Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(10): 18104–18122.
DOI: 10.3934/mbe.2023804
Received: 09 July 2023
Revised: 23 August 2023
Accepted: 05 September 2023
Published: 21 September 2023

Research article

A multi-dimension information fusion-based intelligent prediction approach for health literacy

Xiaoyan Zhao[1,2]*** and Sanqing Ding**[1]

[1] School of Public Policy and Management, China University of Mining and Technology, Xuzhou 221116, China

[2] School of Marxism, Bengbu Medical College, Bengbu 233030, China

* **Correspondence:** Email: zhaoxy2001@163.com.

Abstract: Health literacy refers to the ability of individuals to obtain and understand health information and use it to maintain and promote their own health. This paper manages to make predictions toward its development degree in society with use of a big data-driven statistical learning method. Actually, such results can be analyzed by discovering latent rules from massive public textual contents. As a result, this paper proposes a deep information fusion-based smart prediction approach for health literacy. Specifically, the latent Dirichlet allocation (LDA) and convolutional neural network (CNN) structures are utilized as the basic backbone to understand semantic features of textual contents. The feature learning results of LDA and CNN can be then mapped into prediction results via following multi-dimension computing structures. After constructing the CNN model, we can input health information into the model for feature extraction. The CNN model can automatically learn valuable features from raw health information through multi-layer convolution and pooling operations. These characteristics may include lifestyle habits, physiological indicators, biochemical indicators, etc., reflecting the patient's health status and disease risk. After extracting features, we can train the CNN model through a training set and evaluate the performance of the model using a test set. The goal of this step is to optimize the parameters of the model so that it can accurately predict health information. We can use common evaluation indicators such as accuracy, precision, recall, etc. to evaluate the performance of the model. At last, some simulation experiments are conducted on real-world data collected from famous international universities. The case study analyzes health literacy difference between China of developed countries. Some prediction results can be obtained from the case study. The proposed approach can be proved effective from the discussion of prediction results.

Keywords: multi-dimension information fusion; intelligent prediction; health literacy; textual mining

1. Introduction

Health information literacy (HIL) generally refers to individuals' access to, understanding and use of health information and services [1]. It covers a range of skills and knowledge, including understanding the acquisition, evaluation, and application of health information, as well as using this information to make correct decisions [2]. This literacy also includes understanding the limitations and potential risks of health information, as well as seeking professional medical advice when necessary [3]. The status of HIL among different groups and the relationship between its influencing factors is an important exploration field [4]. In real life, different populations may have differences in accessing, understanding, evaluating and using health information, which may be influenced by various factors. Individuals of different age groups may have differences in accessing and using health information. The level of education may affect an individual's ability to understand, evaluate and utilize health information. People with higher levels of education may be more likely to understand complex health information, while those with lower levels of education may face more difficulties. Differences in health status may also affect individuals' ability to access and use health information. With the integration of information technology in various fields, some scholars have proposed HIL [5]. It refers to the ability of users to identify health information needs, obtain health information from reliable information sources and make reasonable health decisions. In this high-tech era, where the amount of information is growing rapidly and the value of information is rising infinitely, few have mastered the methods and skills of information acquisition. And those who have obtained high value-added scientific and technological information will have the first opportunity to take the initiative and play a leading role in social competition [6,7]. Almost all universities in China have carried out various forms of information literacy education, and also conducted unremitting research and discussion on information literacy education [8]. Therefore, the evaluation research of HIL is also rising. European and American countries have developed many HIL assessment tools suitable for their own citizens according to their own language, culture, socio-economic conditions and medical systems [9]. In particular, HIL evaluation tools developed and applied from a clinical perspective have been quite mature [10].

The cognitive strategy of health information literacy refers to learners' processing of information in the current task [11]. At the same time, this cognitive strategy needs to combine the activities and steps taken in the process of completing the learning task [12]. In order to improve the awareness of the use of learning strategies for health information literacy, it is necessary to provide a variety of authentic learning materials [13]. In addition, the improvement of ability does not lie in the number of strategies used, but in the flexible and rational use of strategies [14]. Because the strategy itself is a prior strategy, it is difficult to ensure that the application process has complete matching and novelty to the recipient [15]. The designer has taken into account the background of the object, the relevant environment and the possible results. But after all, health activities are mainly activities between human bodies [16]. It is difficult for designers to fully estimate the problems in the policy application process [17].

For example, problems that can be solved under the original cognitive level may become difficult due to some interference [18]. It may also become too simple due to some inspiration. The teaching process is a process of guiding and encouraging students to grow and develop continuously [19]. There is a relative frame of reference for the level of cognition and the advantages and disadvantages of cognitive strategies [20]. Students who are weak in some aspects may have advantages in other

aspects [21–23]. Students with low passion for learning at one stage may show higher enthusiasm at another stage. Therefore, teaching strategies must be adjusted timely and flexibly in addition to accurate settings [24]. As the society moves towards the information age, the network, as an important carrier of big data, provides people with rich, novel, rapidly updated and diversified information, making it easier for the public to obtain information [25]. However, compared with the research in the field of semi physical teaching abroad, the research in China has just started, and there is less research on this group of college students [26]. Therefore, exploring the relationship between the HIL status quo of this group and the influencing factors will help to deeply understand the characteristics and laws of this group, and have far-reaching significance in improving the HIL level of college students. Therefore, the research on HIL evaluation will help China grasp the correct research direction, integrate with international standards as soon as possible, and provide an important reference for the development of localized HIL evaluation tools suitable for China.

The research on the evaluation of HIL in this paper will help improve understanding of the essence and influencing factors of HIL, and further explore the relationship between HIL and health status. By evaluating HIL, we can better understand individuals' ability to acquire, understand and apply health information, thereby providing them with more precise and personalized health services and interventions. Research contribution points:

- This paper proposes an intelligent prediction method of health literacy based on deep information fusion. Specifically, potential latent dirichlet assignment (LDA) and convolutional neural network (CNN) structures are used as the basic framework for understanding the semantic features of text content.
- This study will fill the gap in the field of HIL evaluation by systematically sorting and performing in-depth analysis of the evaluation methods of HIL, providing a reference for subsequent research.
- This study will also propose HIL evaluation schemes targeting different populations and scenarios through the evaluation and comparison of existing evaluation tools, providing guidance for practical applications.

2. Research method

2.1. Topic modeling method

In this information age, people's life and study are full of all kinds of information, and information literacy is an essential skill for life and study in the era of big data. With the rapid changes of science and social environment, people's living standard and ideology have changed, people's awareness and requirements for health have been constantly improved and their attention to health has also been increased. However, food safety and medical problems caused by low-quality health information have gradually increased. The authors argue that HIL is a subset of information literacy. At the same time, HIL is one of the important influencing factors of health literacy. Improving HIL is of great benefit to the improvement of public information literacy and health literacy.

With the gradual deepening of HIL research, scholars began to realize that the design of HIL evaluation tools for all residents may lack consideration of population specificity. Therefore, scholars have gradually realized that HIL evaluation should be conducted for different groups of people. A topic model is a set of machine learning models that try to find potential topic structures in massive documents [27, 28]. Before the topic modeling of the text, in order to save the storage space of the text

and improve the retrieval efficiency when the model runs, it is necessary to filter out some meaningless words in advance to shorten the text. For example, search words that appear in every article, verbs and nouns that have no actual meaning but appear many times in the results, etc. By integrating the above words, we can get a stoplist of specific knowledge fields.

After text preprocessing, every word needs to calculate its TF-IDF (term frequency-inverse document frequency) weight. For the word w_i in document d_j, its T F can be expressed as:

$$TF_{i,j} = \frac{n_{i,j}}{\sum kn_{k,j}.} \tag{2.1}$$

In the above formula, the numerator $n_{i,j}$ represents the number of times that the word w_i appears in the document d_j, and the denominator $\sum kn_{k,j}$ represents the sum of the number of times that all words appear in the document d_j.The LDA model is a typical Bayesian network structure, which defines that every document is a random mixture of hidden topics, and hidden topics are randomly composed of feature words with a certain probability. Figure 1 shows the LDA probability model, which is divided into document collection layer, document layer and feature word layer, and each layer is controlled by random variables. The Z represents potential themes, and w represents feature words.

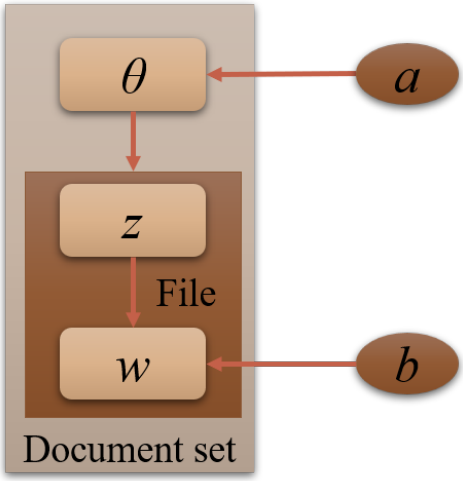

Figure 1. LDA probability model diagram.

Vector a and matrix b define the document set level. The vector a defines the relative strength of potential hidden topics in a document set, the matrix b represents the probability distribution of potential hidden topics in a document set and the element $b_{i,j}$ represents the probability that the j th feature word belongs to the i th hidden topic. According to the above process, the generation probability of the i th feature word w_i in document d is:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j). \tag{2.2}$$

In the formula, $P(w_i \mid z_i = j)$ represents the probability that the feature word w_i comes from the potential topic z_i, and $P(z_i = j)$ represents the probability that the document contains the topic z_i.

Before the advent of the neural network language model, the commonly used model was the n-gram language model trained by natural language networks. The basic idea of this model is: the probability of a word appearing in a text is only related to its first words, and the probability of the word appearing is based on the probability of the words before it [29]. The complexity of parameter selection may increase as the dimension of the input word vector increases. Higher dimensional input word vectors may also lead to overfitting problems, which require more complex regularization techniques to handle. When using input word vectors with higher dimensions, mapping features may become more rich and complex. This complexity is because high-dimensional input word vectors contain more semantic information, which can better capture the subtle differences and semantic relationships between words. This may lead to the model mapping input to output more accurately and producing richer and more accurate prediction results. The sentence T can be expressed as $T = (w_1, \Lambda, w_n)$, where w_i represents the i th word in the sentence T, then the probability of the sentence T appearing in the text can be calculated as:

$$P(T) = \prod^{N} P\left(w_i / w_{i-n-1}, \Lambda, w_{i-1}\right). \tag{2.3}$$

This not only reduces the network parameters, but also reduces the complexity of parameter selection, thus making the feature mapping unique and unchangeable. The problem of converting a one-to-many mapping to a one-to-two mapping usually involves classification, which involves dividing an object into one of two categories based on multiple features or attributes. Set a threshold based on the distribution of features or attributes to classify objects that are greater or less than this threshold into different categories. This method is simple and easy to implement. When converting a one-to-many mapping to a one-to-two mapping, there may be a problem of category imbalance, where the number of samples in certain categories is much larger than in others. In this case, it is necessary to adopt some special processing methods, such as Oversampling, undersampling, synthetic minority Oversampling and other methods to improve the classification performance. These advantages will be especially obvious when the input word vector with higher dimension is used.

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \cdot k_{ij}^l + b_j^{l,}\right) \tag{2.4}$$

where x_i^{l-1} is the input characteristic, x_j^l is the output characteristic, k_{ij}^l is the weight, b_j^l is the offset and $f(\cdot)$ is the activation function. According to mathematical analysis, text classification is a process of mapping unknown categories of texts to predefined categories. This mapping can be one-to-one, one-to-two or one-to-many. One-to-many mapping is usually converted into one-to-two mapping problems for analysis.

Then, there is a relationship W between the new text set and the predefined category, which is expressed as:

$$W : I \rightarrow J(5). \tag{2.5}$$

For the document i in $I, W(i)$ is known information, and the text set can be processed by the guided training of the text classification algorithm, and a text classification model R similar to I can be obtained.R is expressed as:

$$R : I \rightarrow J(6). \tag{2.6}$$

This can not only give full play to the advantages of automatic disambiguation and word segmentation by string frequency statistics, but also give full play to the advantages of high efficiency and fast segmentation speed of string matching [30, 31].

Expected cross entropy is the amount of information obtained when a certain feature word appears in the text. Given the feature t and text category c_i, it can be expressed as the distance between $P(C_i \mid t)$ and $P(c_i)$ to represent the value of ECE. The calculation formula is shown as:

$$\mathrm{ECE}(t) = P(t) \sum_{i=1}^{n} P(c_i \mid t) \log \frac{P(c_i \mid t).}{P(c_i)} \tag{2.7}$$

The greater the ECE(t), the greater the influence of feature t on classification. Assuming that the feature t is strongly correlated with the category c_i, then the $P(c_i \mid t)$ value is large, and when the $P(c_i)$ value is small, the feature t has greater influence on classification.

2.2. Automatic document classification program design

This paper selects noun morphemes and specialized noun morphemes for future word segmentation. After screening out other part-of-speech words, there are two main steps to be done next: synonym merging and name recognition. Because the features that appear in too few documents are not universal and can't have a good recognition ability for this category of articles, the words in this case are treated as noise words. According to the current research, there is no universally applicable method to determine this value. Generally, a large number of experiments are used to determine the threshold value, and their effects are used to measure the judgment. In this paper, define acronym method is used in the initial stage of feature dimensionality reduction, in order to reduce a large number of feature sets for noise removal.

Considering that Chinese itself lacks word signs and strict rules, the existing rules of morphology, syntax and combination in the language field are still very general and complicated. Therefore, the words appearing in this page description are more closely related to the article category than the words in the text [32]. In this paper, LDA topic model is applied to the field of text processing, mainly for text similarity calculation, and then the text clustering and text recommendation algorithms are improved. Then the contribution of each keyword in the text is calculated, and TF-IDF method is mostly used to calculate the contribution. This method takes into account the semantics contained in the text and the semantics of each keyword, and avoids the above ambiguity.

The text vector of the potential topic based on LDA is $d = (z_1, \Lambda, z_n)$. T is the number of potential topics. The calculation method of text similarity based on LDA potential topic vector is shown as:

$$\mathrm{Sim}\left(d_i, d_j\right) = \frac{d_i * d_j}{|d_i| * \left|d_j\right|}. \tag{2.8}$$

When we choose the index feature vector to describe the sample, in order to achieve the purpose of no omission, we often describe a property with different names for many times, which will result in overlapping information. According to the domain knowledge or the method of feature variable clustering, we choose the appropriate feature variable set. Or use the following Mahalanobis distance:

$$D(X, Y) = (X - Y)^T S^{(} - 1)(X - Y). \tag{2.9}$$

where S is the covariance matrix of sample matrix A , and X and Y are the covariance estimators of population distribution.

Mahalanobis distance is the improvement of Ming's distance, which is invariant to all linear transformations, and overcomes the disadvantage that Ming's distance is influenced by dimensions. Mahalanobis distance also partially overcomes multiple correlations. The significance of the rough subtraction method proposed in this article and the traditional unsupervised disambiguation algorithm lies in providing a comparison and reference for subsequent experiments. Rough subtraction is an algorithm based on rough set theory that can be used to process imprecise or uncertain data, while traditional unsupervised disambiguation algorithms are based on clustering or classification methods aimed at eliminating ambiguity and uncertainty in the data. The comparison between these two algorithms and CNN can provide researchers with the performance and effect of different algorithms in dealing with health information literacy evaluation problems, and help to deeply understand the advantages and disadvantages of various algorithms and applicable scenarios.

The core idea of rough subtraction is to effectively achieve the definition of system dimension reduction rough set without reducing the classification ability of the system, and the introduction of the approximation concept has brought many advantages. It can manipulate large-scale data, and such data can be inaccurate or ambiguous. Through the derivation of some theories of upper approximation and lower approximation, rough set can obtain the minimum expression of knowledge, which is the theoretical basis of knowledge reduction by rough set. The process is shown in Figure 2.

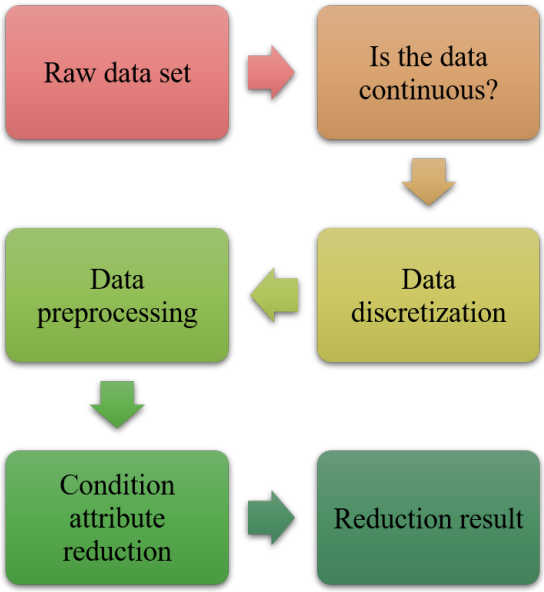

Figure 2. Rough set simplification process.

To use rough set reduction, we need to establish a text model first. Here, this paper chooses to use the Boolean model, which is easy to model, and has very good adaptability to the application fields prepared by this program. Its data requirements are strict and its application scope is limited. However, this simple discretization method has a very good discretization effect when the data distribution

is concentrated and the noise data is little. Then, according to the set number of intervals, the equidistant division is carried out. This method is widely used, and the discretization effect is very good when the data distribution is even. Traditional unsupervised disambiguation algorithms usually use co-occurrence rate to achieve rapid and relatively accurate disambiguation. The co-occurrence rate of words refers to the frequency of two words appearing together in an article. This is used as the basis of disambiguation. For example, if the correct meaning and all wrong meanings of polysemous words do not appear in the full text of the article, the co-occurrence rate of all meanings will be almost equal. It is difficult to solve the ambiguity problem of polysemous words by using co-occurrence rate. The calculation formula of the present rate is shown as:

$$T(w_1, w_2) = \log_2\left(\frac{\sum_{s=1} p(w_1, w_2),}{p(w_1) \cdot p(w_2)}\right) \tag{2.10}$$

where w_1 is a polysemous word and w_2 is a definition of polysemous word. s is the scope of co-occurrence rate, usually the whole statement in which w_1 is located. $p(w)$ indicates the frequency of the word w in the whole text.

In the word forest, word coding is used to calculate similarity and merge. The similarity calculation formula is shown as:

$$S(k_1, k_2) = \frac{\alpha}{\mathrm{Dis}(k_1, k_2).} \tag{2.11}$$

The numerator in the formula is a constant, and the denominator represents the distance between words k_1, k_2. The closer the semantic distance is, the smaller the denominator is, and the greater the calculated similarity is. People generally understand that literature refers to the sum of books, periodicals, papers and other texts that record knowledge. When clustering papers and documents, the biggest difference from clustering data in traditional databases is that the data in traditional databases are structured data, while the text is unstructured data. Regardless of the purpose and means of text mining, the process of text preprocessing includes two basic steps: word segmentation and stop words removal. Figure 3 shows the specific framework of literature analysis based on CNN.

CNN model is a multi-layer neural network model. Each layer of the model is made up of multiple two-dimensional planes, and each plane is made up of multiple independent neurons. Input to the full connection layer to get the final output. The upper layer is usually the full connection layer used as a classifier. In this way, each layer of the convolution upgrade network can obtain the most significant features of the data through digital filters. In the process of back propagation, the weight of the network is adjusted by the error between the actual output and the expected data. After that, the errors of other layers are adjusted by going forward layer by layer. The output of each unit of the hidden layer is:

$$y_k = f\left(\sum_{j=0}^{L-1} V_{ij} h_j + \theta_k\right)^{\cdot} \tag{2.12}$$

The output of each unit of the output layer is:

$$y_k = f\left(\sum_{j=0}^{L-1} W_{ij} h_j + \varphi_j\right)^{\cdot} \tag{2.13}$$

where V_{ij} represents the weights of input layer information i to hidden layer output information j, W_{ij} represents the weights of hidden layer output information j to output layer information k, and θ_k, φ_j are

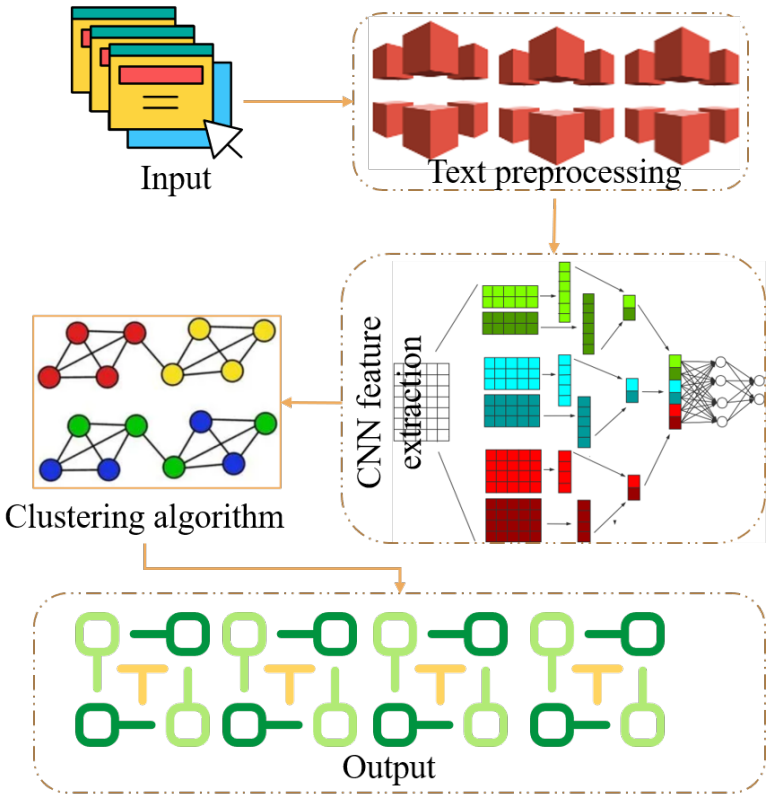

Figure 3. CNN-based document analysis framework.

used to represent the thresholds of output units and hidden layer units, respectively. $f()$ is the activation function.

In feature selection, the degree of independence between feature t and topic class C can be counted by χ^2. The calculation formula is shown as:

$$\chi^2(t,c) = \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D),} \tag{2.14}$$

where A represents the number of documents belonging to category C and containing feature t, B represents the number of documents not belonging to category C but containing feature t, C represents the number of documents belonging to category C among documents not containing feature t, D represents the number of documents not belonging to category C among documents not containing feature t and N represents the total number of training text sets.

The correlation calculation formula is shown as:

$$\text{CoherenceScore } s_{s_j} = \frac{\sum_{i=1,j\neq 1}^{|s|} \left(s_i, s_j.\right)}{|s-1|} \tag{2.15}$$

In the process of iteration, when the change of the center point is less than β_1 , the whole cluster is added to the selected data set and deleted from the sample set, so that only the samples that have not been correctly identified are retained in the original sample data set. The formula for calculating the

change of the center point is:

$$\beta_r = \frac{1}{|T_i|} \sum_{a_i \in T_{r,i}} a_i - \frac{1}{|T_{i-1}|} \sum_{a_j \in T_{r-i,j}} a_j, \tag{2.16}$$

where r is the number of iterations of the algorithm, and $T_{r,i}$ represents the i_{th} category of the r iteration. When $\beta_r \leq \beta_1$ is used, the conditions are met, and other samples are screened until all sample data are correctly identified.

2.3. Multidimensional cognitive information implementation

Compared with the early cognitive radio technology, the biggest difference between cognitive network and cognitive radio technology is that the object of perception and management operations has changed. As the evolution of cognitive radio network, cognitive network is not limited to spectrum resources. In order to meet the end-to-end decision objectives, cognitive networks need to manage and reconfigure the entire network, which means that cognitive networks are managed for the entire network. Therefore, for cognitive network, all the links and factors that can affect the communication target in the whole network are the objects that it perceives, analyzes, manages and configures. The concept of multi-dimensionality needs to be introduced here. From a macro perspective, the multi-dimensional nature of resources refers to the diversity of cognitive network resources. From a micro perspective, the multidimensional nature of resources can also be understood as that for each resource, its performance can be described from multiple perspectives, that is, the diversity of feature parameters. This chapter first introduces the resource analysis process of cognitive networks, and then conducts multi-dimensional representation for different resources.

Multidimensional cognitive information is a quantitative analysis method based on mathematical statistics, which studies the external characteristics of literature. Content analysis is a qualitative method to study the content of literature. In this paper, the evolution of information literacy education research is quantitatively analyzed based on multidimensional cognitive informations, and the content analysis is carried out by combining qualitative analysis with the distribution of information literacy education topics. Using the multidimensional cognitive information analysis software, this paper analyzes the HIL field literature in the Web of Science database from 2010 to 2020 from six dimensions: year, country, author, research institution, keywords and citations. On this basis, the academic level in the field of HIL is studied, and the present situation and law of HIL development are explored.

The common functions of computer-aided multidimensional cognitive information tools are to realize bibliographic information statistics, generate co-occurrence matrices, carry out cluster analysis and network analysis, etc. Some tools can directly realize the visualization of metrological results. CiteSpace is selected as the research tool in this paper. CiteSpace has a built-in data converter, which can process Chinese and English data, and integrates the functions of multidimensional cognitive informations and visual analysis, and supports network analysis of authors, institutions or countries, network analysis of co-occurrence of topics, keywords or disciplines, co-citation analysis of documents, authors or journals, and literature coupling analysis. The fit of research purposes, the degree of resource acquisition and utilization, and the validity of research methods are the main reasons why Citespace is chosen as an auxiliary tool for multidimensional cognitive information analysis.

This study not only uses various multidimensional cognitive information indicators to reveal the characteristics of information ecology, but also combines information visualization techniques such as

social network analysis and scientific knowledge mapping to vividly outline the development trend of information and ecology. So, the most important premise of drawing a subject knowledge map is to construct the co-occurrence matrix of some kind of data.

3. Analysis and discussion of results

Using Bicomb 2.0 to count the literature by time, from 2010 to 2020, the research papers on information literacy education showed an increasing trend year by year. Relevant results are shown in Figure 4. Although the growth was not obvious before 2011, after 2011, the in-depth research and practice of Web 2.0 promoted the rapid development of Library 2.0, and also made the literature added value of information literacy education in university libraries stable. According to the law of literature growth, the curve in Figure 4 shows that the research on information literacy education is gradually maturing. Table 1 and Figure 5 show the distribution of research hotspots of HIL research papers from 2010 to 2020. Figure 5 shows the distribution of research hotspots in HIL research papers from 2010 to 2020. The research results indicate that the distribution of hot topics in papers is relatively chaotic, with a relatively small number of papers in 2020, with a proportion ranging from 70% to 75%.

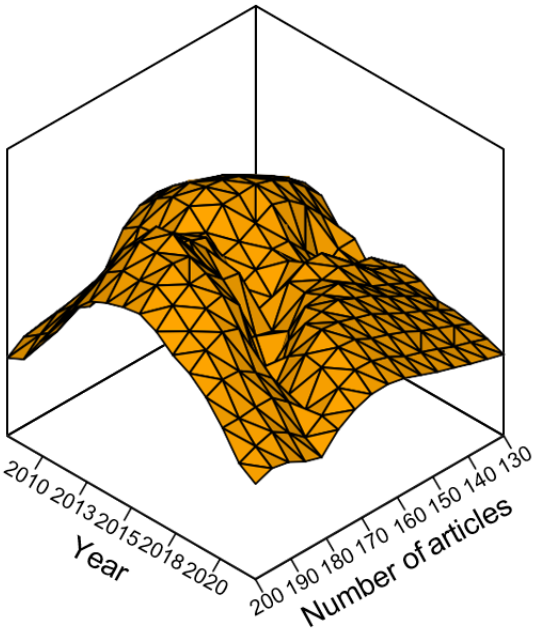

Figure 4. Statistical chart of information literacy education publications from 2010 to 2020.

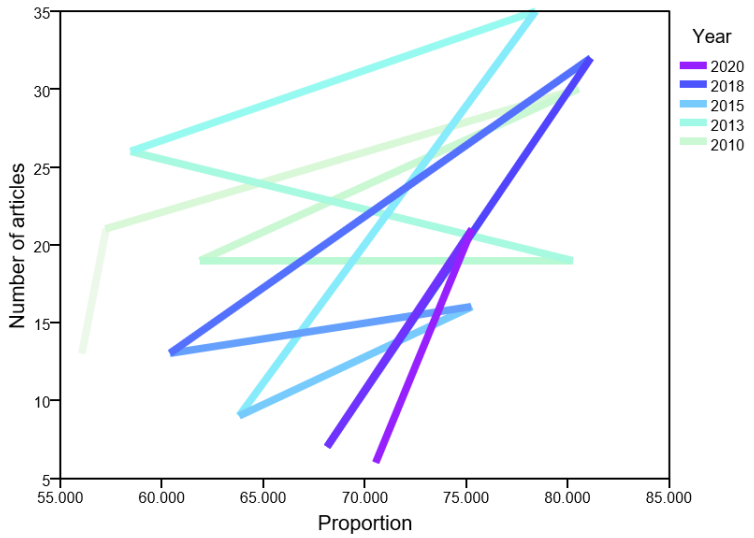

Figure 5. Research hotspot road map.

Table 1. Distribution of research hotspots of HIL research papers in each year.

Year	Research hotspot	Thesis number	Proportion%
2010	Overseas	30	80.541
2011	Domestic	19	61.892
2012	Domestic	19	80.267
2013	Overseas	26	58.448
2014	Overseas	35	78.426
2015	Overseas	9	63.793
2016	Overseas	16	75.266
2017	Overseas	13	60.416
2018	Overseas	32	81.144
2019	Domestic	7	68.155
2020	Domestic	21	75.279

For cognitive network resources, there is no literature to explain them. At the same time, cognitive network resources are far beyond the scope of the cognitive radio system. The analysis of it will be more complex and tedious. In order to describe cognitive network resources in a clear and orderly way, this paper proposes a layer by layer decomposition analysis method based on similar methods, and gradually analyzes and describes cognitive network resources. The resources of a cognitive network are the sum of all perceptible, manageable and operable network components and factors that can affect end-to-end communication objectives in a cognitive network. Based on this positioning of cognitive network resources, taking the network as a whole as the source of resources, the following two-layer decomposition analysis can be carried out: selecting 8 years of research hotspots to establish the corresponding information literacy evalution index system. The research hotspots in seven years are all evaluation criteria, and the research hotspots in the other two years are designated as evaluation

studies because the number of evaluation criteria and related papers in evaluation practice in that year are the same.

From 2010 to 2020, the research literature on HIL was published in many disciplines, among which the top 5 journals are shown in Table 2. On the whole, the research on HIL is published in many journals of library and information science, which shows that libraries are widely involved in HIL research. One of its "Chinese journal of medical library and information science" documents has been cited more than 50 times, which may be the reason for the high total cited frequency. In addition, the number of articles published by library information work is small, but the total frequency of citations from library journals is high. This demonstrates that library information science has a certain influence in the field of HIL research.The country with the largest number of publications is the United States, with a total of 5047 publications, with a center degree of 0.501, as shown in Table 3 and Figure 6.

Table 2. Top 5 journals.

Title	Quantity of documents issued	Total cited frequency
Chinese journal of health education	41	333
Journal of medical informatics	35	225
modern information	15	193
Chinese journal of medical library and information science	11	193
Chinese journal of school health	8	132

Table 3. Ranking of articles issued by each country.

Country	Number of articles	Centrad
United States of America	1729	0.501
Australia	572	0.405
Britain	559	0.469
Canada	557	0.428
Germany	499	0.417
The Netherlands	469	0.397
China	327	0.493
Spain	174	0.509
Sweden	161	0.442

With the rapid development of American technology, the research level of HIL is in the leading position, and it has become the object of competing cooperation among countries, forming a self-centered cooperation network. In contrast, although China ranks high in the number of published articles in the field of HIL, the number of published articles is only 18.9% of that in the United States, with a centrality of 0.493, which shows that the number of research achievements and cooperation in the field of HIL in China are poor. If a certain keyword appears repeatedly in its research field within a certain period of time, or the number of literatures on a certain represented topic suddenly increases, this topic may become a research hotspot within this period of time. Figure 7 is the keyword knowledge map formed after software analysis.

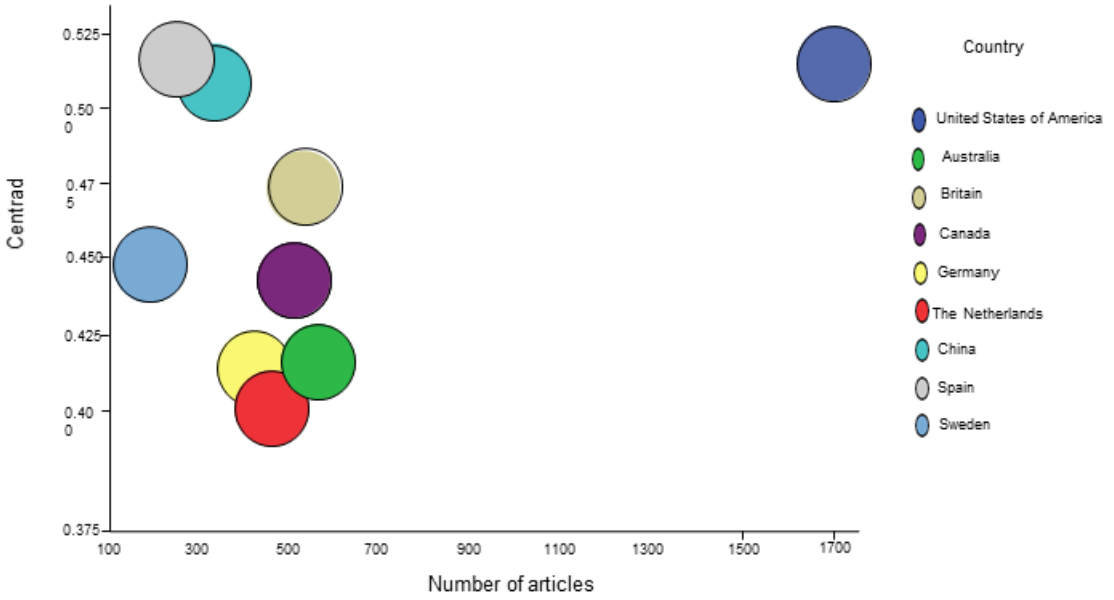

Figure 6. Discrete map of the number of articles issued by each country.

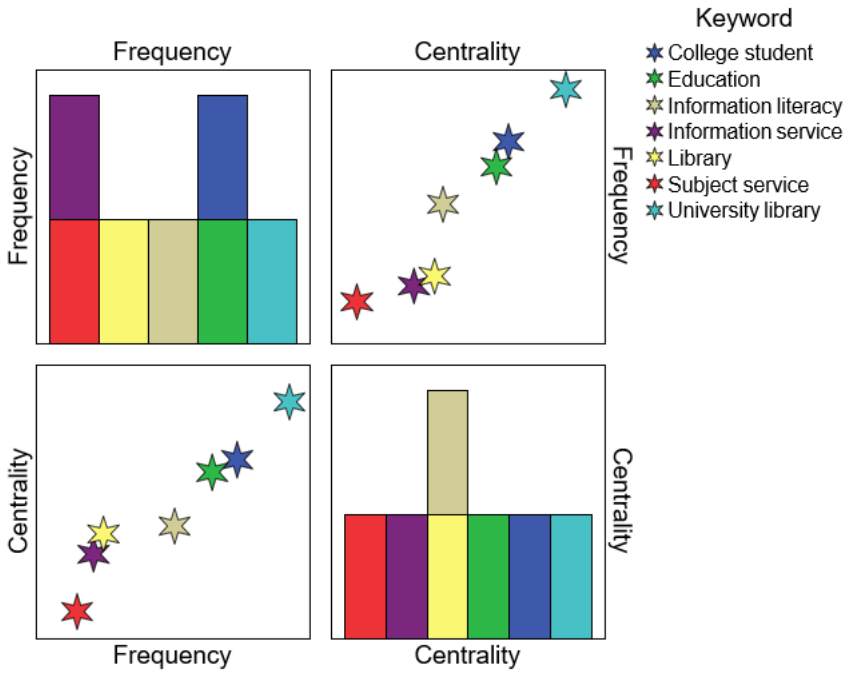

Figure 7. Keywords centrality of knowledge map.

The important link of information literacy education is the cultivation of information acquisition ability, and the cultivation of this ability is mainly reflected in the understanding of information sources,

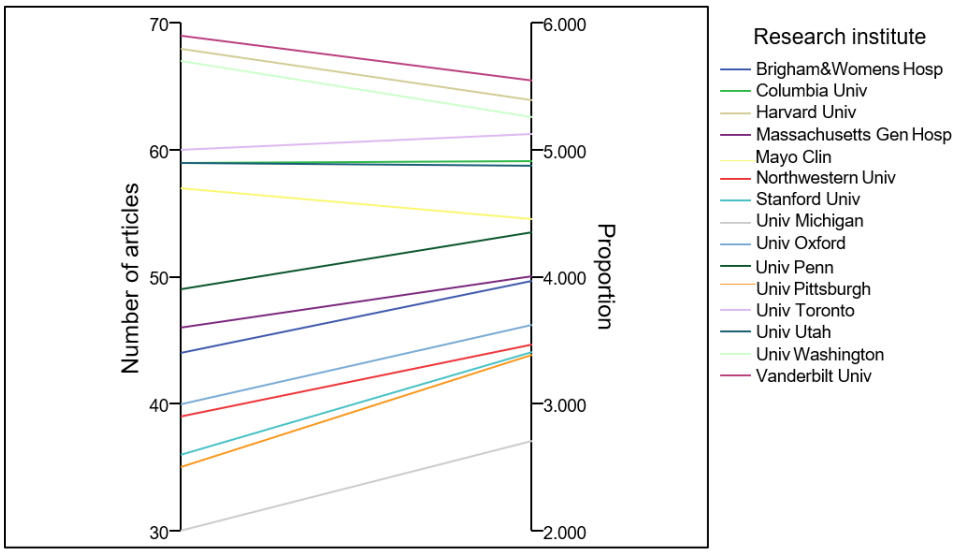

Figure 8. Distribution of major research institutions with the highest number of published articles.

the induction, analysis and utilization of information retrieval tools, retrieval techniques, retrieval strategies and retrieval results. Therefore, the curriculum system around information literacy education, such as the formation of the main branches of the curriculum, the construction of theoretical framework, the construction of teaching materials, the effectiveness of curriculum implementation and the reform of teaching methods, has become one of the research hotspots of information literacy education in recent years. This article refers to the health care big data standard of the Chinese Society of Health Information and Health Care. By organizing specialized training on medical information standards, we hope to help improve the management of health and medical informatization through systematic teaching. International organizations for standardization, such as ISO and HL7, are market-oriented and profitable organizations. They have a strong pursuit for the deepening of standardized services, and therefore they are in the forefront of the world in standardized services. The differences in standardization organizations among countries are mainly reflected in the differences in behavioral roles and their numbers.

According to statistics, there are 952 international health information standard research institutions, mainly universities, research institutes, medical research centers and well-known hospitals. Thirteen of the top 15 research institutions are from the United States. Among them, Vanderbilt University in the United States ranks first with 69 articles, accounting for 5.545% of the total literature. This was followed by Harvard University, which published 68 articles in the field of health information standards, accounting for 5.39% of the total literature. See Figure 8 for details. In addition, according to the statistics of literature languages, there are 1,062 articles in English, accounting for 97.25% of the total literature, and the other literature languages are Spanish, German, Portuguese, French, Italian, etc. The historical posts of the top three high-yield countries are illustrated and put together to analyze the development trends and mutual relations of frontier countries in HIL, as shown in Figure 9. The above figure clearly shows that the research in the field of information ecology in China is on the

rise, the related concepts of HIL are gradually refined, vocabulary research and language evaluation are gradually paid attention to, and the research on language proficiency and HIL in the context of globalization is a hot spot that continues up to now. The focus of HIL research has returned from solving social problems to the enlightenment of bilingual phenomena to HIL. This article echoes the research frontier in literature co-citation analysis.

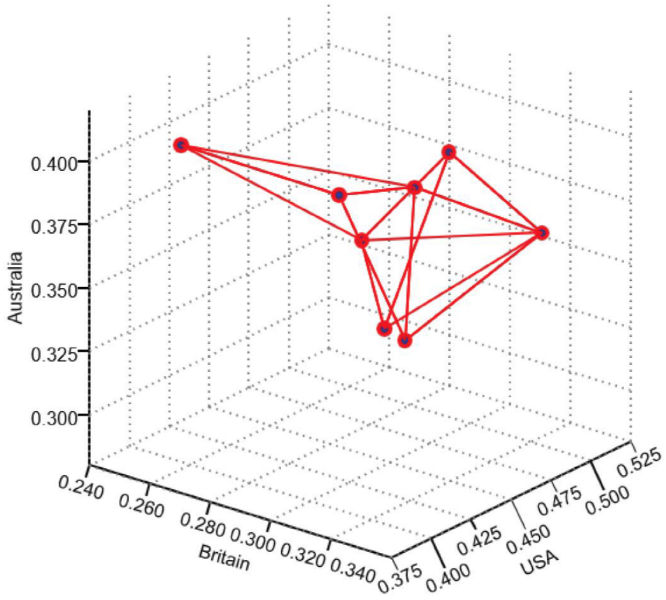

Figure 9. Comparison of historical publication centers in high-yield countries.

4. Conclusions

Based on multi-dimensional cognitive information technology, this paper analyzes the current situation and hot spots of health information literacy at home and abroad. The results show that most of the relevant research in China is based on foreign achievements, and no authoritative national standards have been formed. With the rapid development of American technology, America's HIL research level is in the leading position, becoming the object of competition and cooperation among countries, and forming a self-centered cooperation network. HIL's cognitive strategies refer to learners' processing of information in the current task. The research has improved people's awareness of the use of HIL information vocabulary learning strategies, and it is necessary to provide students with rich, diverse and authentic vocabulary learning materials. In addition, the improvement of HIL capability does not lie in the number of strategies used, but in the flexible and rational use of strategies.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgement

This work was supported by Key Research Project of Humanities and Social Sciences of Bengbu Medical College under grant 2020byzx262sk.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Z. Guo, Y. Shen, S. Wan, W. Shang, K. Yu, Hybrid intelligence-driven medical image recognition for remote patient diagnosis in internet of medical things, *IEEE J. Biomed. Health Inf.*, **26** (2022), 5817–5828. https://doi.org/10.1109/JBHI.2021.3139541
2. Q. Li, L. Liu, Z. Guo, P. Vijayakumar, F. Taghizadeh-Hesary, K. Yu, Smart assessment and forecasting framework for healthy development index in urban cities, *Cities*, **131** (2022), 103971. https://doi.org/10.1016/j.cities.2022.103971
3. N. C. Oleck, A. R. Johnson, B. N. N. Tran, H. S. Ayyala, E. S. Lee, B. T. Lee, A multimetric health literacy analysis of online information for gluteal augmentation with fat grafting, *Ann. Plast. Surg.*, **85** (2020), S97–S101. https://doi.org/10.1097/SAP.0000000000002425
4. L. Yang, Y. Li, S. X. Yang, Y. Lu, T. Guo, K. Yu, Generative adversarial learning for intelligent trust management in 6G wireless networks, *IEEE Network*, **36** (2022), 134–140. https://doi.org/10.1109/MNET.003.2100672
5. L. Zhao, Z. Yin, K. Yu, X. Tang, L. Xu, Z. Guo, et al., A fuzzy logic based intelligent multi-attribute routing scheme for two-layered SDVNs, *IEEE Trans. Network Serv. Manage.*, **19** (2022), 4189–4200. https://doi.org/10.1109/TNSM.2022.3202741
6. G. Kampouroglou, V. S. Velonaki, I. Pavlopoulou, E. Drakou, M. Kosmopoulos, N. Kouvas, et al., Parental anxiety in pediatric surgery consultations: The role of health literacy and need for information, *J. Pediatric Surg.*, **55** (2020), 590–596. https://doi.org/10.1016/j.jpedsurg.2019.07.016
7. C. C. Cutilli, *Health Literacy, Health Disparities and Sources of Health Information in US Older Adults*, Ph.D thesis, Duquesne University, 2015.
8. Z. Guo, K. Yu, N. Kumar, W. Wei, S. Mumtaz, M. Guizani, Deep distributed learning-based poi recommendation under mobile edge networks, *IEEE Internet Things J.*, **10** (2023), 303–317. https://doi.org/10.1109/JIOT.2022.3202628
9. J. Zhang, L. Zhao, K. Yu, G. Min, A. Y. Al-Dubai, A. Y. Zomaya, A novel federated learning scheme for generative adversarial networks, *IEEE Trans. Mobile Comput.*, **2023** (2023). https://doi.org/10.1109/TMC.2023.3278668
10. B. St. Jean, N. Greene Taylor, C. Kodama, M. Subramaniam, Assessing the health information source perceptions of tweens using card-sorting exercises, *J. Inf. Sci.*, **44** (2018), 148–164. https://doi.org/10.1177/0165551516687728

11. B. Parlak, A. K. Uysal, The effects of globalisation techniques on feature selection for text classification, *J. Inf. Sci.*, **47** (2021), 727–739. https://doi.org/10.1177/0165551520930897

12. K. Saho, K. Sugano, M. Kita, K. Uemura, M. Matsumoto, Classification of health literacy and cognitive impairments using higher-order kinematic parameters of the sit-to-stand movement from a monostatic doppler radar, *IEEE Sensors J.*, **21** (2021), 10183–10192. https://doi.org/10.1109/JSEN.2021.3060050

13. Y. J. Yi, B. Hwang, H. Yoon, H. Jeong, Health literacy and health information-seeking behavior of immigrants in south korea, *Library Inf. Sci. Res.*, **43** (2021), 101121. https://doi.org/10.1016/j.lisr.2021.101121

14. Q. He, Z. Feng, H. Fang, X. Wang, L. Zhao, Y. Yao, et al., A blockchain-based scheme for secure data offloading in healthcare with deep reinforcement learning, *IEEE/ACM Trans. Networking*, **2023** (2023). https://doi.org/10.1109/TNET.2023.3274631

15. A. L. Neves, L. Freise, L. Laranjo, A. W. Carter, A. Darzi, E. Mayer, Impact of providing patients access to electronic health records on quality and safety of care: A systematic review and meta-analysis, *BMJ Quality Saf.*, **29** (2020), 1019–1032. https://doi.org/10.1136/bmjqs-2019-010581

16. A. McNeil, R. Arena, The evolution of health literacy and communication: introducing health harmonics, *Prog. Cardiovasc. Dis.*, **59** (2017), 463–470. https://doi.org/10.1016/j.pcad.2017.02.003

17. Z. Guo, D. Meng, C. Chakraborty, X. R. Fan, A. Bhardwaj, K. Yu, Autonomous behavioral decision for vehicular agents based on cyber-physical social intelligence, *IEEE Trans. Comput. Soc. Syst.*, **10** (2022), 2111–2122. https://doi.org/10.1109/TCSS.2022.3212864

18. J. Zhang, Q. Yan, X. Zhu, K. Yu, Smart industrial iot empowered crowd sensing for safety monitoring in coal mine, *Digital Commun. Networks*, **9** (2023), 296–305. https://doi.org/10.1016/j.dcan.2022.08.002

19. A. De Thurah, Sp0008 the future for health professionals in rheumatology, *Ann. Rheum. Dis.*, **76** (2017), 3. https://doi.org/10.1136/annrheumdis-2017-eular.7155

20. Z. Zhou, Y. Su, J. Li, K. Yu, Q. M. J. Wu, Z. Fu, et al., Secret-to-image reversible transformation for generative steganography, *IEEE Trans. Dependable Secure Comput.*, **2022** (2022). https://doi.org/10.1109/TDSC.2022.3217661

21. S. S. Coughlin, J. L. Stewart, L. Young, V. Heboyan, G. De Leo, Health literacy and patient web portals, *Int. J. Med. Inf.*, **113** (2018), 43–48. https://doi.org/10.1016/j.ijmedinf.2018.02.009

22. R. Garcia-Retamero, E. T. Cokely, Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics, *Hum. Factors*, **59** (2017), 582–627. https://doi.org/10.1177/0018720817690634

23. Y. Zhang, Y. Sun, Y. Kim, The influence of individual differences on consumer's selection of online sources for health information, *Comput. Hum. Behav.*, **67** (2017), 303–312. https://doi.org/10.1016/j.chb.2016.11.008

24. H. Wang, N. Wang, M. Li, S. Mi, Y. Shi, Student physical health information management model under big data environment, *Sci. Program.*, **2021** (2021), 1–10. https://doi.org/10.1155/2021/5795884

25. Z. Zhou, X. Dong, R. Meng, M. Wang, H. Yan, K. Yu, et al., Generative steganography via autogeneration of semantic object contours, *IEEE Trans. Inf. Forensics Secu.*,**18** (2023), 2751–2765. https://doi.org/10.1109/TIFS.2023.3268843

26. H. Xue, D. Chen, N. Zhang, H. Dai, K. Yu, Integration of blockchain and edge computing in internet of things: A survey, *Future Gener. Comput. Syst.*, **144** (2023), 307–326. https://doi.org/10.1016/j.future.2022.10.029

27. M. Grene, Y. Cleary, A. Marcus-Quinn, Use of plain-language guidelines to promote health literacy, *IEEE Trans. Prof. Commun.*, **60** (2017), 384–400. https://doi.org/10.1109/TPC.2017.2761578

28. T. P. Liang, Y. H. Liu, Research landscape of business intelligence and big data analytics: A bibliometrics study, *Expert Syst. Appl.*, **111** (2018), 2–10. https://doi.org/10.1016/j.eswa.2018.05.018

29. F. Huang, L. Wang, H. Jia, Research trends for papillary thyroid carcinoma from 2010 to 2019: A systematic review and bibliometrics analysis, *Medicine*, **100** (2021), e26100. https://doi.org/10.1097/MD.0000000000026100

30. M. M. Mirończuk, J. Protasiewicz, A recent overview of the state-of-the-art elements of text classification, *Expert Syst. Appl.*, **106** (2018), 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

31. S. Burkhardt, S. Kramer, Online multi-label dependency topic models for text classification, *Mach. Learn.*, **107** (2018), 859–886. https://doi.org/10.1007/s10994-017-5689-6

32. F. Lei, X. Liu, Z. Li, Q. Dai, S. Wang, Multihop neighbor information fusion graph convolutional network for text classification, *Math. Prob. Eng.*, **2021** (2021), 1–9. https://doi.org/10.1155/2021/6665588

AIMS Press