Spectral Clustering-Based Knowledge Aggregation Methods for Virtual Health Communities: A Postprint

Authors: Zhang Haitao, Song Tuo, Zhou Honglei, Zhang Xinrui

Date: 2023-04-01T16:15:53+00:00

Abstract

[Purpose/Significance] To improve the quality of knowledge aggregation in virtual health communities and provide technical methodological support for virtual health community services. [Method/Process] Spectral clustering methods were employed to extract knowledge from virtual health communities, concept similarity calculation was utilized to obtain a knowledge topic similarity matrix, and spectral clustering was performed based on this similarity matrix. [Results/Conclusion] Information published on the Haodf.com online health consultation platform was used as the data source for method validation. The results demonstrate that when the number of clusters is 5, the proposed method achieves the highest score. By fully exploiting the latent information in virtual health communities through spectral clustering methods, the quality of knowledge aggregation is improved, providing a new approach for knowledge aggregation and knowledge services.

Full Text

Preamble

Vol. 64 No. 8, April 2020, ChinaXiv Cooperative Journal

Research on Knowledge Aggregation Method for Virtual Health Communities Based on Spectral Clustering

Zhang Haitao[1],[2], Song Tuo[1], Zhou Honglei[1], Zhang Xinrui[1]

[1]School of Management, Jilin University, Changchun 130022
[2]Jilin University Information Resource Research Center, Changchun 130022

Abstract: [Purpose/Significance] To improve the quality of knowledge aggregation in virtual health communities and provide technical methodological support

for virtual health community services. [Method/Process] The spectral clustering method was applied to extract knowledge from virtual health communities. Concept similarity calculation was used to obtain a knowledge topic similarity matrix, and spectral clustering was performed based on this similarity matrix. [Result/Conclusion] Information published on the Haodf.com online health consultation platform was used as a data source for method validation. Results show that when the number of clusters is 5, the proposed method achieves the highest score. This spectral clustering approach fully mines potential information in virtual health communities, improves knowledge aggregation quality, and provides a new pathway for knowledge aggregation and knowledge services.

Keywords: virtual health community; knowledge aggregation; spectral clustering; similarity

Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.08.015

The report of the 19th National Congress of the Communist Party of China points out that the principal contradiction in Chinese society has evolved into the contradiction between unbalanced and inadequate development and the people's ever-growing needs for a better life. The people's aspiration for a healthy life is the fundamental driving force behind the development of the health and medical industry [1]. With the continuous growth of residents' disposable income, people have put forward higher requirements for disease diagnosis and treatment, and have new expectations for maintaining health. Fields such as telemedicine, health management, and scientific elderly care have driven continuous innovation in the health and medical industry, while also giving rise to many new phenomena such as Internet healthcare, which have attracted widespread attention from all sectors of society.

With increasing public concern about health issues and the emergence of online communities based on user-generated content, more and more users are utilizing virtual health communities to exchange health information and opinions. Virtual health communities aim to provide better treatment by reducing medical costs, making full use of existing resources, and offering patients more diverse communication channels [2]. Virtual health communities have become the mainstream platform for users to exchange health experiences.

However, current research on knowledge aggregation and services in virtual health communities is limited. Revealing and mining the knowledge embedded in virtual health community posts to achieve user-demand-oriented knowledge aggregation, innovate knowledge service models for virtual health communities, and improve knowledge service capabilities and quality has become the primary problem troubling knowledge service provision in virtual health communities.

This paper addresses the characteristics of virtual health communities, combines relevant domestic and international research findings, adopts the spectral clustering method to aggregate knowledge content in virtual health communities, and validates the proposed method through empirical sample data. Virtual

health communities are developing rapidly and hold great potential. This research provides a new perspective for knowledge aggregation in virtual health communities, which is significant for identifying problems and deficiencies in virtual health communities, improving user satisfaction with virtual health services, and promoting the sustainable development of virtual health communities.

2. Related Concepts and Research

2.1 Knowledge Aggregation

The concept of "aggregation" originates from chemistry, where it refers to the process of assembling dispersed monomer small molecular structures into large molecular structures through linking relationships [3]. Scholars in the library and information science field have also conducted in-depth research on this concept, with research objects gradually transitioning from data to the knowledge domain. Li Yating defined knowledge aggregation from the perspective of the knowledge service process, stating that during knowledge service, the condensation of unordered and dispersed knowledge can reveal associations between knowledge units and form an organic knowledge system [4]. Wang Jingdong considers knowledge aggregation a process of knowledge cluster analysis, after which knowledge becomes more enriched in connotation, making the decision-making process more meaningful [5]. Guan Jun, Bi Qiang, et al. believe that knowledge aggregation aims to construct multi-dimensional yet interrelated knowledge systems, where knowledge units and their intrinsic relationships can be extracted through methods such as data mining and artificial intelligence [6]. Li Jie believes that the knowledge aggregation implementation process includes knowledge gathering and knowledge integration, where massive information resources can be screened and mined through association and clustering to achieve intelligent knowledge fusion [7].

From the above definitions, it can be seen that knowledge aggregation is a process that employs artificial intelligence methods and techniques such as data mining and semantic technology to analyze knowledge characteristics, reorganize and screen unordered and dispersed knowledge, further discover associations between knowledge, and form organic knowledge systems, thereby providing users with targeted, complete, and systematic services that enable efficient knowledge utilization. Different aggregation methods can be used for different knowledge forms. Currently, mainstream knowledge aggregation methods include semantic-enhanced knowledge aggregation methods, multi-dimensional knowledge aggregation methods, and clustering-based knowledge aggregation methods [8]. Semantic-enhanced aggregation methods can solve the problem of semantic loss during knowledge aggregation, generally through concept association or semantic tags. Multi-dimensional knowledge aggregation methods utilize multi-dimensional division of "user-resource-tag," with co-occurrence analysis being a commonly used multi-dimensional knowledge aggregation method. Clustering-based knowledge aggregation methods associate and aggregate knowledge according to the degree of knowledge relevance, such as text clustering and

tag clustering, which are commonly used aggregation methods.

2.2 Virtual Health Community Knowledge Aggregation

Network community knowledge aggregation has its developmental origins, presenting a logical sequence over time that gradually deepens in aggregation level (from information aggregation to knowledge aggregation) and extends from special to general aggregation scenarios (from collection resource knowledge aggregation to academic community knowledge aggregation, and then to general network community knowledge aggregation). The deepening of research levels and extension of research scenarios make knowledge aggregation for network communities both necessary and well-founded [8].

Zhang Lianfeng et al. established an integrated model framework for virtual academic community knowledge aggregation that incorporates both topics and the SECI model, based on analysis of relevant knowledge needs of academic community users [9]. Hu Yuan et al. designed a digital library community service push system based on knowledge aggregation, grounded in user communication behaviors and needs within communities [10]. Shang Xianli et al. designed a method for aggregating knowledge resources in academic blogs based on tag co-occurrence [11]. K. Liang et al. analyzed the characteristics of fragmented learning behavior and re-aggregated knowledge in online education according to learners' individual learning needs, thereby guiding learners to make full use of fragmented time to obtain accurate and meaningful knowledge content [12]. V. Tarko et al. introduced a knowledge aggregation and integration method based on processes and designed an aggregation system dependent on meta-experts and computer algorithms based on aggregation mechanisms, providing tools for knowledge aggregation and discussing the possibility of building a "virtual think tank" [13]. M. Ritou et al. proposed a knowledge-based multi-level aggregation strategy to support decision-making, intelligently generating meaningful data through knowledge aggregation methods, and verified the usefulness of the strategy using data from manufacturing processes in the aviation industry to aid manufacturing decisions [14]. J. Oosterman et al. studied how different knowledge extraction and aggregation configurations affect the identification of artwork annotations, using crowdsourcing methods to automatically aggregate local annotations of artworks to facilitate access and retrieval [15].

Virtual health communities contain a large number of knowledge units, with potential connections and influences existing among various knowledge units. Revealing and discovering the associated knowledge in user-generated answers is fundamental to the effective organization, management, and knowledge discovery of answer knowledge.

2.3 Spectral Clustering

Clustering is an unsupervised learning method and an effective means of discovering and exploring intrinsic connections among things, widely applied in

various fields. Clustering does not require prior knowledge; through cluster analysis, similar objects can be divided into clusters such that objects within clusters are as similar as possible while objects between clusters are as different as possible. Clustering can differentiate different types of knowledge, and after dividing knowledge into clusters, users can obtain more meaningful knowledge through cluster analysis.

Co-occurrence analysis is a commonly used multi-dimensional knowledge aggregation method. Clustering methods such as the PF algorithm, SM algorithm, and NJW algorithm are classical clustering methods that can effectively partition spherical clusters, but they are not suitable for non-convex shaped clusters and tend to fall into local optimal solutions. Spectral clustering, as a graph theory-based clustering method, can effectively discover clusters of arbitrary shapes and converge to global optimal solutions. The spectral clustering algorithm treats each data point as a vertex in a graph, uses similarity as the weight connecting vertices, calculates the adjacency matrix and similarity matrix of vertices, transforms it into a Laplacian matrix, and then obtains eigenvalues and their corresponding eigenvectors to achieve data dimensionality reduction and partitioning. R. Janani et al. combined spectral clustering with swarm optimization to process massive text files, validated it through standard datasets, and compared it with spherical k-means, expectation maximization, and standard particle swarm algorithms, finding that this algorithm has better clustering accuracy than other clustering algorithms [16]. X. Li improved spectral clustering using eigenvalue differences and orthogonal eigenvectors, achieving automatic determination of the number of clusters. This algorithm was used to cluster users and items in a two-dimensional rating matrix, and the clustered rating matrix was decomposed to obtain a shared rating matrix. Simulation results showed that compared with eight other traditional collaborative filtering methods, this method can effectively improve recommendation accuracy and generalization ability [17].

The posts published in virtual health communities contain a large amount of medical knowledge, and internal connections exist among the various posts. Current virtual community aggregation methods adopt traditional clustering approaches that result in aggregated knowledge lacking semantics, or they require constructing ontologies, which consumes substantial effort. The spectral clustering method, however, can cluster posts based on knowledge similarity, thereby enhancing semantic associations among knowledge. Clustering posts in virtual health communities can effectively establish relationships between posts, perform cluster analysis on prior knowledge, and discover knowledge contained in various documents, making aggregation results more enriched. Aggregating knowledge in virtual health communities aims to meet users' knowledge needs. Using relevant computer methods to mine and extract knowledge units in discrete distribution states in answers and their relationships achieves close connections and orderly organization of associated knowledge units in communities. This approach can provide virtual health community users with knowledge recommendation and discovery services that meet personalized needs, further

improving service quality and user experience.

3. Calculation Method and Process

Knowledge acquisition is the prerequisite for knowledge aggregation, and concepts are the core units of knowledge. During virtual community knowledge aggregation, text data needs to be preprocessed to achieve mathematical representation of knowledge. In the knowledge aggregation process, knowledge hidden in documents should be fully mined to discover unique associations among knowledge. The spectral clustering method can partition text content and discover relationships between texts and knowledge contained in document content. This paper extracts text feature words as concepts, calculates concept similarity, uses an improved semantic similarity matrix to replace the spatial vector model, constructs a text similarity matrix through conceptual semantic similarity, uses it as the input matrix for spectral clustering, and employs spectral clustering as the knowledge aggregation method to reduce matrix dimensionality and improve clustering result accuracy. The virtual health community knowledge aggregation method model constructed in this paper is shown in Figure 1 [Figure 1: see original paper].

As shown in Figure 1, the virtual health community knowledge aggregation method model consists of three layers: data layer, calculation layer, and application layer. The data layer crawls topic posts from virtual health communities through a crawler program and saves them in text form in a database. Word segmentation software is used to segment content and calculate word frequency. By screening, feature keywords that can represent post content—i.e., concepts—are obtained, completing the mathematical representation of knowledge. In the calculation layer, similarities between various concepts are calculated, and then knowledge topic similarity is obtained. This serves as the similarity matrix for calculating the Laplacian matrix and performing spectral clustering. Through spectral clustering, semantic associations hidden in knowledge resources can be effectively discovered. In the application layer, the spectral clustering method based on knowledge topic similarity is used to aggregate knowledge in texts, thereby discovering knowledge within various documents.

In virtual health community knowledge aggregation methods, concepts from virtual health community posts are used for measurement. In Formula (1), K1 K2 represents the number of times concepts co-occur in posts, K1-K2 represents the number of times K1 appears but K2 does not, and K2-K1 represents the number of times K2 appears but K1 does not. For simplicity, coefficients α and β are both set to 0.5.

3.1 Concept Similarity Calculation

A concept refers to a word or phrase extracted from a document that is specific and can reflect the document's theme. Such words can not only embody the core and thematic knowledge of the document but also cover its content,

facilitating user indexing and searching. Through concept extraction, users can more clearly and directly understand the content and overall overview of textual knowledge. Therefore, this paper extracts keywords as concepts contained in virtual health community posts for representation and calculation. Concepts are the core units of knowledge. During the virtual community knowledge aggregation process, text data needs to be preprocessed to achieve mathematical representation of knowledge. In knowledge aggregation, knowledge hidden in documents should be fully mined to discover unique associations among knowledge. The spectral clustering method can partition text content and discover relationships between texts and knowledge contained in document content. This paper extracts text feature words as concepts, calculates concept similarity, uses an improved semantic similarity matrix to replace the spatial vector model, constructs a text similarity matrix through conceptual semantic similarity, uses it as the input matrix for spectral clustering, and employs spectral clustering as the knowledge aggregation method to reduce matrix dimensionality and improve clustering result accuracy.

In virtual health community knowledge aggregation methods, concepts from virtual health community posts are used for measurement. In Formula (1), K1 K2 represents the number of times concepts co-occur in posts, K1-K2 represents the number of times K1 appears but K2 does not, and K2-K1 represents the number of times K2 appears but K1 does not. For simplicity, coefficients α and β are both set to 0.5.

3.2 Knowledge Topic Similarity Calculation

Knowledge topic similarity refers to the degree of similarity in theme or content between texts, generally calculated by extracting feature keywords or concepts from texts [19]. Similar to the concept of text similarity, when calculating knowledge topic similarity between two posts, concepts or feature keywords can be extracted from posts to represent posts as collections of concepts, i.e., in vector form [20], and then described through contained concepts to facilitate similarity calculation.

When calculating knowledge topic similarity, the adjacency matrix between knowledge topics needs to be calculated. Typically, points with greater distance have lower weight values, while points with closer distance have higher weights. This allows weights to be used as similarity measures between points. Methods include full connection and nearest neighbor connection. This paper calculates similarity between virtual health community posts to construct a similarity matrix. This similarity matrix is a symmetric matrix. This further yields the degree matrix. Through calculation, the Laplacian matrix L can be obtained. The k eigenvalues of L are obtained, and eigenvectors V are constructed by sorting according to eigenvalue magnitude. Each row of V is treated as new data, allowing clustering methods to be used for partitioning to obtain clustering results C(C1, C2, ..., Cn). Spectral clustering only requires a similarity matrix between data, which actually performs dimensionality reduction on the

matrix, facilitating the processing of sparse data.

The formula for calculating the distance between posts is as follows:

Dist(dx, dy) = Dist(Λ_i=1^n Kxi, Λ_j=1^m Kyj) = Σ_i=1^n Σ_j=1^m fi × fj × (1 + Sim(Ki, Kj))

Where dx and dy are two different posts, xi and yj are concepts contained in posts dx and dy respectively; fi and fj represent the occurrence frequencies of concepts xi and yj in posts dx and dy. n and m are the numbers of concepts contained in the two posts respectively. d is the sum of the numbers of concepts appearing in both posts. The purpose of using d here is to consider that if a concept appears too many times in a post, it leads to excessive semantic distance between posts, so d is used to normalize this distance.

Formula (3) sim(dx, dy) = 1 / (1 + Dist(dx, dy))

From formula (3), it can be seen that the larger the semantic distance, the smaller the similarity between knowledge topics.

3.3 Knowledge Aggregation Algorithm Based on Spectral Clustering

This paper proposes a virtual health community knowledge aggregation method based on the similarity matrix spectral clustering algorithm. This method extracts keywords from virtual health community posts as concepts of the virtual health community. Knowledge topics are represented by concept lists, and similarity between two knowledge topics can be transformed into solving similarity between concepts. By calculating similarity between concepts, similarity between two posts in the virtual health community is obtained. This serves as the similarity matrix for calculating the Laplacian matrix and performing spectral clustering. Through spectral clustering, semantic associations between virtual health community posts can be effectively discovered.

Virtual health community knowledge aggregation algorithm description:
Input: n data points, number of clusters K
Output: Clustering results C(C1, C2, …, Cn)

Begin
Construct similarity matrix W R^(n×n); construct degree matrix D R^(n×n);
Transform Laplacian matrix L = D - W;
Obtain the first k eigenvalues of L and their corresponding eigenvectors, sort k by eigenvalue magnitude; and construct eigenvector V;
Treat V as a vector in k-dimensional space, where y_{ij} = v_{ij} / (j v{ij}^2)(1/2), use clustering methods for clustering.

4. Experimental Process and Results

Haodf.com is a well-known Internet medical platform trusted by patients. While ensuring the provision of standardized high-quality medical services, it integrates Internet thinking and technology, exploring a medical service model

that "synchronizes online consultation with Q&A, combines online referral with follow-up visits, and matches expert outpatient appointments with private doctor contracts." This not only facilitates communication between doctors and patients but also effectively expands its influence and authority, providing a new direction for alleviating current tense doctor-patient relationships. Through experience exchange and sharing on the platform, users have gradually formed an influential medical academic forum, laying a solid foundation for further improving medical service quality and strengthening the integration of online and offline services.

Therefore, this paper uses Haodf.com data for algorithm validation, crawling a total of 800 articles under common cardiovascular disease tags through Python programming. Data preprocessing is the process of simplifying data, including removing stop words, word segmentation, noise reduction, etc., to extract data content in the required format. Python's Jieba function is used for word segmentation and word frequency statistics, similarity values are calculated, and clustering is finally performed.

4.1 Concept Extraction and Similarity Calculation

Through natural language processing, knowledge can be extracted from text data. Since this knowledge often has specific structures and patterns, it can be used as concepts for calculation [21-22]. In the knowledge aggregation process, concepts provide the finest-grained knowledge units [23]. In health virtual communities, user communication content often revolves around specific domain problems, and concepts extracted from content can often represent domain knowledge. Aggregating these concepts can yield similar knowledge. Before knowledge aggregation, extracted concepts need to be organized, such as using domain relevance and consistency calculation formulas to eliminate irrelevant or meaningless concepts [24]. The domain relevance calculation formula is as follows:

DR(ti, D) = P(ti|D) / Σ_i=1^n P(ti|D) where P(ti|D) = freq / Σ_i=1^n freq_i, freq is the frequency of candidate concept occurrence. The domain consistency calculation formula is as follows:

DC(ti, D) = Σ_i=1^n P(ti|D) × log P(ti|D)

The concept extraction formula is as follows:

TWi = α × DR(ti, D) + β × DC(ti, D)

For calculation simplicity, α and β are set to 0.5, allowing organization of extracted concepts and obtaining relevant concepts.

Haodf.com includes columns such as disease introduction, etiology and symptoms, prevention and examination, disease diagnosis and treatment, medical guide, and nursing care. This paper selects content from all tags for similarity calculation, using Python language programming for preprocessing and calcu-

lation. The calculation results are shown in Figure 2 [Figure 2: see original paper].

4.2 Knowledge Topic Clustering Experiments and Visualization Results

How to create a similarity matrix that more truly reflects the approximate relationships between data points, making similarity higher between nearby points and lower between distant points, is a problem that spectral clustering algorithms must solve. The Gaussian similarity function is a common method for calculating similarity between two points in classical spectral clustering algorithms. When using the Gaussian kernel function, the similarity matrix and adjacency matrix are identical. When implementing spectral clustering algorithms in Python, the Gaussian kernel function can also be selected. Generally, parameters n_{clusters} and gamma in the Gaussian kernel function need to be tuned to select appropriate parameter values. In this method, four cases are considered where the number of clusters n_{clusters} = 3, 4, 5, 6, and gamma selects four values: 0.01, 0.1, 1, 10. The specific calculated scores are shown in Figure 3 [Figure 3: see original paper].

For different clustering results, the highest score is 234.67, achieved when n_{clusters} is 5 and gamma is 1 or 0.1.

This paper's knowledge aggregation results are shown in Figure 4 [Figure 4: see original paper].

As shown in Figure 4, documents are numbered. To better distinguish documents, the format "letter + number" is adopted. After document clustering division, brackets are used for distinction. During word segmentation, it is often necessary to segment according to the user's domain and include domain terminology. Jieba segmentation often suffers from over-segmentation, and stop words can be used to improve search efficiency in information retrieval and save storage space. Stop words are generally manually input, targeting domain-specific terminology, ultimately constructing a stop word list. This paper uses the library function jieba.load_{userdict}(file_{name}) in Python to load the stop word list.

Concept extraction is the foundation of knowledge aggregation. After obtaining concepts, they are processed with extracted concepts as objects and attribute relationship-based similarity as the basis for knowledge aggregation calculation to achieve knowledge aggregation in virtual health communities. To better display aggregation effects after aggregation, this paper displays virtual health community knowledge aggregation results through word clouds, as shown in Figure 5 [Figure 5: see original paper].

According to the knowledge aggregation results related to "cardiovascular disease" in Figure 5, dividing it into 5 categories is more reasonable. Related knowledge topics include valvular disease, hypertension, arrhythmia, congenital

heart disease, and angina, through which related knowledge can be further discovered. For example, in the hypertension knowledge topic tag, hypertension is related to vascular wall pressure; from the arrhythmia knowledge topic tag, words related to arrhythmia include frequency, rhythm, origin, abnormality, etc., and arrhythmia also includes phenomena such as sinus, escape, and ectopic; coronary heart disease angina refers to insufficient coronary blood supply caused by coronary atherosclerotic stenosis; etc.

In the Haodf.com website, subject navigation is also used to classify diseases. However, these classifications are relatively scattered and cannot effectively focus on topics. When users browse related diseases, they spend considerable energy searching for diseases relevant to their needs, causing users to get lost in massive knowledge. This paper's knowledge aggregation results are classified by knowledge topic and displayed through concept word frequency, which can serve as a basis for selecting related topics. For example, symptoms that may appear in coronary heart disease angina include cardiac arrest, spasm, or syncope, and causes may include obesity, myocardial ischemia, etc. Additionally, overeating and hypoxia may also cause such diseases. Users can select topics they want to understand based on appearing concepts. Moreover, in these aggregation results, concepts related to a certain domain have been completely displayed, allowing users to obtain more choices according to their needs. If users have indicated interest in a certain topic, more topics can be recommended to them through aggregation results. For example, other domain concepts related to heart rhythm could be arrhythmia or cardiac arrhythmia. If users follow the domain concept of heart rhythm, related domain concepts or topics can be recommended to them based on aggregation results, thereby providing more targeted services.

Through the above research, it can be found that the virtual health community knowledge aggregation method based on spectral clustering has certain feasibility and effectiveness, can help users understand relevant knowledge and topics in the topic, and users can quickly find relevant knowledge content through topic clusters. Through knowledge aggregation methods, virtual health communities can be helped to further improve knowledge retrieval, knowledge discovery, knowledge navigation, and other services, and can also implement knowledge recommendation, knowledge graph, and other functions based on this method.

5. Conclusion

This paper proposes a knowledge aggregation method for virtual health communities based on spectral clustering. First, content from the Haodf.com virtual health community website is crawled, and the text is preprocessed through the Jieba word segmentation software to extract concepts that can represent virtual health community knowledge. The similarity calculation formula is used to calculate concept similarity, based on which the similarity matrix of virtual health community post knowledge topics is constructed. The similarity is normalized, and the Laplacian matrix is calculated. The first k eigenvalues are obtained,

and eigenvector V is constructed by sorting according to k magnitude. Each row of V is treated as new data, allowing clustering methods to be used for partitioning to obtain knowledge aggregation results. Using spectral clustering to aggregate knowledge in virtual health communities can help virtual health community users quickly understand relevant knowledge topics and content, and can provide targeted knowledge services for virtual health community users, thereby helping virtual health communities effectively improve user experience and service quality.

References

[1] Xinhua News Agency. Behind the transformation of China's principal social contradiction [EB/OL]. [2019-06-26]. http://cpc.people.com.cn/19th/n1/2017/1021/c414305-29600806.html.
[2] HAJLI M N. Developing online health communities through digital media [J]. International journal of information management, 2014, 34(2): 311-314.
[3] Bi Qiang. Digital resources: transformation from integration to aggregation [J]. Digital Library Forum, 2014(6): 1.
[4] Li Yating. Review of knowledge aggregation research [J]. Library and Information Service, 2016, 60(21): 128-136.
[5] Wang Jingdong. Digital library information intelligent retrieval model based on knowledge aggregation [J]. Library Science Research, 2014(21): 72-76, 71.
[6] Guan Jun, Bi Qiang, Zhao Yiping. Research progress on knowledge aggregation and discovery based on linked data [J]. Information and Documentation Services, 2015(3): 15-21.
[7] Li Jie. Research on visualization of knowledge aggregation of library digital resources based on SNA [D]. Changchun: Jilin University, 2016.
[8] Chen Guo, Zhu Qianling, Xiao Lu. Knowledge aggregation for network communities: development, research foundation, and prospects [J]. Journal of Intelligence, 2017, 36(12): 193-197, 192.
[9] Zhang Lianfeng, Li Hui, ! Yunhe. Research on construction of knowledge aggregation model for virtual academic communities [J]. Information Science, 2019, 37(6): 55-60, 74.
[10] Hu Yuan, Diao Shouqi, Zhu Yiping, et al. Design and implementation of digital library community service push system based on knowledge aggregation [J]. Information Science, 2017, 35(11): 72-77.
[11] Shang Xianli, Wang Xuedong, Zhang Yuxuan. Research on knowledge resource aggregation in academic blogs based on tag co-occurrence [J]. Information Science, 2016, 34(5): 125-129.
[12] LIANG K, WANG C, ZHANY Y. Knowledge aggregation and intelligent guidance for fragmented learning [J]. Procedia computer science, 2018, 131(4): 656-664.
[13] TARKO V, DRAGOS ALIGICA P. From "Broad Studies" to Internet-based "Expert Knowledge Aggregation". Notes on the methodology and technology of knowledge integration [J]. Futures, 2011, 43(9): 986-995.
[14] RITOU M, BELKADI F, YAHOUI Z, et al. Knowledge-based multi-level

aggregation for decision aid in the machining industry [J]. CIRP annals, 2019, 68(1): 475-478.
[15] OOSTERMAN A J, YANG J, ALESSANDRO B, et al. On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks [J]. Computer networks, 2015, 90(29): 133-149.
[16] JANANI J, VIJAYARANI S. Text document clustering using spectral clustering algorithm with particle swarm optimization [J]. Expert systems with applications, 2019, 22(2): 633-647.
[17] LI X, WANG Z J, HU R L, et al. Recommendation algorithm based on improved spectral clustering and transfer learning [J]. Expert systems with applications, 2019, 134(15): 192-200.
[18] Li Hang. Statistical Learning Methods [M]. 2nd ed. Beijing: Tsinghua University Press, 2019.
[19] Wang Chunliu, Yang Yonghui, Deng Fei, et al. Review of text similarity calculation methods [J]. Information Science, 2019, 37(3): 158-167.
[20] Li Fenglin, Ke Jia. Text representation method based on deep learning [J]. Information Science, 2019, 37(1): 156-164.
[21] LIU K H, HOGAN W R, REBECCA S. Crowley. Natural language processing methods and systems for biomedical ontology learning [J]. Journal of biomedical informatics 2011, 44(1): 163-179.
[22] ANDRÉS PAREDES-VALVERDE M, ÁNGEL RODRÍGUEZ-GARCÍA M, RUIZ-MARTÍNEZ A, et al. ONLI: an ontology-based system for querying DBpedia using natural language paradigm [J]. Expert systems with applications, 2015, 42(12): 5163-5176.
[23] FRANTZI K T, ANANIADOU S. The C-Value/NC-Value domain independent method for multi-word term extraction [J]. Journal of natural language processing, 2008, 6(3): 145-179.
[24] Liao Fuyan. Research on concept and relation acquisition methods in ontology construction [D]. Xi'an: Xi'an University of Architecture and Technology, 2011.

Author Contributions

Zhang Haitao: Research idea and method formulation, data analysis, paper revision;
Song Tuo: Data collection, analysis and processing, initial paper drafting;
Zhou Honglei: Data collection and organization;
Zhang Xinrui: Paper revision.

Abstract

[Purpose/Significance] To improve the quality of knowledge aggregation in virtual healthy communities and provide technical method support for virtual healthy community services. [Method/Process] The method of spectral clustering was applied to extract knowledge in the healthy virtual community, and the semantic similarity matrix of the text was obtained by using the keyword

co-occurrence. The spectral clustering was performed according to the text semantic similarity matrix, and the text was aggregated into text clusters. [Result/Conclusion] The information published by the doctor's online health consultation platform was used as a data source for method validation. The results show that when the number of clusters is 5, the proposed method has the highest score. This method of spectral clustering considers the semantic relationship between words, fully exploits the potential information of virtual healthy community, improves the quality of knowledge aggregation, and provides a new way for knowledge aggregation and knowledge service.

Keywords: virtual healthy community; knowledge aggregation; spectral clustering; similarity

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