

# Exploring Wetland Health Evaluation Indicators Based on Text Mining Technology

Sheng Miao<sup>1</sup>, Guoqing Ni<sup>1</sup>, Lan Chen<sup>2</sup>, Ruolan Mu<sup>2</sup>, Yansu Qi<sup>3,\*</sup>, and Chao Liu<sup>2</sup>

## Abstract

Wetlands are one of the important ecosystems on Earth, with necessary functions such as regulating climate, providing water, purifying water quality, and protecting biodiversity. At present, wetland health assessment has become a major direction in wetland research, and optimizing wetland assessment is of great significance for global sustainable development. However, most wetland assessments do not have agreed upon indicators and standards. In recent years, many achievements have been made in wetland health research. Therefore, the aim of this study is to use data mining techniques to explore and evaluate the main factors in relation to wetlands. This article uses 100 wetland health papers on Web of Science as the corpus, proposes an indicator extraction method based on natural language processing and text mining technology, explores the interrelationships between the extracted indicators, establishes a wetland indicator evaluation system to evaluate the wetland health status, and explores new ideas for wetland health evaluation.

## Keywords

Agglomeration and Hierarchical Clustering, Natural Language Processing, Text Mining, Wetland Evaluation

## 1. Introduction

Wetlands are among the world's most productive ecosystems due to their unique ecological characteristics. However, rapid social and economic development and population growth have accelerated the destruction of wetlands. This degradation leads to the loss of wetland ecosystems, reduced biodiversity, deteriorating water quality, and other negative impacts on both human society and the natural environment. Healthy wetland ecosystems can maintain normal material circulation and energy flow, and provide economic, social, and ecological benefits. Assessing wetland health involves scientifically evaluating and monitoring the condition of these ecosystems. Such assessments help identify signs and causes of degradation, offering a scientific basis for developing protection and recovery strategies. Currently, the methods for extracting wetland health indices are tailored to specific research areas, limiting their broader applicability. The widespread use of various devices has generated vast amounts of data. Data mining technology, which has become increasingly sophisticated, can process large volumes of unstructured data. Text mining, a branch of data mining, utilizes natural language processing to filter out valuable

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Manuscript received February 5, 2024; first revision July 7, 2024; accepted August 27, 2024.

\* **Corresponding Author:** Yansu Qi (qiyansu0909@hotmail.com)

<sup>1</sup> School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China (smiao@qut.edu.cn, n1633703426@outlook.com)

<sup>2</sup> School of Environmental and Municipal Engineering, Qingdao University of Technology, Qingdao, China (liuchao@qut.edu.cn, chenlan24@outlook.com, muruolan@hotmail.com)

<sup>3</sup> School of Civil Engineering, Qingdao University of Technology, Qingdao, China (qiyansu0909@hotmail.com)

information from text data. This technique is now widely used in fields such as finance, medicine, education, information science, and architectural design [1–3].

The purpose of this research is to use text mining technology to extract an evaluation index system for wetland health evaluation based on wetland health research papers. This study captured 100 papers on wetland health as the corpus, and then preprocessed the text corpus using natural language processing technology. In the text mining corpus, 200 words that are most relevant to the word “wetland” were selected according to the relationship between word vectors, and finally retained 85 wetland health evaluation indicators after screening. Subsequently, hierarchical cluster analysis (HCA) was used to explore the interrelationships between indicators for wetland health, and hierarchical clustering diagrams and heat maps were used to visualize the interrelationships of indicators. The main contributions are as follows:

- Use research papers as text corporate and objective evaluation indicators through text mining.
- Through natural language processing technology, explore the relationship between extracted indicators and form a wetland evaluation index system.
- The accuracy and effectiveness of the wet assessment can be improved and the assessment results made more directive by improving the basic assessment indicators.

## 2. Related Works

Taking wetland health as the core concept, the existing academic resource database was carefully screened and the researchers' established index system was organized. Bentley et al. [4] used a stress state response model to evaluate the response of ecosystems to environmental stress, collecting various ecological data such as wetland vegetation to assess ecosystem health. Chen et al. [5] presented a comprehensive wetland health evaluation system with 13 indicators. Wu and Chen [6] established a wetland indicator system including remote sensing data and spatial data. The evaluation index system is an integration, with multiple indicators representing various aspects of the evaluation object and their interrelationships. Therefore, providing a specific definition for the wetland health evaluation index system becomes challenging. At present, existing research papers have proposed an indicator system for wetland health evaluation based on the construction of a wetland health evaluation index system, drawing on relevant domestic and foreign cases. Therefore, each researcher has constructed their own indicator system based on their professional knowledge, leading to subjectivity in the proposed evaluation indicator system, which may raise concerns about its accuracy.

With the development of internet technology, now is the era of big data. Text mining is a branch of data mining that involves extracting useful information from unstructured data. Hilal et al. [7] proposed a decision and evaluation ranking system SentiBank based on text mining, and then used Neutro VADER to perform emotional ratings on each evaluation of product functionality. SentiBank is a ranking system based on pre trained language models, used to estimate the polarity of user comments on various topics. Use Neutro-VADER to create a simplified neutron number for each feature, and then use the weighted operation of the simplified neutron number to determine the total simplified neutron value for each equivalent option. Use cosine similarity measurement and scoring formula to rank alternative products. Qiu et al. [8] used text mining based on coal mine accident investigation reports, combined with chi-square statistical dimensionality reduction and word cloud analysis methods, to determine the main cause

of coal mine accidents. By constructing and analyzing the network of coal mine accident causes, eight core causes and their related cause sets have been identified. Gurbuz and Uluyol [9] used web crawling techniques to obtain a corpus, preprocessed it, and performed feature selection steps. Subsequently, they used naive Bayes, random forest, and support vector machine algorithms to classify articles based on their subject areas. Based on text mining methods and temporal convolutional networks (TCN), Fang et al. [10] added multimodal attention mechanism (AM) and cross modal transformer structure to construct a TCN model based on AM-TCN, and analyzed the multimodal emotions of network product marketing information [10]. Text mining technology has been applied in various fields and achieved good results, but current research mainly uses supervised learning methods to classify text data, lacking research on its relevance. Therefore, this paper conducted text mining on existing wetland health research papers, extracted wetland health evaluation indicators using unsupervised methods, established a wetland health evaluation system, and explored the relationship between different indicators. In addition, as text mining methods are based on machine learning algorithms and mathematical models, they are not affected by subjective factors and to some extent eliminate subjectivity.

### 3. Materials and Methods

The focus of text mining is on dealing with unstructured data, which does not have a fixed structure and is often difficult for computers to understand. Text mining relies on natural language processing techniques and machine learning algorithms to analyze and model text in order to extract useful information and knowledge from it. It can build classifiers and clusters to analyze the relationships between textual data, and so on, as well as recognize concepts, emotions, keywords, etc., in the text. This article aims to extract an evaluation index system related to wetland health assessment, therefore using word2vec word embedding model and HCA is an option. The word2vec model is a method of transforming text data into vectors that can be processed by computers. HCA is an unsupervised learning method widely used for analyzing and processing unknown types of data, which can discover potential relationships between data. The research process is shown in Fig. 1, which includes five key research stages: corpus collection and establishment; data preprocessing; conversion word embedding; indicator extraction; and classification and visualization of indicators.

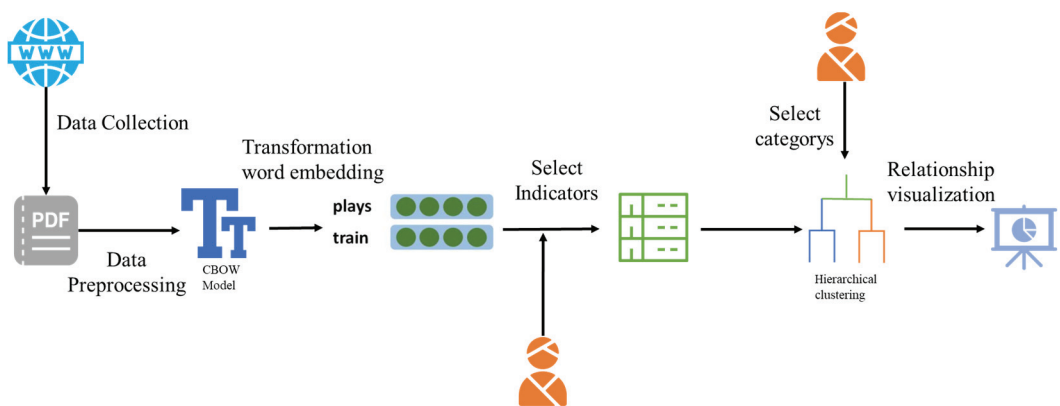


Fig. 1. Research procedures.

### 3.1 Data Collection and Pre-process

When collecting data, the appropriate methods, channels and even distribution of data can affect the final results of the analysis. In recent years, with the increasing interest of researchers in wetland health, many research results have been achieved. However, existing wetland health research results are often more related to the research area, but there is no universal evaluation system applicable to all wetlands. Therefore, the focus of this study is on these research results, and text mining is carried out to integrate their results to establish a universal wetland health evaluation system.

The Web of Science is an English literature database that includes various authoritative and influential international academic journals from around the world. Therefore, this research collected 100 research papers on the Web of Science using the keyword “wetland health” as the corpus for text mining in this study. These 100 papers can cover different functional wetlands (such as urban wetlands, natural wetlands), wetlands in different regions (such as coastal wetlands, lake wetlands, swamp wetlands, etc.), and other aspects. Through this approach, corpora can provide diverse text data, which helps to build comprehensive text mining models. Therefore, these 100 papers may provide a sufficient sample size. These 100 articles can be roughly divided into three categories: wetland evaluation research, review, and actual case studies, with a ratio of approximately 5:1:4. Wetland evaluation research mainly proposes a new wetland evaluation system, review mainly summarizes and analyzes wetland evaluation systems proposed in recent years, and actual case studies mainly apply a new wetland evaluation system to a certain wetland.

Data pre-processing plays an important role in the data mining process. Real-world collections of real data are often incomplete, contain noise, dimensional disasters, and redundant instances, and are inconsistent. In this study, there is no doubt that the data excepted noise, because the text is generated by specialized researchers with a high degree of expertise. However, there will still be a certain amount of text data collected on dimensional disasters and redundancies, including prepositions, conjunctions, and numbers. These words are not very meaningful for the study, and their presence increases the computational load of the subsequent processing. Therefore, this research removes them from the pretreatment process. Another type of word without valuable meaning is common words, which are too broad to provide useful information for text mining and will increase the computational effort. The NLTK module is used to conduct the study, which can obtain stop words while removing redundancy in the text corpus in combination with topic experience. Also, this operation reduces the dimensionality of the word vector in the next process.

In addition, the normalization of words is an important step in the text-mining process. In this study, the different forms of the same word in the text corpus are unified by preprocessing, since there are many words in the discourse corpus that express the same meaning but have different forms. Thus, this study unified these lowercase words, singular words and main tense words in the text corpus. In addition, NLTK can automatically complete spell-checking and word root extraction. Considering that the words used for wetland evaluation indicators in research papers on wetland health may be in the form of adjectives, while the final study should use nouns, this study extracted word roots in the pre-processing stage. All these pre-processing steps not only reduce the number of dimensions and redundant words and information loss, but also improve the value density of the text corpus for the subsequent mining process.

### 3.2 Word Embedding

Text data cannot be directly calculated on computers. This research needs to use the word2vec model, which can convert words from natural language into language that computers can understand. It simplifies text content processing into vector operations, and then calculates the similarity in vector space to represent the semantic similarity of the text. The distance between the word vectors generated by this mapping represents the correlation between words. There are many ways to use the word2vec model, and this article uses the continuous bag-of-words (CBOW) model, which predicts the current word based on context.

The CBOW model consists of an input layer, a hidden layer, and an output layer. Adjacent neural network layers are connected by a weight matrix and named as  $\omega$  and  $\varphi$ . The CBOW model uses  $n$  words before and after a word to predict it. The predicted word is called the center word, represented by  $c$ , and the words in the context are called surrounding words, represented by  $s$ . In addition, we also refer to the number of surrounding words as window size. In this study, the window size range is represented as  $2n$ , the number of different words contained in the preprocessed corpus is represented as  $V$ , and the embedding dimension of the generated words is represented as  $N$ . The first step in CBOW modeling is to use one-hot encoding to encode all words in the preprocessed corpus. All words are represented as vectors of length  $V$ , and the one-hot encoding of surrounding words is used as input, passing through an input weight matrix, calculate hidden layer features using the following equation:

$$HL = \frac{1}{n} \omega \sum_{i=1}^n X_i \quad (1)$$

where  $\omega$  is input weight matrix,  $X_i$  is surround word's one-hot encoding.

In this step, all words in the text corpus are passed through a weight matrix  $\omega$  map to an  $N$ -dimensional vector, and then use the vector matrix HL of the hidden layer to pass through the weight matrix  $\varphi$  to calculate the output layer  $\mu$ , the expression is as follows:

$$\mu = \varphi \cdot HL \quad (2)$$

For the output of the predicted values, the softmax activation function calculates the probability of each predicted word, the probability that the central word  $c$  will appear for a particular surrounding word: using the following equation:

$$p(c/s) = y_i = \frac{\exp(\mu_i c)}{\sum_{j=1}^V \exp(\mu_j c)} \quad (3)$$

where  $\mu_j$  is  $j$  column of matrix  $\phi$ ,  $y$  is the probability vector of output.

According to the above equation, it can be seen that, given the surrounding word  $s$ , the weight matrix is determined by training a neural network  $\omega$  and  $\varphi$ . To increase the probability of predicting the central word  $c$ , and to establish a CBOW model for forward calculation, this training is performed through a very common feedforward propagation algorithm.

The setting of parameters is crucial in the CBOW algorithm. The dimension of word vectors and the width of surrounding word sequences affect the accuracy of training, and the value of dimension is generally related to the size of the corpus. The larger the embedding dimension of the generated word, the better the feature representation of the word. However, larger scales require more training data. In

this study, through multiple different experimental tests, the final size of word embedding was determined to be 300. We calculated the length of sentences in the corpus and set the width of the surrounding word window to 7. In this study, we used multivariate cross entropy as the loss function, as shown in formula (4). After 30 rounds of training, the loss function of the word2vec model remained basically unchanged, and the model has achieved convergence.

$$\hat{y}_i = \text{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, \quad (4)$$

$$L_{\text{cross-entropy}}(\hat{y}, y) = - \sum y_i \log(\hat{y}_i). \quad (5)$$

where,  $z_j$  represents the model output of the category,  $K$  represents the length of the one hot word vector, and in formula (5),  $y$  represents the label and model prediction.  $\hat{y}$  represents the model prediction result.

In the experiment, 10 independent training sessions can predict highly similar vectors of three commonly used words in wetland health assessment indicators: vegetation, plants, and species. This demonstrates the ability of the word2vec model to efficiently and accurately convert words from a text corpus into vectors.

### 3.3 Word Clustering

Wetland health evaluation indicators can be divided into different categories based on their evaluation fields, and clustering algorithms are an independent tool in data mining, which is precisely suitable for this research plan. Its purpose is to divide the dataset into several different subsets based on a certain similarity measure, which is called clustering, to reveal the potential relationships between elements in the dataset. In this study, words in the text were transformed into computer-understandable word embeddings, and the Ward link method's hierarchical clustering algorithm was used to calculate similarity. The Ward link method defined the distance based on the difference in the sum of squared variances within the merged population, the formula is:

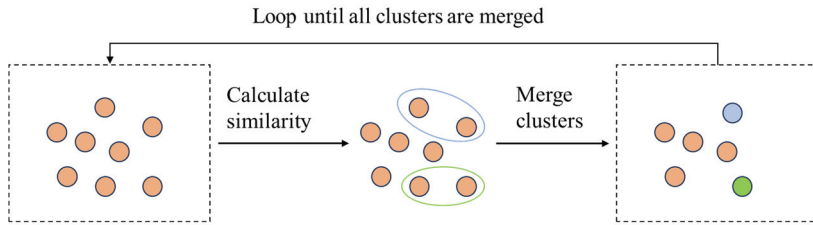
$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B), \quad (6)$$

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A), y_i \in A, \quad (7)$$

$$\bar{y}_{AB} = \frac{\sum_{i=1}^{n_A} y_i + \sum_{j=1}^{n_B} y_j}{n_A + n_B}, \quad (8)$$

where  $I_{AB}$  is the difference between the sum of squared variances of population A and population B after merging, SSE is error sum of squares and  $y_i$  is the element within the population and  $\bar{y}$  is the average value of the population.

This paper uses a word2vec model trained on a corpus to generate word embeddings for wetland evaluation metrics, which can express semantic relationships. Therefore, the distance between embeddings can express how they correlate in the corpus. This research decided to use hierarchical clustering to determine the main wetland health evaluation indicators by clustering highly correlated words together, as prior knowledge cannot determine the number of clusters. The process of the algorithm is shown in Fig. 2.



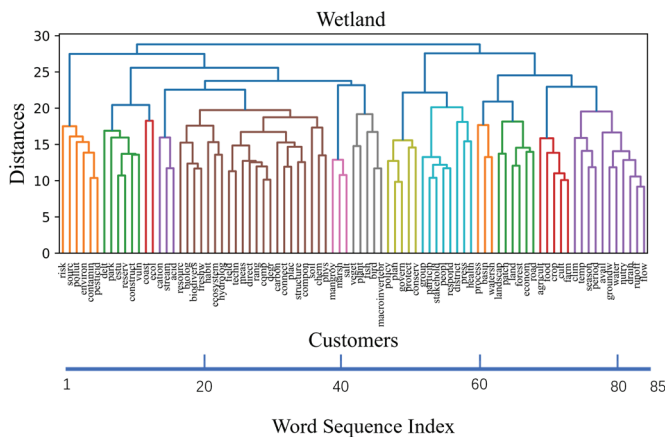
**Fig. 2.** HCA procedure.

## 4. Results and Discussion

### 4.1 Result

Wetland health evaluation is an important aspect of wetland protection and management, and the selection of evaluation indicators determines the accuracy and operability of the evaluation results. This article conducts clustering analysis on high-frequency words used in the study of wetland health evaluation indicators. By setting a threshold of group distance  $\geq 20$ , 13 groups of clustering words were obtained, as shown in Fig. 3. From the graph, it can be seen that currently, research on wetland health evaluation indicators mainly focuses on environmental pollution indicators (risk, source, pollution, environment, pollution, pest), wetland types (delta, park, esou, reserve, construct, vuln), biodiversity indicators (get, plant, fish, bird, macroinvertebr), policy management indicators (policy, plan, government, protect, preserve) Landscape indicators (landscape, patch, land, forest, economy, road), climate (clim, temp, season, penod, avai, etc.), agriculture (agriculture, food, crop, cult, farm), etc.

In the field of wetland health assessment, many studies have proposed various indicators to evaluate the health status of wetland ecosystems. For example, Chen et al. [5] used 19 wetlands in the Beijing-Tianjin-Hebei region of China as the research object, constructed a wetland ecological health evaluation index system, and selected 13 indicators. Wu and Chen [6] took Hongze Lake as the research object and constructed a wetland health indicator system consisting of 12 indicators based on remote sensing technology and landscape index. These two sets of indicators are shown in Table 1.



**Fig. 3.** Result of hierarchical clustering.

**Table 1.** Wetland evaluation indicators comparison

| Indicators                         |                             |                            |
|------------------------------------|-----------------------------|----------------------------|
| Chen et al. [5]                    | Wu and Chen [6]             | Proposed                   |
| Water quality                      | Water quality               | Water                      |
| Soil heavy metal content           | Soil heavy metal content    | Pollution, soil            |
| Soil pH                            | Soil pH value               | Chem, acid                 |
|                                    | Normalized vegetation index | Plant, forest              |
| Wildlife habitat suitability index | Wildlife habitat index      | Habit                      |
| Protection awareness               | Wetland protection          | Protect                    |
| Land-use intensity                 | Land-use intensity          | Construct, structure, park |
| Population density                 | Population density          | People                     |
| Wetland area change rate           | Wetland area change rate    | Range                      |
| Biodiversity index                 |                             | Biodivers, invertebr       |
| Welfare index                      | Welfare index               | Agricult, farm             |
| Water supply guarantee rate        | Patch density               | Season, stream             |
| Alien species invasion             | Normalized mud index        | Econom                     |
| Soil moisture                      |                             | Hydrolog                   |
|                                    |                             | Policy                     |
|                                    |                             | Resource                   |

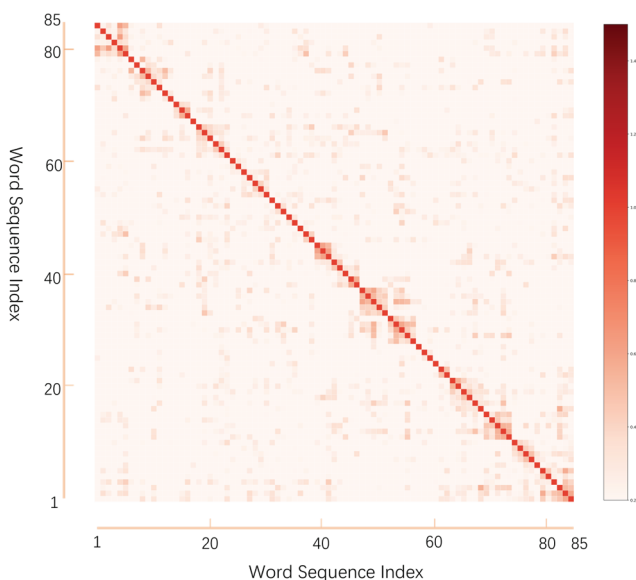
The indicators in the first 10 rows of Table 1 are in common for all three journals, while the indicators in the last 7 rows of the Table 1 are unique for each paper. The above indicators are obtained to evaluate the health status of wetlands. In traditional wetland health assessment, indicators are often obtained by combining the wetland's own condition with historical records, field investigations, expert opinions, etc. The cost of collecting indicators is high and may be limited by time and resources. However, text mining can quickly collect a large amount of information, provide a wide range of perspectives and global data, and has lower costs and efficiency. Due to the ability of text mining technology to process large-scale data and extract multiple indicators, it can provide global wetland health assessment information in a short period of time. Although text mining may not provide sufficiently accurate data for certain specific fine-grained metrics. But it can provide a broad perspective and global data for global wetland health assessment, while quickly identifying potential wetland health assessment indicators. It provides auxiliary decision-making for the systematic evaluation of wetland health.

By analyzing wetland health indicators, we are able to gain a deeper understanding of the problems faced by wetlands, such as the accumulation of hazardous substances and the degree of pollution of water bodies and soils. These indicators not only provide reliable quantitative data support for wetland health assessment, but also reveal the complexity and vulnerability of wetland ecosystems. It is worth noting that different types of wetlands contribute differently to the environment due to their unique ecological characteristics and functions. Everglade wetlands and lake wetlands play an important role in purifying water bodies and storing carbon, while riverine wetlands play a key role in water quality maintenance and providing biological habitats. Therefore, when assessing wetland health, it is important to select assessment indicators that match the characteristics and functions of each type of wetland. For example, when assessing the health status of lakes and marshes, water quality parameters such as redox potential and nutrient content should be the focus of attention. These parameters can intuitively reflect the quality status of wetland water bodies and provide important references for wetland ecological health. From the results of text mining, policy and economy also have considerable influence on wetland health. With the economic development, the city gradually expands, and a large number of wetlands are converted into residential and industrial land, resulting in great



changes in wetland habitats. In addition, in the process of urbanization, industrial wastewater and domestic sewage are discharged into wetlands, leading to pollution of wetland water bodies and destroying the balance of the ecosystem. Therefore, wetland health assessment should consider several aspects in order to evaluate the health of wetlands more comprehensively.

By analyzing environmental pollution indicators, we can better understand the environmental problems that exist in wetlands, such as the accumulation of harmful substances, the degree of pollution in water and soil, etc. These indicators can provide quantitative data support for wetland health assessment. Different types of wetlands have varying contributions and functions to the environment. Swamp lands and lakes can purify water bodies and store carbon, while river wetlands can maintain water quality and provide habitats. Therefore, in wetland health assessment, it is necessary to select appropriate evaluation indicators based on the characteristics and functions of wetland types. For example, in the health assessment of lakes and wetlands, attention can be paid to water quality parameters such as redox potential, nutrient content, etc. In addition, agriculture also has an impact on the ecological health of wetlands, and agricultural land reclamation and water conservancy construction may lead to the destruction of wetlands and the loss of wetland biological habitats. The use of pesticides, fertilizers, and agricultural wastewater may also cause pollution to wetland water bodies and soil. In addition, farmland drainage may alter the water supply and quality of wetlands, affecting their ecological processes. Therefore, evaluating the impact of agriculture on wetlands is an important aspect of wetland health assessment. Therefore, wetland health assessment should consider multiple aspects to evaluate the health status of wetlands more comprehensively.



**Fig. 4.** Heatmap of evaluation index.

With the help of heat maps, we can clearly see the relationships between words of interest, while exploring the correlations between words, further exploring semantic relationships and underlying knowledge. The semantic distance between words is expressed as cosine similarity. Corresponding to the horizontal and vertical axes, each color block in the heatmap represents the cosine similarity between

words. Different colors correspond to different correlation coefficients. When the semantic relationship is closer, the squares between the words are colored darker. In this study, wetland evaluation indicators extracted through text mining were arranged according to the clustering results to draw a heatmap, as shown in Fig. 4. In the heatmap, the area formed by the squares around the diagonal is darker in color, indicating that the population formed by clustering has a better clustering effect at a certain distance.

## 4.2 Discussion

Wetlands have important ecological functions and environmental services, including water quality purification, flood control and storage, and biodiversity maintenance. In recent years, wetland health issues have aroused widespread interest among researchers. Therefore, this study provides a method for extracting wetland health evaluation indicators based on text mining and natural language processing, which can be used for monitoring wetland ecological conditions. By collecting and analyzing relevant indicator data, it is possible to timely understand the changes and trends of wetland ecosystems, and thus take necessary protection and management measures.

This study selected 100 wetland health papers from the Web of Science as the corpus. As these papers are specifically focused on wetland health, the frequency of wetland health assessment indicators is relatively high, and these indicators can be considered as high-frequency words. In order to capture the semantic relationship between these indicators, a word2vec model based on the CBOW algorithm is used to generate word vectors, extract wetland health evaluation indicators through semantic relationships, and use unsupervised algorithms to analyze the relationship between wetland health evaluation indicators. Compared with existing text mining, the method proposed in this paper does not require a large amount of annotated data and can easily extract indicators.

In addition, the extracted wetland health assessment indicators can be further divided into different assessment areas, such as ecological models, ecological structures, etc. This classification can help professionals better understand and evaluate the different aspects of wetland health. Generate clustering result graphs through clustering algorithms to display the relationships between indicators of different categories. To further verify the effectiveness of the clustering results, the relationship between word vectors was used to create a heatmap showing indicator similarity. According to the results of hierarchical clustering and heatmap, it can be seen that the clustering results using the variance difference before and after word vector merging as feature representation are consistent with the hierarchical clustering results based on Euclidean distance. This indicates that the extracted indicators have a certain degree of semantic similarity, and the clustering results are relatively reliable. Compared to previous wetland health research papers, our research method combines wetland health assessment papers of various types of wetlands, eliminating the subjectivity of individual research papers. The indicator extraction methods based on text mining and natural language processing used in this study can also be applied to other fields.

## 5. Conclusion

This article uses text mining technology to analyze 100 wetland health papers on the Web of Science, extracts wetland health evaluation indicators, and provides valuable opinions on wetland health protection decisions. The results have certain value and effectiveness, but there are still shortcomings that

need further research. Firstly, all descriptions of wetland health were collected from a single source scientific website. Although it is internationally recognized as an important database for obtaining academic information, it has many unpublished journals, resulting in a limited number of papers being reviewed. In future work, the credibility of research results can be improved by including more data sources and establishing a complete text corpus. In addition, due to the focus of this study on British literature, the generalizability of the text-mining results is poor. Therefore, the method of this study has been improved in terms of language, such as multilingual texts, to form a complementary approach. In addition, based on the results of text mining, we found that some professional vocabulary was incorrectly identified, so we promoted text mining based on topic-specific dictionaries.

## Conflict of Interest

The authors declare that they have no competing interests.

## Acknowledgements

This paper is the extended version of “Utilizing Text Mining to Extract Critical Indicators for Wetland Health Evaluation,” in the 15th International Conference on Computer Science and its Applications (CSA 2023) held in Nha Trang, Vietnam, dated December 18-20, 2023.

## Funding

None.

## References

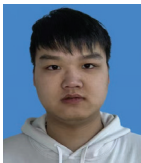
- [1] K. H. Goh, L. Wang, A. Y. K. Yeow, Y. Y. Ding, L. S. Y. Au, H. M. N. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, “Prediction of readmission in geriatric patients from clinical notes: retrospective text mining study,” *Journal of Medical Internet Research*, vol. 23, no. 10, article no. e26486, 2021. <https://doi.org/10.2196/26486>
- [2] F. Bao, W. Xu, Y. Feng, and C. Xu, “A topic-rank recommendation model based on microblog topic relevance & user preference analysis,” *Human-centric Computing and Information Sciences*, vol. 12, article no. 10, 2022. <https://doi.org/10.22967/HICIS.2022.12.010>
- [3] D. Zhao, Y. Liu, G. Zeng, X. Wang, S. Miao, and W. Gao, “A knowledge-based human-computer interaction system for the building design evaluation using artificial neural network,” *Human-centric Computing and Information Sciences*, vol. 13, article no. 2, 2023. <https://doi.org/10.22967/HICIS.2023.13.002>
- [4] S. B. Bentley, S. A. Tomscha, and J. R. Deslippe, “Indicators of wetland health improve following small-scale ecological restoration on private land,” *Science of the Total Environment*, vol. 837, article no. 155760, 2022. <https://doi.org/10.1016/j.scitotenv.2022.155760>
- [5] W. Chen, C. Cao, D. Liu, R. Tian, C. Wu, Y. Wang, Y. Qian, G. Ma, and D. Bao, “An evaluating system for wetland ecological health: case study on nineteen major wetlands in Beijing-Tianjin-Hebei region, China,” *Science of the Total Environment*, vol. 666, pp. 1080-1088, 2019. <https://doi.org/10.1016/j.scitotenv.2019.02.325>

- [6] C. Wu and W. Chen, "Indicator system construction and health assessment of wetland ecosystem: taking Hongze Lake Wetland, China as an example," *Ecological Indicators*, vol. 112, article no. 106164, 2020. <https://doi.org/10.1016/j.ecolind.2020.106164>
- [7] A. M. Hilal, J. S. Alzahrani, H. Alsolai, N. Negm, F. M. Nafie, A. Motwakel, I. Yaseen, and M. A. Hamza, "Sentiment analysis technique for textual reviews using neutrosophic set theory in the multi-criteria decision-making system," *Human-centric Computing and Information Sciences*, vol. 13, article no. 24, 2023. <https://doi.org/10.22967/HICIS.2023.13.024>
- [8] Z. Qiu, Q. Liu, X. Li, J. Zhang, and Y. Zhang, "Construction and analysis of a coal mine accident causation network based on text mining," *Process Safety and Environmental Protection*, vol. 153, pp. 320-328, 2021. <https://doi.org/10.1016/j.psep.2021.07.032>
- [9] T. Gurbuz and C. Uluyol, "Research article classification with text mining method," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 1, article no. e7437, 2023. <https://doi.org/10.1002/cpe.7437>
- [10] Z. Fang, Y. Qian, C. Su, Y. Miao, and Y. Li, "The multimodal sentiment analysis of online product marketing information using text mining and big data," *Journal of Organizational and End User Computing (JOEUC)*, vol. 34, no. 1, pp. 1-19, 2022. <https://doi.org/10.4018/JOEUC.316124>



**Sheng Miao** <https://orcid.org/0000-0001-6176-3624>

He is an associate professor in the School of Information and Control Engineering, Qingdao University of Technology. He received his Ph.D. degree from Towson university, Maryland, USA in 2017. His research interests include machine learning, human computation, smart healthcare, and intelligence systems.



**Guoqing Ni** <https://orcid.org/0009-0006-8763-8710>

He obtained a Bachelor's degree in Computer Science and Technology from Henan University of Urban Construction in 2021. Since September 2022, he has been pursuing a degree in Computer Science and Technology at Qingdao University of Technology. His current research interest is remote sensing data analysis.



**Lan Chen** <https://orcid.org/0009-0002-9366-3809>

She received B.S. degree in School of Petroleum and Environmental Engineering from Yan'an University in 2021. Since September 2022, she has been pursuing her M.S. degree in environmental engineering from Qingdao University of technology. Her current research interest is ecosystem analysis.



**Ruolan Mu** <https://orcid.org/0009-0004-5286-9573>

She received her B.S. degree in the College of Environmental and Municipal Engineering from Qingdao University of Technology in 2022. Since September 2022, she pursued her M.S. degree in Civil Engineering and Water Resources from Qingdao University of technology. Her research interest is ecosystem analysis.



**Yansu Qi** <https://orcid.org/0000-0003-4316-696X>

She was born in Xuzhou, China, in 1990. She is a Ph.D. student in School of Environmental and Municipal Engineering, Qingdao University of Technology. Her research interests include the urban thermal environment and outdoor thermal comfort.



**Chao Liu** <https://orcid.org/0009-0001-7842-9064>

She received Ph.D. degree in municipal engineering from the Qingdao University of technology, in 2018. She is currently an associate professor in School of Environmental and Municipal Engineering, Qingdao University of Technology, China.