Hindawi
Complexity
Volume 2021, Article ID 8825947, 16 pages
https://doi.org/10.1155/2021/8825947

Research Article

An Improved PageRank Algorithm Based on Text Similarity Approach for Critical Standards Identification in Complex Standard Citation Networks

Yongchang Wei, Feiteng Yi, Xufeng Cui, and Fangyu Chen

Zhongnan University of Economics and Law, Wuhan 430073, China

Correspondence should be addressed to Fangyu Chen; cfyconan@126.com

Received 18 August 2020; Revised 27 January 2021; Accepted 18 March 2021; Published 28 March 2021

Academic Editor: Yi Su

A standard system, which is a powerful tool in maintaining the normal operations and development of a specific industry, is intrinsically a complex network composed of numerous standards which coordinate and interact with each other. In a networked standard system, the identification of critical standards is of great significance when drafting and revising standards. However, a majority of the existing literature has focused on the citation relationships between standards while ignoring the intrinsic interdependent relationships between the contents of standards. To overcome this limitation, we utilize the text similarity approach (TSA) to quantify the relationship intensity between each pair of standards, in order to generate a directed weighted network. The critical contribution of this study is that the similarity computed by the TSA is incorporated into the traditional PageRank algorithm for the identification of critical standards. The improved algorithm comprehensively considers the quantity and importance of neighboring standards and the citation intensity, as quantified by TSA. The algorithm is finally validated using the Chinese environmental health standards through comparison with the traditional PageRank algorithm and different classic measurements.

1. Introduction

Standards refer to the normative documents formulated by consensus and issued by recognized institutions, in order to achieve the best operations within a certain scope. In the current knowledge economy era, the orderly operation of any industry is basically inseparable from standards, where the quality of standards reflects a country's economic and technological level to a large extent [1–4]. Standards often play a synergistic role in the form of a standard system, which refers to an organic whole composed of numerous standards that interact and coordinate with each other. The specific manifestation of interaction and coordination is the citation between standards, which maps the transmission of information and knowledge carried by standard documents. Therefore, a standard citation network is also a transmission network of information [5].

A networked standard system has the scale-free characteristic, in which a minority of standards play dominant roles in controlling the transmission of information and knowledge [6]. In standardization activities, it is very important to identify the critical standards, in order to assist decision-making in revising or drafting standards. Specifically, we should optimize the structure of existing standard systems in diversified application areas by creating effective and reasonable links between newly drafted standards and existing standards, with the aim of enhancing the systematicness, compatibility, and coordination. The newly created links, in the form of citations, should reflect the essential relationships between standards. In this situation, ignoring critical standards may lead to poor system structure or standard conflicts. In the process of revising standards, we also should pay attention to the most significant standards, in order to maximize their economic or social effects, due to limited resources (i.e., in terms of people, budgets, and time.)

However, it is difficult and/or time-consuming to identify the critical standards in massive standard documents when we consider only individual standards while

ignoring their interactions. In addition to network methods, the criticality of standards can be assessed by many traditional methods, such as AHP, ANP, and TOPSIS [7–9]. However, these methods can hardly address the interactions between standards and are subjective in selecting measurements and determining weights. In recent years, the literature has begun to pay attention to identifying the critical standards in directed standard citation networks, which are useful in disclosing the evolution of standard networks [10, 11]. However, in the previous literature, the citations between standards have been represented by adjacent matrix without weights, which provide limited information. In reality, the citations between standards differ in intensity, which should be determined by assessing the intrinsic relationships between the contents of each pair of standards. To overcome this shortcoming, the text similarity approach (TSA) is introduced to quantify the relationship between each pair of standards and further form a directed weighted network. The similarity degree metric as computed by the TSA is used to improve the traditional PageRank algorithm for identification of critical standards.

The PageRank algorithm was originally used by Google to rank web-pages [12] embedded in search engines. As web page networks and citation networks have similar structures, many scholars have introduced this algorithm for the identification of important nodes [13–16]. However, the traditional PageRank algorithm also has some limitations. For example, in each iteration of this algorithm, every node distributes its PageRank (PR) value equally to its immediate downstream nodes, which makes it incapable of reflecting the fact that important nodes often have a stronger ability to build influence. Therefore, we argue that the PR value distribution is unbalanced; for example, a standard cites two other standards, then the standard with a more similar topic should be assigned a higher PR value than the other in each iteration. So, in order to overcome this limitation, we propose an improved PageRank algorithm based on TSA. In this new algorithm, PR values quantified by TSA are assigned as the weights in each iteration. The improved algorithm comprehensively considers the quantity and importance of neighboring standards and the citation intensity between each pair of standards, as quantified by the TSA. Our experimental results show that the improved PageRank algorithm behaves better than the traditional PageRank algorithm.

2. Literature Review

This study focuses on the identification of critical standards using an improved PageRank algorithm based on a network model. The network model was developed by integrating the information with respect to citations and similarity, as quantified by TSA. In the following, we review the two main streams of literature on citation network analysis and the application of TSA.

2.1. Citation Network Analysis. To date, citation network analysis has been widely used in different kinds of citation networks, such as journal citation networks, paper citation networks, and patent citation networks. However, its application in standard systems is still in its infancy. Based on citation networks, the collaborations between authors, institutions, and countries can be investigated by constructing collaboration networks. The theories and methods used in other citation networks are valuable for our investigation of standard citation networks, due to their similarities.

Garfield [17] originally proposed the evaluation of journal articles through citation criteria. The network analysis approach was introduced into the area of citation analysis by Small [18], in which co-citation was used to quantify the relationships between articles. In 1985, Doreian [19] proposed the concept of stratified journal network where the nodes are journals and the relations are citations aggregated over the articles in these journals. Calero-Medina et al. [20] constructed a paper citation network by collecting citation data between papers in specific fields from 1990 to 2005 and analyzed the development path of future research. With the pure number of citations becoming one of the most important factors of measuring the scientific impact and quality of authors, Fister et al. [21] established a three-layer network and proposed two scenarios with SPARQL queries to find citation cartels, which boldly revealed a common academic phenomenon. In 2014, Tobias et al. [22] analyzed the memes in scientific literature and defined the product of the frequency of occurrence and the propagation score along the citation graph as the meme score; this research provided new directions and inspiration for citation networks. In addition, many scholars have applied the PageRank algorithm to paper citation networks. Singh et al. [23] used the PageRank algorithm to rank papers based on a citation network. Qiao et al. [24] defined the value of each paper by using an improved weighted PageRank algorithm in combination with journal impact factors. Garcia et al. [25] transformed the rules of social impact of papers in a citation network into a mathematical equation, in order to predict the level of social impact of academic papers in the citation network, by using cooperative game theory.

Narin [26] was the first scholar to apply the citation network analysis method to the field of patents, opening up a new direction of patent literature measurement. Subsequently, Li et al. [27] analyzed the topological structure of a nanotechnology patent citation network through social network analysis (SNA). Cho and Shi [28] selected 42,650 patent documents issued by Taiwan from 1997 to 2008 and identified five core technologies and emerging technologies that affected Taiwan's economy by analyzing the associated patent citation network. Fujita et al. [29] established a directed weighted network with patent timeliness and text relevance between patents as mixed parameters, aiming to assess the limitations of undirected citation networks. Beltz et al. [30] adopted a sorting algorithm based on reinforcement learning to rank patents in the patent citation network, comparing it with the traditional PageRank algorithm. Their results showed that this algorithm behaves better than the traditional PageRank algorithm.

In addition, previous research on standards has mainly focused on standardization activities and the impact of

standard implementation on social and economic benefits [31–34]. As introduced previously, many standards compiled together with complicated interactions form a complex standard network. The structure plays a fundamental role in achieving the objectives of standard systems. However, the previous research has remained incapable of addressing such structural problems. In recent years, some scholars have begun to pay attention to the topological characteristics of standard citation networks. For example, in 2018, Wei et al. [10] established a dynamic standard citation network model and analyzed the structural problems existing in the standard system of China's automobile industry. To identify important standards, Wei et al. [11] proposed an integrated multi-criteria decision-making method, which combined the entropy weight (EW) method and a TOPSIS method based on traditional node measurements, such as degree centrality and betweenness centrality. However, their existing research focused on simple 0-1 citation relationships, while ignoring the magnitude information of the relationships between pairs of standards, which can be reflected by their connections in terms of document contents.

2.2. Application of Text Similarity Approach. TSA, which is a type of text mining method [35–37], is employed to quantify the relationship between standards in this study. At present, the literature on TSA is extensive and many algorithms have been proposed, including the cosine similarity algorithm, Levenshtein distance algorithm, SimHash algorithm, and Euclid distance algorithm. These algorithms generally include four main steps: text segmentation, extracting text keywords, converting spatial vectors using the word frequency, and calculating the text similarity using the relevant algorithm.

The theories and methods of TSA have been widely used in various research fields. Originally, Salton et al. [38] proposed the VSM (vector space model) in the 1970s. The core idea of VSM is to simplify the processing of text content into vector operations in a vector space, as well as expressing semantic similarity using spatial similarity. This intuitive and understandable form is the most widely used method for text association analysis. Andy and Alice [39] designed an algorithm to extract the core content of the design model using the basic idea of text mining and the Bayesian belief network model, in order to promote the sharing of information among designers. In 2007, Tseng et al. [40] proposed a set of perfect text mining technologies to assist patent analysts in carrying out various types of data analysis, the specific algorithm used in the present paper was greatly inspired by this research. In addition, some scholars have attempted to introduce the TSA into the traditional PageRank algorithm, in order to develop a better algorithm for web page ranking [41, 42]. However, from the application perspective, to the best of our knowledge, the TSA or PageRank algorithm has seldom been used in analyzing standard documents.

In summary, the current research on standard citation networks is still at an early stage. The previous literature has failed to consider the intrinsic relationships with respect to content between standards, rather than just 0-1 citation scoring. Consequently, the existing network models are represented as adjacency matrices with only two elements: 0 and 1. This is highly insufficient for measuring the significance of standards. On the other hand, text mining has provided fruitful algorithms for quantifying the similarity between different kinds of documents, which motived us to use TSA for the identification of critical standards in standard networks.

3. Network Model Construction Based on TSA

In this paper, TSA is employed to quantify the relationship between each pair of standards and, further, to form a directed weighted network, which lays the foundation for designing an improved PageRank algorithm. The overall procedure for critical standard identification is shown in Figure 1. It is noteworthy to emphasize that the TSA is only applied to the pairs of standards with citation relationships, in order to improve the computational efficiency. This is particularly important in the situation where the number of standards is very large.

A directed unweighted standard citation network is firstly established. After obtaining the text similarities between pairs of standards with citation relationships, a directed weighted network based on text similarity is established. Finally, the improved PageRank algorithm is used for the identification of critical standards.

3.1. Quantifying the Citation Intensity with TSA. The key part of network construction lies in the calculation of similarity values between each pair of standards. In the following, the implementation of the text similarity algorithm is described. The source code for the TSA method is included in the supplementary material (available here).

This paper mainly combines the vector space model (VSM), the TF-IDF (term frequency-inverse document frequency) method, and the cosine similarity algorithm to calculate the text similarity value between each pair of standards with a citation relationship. The TF-IDF method is a common weighting technique used for information retrieval and data mining [43, 44]. After processing, the standard document is transformed into an n-dimensional vector composed of keyword frequencies. Further, the keyword frequencies are converted into keyword values by weighting with the TF-IDF method. Then, the text similarity values between standards are calculated using the cosine value algorithm.

The cosine value of two n-dimensional spatial vectors $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$, $\cos\beta$, is calculated as

$$\cos\beta = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}. \quad (1)$$

The specific process of the TSA is depicted as Figure 2.

Step 1 (content extraction from standard documents): content extraction lays the foundation for similarity

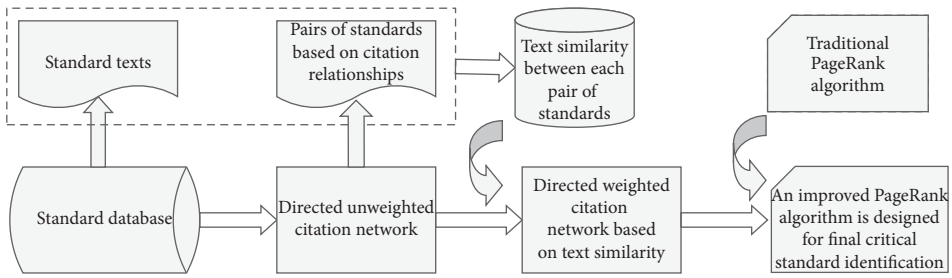

Figure 1: The overall procedure of critical standards identification.

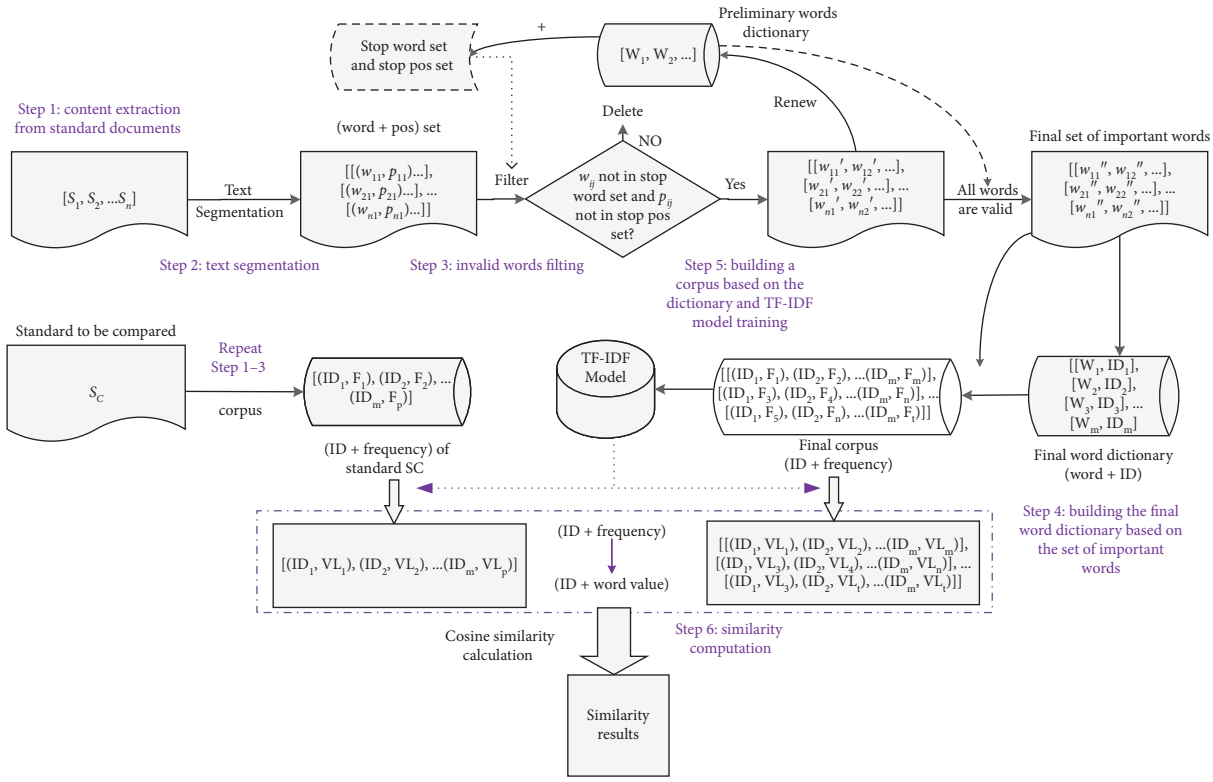

Figure 2: The specific process of text similarity analysis.

analysis. In this study, we do not extract all the sentences of an original standard document for similarity calculation, as the use of too many useless sentences greatly increases the computational time and may lead to large deviations in the similarity results. In this study, we manually reviewed all the standard documents and selected the most significant sentences that matched the theme of each standard. In particular, all the titles were included, whereas all the tables and figures were excluded from analysis.

The process includes six steps: content extraction, text segmentation, invalid words filting, building word dictionary, building corpus and TF-IDF model, and similarity computation.

Step 2 (text segmentation): text segmentation is an important step in the preprocessing of text mining. In this step, each original standard text is transferred into a set of "word + part of speech (pos)."

Step 3 (invalid words filtering): invalid words are automatically filtered out, in order to save storage space and increase computational efficiency during the text similarity calculation. In this step, we use a stop word set and a stop pos set to filter the invalid words. Invalid words mainly include auxiliary words and tone words. In this step, we have a list of invalid words and a list of corresponding pos (called stop word set and stop pos set). The words in the stop word set mainly include auxiliary words and tone words. As we cannot enumerate all the words, we think to be invalid at once, in order to increase the efficiency of filtration, we add an extra stop pos set including all the invalid, such as conjunctions and adverbs. Then, all the "word + pos"

sets obtained in the previous step are double-conditionally filtered, where the words are retained only if the words and corresponding pos are neither in the stop word set nor the stop pos set. At the same time, the filtered results form a preliminary word dictionary. We check this dictionary carefully and add the missing words which we think are invalid into the list of invalid words, followed by filtering again. The above process is repeated until the word dictionary includes all the important words; this process also makes the final calculation results as scientific as possible.

Step 4 (building the final word dictionary based on the set of important words): the main work of this step is to assign a specific ID to all the important words for subsequent quantitative processing.

Step 5 (building a corpus based on the dictionary and TF-IDF model training): the corpus is a list of sparse vectors. After this step, each standard document is transferred into an important word list, stored in the form of "word ID + word frequency." Then, TF-IDF model training is in process. It can be understood that the TF-IDF model is used to measure a word's value, relative to an article, based on the word's word frequency. After obtaining the TF-IDF model, all the important word lists above are stored in the form of "word ID + word value."

Step 6 (similarity computation): the procedure from Step 1 to Step 3 is repeated for all the standards to be compared. Then, the standard contents are transferred into a list in the form of "word ID + word value" based on the corpus and TF-IDF model obtained in Step 5. Further, the text similarities between this standard and the other standards are calculated using formula (1), which gives the final standard's similarity matrix. After repeating all the steps for each standard, we finally obtain the final similarity result between each standard.

3.2. Network Model Construction. After calculating the text similarity of standards, we can quantify the citation intensity between each pair of standards. A weighted and directed standard citation network can be represented as a set composed of three subsets: $\mathscr{G} = \{\mathscr{N}_s, \mathscr{E}_c, \mathscr{T}_t\}$, where the node set $\mathscr{N}_s = \{s_1, s_2, \ldots, s_n\}$ includes n standards, $\mathscr{E}_c = \{\langle s_i, s_j\rangle > |s_i, s_j \in \mathscr{N}_s\}$ is the set of directed edges based on the citation relationship, and $\mathscr{T}_t = \{\xi_{ij}|\langle s_i, s_j\rangle \in \mathscr{E}_c\}$ is the set comprised of the text similarity of each pair of standards $\langle s_i, s_j\rangle$ corresponding to the edges in $\mathscr{E}_c$, as computed by the TSA algorithm described above. From the view of the model, the set $\mathscr{T}_t$ reflects the substantial difference between the model herein and the citation network model of standards in the existing literature, in which only the citation relationship has been considered [10, 11]. By incorporating the similarity relationship, the standard system becomes a directed and weighted network, which subsequently changes the network measurements. For ease of computation, the citation relationships between standards can be represented as a weighted adjacency matrix $\mathbf{A}^w = (a_{ij}^w)^{n\times n}$, $i, j \in \mathscr{N}_s$, in which $a_{ij}^w = \xi_{ij}$ if standard i cites standard j, and $a_{ij} = 0$ otherwise.

4. Critical Standards Identification with an Improved PageRank Algorithm

The standard network model provides an important foundation for the identification of critical standards. However, in this study, as the standard system is represented as a weighted network with similarities as weights, the network measurements are substantially different from those of traditional unweighted networks. In this study, we focus on the well-known PageRank algorithm, which has been used by Google for their search engine and has been widely applied in citation networks for node evaluation. We note that the traditional PageRank algorithm equally distributes the PR values to its downstream nodes, which can be improved by considering the importance difference of downstream nodes as characterized by the similarity metric. In this section, we start by analyzing the limitations of the traditional PageRank algorithm and then propose the improved PageRank algorithm for critical standard identification.

4.1. Traditional PageRank Algorithm. The PageRank algorithm was originally used by Google to evaluate the importance of web-pages. In a standard citation network, the higher PR value a standard has, the greater importance the standard has. Two basic assumptions are made when applying the PageRank algorithm to standard citation networks:

(1) *Quantity Assumption*. A standard cited by more other standards is more likely to be important.

(2) *Quality Assumption*. A standard cited by higher quality standards is more likely to be important.

The quantity assumption is usually reflected by other network measurements, such as degree centrality and entropy [45]. As the concept of quality is abstract, there are various quality or importance measurements for network nodes. In the PageRank algorithm, the quality or importance of a standard in a network is measured by its PR value. The PR value of a node can be transmitted to its neighboring nodes.

The process of the PageRank algorithm is described as follows. In the initial stage, each standard is evenly assigned a PR value $1/n$, where n represents the number of standards in the citation network. Then, a new round of PR value allocation follows, with each standard distributing its current PR value equally to the standards cited by this standard, based on its out-degree. Each cited standard obtains the corresponding PR value from the upstream citation standards. Then, each standard sums all the incoming PR values from the upstream citation standards, in order to update a new PR value. When each standard has an updated PR value, a round of PageRank calculations is completed. After some iterations, the PR value for each standard tends to a stable value,

which is the final PR value of each standard. The specific formula of the algorithm is as follows:

$$\mathrm{pr}_t^i = (1-\gamma)\frac{1}{n} + \gamma \sum_{s_j \in \mathscr{M}(s_i)} \frac{\mathrm{pr}_t^j}{L(s_j)}, \tag{2}$$

where pr_t^i is the PR value of standard s_i, $\mathscr{M}(s_i)$ is the set of upstream standards that cite standard s_i, $L(s_j)$ is the number of standards that standard s_j cites, and γ represents the damping factor.

4.2. An Improved PageRank Algorithm. The traditional PageRank algorithm was originally designed to rank the importance of web-pages. However, there exist some differences between standard citation networks and the web page networks, in terms of network structure and topological properties. In the following, we introduce the limitations in the case where the algorithm is directly applied to the critical standards identification.

The major limitation is related to the distribution mechanism of the PR values. For a web page in a web page network, the quantity and quality of the web-pages linked to this web page determine the page's final PR value. The traditional PageRank algorithm is an unbiased allocation algorithm, namely, the PR value of a page is distributed equally, according to the number of pages linked to this page, in each iteration. However, in practice, the pages differ in information quality, and the quality of the pages linked to a page also highly differ. Hence, the final ranking results may be inconsistent with the real influence of web-pages, without considering the importance of linked pages and redesigning the distribution mechanism of PR values.

If a standard in a standard citation network distributes its PR value equally (i.e., only according to its out-degree), the final ranking results may also not be rational. This is because the information transmission between standards that are closely related to each other plays a critical role in forming the structure of the standard systems. In other words, we should take the citation quality between standards into account. To deal with this problem, we use the text similarity to reflect the citation quality and to distribute the PR values. When distributing the PR values in each iteration, the standards whose topics are close to that of standard s_j in $\mathscr{M}(s_i)$ obtain high PR values from standard s_j.

A minor limitation is related to the damping factor γ. The damping factor γ indicates the probability that an user will continue to browse a web page, and $1-\gamma$ indicates the probability that the user will reopen a new page to browse. In the traditional PageRank algorithm, the value of γ is set as 0.85 [46]. Chen et al. [47] found that the citation distance of the paper citation network is smaller than that in web page networks, where the average is 2. Therefore, when the PageRank algorithm is applied to a paper citation network, the damping factor has been suggested set as 0.5. As standard citation network and paper citation network have similar network structure and topological properties, we also set γ as 0.5.

4.3. An Improved PageRank Algorithm Based on Text Similarity Analysis. Based on the above description, we propose an improved PageRank algorithm based on TSA. The source code for the improved PageRank algorithm is included in the supplementary material. In this algorithm, the damping factor γ is set to 0.5, and every standard distributes its PR value using the weights measured by the TSA. The specific formula of the improved PageRank iteration algorithm is expressed as

$$\mathrm{pr}_m^i = (1-\gamma)\frac{1}{n} + \gamma \sum_{s_j \in \mathscr{M}(s_i)} \mathrm{pr}_m^j \times \frac{\xi_{ji}}{\sum_{s_k \in \mathscr{D}(s_j)} \xi_{jk}}, \tag{3}$$

where $\mathscr{D}(s_j)$ is the downstream standard set of standard s_j and ξ_{jk} is the text similarity value between standard s_j and standard s_k. Compared with the traditional PageRank algorithm, the substantial difference of our improved PageRank algorithm lies in the allocation mechanism of PR values, with consideration of the association degree between upstream and downstream standards, in terms of similarity.

5. Empirical Analysis

The area of environmental health focuses on all aspects of the natural and built environment affecting human health. Thus, improving environmental health is an important foundation for guaranteeing human safety and realizing sustainable development. The standards associated with environmental health have played effective roles in controlling the environmental factors and mitigating the related risks that affects public health through a series of standardization activities. Due to the interdependence of individual standards, an environmental health standard system is a typical complex network, which should be established with desired performance with respect to systematicness, compatibility, and coordination, which are intrinsically determined by the network structure. In the process of drafting and revising standards, we must pay attention to the critical standards, in order to optimize the system structure. Therefore, effective algorithms or measurements for the identification of critical standards are highly important.

In this section, we focus on demonstrating the effectiveness of the proposed improved PageRank algorithm, in contrast to the traditional one, with realistic data of environmental health standards. Since the existing measurements or algorithms differ from each other, it is difficult to compare them directly. Therefore, we propose two assumptions for indirect validation. Firstly, a strong algorithm should be closely related to classical measurements. Secondly, a strong algorithm should generally identify more common critical standards than classical measurements. These two assumptions are justified by the fact that different measurements or algorithms share a common objective of assessing node importance despite their distinctions. The validation process was carried out in two steps. In the fist step, we assessed the correlation differences between the two PageRank algorithms and classical measurements. In the second step, we compared the differences with respect to the capability of critical standards identification of these two

algorithms. This was performed by analyzing the proportion of common critical standards identified simultaneously by the different classical measurements and each PageRank algorithm.

5.1. Data Collection and Network Visualization. We selected 103 standards on environmental health from the category of health-related standards from *National Health Commission of the People's Republic of China* (http://www.nhc.gov.cn/wjw/pgw/wsbz.shtml). The corresponding citation relationship was established by viewing the normative document of each standard and confirmed on the website of the *National Public Service Platform for Standards Information* (http://std.samr.gov.cn/gb) to ensure the accuracy of the data. After obtaining each standard's citation data, each standard was assigned with a unique ID from "S1" to "S103," and an initial adjacency matrix with 0-1 elements of the citation network, as established in the EXCEL 2019. The adjacency matrix of the standard citation network was then imported into Gephi, an open-source visualization software for network analysis, in order to visualize the network structure, in which each node represents a standard and each directed edge represents a citation relationship, as shown in Figure 3. The classical node measurements, including degree centrality, eigenvector centrality, eccentricity, and betweenness centrality, were also computed by Gephi. Then, the TSA was performed to update all the "1" elements in the adjacency matrix to form a weighted and directed network. After that, the improved PageRank algorithm was programmed in Python 3.7.3 to rank the importance of standards. The detailed information of standards mentioned in this paper is shown in Table 1. In the following, we represent the results of the classical measurements, in order to validate the effectiveness of our improved PageRank algorithm for critical standards identification.

Each node represents a standard; the directed edges represent the citation relationship. The size of nodes is displayed according to its in-degree, and the nodes holding the same out-degree share the same color.

5.2. Results of Classical Measurements

5.2.1. Selection of Classical Measurements. Extensive literature in the field of network analysis has been devoted to the concept of centrality, which aims at answering a fundamental question: which nodes occupy the core positions of a network? Some centrality measurements have been proposed to evaluate the importance of nodes [48–51]. We selected four widely investigated measurements, namely, degree centrality, eigenvector centrality, eccentricity, and betweenness centrality, to evaluate the importance of China's environmental health standards system and to validate the improved PageRank algorithm.

It should be noted that the weighted adjacency matrix constructed in this paper was only used in our improved PageRank algorithm. The citation relationship between standards here was represented as an unweighted adjacency matrix $A^u = (a_{ij}^u)^{n\times n}$, $i, j \in \mathcal{N}_s$, in which $a_{ij}^u = 1$ if standard i cites standard j, and $a_{ij}^u = 0$ otherwise.

(1) Degree Centrality. In a directed standard citation network, degree centrality is divided into in-degree and out-degree. In general, a standard with high in-degree reflects the fact that many standards cite it, indicating its strong impact across the whole network [11]. In contrast, the relationship between out-degree and node importance is implicit. The in-degree of standard $i \in \mathcal{N}_s$, $\mathrm{dc}_{\mathrm{in}}^i$, is defined as $\mathrm{dc}_{\mathrm{in}}^i = \sum_{j\in\mathcal{N}_s} a_{ji}^u$, which reflects the number of standards that cite the concerned standard. Similarly, the out-degree of standard $i \in \mathcal{N}_s$, $\mathrm{dc}_{\mathrm{out}}^i$, is defined as $\mathrm{dc}_{\mathrm{out}}^i = \sum_{j\in\mathcal{N}_s} a_{ij}^u$, which reflects the citation count of standard i.

(2) Eigenvector Centrality. Degree centrality has the limitation of only considering the number of neighboring standards, while the eigenvector centrality, which is similar to the PageRank algorithm, overcomes this problem by considering the importance or quality of neighboring standards. The eigenvector centrality gives a relative score to each node in the network, as proportional to the sum of the eigenvector centrality of all the nodes connected to it [52]. In a standard citation network, the higher eigenvector centrality a standard has, the more important the standard is. The eigenvector centrality of standard $i \in \mathcal{N}_s$, ec^i, is represented as

$$\mathrm{ec}^i = \frac{1}{\lambda} \sum_{j\in\mathcal{N}_s} a_{ji}^u \mathrm{ec}^j, \tag{4}$$

where λ represents the eigenvalue of adjacency matrix $\mathbf{A}$.

(3) Eccentricity. Eccentricity is a different measurement, which defines the importance of a standard by describing its "influence transfer depth" in the network. Eccentricity refers to the maximal distance from a node to other nodes in a network [50]. In a standard citation network, a standard with high eccentricity has a persistent impact on other standards. The eccentricity of a standard $i \in \mathcal{N}_s$, et^i, is represented as

$$\mathrm{et}^i = \max\left(\mathrm{len}_{ji}\right)_{j\in\mathcal{R}_i}, \tag{5}$$

where $\mathcal{R}_i$ is the set of standards that standard i can transmit information to a certain path in the network and len_{ji} is the length of the shortest path between standard j and standard i.

(4) Betweenness Centrality. The betweenness centrality is a measure of centrality in a network based on the shortest paths, which refers to the ratio of the number of shortest paths passing through a node to the total number of all the shortest paths in the network [53]. Consequently, a standard with high betweenness centrality plays a hub-role in information transmission. The betweenness centrality of a standard $i \in \mathcal{N}_s$, bc^i, is represented as

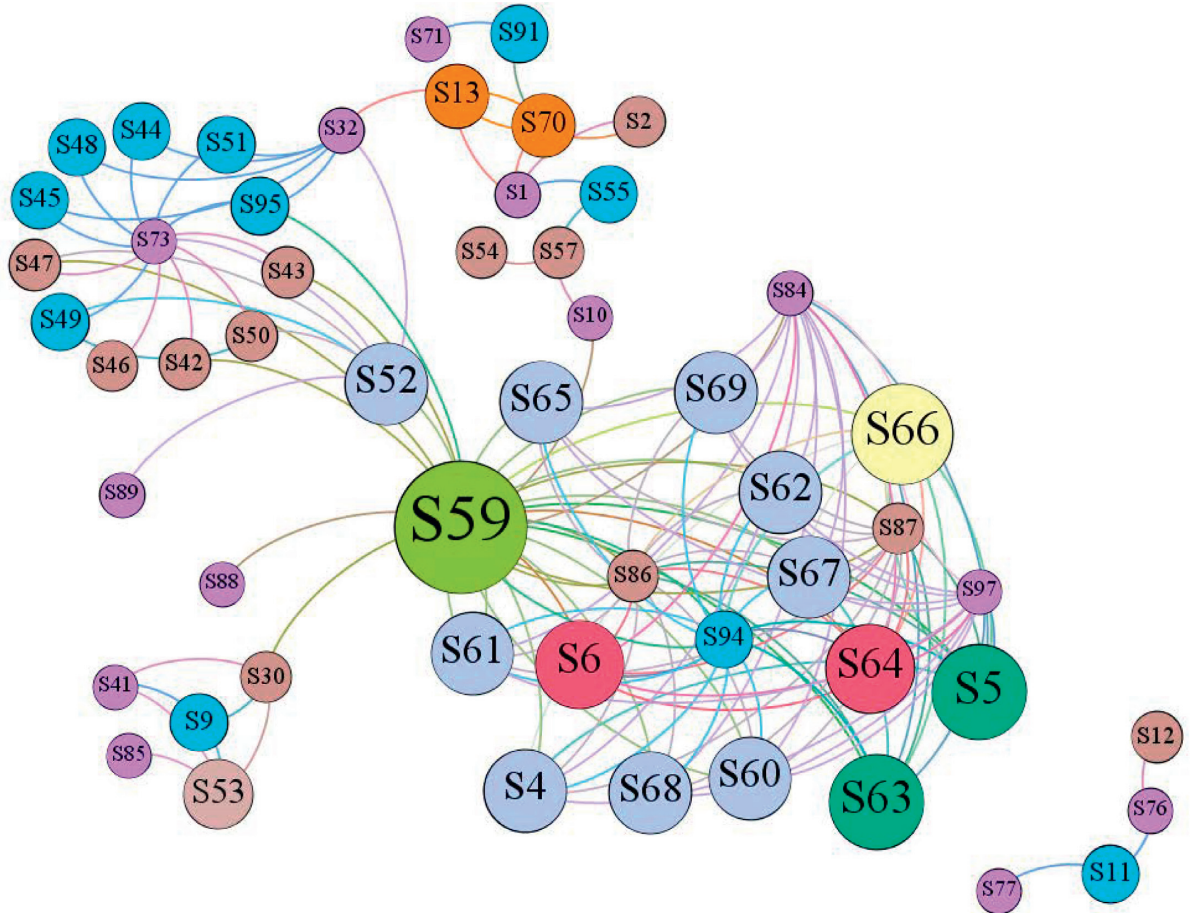

Figure 3: The China's environmental health standard citation network.

Table 1: The detailed information of the standards mentioned in this paper.

Id	Standards name
S59	Hygienic standards for drinking water
S66	Standard test method for drinking water: indicators of disinfection by-products
S5	Standard test method for drinking water: indicators of organic matter
S63	Standard test method for drinking water: sensory properties and physical indicators
S64	Standard test method for drinking water: indicators of inorganic nonmetal
S6	Standard test method for drinking water: indicators of metal
S4	Standard test method for drinking water: indicators of pesticide
S67	Standard test method for drinking water: indicators of disinfectant
S61	Standard test method for drinking water: collection and preservation of water samples
S60	General principles of hygienic standards for drinking water
S52	Hygienic standards for hotel
S30	Hygienic standards for planning of village and town
S95	Hygienic standards for restaurants
S94	Hygienic standards for secondary water supply facilities
S70	Indoor air quality standards
S47	Hygienic standards for gymnasiums
S91	Hygienic standards for inhalable particulate in indoor air
S53	Hygienic requirements for innocuous feces
S87	Safety evaluation standards for drinking water distribution equipment and protective materials
S97	Quality assurance specification for biological monitoring
S84	Standard for determining the chronic arsenic poisoning area of residents caused by environmental arsenic pollution
S86	Sanitary safety evaluation of chemical treatment agent for drinking water

$$\mathrm{bc}^i = \sum_{m,n \in \mathcal{N}_s} \frac{\sigma(m,n|i)}{\sigma(m,n)}, \quad (6)$$

where $\sigma(m,n)$ is the number of all the shortest paths from standard m to standard n and $\sigma(m,n|i)$ is the number of those paths passing through standard i other than standard m or standard n. It is important to notice that standard i cannot be either the starting or ending node, that is, $i \cap (m,n) = \varnothing$.

It should be noted that this paper is not a denial of these classical measurements but, instead, inheriting and improving them. The improved PageRank algorithm not only comprehensively considers the quantity and quality of neighboring standards but also highlights the quality of the citation between the standards as quantified by TSA.

5.2.2. Results of Classical Measurements. In view of the large number of standards, we only represent the top ten standards in criticality identification in this paper, as shown in Table 2. As the in-degree can better reflect the importance of the standard than out-degree, we mainly analyzed in-degree centrality. As can be seen from Table 2, the top 10 standards are almost all standards relating to drinking water methods, indicating that the standards of drinking water occupy a very important position in the whole environmental health standards system. In addition, we found that, in the top 10 standards ranked by out-degree, only S59 and S66 are in the list of top 10 standards ranked by in-degree. It means that there exists a trade-off between in-degree and out-degree measurements.

The results of eigenvector centrality in Table 2 are normalized, which show that the top 10 standards, with respect to eigenvector centrality, are still the standards relating to drinking water; again verifying the importance of drinking water standards in China's environmental health standard system. In addition, S59 ranks first in both in-degree and out-degree, but drops to sixth in terms of eigenvector centrality. This phenomenon is due to the core idea of eigenvector centrality, which not only considers the number of neighboring standards, but also highlights the quality of them.

As can be seen from Table 2, the eccentricity of each standard in the network is not high, with a maximum of 3. As the eccentricity of a standard defines the importance of a node by describing the "transfer depth of influence" of a node in the network, in the citation network, the information transmission of a standard is interrupted by "three layers" at most. This, to some extent, indicates that the current environmental health standards do not have far-reaching impacts. In addition, it is noteworthy that most of the top 10 standards, in terms of eccentricity, are also about drinking water.

As we can see from Table 2, among the top 10 standards, the betweenness centrality values vary from each other. The two standards S59 and S52 are the most prominent, in terms of betweenness centrality, with values of 256.5 and 62.9, respectively. As can be seen from Figure 3, these two standards are indeed information hubs of the network. If these two nodes were removed, the entire network would be broken down into several isolated components, inevitably interrupting the information transmission between standards.

Table 2: Results of critical standards based on classical measurements.

Rank	Id	$\mathrm{dc}^i_{\mathrm{in}}$	Id	$\mathrm{dc}^i_{\mathrm{out}}$	Id	ec^i	Id	et^i	Id	bc^i
1	S59	14	S59	16	S66	1	S66	3	S59	256.5
2	S66	9	S94	14	S5	0.889	S6	3	S52	62.9
3	S5	8	S86	14	S63	0.888	S91	3	S30	17
4	S63	8	S87	14	S64	0.747	S63	3	S95	11.9
5	S64	7	S84	14	S6	0.673	S64	3	S66	9
6	S6	7	S97	13	S59	0.598	S5	3	S94	6
7	S4	6	S73	12	S61	0.452	S67	3	S70	5
8	S67	6	S32	7	S60	0.452	S61	3	S64	4
9	S61	6	S1	4	S4	0.452	S62	3	S61	4
10	S60	6	S66	3	S67	0.452	S60	3	S47	3.4

5.3. Results of Improved PageRank and Traditional PageRank

5.3.1. An Example. In this section, we compare the results of the two PageRank algorithms, in order to assess their differences with two exemplified standards. Table 3 shows that the PR values of the top five standards under the two PageRank algorithms are basically the same. However, the four standards S67, S87, S61, and S62 exhibit higher significance in the improved PageRank algorithm than the traditional PageRank algorithm. The four standards S5, S70, S52, and S53 are evaluated as more important by the traditional PageRank algorithm.

To explain the essential differences between the two PageRank algorithms, we selected two standards S5 and S67 due to their large ranking gaps in the two algorithms. Specifically, S5 ranks fourth in the traditional PageRank algorithm and 13^{th} in the improved PageRank algorithm, while S67 ranks 20^{th} in the traditional PageRank algorithm but rises to seventh in the improved PageRank algorithm. As a standard's PR value is derived from its upstream standards that cite it, we also analyze the upstream standards of the two standards. Table 4 shows that their upstream standards are almost the same; that is, S87, S97, S59, S84, and S86 all cite S5 and S67. Even if S5 has one more upstream standard S4 which ranks 16^{th} and 28^{th} in the improved PageRank and traditional PageRank algorithms, respectively, it does not have a significant impact on the PR values of S5 and S67. In addition, S5 has more upstream standards to distribute PR values than S67, but its final PR value is lower than S67, which is counterintuitive. Therefore, we can judge that the distribution mechanism of PR values of the two different PageRank algorithms leads to the substantial differences in the results presented in Table 3.

To analyze the interdependent relationships between standards, Figure 4 shows the word cloud diagrams of seven standards with the font size of keywords positively correlated to the keyword count in each standard. The cloud diagrams show that S59, S86, and S87 are closely related with each

Table 3: Results of critical standards based on two PageRank algorithms.

Rank	Id	pr_t^i	Id	pr_m^i
1	S59	0.100829319	S59	0.107478692
2	S66	0.069139751	S6	0.072567929
3	S6	0.048371879	S66	0.065847632
4	S5	0.044023854	S63	0.058175466
5	S63	0.044021582	S64	0.049741324
6	S70	0.04191036	S91	0.047526619
7	S64	0.036734466	S67	0.033505722
8	S52	0.033686738	S87	0.032926043
9	S91	0.032476497	S61	0.032181632
10	S53	0.030890666	S62	0.029800786

Table 4: Upstream standards of S5 and S67.

Id	Upstream standards
S5	S87, S97, S59, S84, S86, S4
S67	S87, S97, S59, S84, S86

other in terms of drinking water. The three standards S5, S59, and S67 simultaneously concern indicators. Their similarity values are further shown in Figure 5, which shows that the similarity between S67 and each upstream standard is generally higher than S5, implying that S67 receive more PR values from S87, S97, S59, S84, and S86 than S5. According to the improved PageRank algorithm, S67 has a stronger ability to acquire PR values than S5, due to having closer relationship with its upstream standards. Therefore, it is rational that S67 is identified as a more important standard than S5 by the improved PageRank algorithm. The above example demonstrates the superiority of our improved PageRank algorithm, with respect to the two exemplified standards.

Word cloud diagram presents the topics of a document. The standards S59, S86, and S87 are closely related with each other in terms of "drinking water," while the three standards S5, S59, and S67 simultaneously concern the "indicators."

The five upstream standards S59, S84, S86, S87, and S97 simultaneously cite S5 and S67; these five standards contribute to the final PR value of S5 and S67. The similarity between S67 and each upstream standard is generally higher than S5, which imply that S67 will receive more PR value from the five common standards than S5.

5.3.2. Performance Comparison of the Identification Capability of Critical Standards. The previous analysis demonstrates that the improved PageRank algorithm has a significant impact on the importance assessment of standards. In the following, we focus on exploring the relationships between the classical node measurements and the two PageRank algorithms by correlational analysis and by varying the parameter of the predetermined proportion of critical standards.

Each node in the graph represents a standard with data under the corresponding measurement. The left four figures show the correlation between the improved PageRank algorithm and classical measurements; the right ones show the correlation between the traditional PageRank algorithm and classical ones. γ represents the correlation coefficients between the corresponding measurements.

As shown in Figure 6, the left figure in each row represents the correlation between the improved PageRank algorithm and classical measurements, and the right figure in each row shows the correlation between the traditional PageRank algorithm and classical measurements.

By comparing the correlation coefficients of γ, we find that the correlations of the improved PageRank algorithm and the classical measurements are stronger than those of the traditional PageRank algorithm. This proves the assumption proposed previously that the improved PageRank algorithm is more comprehensive in the identification of critical standards. In addition, we can also see a phenomenon occurring with the improved PageRank algorithm, where the points with high PR values and high classical measurements data values are denser than the traditional PageRank algorithm. This means that the improved PageRank algorithm behaves better than the traditional algorithm in terms of differentiating capability.

In order to further demonstrate the efficacy of our improved PageRank algorithm, we analyze the proportion of common critical standards identified by the different classical measurements and each PageRank algorithm. We define a parameter $\mathrm{pc}_{\hbar m}$ to represent the proportion of critical standards, which is calculated as

$$\mathrm{pc}_{\hbar m} = \frac{n_{\mathrm{cs}}}{|\mathcal{N}_s|}, \tag{7}$$

where n_{cs} is the number of top critical standards selected and $|\mathcal{N}_s|$ is the total number of standards in the citation network $\mathcal{G}$.

Meanwhile, we define the proportion of common top critical standards identified by classical measurements, $\mathrm{pc}_{\hbar c}$, as

$$\mathrm{pc}_{\hbar c} = \frac{n_{\mathrm{cs}}^{\chi}}{|\mathcal{N}_s|}, \tag{8}$$

where n_{cs}^{χ} is the number of common top critical standards identified by classical measurements and the PageRank algorithm.

The ideal function relationship between $\mathrm{pc}_{\hbar m}$ and $\mathrm{pc}_{\hbar c}$ is $y = x$, that is, the ranking results of standards between the improved PageRank algorithm and classical measurements are absolutely the same. As we can see from Figure 7, the relation curves between $\mathrm{pc}_{\hbar m}$ and $\mathrm{pc}_{\hbar c}$ with respect to the improved PageRank algorithm under different classical measurements is closer to $y = x$. In particular, in the interval of $(0.6, 1.0)$, the relation curve of the improved PageRank algorithm and the four classical measurements completely coincides with $y = x$, while that with the traditional PageRank algorithm deviates greatly.

A strong algorithm will generally identify more common critical standards as classical measurements. The standard ranking results obtained by the two PageRnk algorithms are fitted with the four classical measurements in the proportion

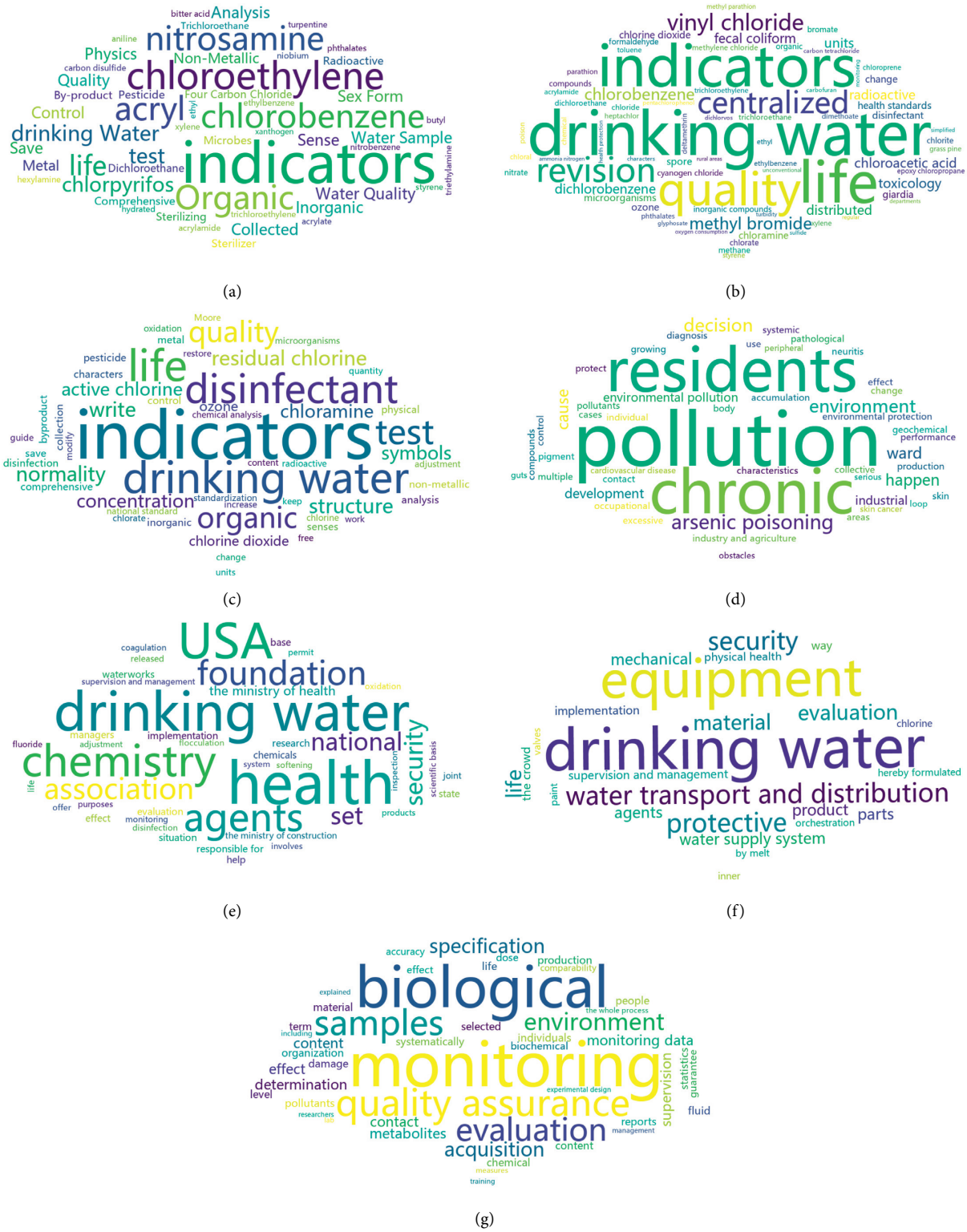

FIGURE 4: The word cloud diagram of seven standards. (a) S5. (b) S59. (c) S67. (d) S84. (e) S86. (f) S87. (g) S97.

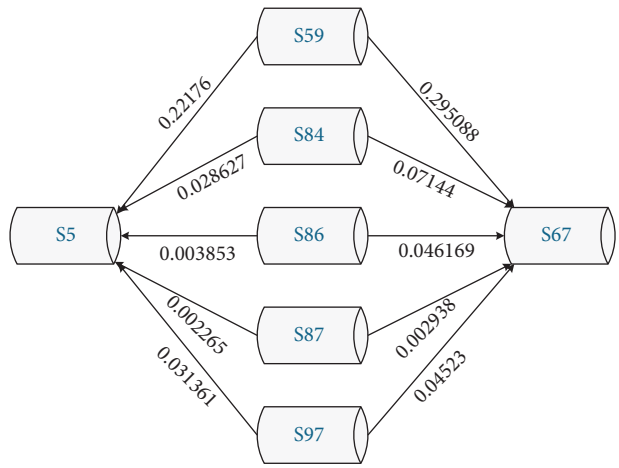

Figure 5: The text similarity of S5 and S67 with common upstream standards.

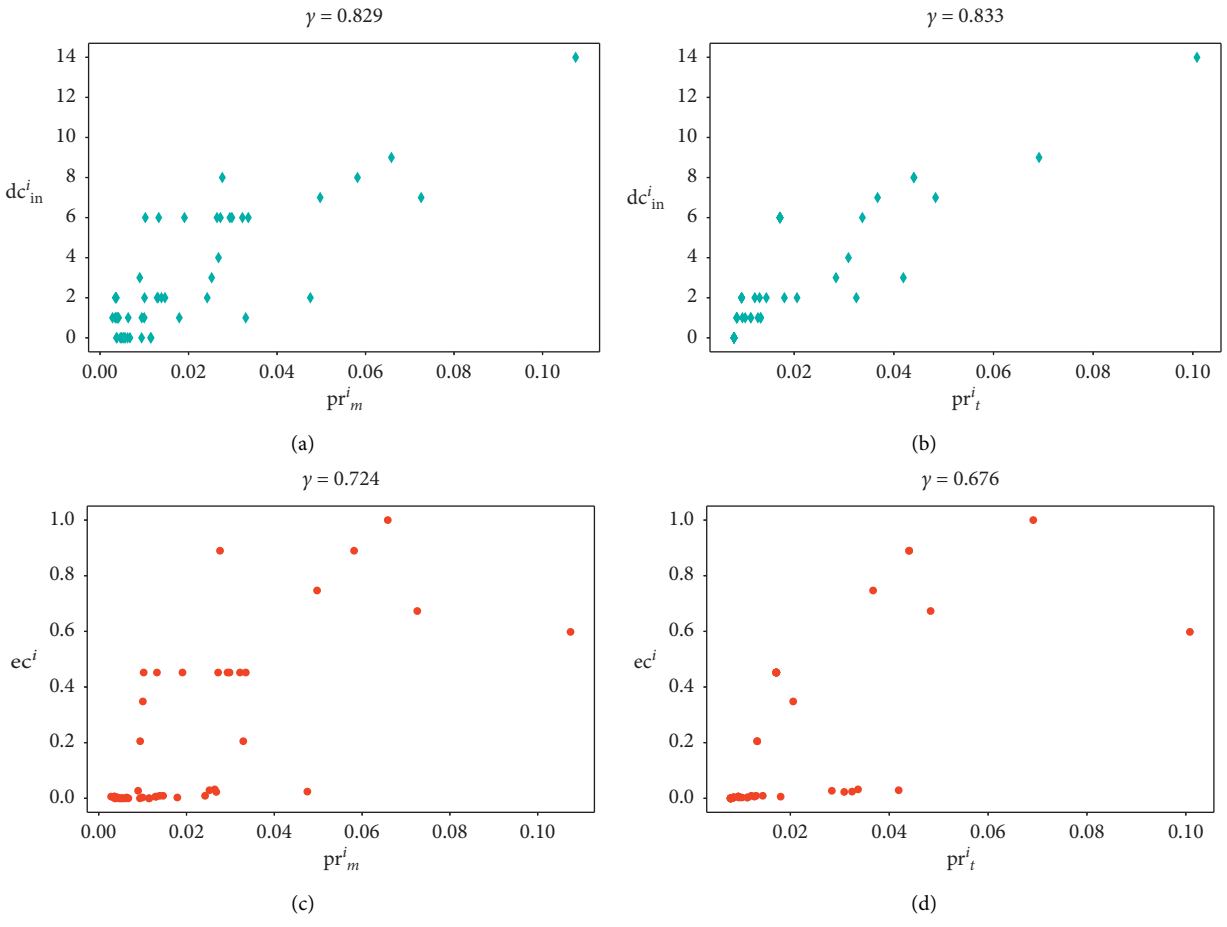

Figure 6: Continued.

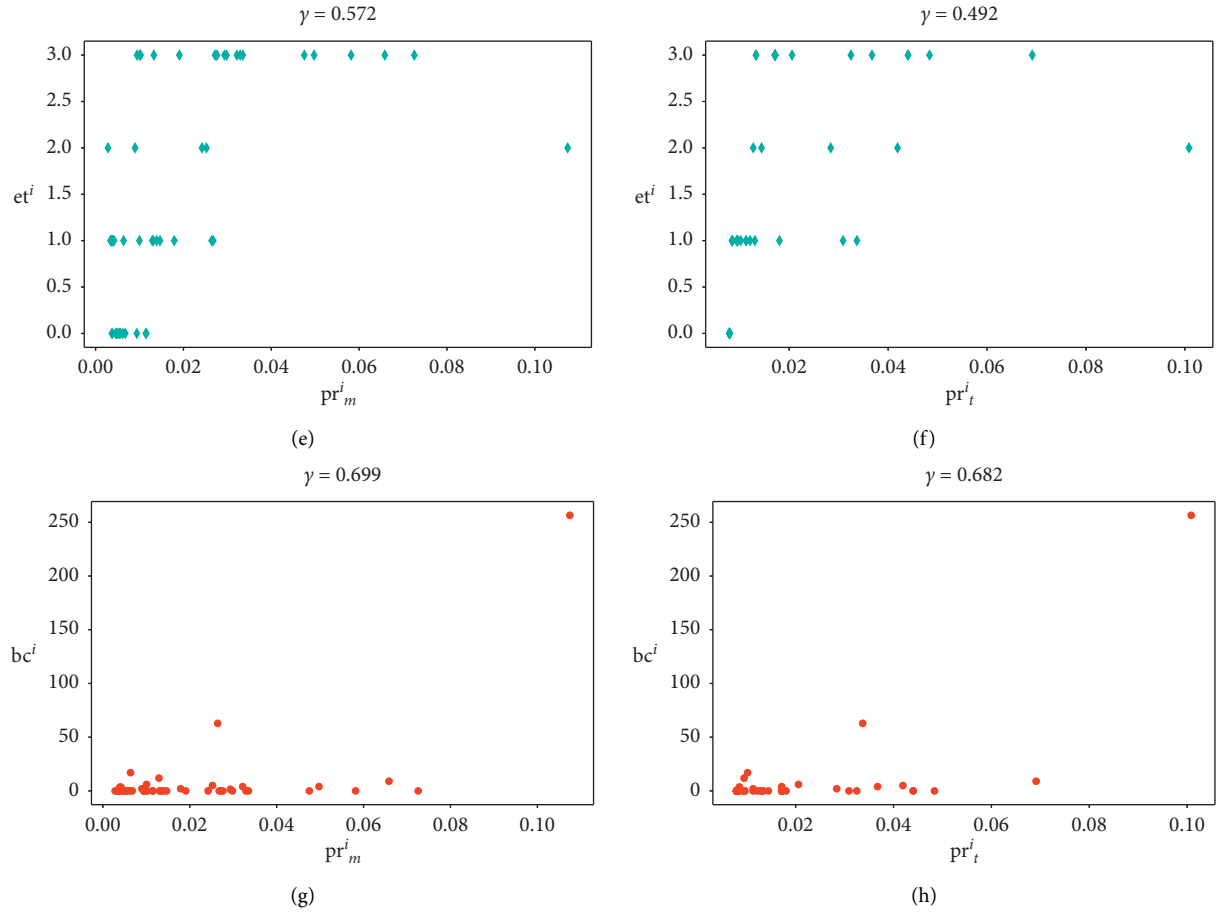

Figure 6: Correlation between two PageRank algorithms and classical measurements. Correlation between (a) pr^i_m prim and dc^i_{in}, (b) pr^i_t prim and dc^i_{in}, (c) pr^i_m prim and ec^i, (d) pr^i_t prim and ec^i, (e) pr^i_m prim and et^i, (f) pr^i_t prim and et^i, (g) pr^i_m prim and bc^i, and (h) pr^i_t prim and bc^i.

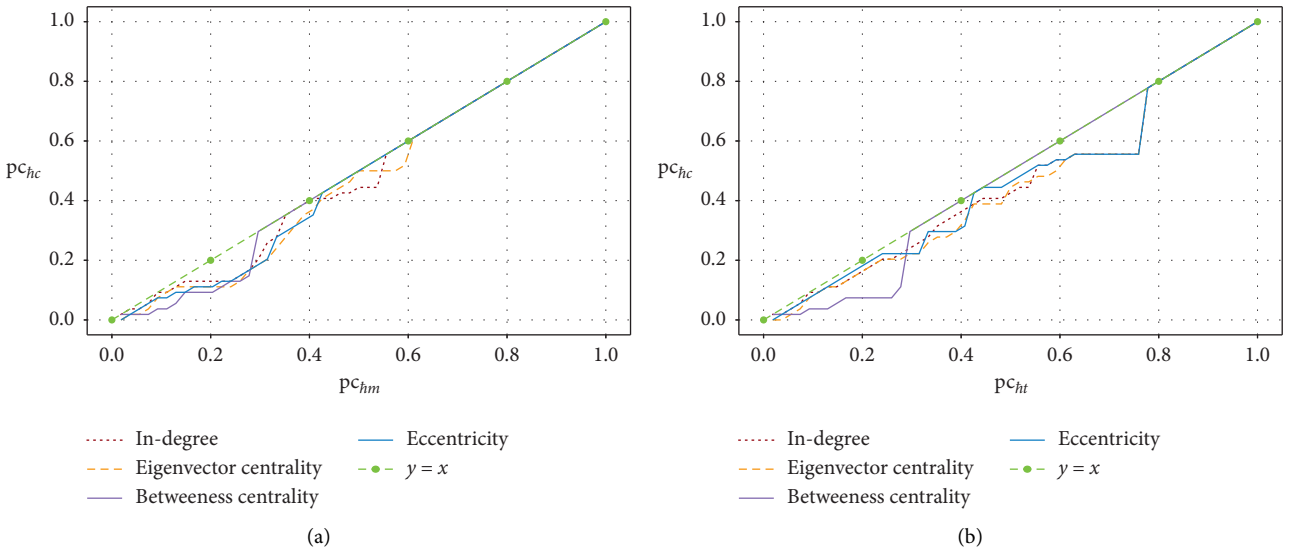

Figure 7: Proportion of common critical standards identified by different classical measurements and each PageRank algorithm. (a) Improved PageRank algorithm. (b) Traditional PageRank algorithm.

Table 5: Mean and variance of identification capability of critical standards of two Pagerank algorithms.

$(\mu_{ic}, \sigma_{ic}^2)$	In-degree centrality	Eigenvector centrality	Betweenness centrality	Eccentricity
Traditional PageRank algorithm	(2.519, 6.934)	(3.000, 7.170)	(1.426, 7.834)	(2.352, 7.440)
Improved PageRank algorithm	(1.259, 3.630)	(1.759, 4.715)	(1.130, 4.568)	(1.370, 4.162)

$(0.0, 0.2, 0.4, 0.6, 0.8, 1.0)$; dotted line $y = x$ represents the perfect fit. Compared with traditional PageRank algorithm, our improved PageRank algorithm does not show obvious advantage in proportion $(0.0, 0.6)$, while behaves much better in proportion $(0, 6, 1.0)$.

To quantify the distance between the relationship curve between pc_{hm} and pc_{hc} and $y = x$, we here define the two measurements of identification capability: the mean and variance of discrete deviations, μ_{ic} and σ_{ic}^2, which are represented as

$$\mu_{ic} = \frac{\sum_{n_{\mathrm{cs}}=1}^{|N_s|} |n_{\mathrm{cs}} - n_{\mathrm{cs}}^{\chi}|}{|N_s|}, \tag{9}$$

$$\sigma_{ic}^2 = \frac{\sum_{n_{\mathrm{cs}}=1}^{|N_s|} \left(|n_{\mathrm{cs}} - n_{\mathrm{cs}}^{\chi}| - n_{\mathrm{cs}}\right)^2}{|N_s|}. \tag{10}$$

A PageRank algorithm with lower μ_{ic} is associated with stronger identification capability, regardless of the predefined proportion of critical standards pc_{hm}. Further, σ_{ic}^2 is associated with the stability or the variability of a measurement. The computational results of μ_{ic} and σ_{ic}^2 with respect with each class measurement are shown in Table 5. It can be seen that the improved PageRank algorithm is more capable of identifying the critical standards than the traditional PageRank algorithm.

6. Conclusion

This paper aimed to overcome the limitations of traditional PageRank in the context of standard citation networks, by incorporating the TSA. Specifically, an improved PageRank algorithm, which incorporates a different distribution mechanism for PR values, was proposed for the identification of critical standards. To demonstrate the effectiveness of the improved algorithm, we used standard data from the National Health Commission of the People's Republic of China. The verification process was carried out by analyzing the correlations between the two PageRank algorithms and classical measurements. We also defined two mean- and variance-based metrics, in order to evaluate the identification capability by computing the common proportion of critical standards identified by classical measurements and PageRank algorithms simultaneously. Our results show that the improved PageRank algorithm is superior to the traditional PageRank algorithm. Due to the substantial differences between weighted and unweighted networks, the network model is highly valuable in designing measurements for the importance evaluation of standard nodes.

It should be noted that there are some shortcomings in this paper, which deserve further exploration in our ongoing study. In addition to VSM, other TSA algorithms can also be applied to quantify the connectivity of standards; the relevant researches Onan et al. have done can provide great inspiration [35–37, 54]. Meanwhile, other network measurements can be proposed for the weighted standard citation network. A meaningful research direction would be to compare the differences of the results of critical standards between identification obtained by different network models. Finally, the improved PageRank algorithm is based on a small-scale standard citation network. The research idea proposed in this study can be employed to other large-scale standard citation networks in different application areas.

Data Availability

The data used to support this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 71701213).

Supplementary Materials

The source code for the TSA method 180: the source code for the improved PageRank algorithm. (*Supplementary Materials*)

References

[1] J. S. Smolen, G. Steiner, and E. M. Tan, "Standards of care: the value and importance of standardization," *Arthritis & Rheumatism*, vol. 40, no. 3, pp. 410–412, 1997.

[2] T. E. Elder, "The importance of relative standards in ADHD diagnoses: evidence based on exact birth dates," *Journal of Health Economics*, vol. 29, no. 5, pp. 641–656, 2010.

[3] D. Rae, "Fit for purpose: the importance of quality standards in the cultivation and use of live plant collections for conservation," *Biodiversity and Conservation*, vol. 20, no. 2, pp. 241–258, 2011.

[4] S. C. Kirshblum, F. Biering-Sorensen, R. Betz et al., "International standards for neurological classification of spinal cord injury: cases with classification challenges," *The Journal of Spinal Cord Medicine*, vol. 37, no. 2, pp. 120–127, 2014.

[5] G. Fuchsbauer, J. Katz, and D. Naccache, "Efficient rational secret sharing in standard communication networks," in *Proceedings of the Theory of Cryptography Conference*, Zurich, Switzerland, February 2010.

[6] J. Wu, Z. Gao, H. Sun, and H. Huang, "Urban transit system as a scale-free network," *Modern Physics Letters B*, vol. 18, pp. 1043–1049, 2004.

[7] T. L. Saaty, "What is the analytic hierarchy process?" in *Mathematical Models for Decision Support*Springer, Berlin, Germany, 1988.

[8] T. L. Saaty and L. G. Vargas, *Decision Making with Dependence and Feedback: The Analytic Network Process*, Springer, Berlin, Germany, 2006.

[9] H.-S. Shih, H.-J. Shyur, and E. S. Lee, "An extension of topsis for group decision making," *Mathematical and Computer Modelling*, vol. 45, no. 7-8, pp. 801–813, 2007.

[10] Y. Wei, F. Chen, H. Xue, and L. Wang, "Research on structural dynamics in Chinese automobile standard citation network," *Neural Computing and Applications*, vol. 32, no. 1, pp. 31–39, 2018.

[11] Y. Wei, L. Chen, Y. Qi et al., "A complex network method in criticality evaluation of air quality standards," *Sustainability*, vol. 11, no. 14, p. 3920, 2019.

[12] R. S. Wills, "Google's Pagerank," *The Mathematical Intelligencer*, vol. 28, no. 4, pp. 6–11, 2006.

[13] D. Fiala, "Time-aware Pagerank for bibliographic networks," *Journal of Informetrics*, vol. 6, no. 3, pp. 370–388, 2012.

[14] F. A. Massucci and D. Docampo, "Measuring the academic reputation through citation networks via Pagerank," *Journal of Informetrics*, vol. 13, no. 1, pp. 185–201, 2019.

[15] Y. Ding, E. Yan, A. Frazho, and J. Caverlee, "Pagerank for ranking authors in co-citation networks," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2229–2243, 2009.

[16] P. Bruck, I. Réthy, J. Szente, J. Tobochnik, and P. Érdi, "Recognition of emerging technology trends: class-selective study of citations in the U.S. patent citation network," *Scientometrics*, vol. 107, no. 3, pp. 1465–1475, 2016.

[17] E. Garfield, "Citation indexes for science: a new dimension in documentation through association of ideas," *International Journal of Epidemiology*, vol. 35, no. 5, pp. 1123–1127, 2006.

[18] H. Small, "Co-citation in the scientific literature: a new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[19] P. Doreian, "A measure of standing of journals in stratified networks," *Scientometrics*, vol. 8, no. 5-6, pp. 341–363, 1985.

[20] C. Calero-Medina and E. C. M. Noyons, "Combining mapping and citation network analysis for a better understanding of the scientific development: the case of the absorptive capacity field," *Journal of Informetrics*, vol. 2, no. 4, pp. 272–279, 2008.

[21] I. Fister Jr., I. Fister, and M. Perc, "Toward the discovery of citation cartels in citation networks," *Frontiers in Physics*, vol. 4, p. 49, 2016.

[22] T. Kuhn, M. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," *Physical Review X*, vol. 4, no. 4, Article ID 041036, 2014.

[23] A. P. Singh, K. Shubhankar, and V. Pudi, "An efficient algorithm for ranking research papers based on citation network," in *Proceedings of the 2011 3rd Conference on Data Mining and Optimization (DMO)*, Selangor, Malaysia, June 2011.

[24] H. Qiao, Y. Wang, and Y. Liang, "A value evaluation method for papers based on improved pagerank algorithm," in *Proceedings of the 2012 2nd International Conference on Computer Science and Network Technology*, Changchun, China, December 2012.

[25] J. A. García, R. Rodriguez-Sánchez, and J. Fdez-Valdivia, "Social impact of scholarly articles in a citation network," *Journal of the Association for Information Science and Technology*, vol. 66, no. 1, pp. 117–127, 2015.

[26] F. Narin, "Patent bibliometrics," *Scientometrics*, vol. 30, no. 1, pp. 147–155, 1994.

[27] X. Li, H. Chen, Z. Huang, and M. C. Roco, "Patent citation network in nanotechnology (1976–2004)," *Journal of Nanoparticle Research*, vol. 9, no. 3, pp. 337–352, 2007.

[28] T.-S. Cho and H.-Y. Shih, "Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008: 2011–2008," *Scientometrics*, vol. 89, no. 3, pp. 795–811, 2011.

[29] K. Fujita, Y. Kajikawa, J. Mori, and I. Sakata, "Detecting research fronts using different types of weighted citation networks," *Journal of Engineering and Technology Management*, vol. 32, pp. 129–146, 2014.

[30] H. Beltz, A. Fülöp, R. R. Wadhwa, and P. Érdi, "From ranking and clustering of evolving networks to patent citation analysis," in *Proceedings of the 2017 International Joint, Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA, May 2017.

[31] P. R. Portney, I. W. H. Parry, H. K. Gruenspecht, and W. Harrington, "Policy watch: the economics of fuel economy standards," *Journal of Economic Perspectives*, vol. 17, no. 4, p. 203, 2003.

[32] S. Bergström, "Value standards in sub-sustainable development. on limits of ecological economics," *Ecological Economics*, vol. 7, no. 1, pp. 1–18, 1993.

[33] R. K. Chellappa and S. Shivendu, "Economics of technology standards: implications for offline movie piracy in a global context," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, Big Island, HI, USA, January 2003.

[34] S. Buckles and M. Watts, "National standards in economics, history, social studies, civics, and geography: complementarities, competition, or peaceful coexistence?" *The Journal of Economic Education*, vol. 29, no. 2, pp. 157–166, 1998.

[35] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[36] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, no. 2, pp. 330–348, 2017.

[37] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, pp. 1–16, 2016.

[38] G. Salton, "Automatic text analysis," *Science*, vol. 168, no. 3929, pp. 335–343, 1970.

[39] A. Dong and A. M. Agogino, "Text analysis for constructing design representations," *Artificial Intelligence in Engineering*, vol. 11, no. 2, pp. 65–75, 1997.

[40] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, vol. 43, no. 5, pp. 1216–1247, 2007.

[41] S. Poomagal and T. Hamsapriya, "Cosine similarity-based Pagerank calculation," *International Journal of Web Science*, vol. 1, no. 1/2, pp. 142–159, 2011.

[42] S. Hatakenaka and T. Miura, "Ranking documents using similarity-based pageranks," in *Proceedings of the 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, BC, Canada, August 2011.

[43] A. Aizawa, "An information-theoretic perspective of TF-IDF measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[44] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1–37, 2008.

[45] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[46] S. Brin and L. Page, "Reprint of: the anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[47] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with Google's PageRank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[48] D. Chen, X. Ren, Q. Zhang, Y. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Physics Reports*, vol. 650, pp. 1–63, 2016.

[49] S. Gago, J. Hurajová, and T. Madaras, "Notes on the betweenness centrality of a graph," *Mathematica Slovaca*, vol. 62, no. 1, pp. 1–12, 2012.

[50] P. Hage and F. Harary, "Eccentricity and centrality in networks," *Social Networks*, vol. 17, no. 1, pp. 57–63, 1995.

[51] B. Ruhnau, "Eigenvector-centrality-a node-centrality?" *Social Networks*, vol. 22, no. 4, pp. 357–365, 2000.

[52] L. Solá, M. Romance, R. Criado, J. Flores, A. García del Amo, and S. Boccaletti, "Eigenvector centrality of nodes in multiplex networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 23, no. 3, Article ID 033131, 2013.

[53] L. Leydesdorff, "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 9, pp. 1303–1319, 2007.

[54] A. Onan, "Mining opinions from instructor evaluation reviews: a deep learning approach," *Computer Applications in Engineering Education*, vol. 28, no. 1, pp. 117–138, 2020.