
Alzheimer's Disease
PMC4718516
Network Topology Analysis of Post-Mortem Brain Microarrays Identifies More Alzheimer's Related Genes and MicroRNAs and Points to Novel Routes for Fighting with the Disease
Alzheimer's Disease
PMC5508523
NRF2 deficiency replicates transcriptomic changes in Alzheimer's patients and worsens APP and TAU pathology
Neurodegenerative Disease
Alzheimer's Disease
Nfe2l2 (housemouse)
Nfe2l2 (Norway rat)
NFE2L2 (human)
PMC6977098
NRF2 as a Therapeutic Target in Neurodegenerative Diseases
Local Concept Level
Iron
Alzheimer Disease
Neurodegenerative Diseases
Alzheimer Disease
PCC
HIP
Common Variable Immunodeficiency
Neurodegenerative Diseases
Alzheimer Disease
BACE1
CD4
Psen1
PSEN1
CXCL8
DCN
Alzheimer Disease
Nfe2l2 house mouse
NFE2L2 human
Parkinson Disease
Stroke
Nfe2l2 Norway rat
NFE2L2 human
Alzheimer Disease
Nfe2l2 house mouse
Multiple Sclerosis
Domain Concept Level
Iron
Alzheimer Disease
Neurodegenerative Diseases
Alzheimer Disease
Iron
HIP
PCC
CD4
CXCL8
Neurodegenerative Diseases
HIP
PCC
Alzheimer Disease
Iron
DCN
Psen1
PSEN1
BACE1
Iron
Common Variable Immunodeficiency
BACE1
PSEN1
Psen1
DCN
CXCL8
CD4
HIP
PCC
Alzheimer Disease
Neurodegenerative Diseases
NFE2L2 human
Nfe2l2 house mouse
Iron
Common Variable Immunodeficiency
PCC
HIP
Parkinson Disease
Iron
Neurodegenerative Diseases
Multiple Sclerosis
Alzheimer Disease
Stroke
BACE1
PSEN1
Psen1
Common Variable Immunodeficiency
DCN
CD4
CXCL8

Fig. 3. Knowledge representation of one knowledge evolution chain of AD. Paper level shows the PMCID, title of each paper, and entities in the citation context. AD is the ego of the egocentric networks in local and domain concept level, presented as a red rectangle node. Blue, green, and red nodes denote chemicals, genes, and diseases, respectively. In domain concept level, nodes appearing in the previous networks are grayed out in the current network (Color figure online)

mouse (Mus musculus). New genes, Nfe2l2 (house mouse) and NFE2L2 (human), are added to DCL after the addition of PMC5508523 to the chain. Finally, PMC6977098 is a review paper that summarized the neurodegenerative diseases related to NRF2 gene, including AD, Parkinson's disease, multiple sclerosis, and stroke. The review article connects several diseases due to their common relationship with NRF2 genes.

Overall, this chain focuses on the mechanisms of AD and its pathogenesis, with a common theme on oxidative stress and AD. It can be seen from Fig. 3 that the knowledge networks in LCL reveal the prominent knowledge structure within a paper and DCL depicts the accumulation of knowledge. The combined information from three layers not only reveals the direction of knowledge flow and evolution, but also details the content of knowledge creation, usage, and evolution. It should be noted that this is only one of the knowledge evolution chains extracted from the AD related literature in PMC-OA. The new entities and relationships in this chain may have been covered by or imported from other chains in the citation network. The representation of the model also allows large scale automated analysis on multiple chains in addition to the demonstration of a case here.

7 Discussion and Future Directions

The purpose of this study is to propose a knowledge representation model that facilitates studies on knowledge creation, usage, and evolution. The three-layer network structure of the model includes: 1) a Paper Level (PL) for the citation relationship, as well as semantic relationship, among papers; 2) a Local Concept Level (LCL) for the internal concept structure within papers; 3) a Domain Concept Level (DCL) for accumulative knowledge in a domain. The PL in our model is a citation network with added citation contexts. It is comparable with what most existing studies use for main path analysis to reveal knowledge development [29]. The uniqueness of the proposed model is to integrate citation relationship and content in a multi-layer network model. The resulting model not only allows to trace the knowledge flow between cited and citing papers, but is able to specify how the knowledge content evolves as new papers are added to a citation network. This becomes clear in the case demonstration where the PL shows the direction and path of the knowledge diffusion, the LCL elaborates the content in each paper, and the DCL documents the contribution of each paper to the accumulated knowledge in a chain (in this case, the domain is defined narrowly to a chain). In addition, the principled network model allows large scale automated analysis on knowledge creation, usage, and evolution. The current study applies it on a data set in PMC-OA and shows a case of one chain. It demonstrates the feasibility of the model and its possible use.

The model has implications for science of science studies [41] in that it is focused on the knowledge production process, and the model accommodates for large scale analysis. In its current form, the model can be used to study formally documented flow of ideas and idea interactions. This can help discover where novelty arises and what contributes to the process. When combined with additional information on authors and institutions, the model can help study the interaction between knowledge representations and social structures.

7.1 Future Directions

In the process of carrying out this study, we have also recognized some limitations that point out the future directions of the work: First, while Pubtator Central is the state-of-the-art system to extract biomedical entities from PubMed articles, we still identified several errors in our experiment. For example, a general word may be recognized as a gene entity when it matches the abbreviation of a gene, such as tea and gene Slc7a2 (also known as Tea). Also, Pubtator Central only identifies biomedical entities that fall in the category of gene, disease, chemical, mutation, species, and cell line. Some areas of study, such as food and nutrition, are inadequately covered. Therefore, it is beneficial to improve the accuracy and coverage of entity recognition in our future studies. One way to do that is to integrate external knowledge systems, e.g. PubMed knowledge graph [42], and develop in-house machine learning models. Second, we only calculated preliminary network metrics on the knowledge evolution chains and chose one of them as a case study. Future work will further investigate how to characterize knowledge creation, usage, and evolution using quantitative measures to reveal patterns and regularities in the knowledge production process. Third, while PMC-OA is an ideal dataset for full-text mining, it is obvious that many citations are excluded since they point to articles outside PMC-OA. The knowledge evolution pathways may be skewed due to the problem of incomplete data. Over 90% of the referenced articles in PMC-OA have PMIDs, which means that their abstracts and metadata are available from PubMed. However, how to align these abstract-only articles with full-text articles and trim the current procedure to identify conceptual citations from them remains a challenge.

Despite the limitations, the proposed multi-layer knowledge representation model serves as a powerful infrastructure for various applications. This model also has an extensible architecture that allows for different approaches to construct the inner networks. For instance, we are working on a solution to refine the internal concept structure of individual papers, which focuses on the main topics of the document and filters background and literature review that casts broad connections. This will result in more robust LCL networks. Future studies could also seek an optimal network structure for their applications, ranging from pathway analysis to network simulation.

Acknowledgements. This study was partially funded by the National Natural Science Foundation of China (NSFC) Grant Nos. 71804135 and 71921002.

References

1. Landhuis, E.: Scientific literature: information overload. Nature **535**(7612), 457–458 (2016)
2. de Solla Price, D.J.: Networks of Scientific Papers. Science **149**(3683), 510–515 (1965)
3. Garfield, E., Pudovkin, A.I., Istomin, V.S.: Why do we need algorithmic historiography? J. Am. Soc. Inf. Sci. Technol. **54**(5), 400–412 (2003)
4. van Eck, N.J., Waltman, L.: CitNetExplorer: a new software tool for analyzing and visualizing citation networks. J. Informetr. **8**(4), 802–823 (2014)
5. Kuhn, T., Perc, M., Helbing, D.: Inheritance patterns in citation networks reveal scientific memes. Phys. Rev. X **4**(4), 41036 (2014)
6. Liang, Z., Mao, J., Cao, Y., Li, G.: Idea diffusion patterns: SNA on knowledge meme cascade network. In: Proceedings of ISSI 2019, Rome, Italy, pp. 2612–2613 (2019)
7. Mao, J., Liang, Z., Cao, Y., Li, G.: Quantifying cross-disciplinary knowledge flow from the perspective of content: introducing an approach based on knowledge memes. J. Informetr. **14**(4), 101092 (2020)
8. Bates, M.J.: Defining the information disciplines in encyclopedia development. Inf. Res. **12**(4), 29 (2007)
9. Weller, K.: Knowledge Representation in the Social Semantic Web. Walter de Gruyter, Germany (2010)
10. Hjørland, B.: Knowledge organization (KO). KO Knowl. Organ. **43**(6), 475–484 (2016)
11. Weller, K.: Folksonomies and ontologies: two new players in indexing and knowledge representation. In: Proceedings of the Online Information Conference, London, Great Britain, pp. 108–115 (2007)
12. Leydesdorff, L., Rafols, I.: A global map of science based on the ISI subject categories. J. Am. Soc. Inf. Sci. Technol. **60**(2), 348–362 (2009)
13. Callon, M., Courtial, J.-P., Turner, W.A., Bauin, S.: From translations to problematic networks: an introduction to co-word analysis. Soc. Sci. Inf. **22**(2), 191–235 (1983)
14. Lu, K., Wolfram, D.: Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches. J. Am. Soc. Inf. Sci. Technol. **63**(10), 1973–1986 (2012)
15. White, H.D., McCain, K.W.: Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. J. Am. Soc. Inf. Sci. **49**(4), 327–355 (1998)
16. Kessler, M.M.: Bibliographic coupling between scientific papers. Am. Doc. **14**(1), 10–25 (1963)
17. Perianes-Rodríguez, A., Olmeda-Gómez, C., Moya-Anegón, F.: Detecting, identifying and visualizing research groups in co-authorship networks. Scientometrics **82**(2), 307–319 (2010)

18. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the fuzzy sets theory field. J. Informetr. **5**(1), 146–166 (2011)
19. Markman, A.B.: Knowledge Representation. Psychology Press, Mahwah (1999)
20. van Harmelen, F., Lifschitz, V., Porter, B.: Handbook of Knowledge Representation. Elsevier, Amsterdam (2008)
21. Cronin, B.: The Citation Process: The Role and Significance of Citations in Scientific Communication. Taylor Graham, London (1984)
22. Lutz, B., Hans-Dieter, D.: What do citation counts measure? A review of studies on citing behavior. J. Doc. **64**(1), 45–80 (2008)
23. White, H.D.: Reward, persuasion, and the Sokal Hoax: a study in citation identities. Scientometrics **60**(1), 93–120 (2004)
24. Frandsen, T.F., Nicolaisen, J.: Citation behavior: a large-scale test of the persuasion by name-dropping hypothesis. J. Assoc. Inf. Sci. Technol. **68**(5), 1278–1284 (2017)
25. Sugimoto, C.R.: Theories of Informetrics and Scholarly Communication. Walter de Gruyter, Germany (2016)
26. Garfield, E., Sher, I.H., Torpie, R.J.: The use of citation data in writing the history of science. Institute for Scientific Information, Philadelphia (1964)
27. Hummon, N.P., Dereian, P.: Connectivity in a citation network: the development of DNA theory. Soc. Network. **11**(1), 39–63 (1989)
28. Lucio-Arias, D., Leydesdorff, L.: Main-path analysis and path-dependent transitions in HistCite$^{\mathrm{TM}}$-based historiograms. J. Am. Soc. Inf. Sci. Technol. **59**(12), 1948–1962 (2008)
29. Liu, J.S., Lu, L.Y.Y., Ho, M.H.-C.: A few notes on main path analysis. Scientometrics **119**(1), 379–391 (2019)
30. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zhai, C.: Content-based citation analysis: the next generation of citation analysis. J. Assoc. Inf. Sci. Technol. **65**(9), 1820–1833 (2014)
31. Chen, C., Li, Q., Chiu, K., Deng, Z.: The impact of Chinese library and information science on outside disciplines: a citation analysis. J. Librariansh. Inf. Sci. **52**(2), 493–508 (2019)
32. Wang, F., et al.: Exploring all-author tripartite citation networks: a case study of gene editing. J. Informetr. **13**(3), 856–873 (2019)
33. Walter, C., Ribière, V.: A citation and co-citation analysis of 10 years of KM theory and practices. Knowl. Manag. Res. Pract. **11**(3), 221–229 (2013)
34. Shiau, W.-L., Dwivedi, Y.K.: Citation and co-citation analysis to identify core and emerging knowledge in electronic commerce research. Scientometrics **94**(3), 1317–1337 (2013)
35. Boccaletti, S., et al.: The structure and dynamics of multilayer networks. Phys. Rep. **544**(1), 1–122 (2014)
36. Scott, J., Carrington, P.J.: The SAGE Handbook of Social Network Analysis. SAGE, Thousand Oaks (2011)
37. Criado, R., Romance, M., Vela-Pérez, M.: Hyperstructures, a new approach to complex systems. Int. J. Bifurc. Chaos. **20**(03), 877–883 (2010)
38. Kim, H.J., Jeong, Y.K., Song, M.: Content- and proximity-based author co-citation analysis using citation sentences. J. Informetr. **10**(4), 954–966 (2016)
39. Colavizza, G., Boyack, K.W., van Eck, N.J., Waltman, L.: The closer the better: similarity of publication pairs at different cocitation levels. J. Assoc. Inf. Sci. Technol. **69**(4), 600–609 (2018)
40. Wei, C.-H., Allot, A., Leaman, R., Lu, Z.: PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. **47**(W1), W587–W593 (2019)
41. Fortunato, S., et al.: Science of science. Science **359**(6379), eaao0185 (2018)
42. Xu, J., et al.: Building a PubMed knowledge graph. Sci. Data. **7**, 1–19 (2020)