Research on Intelligent Construction Algorithm of Subject Knowledge Thesaurus Based on Literature Resources

Xiaoxia Wang¹,², Xiaozhong Xu³, Jiarui Zhang¹,², Yue Zhu¹,², Yuhang Fan¹,² and Pengjing Xu¹,²

¹ Shanghai Technology Development Co. Ltd., Shanghai China, 200062
² Shanghai-Warsaw Joint Lab on Artificial Intelligence, Shanghai, People's Republic of China, 200062
³ College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China, 200062

* Corresponding author’s e-mail: pjxu@sstir.cn

Abstract. The implementation of National Science and Technology Innovation Strategy demands exponential growing in knowledge services on literature information institutions. It is the most important knowledge organization tool for Information Retrieval, which can be widely used for semantic citation, organization and retrieval of literature resources. This study aims to develop an innovative algorithm for constructing subject thesaurus based on massive literature resource data and mining academic neologisms, also the semantic relationship between academic neologisms and subject system. We firstly collect a dataset of literature corpus, corresponding work for data pre-processing carried out. Then using the FastText model to complete academic neologisms mining, we construct an automatic categorization model of academic neologisms based on the Bert and TextCNN algorithm. The algorithm proposed in this study is validated by 8.1 million multi-source and heterogeneous literature data in the field of marine disciplines. The result shows that the algorithm can effectively replace 90% of the manual annotation volume, mine a large number of high-quality marine neologisms and successfully build the marine science knowledge base with a pass rate of 82.6% reviewed by expert, which present high accuracy and certain engineering application prospects.

1. Introduction
In recent years, major countries around the world such as China, the United States and Japan, are improving their core competitiveness in the field of science and technology innovation, in which the integration and organization of science and technology resources and knowledge services are prominent issues that must be addressed. Subject knowledge thesaurus refers to the classification system used to represent the subject knowledge framework in a specific subject field, the terminologies representing subject topics, and the association mapping relationship oriented to the knowledge system. It is an important semantic tool in the field of scientific and technological information service, the premise and foundation of scientific and technological knowledge mining. Specifically, based on knowledge points and relations in the thesaurus, it supports academic services and applications for scientific research through the integration, mining and association of deep knowledge contents of scientific and technological literature information resources, and helps scientific and technological information service organizations complete semantic annotation, semantic
retrieval, knowledge browsing and association discovery, etc. It has a very wide application value. Therefore, it is necessary to build knowledge lexicon oriented to the discipline field.

In current research, due to the immaturity of the previous big data algorithm technology, researches on the automatic construction technology of subject thesaurus are rarely carried out. Such works are mostly completed by manual collation, which requires a lot of human, material and financial resources. For example, the Science and Technology Knowledge Organization System (STKOS) of the Documentation and Information Center of the Chinese Academy of Sciences (CAS) also has problems of weak timeliness and correlation mapping ability, high update and maintenance cost, and the lack of uniformity in the depth of different knowledge system layers which is not conducive to knowledge organization and management. Merley da Silva Conrado[1] et al. proposed a method to automatically extract term words by machine learning. This research has made some contributions, but lacks of mining the association between words and knowledge system. Yunxiao Zhou[2] et al. used supervised text classification algorithm to classify the semantic relationship of Chinese words through the embedded features of word level and character level. The results show that the prediction effect of this method is higher than the average level, but the categories are only antonym, synonym, meronym and hyponym. Without a complete knowledge system, it is impossible to form a systematic and structured knowledge thesaurus. In order to solve these problems, this paper focuses on the intelligent construction of disciplinary knowledge lexicon based on literature resources. In recent years, natural language processing technology[3] has been developed rapidly, Armand Joulin[4] studied the fastText model, using hierarchical softmax to improve the efficiency of the model, and using n-gram features to maintain the word order information, which is very suitable for new word mining. Yoon Kim[5] proposed a new approach for sentences classification by convolutional neural network, and extract the key information of n-gram effectively. However, word2vec[6] is used in the pre training word vector model, so Out-of-vocabulary (OOV) is the problem of word embedding features. Jacob Devlin[7] et al. from Google introduced a new language representation model called BERT. The model use an improved bidirectional transformer[8] structure, and use more powerful machines to train more large-scale data to make the results of Bert reach a new height and perform well in many tasks. But this model is more suitable for word vector representation, and the effect of direct classification is weak. Consequently, in this paper we firstly need to carry out data preparation and preprocessing of massive domain-specific literature corpus data as well as seed words to provide high-quality data guarantee for the algorithm. Next, we use FastText algorithm to construct word embedding space model, search for new academic words based on seed words, and in the end, the candidate academic vocabulary set is generated. Finally, we semantically expand the academic vocabulary, generate semantic vectors according to the Bert model, and use the TextCNN model to classify the academic vocabulary and form a hierarchical association relationship between the terminology words and the system. The proposed method constructs a thesaurus entirely based on disciplinary literature data with academic knowledge structure and unified and complete knowledge system hierarchy, authoritative and practical. It can complete the automatic construction and regular update of knowledge lexicon with the most advanced artificial intelligence technology and strong core semantic association mapping capability.

2. Methodology

2.1. FastText

FastText[4] is a text classification tool with a typical application scenario of supervised text classification problems, combining the most successful ideas from natural language processing and machine learning.

N-gram model is an algorithm based on language model. Its basic idea is to put the text in a sliding window of size N in order, and form a list of pieces of text of length N. FastText decomposes each word of the input context based on this model, and gathers all of the pieces of the original word as the semantic information representing the context. Compared to Word2Vec which treats each word in the
corpus as an atom, the FastText model splits words at a finer granularity, using character-level n-grams to represent a word. Introducing subword information, the model can acquire more fine-grained features so that word morphological changes have little affect on the semantics, thus generating semantically more accurate and higher quality word vectors.

Figure 1 shows the model architecture of FastText.

FastText uses word-bags as well as n-gram bags to characterize utterances, the obtained subword information as input. The algorithm uses the softmax function f to generate a probability distribution of predefined classes, i.e., for a set of documents containing N documents, the goal of the model is to solve for the class category $y_n$ corresponding to the nth sample such that it satisfies the following equation (1), where n is the number of samples and $x_n$ is the subword information of the nth document feature.

$$
\min_{y_n} \frac{1}{N} \sum_{n=1}^{N} y_n \log f(x_n)
$$

2.2. Bidirectional encoder representation from transformers (Bert)

The BERT[7] model is a deep bi-directional full connected structure consisting of bi-directional transformer encoders, which essentially learns a good feature representation for words by running a self-supervised learning method on the basis of a large corpus. The encoder corresponding to each word in the BERT structure receives information from both of two sides with all information, thus allowing for a representation of the entire sentence. At the word level, the BERT model uses a masked language model that randomly masks 15% of the words in the sentence and predicts the masked words in order to solve the problem that the information from the predicted unit traverses into the input through the bidirectional propagation structure during the training process. At the sentence level, the BERT model introduces the "next sentence prediction" method in order to allow the model to learn the connection between two sentences better.

The structure of the BERT model is shown in Figure 2, where the input vectors of the model are $E_1$, $E_2$, ..., $E_n$, each TRM corresponding to a Transformer Block. With parameters obtained by training on large-scale unlabeled data, the word vector representation of the input sequence is inferred, i.e., $T_1, T_2, ..., T_n$, so that the word vector has a strong semantic representation capability.
2.3. Text convolutional neural networks (TextCNN)

The TextCNN[5] model is a text classification model based on convolutional neural network, which is mainly composed of input layer, convolutional layer, pooling layer, and output layer. The basic model structure diagram is shown in the Figure 3:

![Figure 3. TextCNN.](image)

The input layer is composed of word vector matrix \( X \in (m,n) \), \( m \) is the size of the dictionary, \( n \) is the dimension of the word vector; the convolutional layer extracts n-gram information between sentences through convolution operations, which usually set different size of convolution kernels to extract more depth of semantic information. The size of convolution kernel can usually be set to \( (1,n),(2,n),..., (p,n) \). The operation of the convolution layer is equation (2):

\[
C = f(X \otimes W + b)
\]  

Among them, \( W \) is the weight matrix of the neural network connection; \( b \) is the bias; \( \otimes \) is the convolution operation. Then the results of the convolution operation are pooled, which mainly including maximum pooling and average pooling. Some pooling results are concatenated, and finally the concatenated vector is fully connected to obtain the classification result.

The TextCNN model easily captures the semantic information between sentences through convolution operations, therefore, the accuracy of the model is guaranteed. The internal parameters of the model are shared. The model uses pooling operation, which is equivalent to a feature dimensionality reduction operation, so some unnecessary parameters are reduced. Therefore, the training speed of the model is faster.
2.4. The proposed method

Based on the above research theory, this paper proposes to use large-scale literature resource data for semantic analysis and data mining algorithm research to complete the automatic construction of subject knowledge thesaurus. The flowchart of the proposed method is shown in Figure 4.

Figure 4. The flowchart of the proposed method.

The whole construction process of thesaurus mainly includes three modules:

1) Build a basic Thesaurus. The seed words are manually labeled by experts in the subject field and the thesaurus classification system is created in this module. The results will be used as the template of model learning and training.

2) Build a professional corpus. Firstly, the source of literature data acquisition is determined, such as Wanfang database, HowNet database and Encyclopedia database. Secondly, the critical data such as literature abstracts and keywords are acquired, and the high-quality professional corpus is screened through seed words. Finally, the multi-source heterogeneous data is integrated, cleaned and preprocessed to generate professional corpus.

3) Build a Knowledge Thesaurus. In this part, the main task is to discover new subject words and classify them automatically. Firstly, it is essential to segment words in professional corpus so that we can use FastText algorithm to train the word embedding model. And in order to find new academic words based on seed words, the semantic similarity is calculated. Secondly, the semantic expansion of academic words is carried out, then the training of multi-label classification algorithm and the automatic classification inference of new academic words are completed based on Bert and TextCNN algorithm. Finally, the domain experts review the generated knowledge thesaurus.

The proposed model structure is shown in Figure 5. The model consists of FastText, Bert and
Figure 5. The proposed model.

TextCNN as basic units. First, all the corpus in fastText unit are used as input data to represent the words by vector in the word embedding layer. The similarity layer is calculated to search the similar words centered on the seed words. Secondly, in Bert unit, the complete semantic expansion of words is implemented to form short sentences containing words. The vector coding of short sentences is carried out in the Bert semantic embedding layer to generate token embeddings, segment embeddings and position embeddings respectively, then the semantic vector of Bert is generated through the pre-training model. Finally, in TextCNN unit, bert semantic vector is used as model input, convolution, pooling and flatten operation are carried out in turn. The softmax classification results of the semantic extension short sentences where the words are located are obtained as the final model output.

This paper proposes an intelligent knowledge base construction algorithm based on academic data, which can automatically construct the knowledge structure of the subject from a large number of documents. The algorithm has a high accuracy that significantly improves the efficiency of knowledge organization and reduces the maintenance cost.

3. Case Study

3.1. Data description
In order to verify the performance and effectiveness of the algorithm, this paper takes the field of marine science as the research case to construct the knowledge thesaurus in this field. 8.1 million
heterogeneous marine literature data from Wanfang, HowNet, Baidu Encyclopedia and Wikipedia are used as the experimental original data set, in which 5.66 million key words are used as self-defined word segmentation dictionary. The construction of knowledge classification system of thesaurus comes from the National Natural Science Foundation of China (NSFC), as shown in the Figure 6, which contains 15 secondary subject codes and 143 tertiary codes. In addition, there are 3361 seed words marked by marine experts. The data of the academic words classification model is 980000, and the data are divided into training set and test set according to the ratio of 8:2.

3.2. Data Cleaning and preprocessing
It is common knowledge that data cleaning and preprocessing process is essential in the case study. The data comes from multiple literature databases, there are problems of different data storage formats, duplicate data and text containing null values, garbled codes and other symbols. Therefore, it is necessary to integrate, eliminate duplications and clean multi-source heterogeneous data to ensure the accurate representation of semantic vectors and the effect of model training.

Besides, word segmentation is the basis of Chinese text processing, and the quality of word segmentation results directly affects the effect of academic new word mining. Therefore, before training the fast text model, we need to segment the corpus based on spark big data distributed computing firstly, and use the user-defined stopword list to remove irrelevant words that have a negative effect on academic semantic analysis.

In the semantic extension layer of the model, the training samples of new word classification algorithm need to be formatted, which mainly includes extending the semantic short sentences of each seed word to 300 sentences at most, intercepting the length of each sentence to be less than 128 characters, which is the input data of the next classification model. The specific formatted data set is shown in Table 1.

| Class            | Seed word       | Semantic extension sentence                                           |
|------------------|-----------------|---------------------------------------------------------------------|
| Storm Surge       | Typhoon storm   | The typhoon storm surge and the storm current are frequently         |
| Disaster          |                 | natural hazards for the ocean engineering                            |
| Red Tide/Green    | Enteromorpha    | Distribution of temperature, salinity, dissolved oxygen, nutrients   |
| Tide Disasters    |                 | and their relationships with green tide in Enteromorpha prolifera    |
|                   |                 | outbreak area of the Yellow Sea                                      |
| Red Tide/Green    | Green tide      | As a result, the strong reproductive capacity of floating green      |
| Tide Disasters    | algae           | tide algae Ulva prolifera might be the main reason of green tide     |
| Tide And Tidal    | Tidal prediction| The work of tidal prediction is very important to understand and      |
| Current           |                 | make use of the ocean, it is indispensable not only in citizens      |
|                   |                 | life but also in military                                           |
| Ocean Waves       | Typhoon waves   | The simulated result of typhoon waves simulation and prediction      |
|                   |                 | will be more effective with the surface wind provided by a mesoscale |
|                   |                 | tropical cyclone numerical model, than by the empirical wind mode    |
| Ocean Mixing      | Water salinization | Studies on the changes of several ecological factors during the     |
|                   |                 | process of water salinization in Daihai lake, China                 |

3.3. Results and analysis
In our experiment, we used the NVIDIA GPU serve to train our model. The following optimal parameters are obtained through multiple experimental validations. In FastText units, the size of word vector is set by 200, the max length of word n-gram character is set by 3 and the context window size
is set by default value 5. In Bert units, the length of max sequence is set by 128. In TextCNN units, the filter size \( \in \{1,2,3\} \), the numbers of kernel is 128 and the activation is relu in convolution layer. We use the dropout layer to avoid overfitting and the rate is 30%. We use the Focal Loss function to calculation the model loss. Finally, we use two folds cross validation to train model, which is training 10 rounds for each fold, then the batchsize is set by 32.

Semantic vector is trained by the FastText model through the prepared marine discipline literature corpus, and further mining academic new words. We use the t-SNE tool to project the trained high-dimensional space word vectors into the two-dimensional space coordinates, and then according to calculation the similarity, the discovery results of academic space words centered on the seed words are intuitively analyzed. As the seed word shown in the Figure 6 is “Submarine pumice”, and the most similar words are “Submarine fault”, “Submarine Highland”, “Submarine trench” and so on, the similarity is 0.90, 0.88, 0.87, which is in our judgement limit, we join the words into the marine new word candidate set. If the word is so far away from the seed word, such as “Gully terrain”, “Ecological zone”, and “Primitive ocean” are not new marine academic terms that we have discovered.

Figure 6. 2D space vector model.

In the process of classifying marine words into the knowledge system, we trained the Bert_TextCNN model to complete the multi-label classification task by giving the sample data after the semantic extension through seed words. The optimizer is set by “Adam”, the learning_rate is set by 1e-5, the loss is set by Focal Loss, and the metrics is set by “acc”. The loss and the accuracy during the training process is shown in Figure 7.

Figure 7. The loss and accuracy.
As we can see from the figure, as the number of epochs are increasing, the loss function values are gradually decreasing, and accuracy is gradually increasing. When the epoch is 6, the loss function value and accuracy keep unchanged, and the model reaches convergence. The training time for each epoch is about 5320s, and the 32GB GPU memory occupancy rate is 50%. Part of the prediction results of marine words are shown in the Table 2. It can be seen from the table that the category inference of candidate marine new words has a certain degree of accuracy. There is a strong semantic correlation between the new words and the predicted categories, which fully refers to the academic words semantic expansion information.

| word                      | class_code1 | Class_name1 | class_code2 | Class_name2                      |
|---------------------------|-------------|-------------|-------------|----------------------------------|
| Tidal current and sediment| D060103     | Tide And Tidal Current | D060303     | Marine Sedimentary Dynamics      |
| Glacial facies            | D061506     | Polar Cryosphere | D060101     | Descriptive Physical Oceanography|
| Wave coupling             | D060102     | Ocean Waves  | D060107     | Marine Power                     |
| Laminar boundary layer    | D060107     | Marine Power  | D060105     | Ocean Mixing                     |
| Intertidal shoal          | D060301     | Marine Geomorphology | D060303     | Marine Sedimentary Dynamics      |

Through deep mining of marine literature data, AI algorithm technology is used to construct a marine subject knowledge thesaurus, which is finally expanded to 40000 words through 3000 seed words, as shown in Figure 8. It can be seen from the figure that the entire knowledge thesaurus is composed by a three-layer knowledge system and the underlying terminology. The knowledge system represents the subdivision of the subject field, and the underlying terminology represents the subject direction or topic in the subdivision field. The relationship is many-to-many, that is, a classification system has multiple term words, and a term word may belong to multiple classification systems. The knowledge thesaurus has the characteristics of hierarchical and academic, which can be used for knowledge management in the field of science and technology services, etc., and it has strong practicability.

Figure 8. marine subject knowledge thesaurus.
4. Conclusion
In the era of knowledge explosion, Subject Thesaurus is an important tool for knowledge organization 
and management. In the past, it was usually finished manually, so there are problems of low efficiency 
and high cost. In order to solve these problems, this paper uses artificial intelligence algorithm to carry 
out a research based on the massive literature resource data of thesaurus construction algorithm 
innovatively, which can replace 90% of the manual workload and the remaining 10% is reviewed by 
experts, this will ensure the quality. We use marine science to test and classify new words 
automatically by large-scale corpus. The results show that the model proposed in this paper is effective 
and accurate. And the final expert audit pass rate is 82.6%, which greatly improves the efficiency of 
word base construction. The Subject Thesaurus constructed efficiently and accurately by this method 
can effectively help the domain of knowledge service for semantic indexing, organization and retrieval 
of literature resources. In the future work, we will further optimize the model and consider adding 
crowdsourcing technology to improve the efficiency and results of expert audit.

Acknowledgments
This research was supported by Shanghai Science and Technology Innovation Resources 
Center(SSTIR), and Shanghai-Warsaw Joint Lab on Artificial Intelligence.

References
[1] Conrado M, Pardo T, Rezende S O. (2013) A machine learning approach to automatic term 
 extraction using a rich feature set. In: Proceedings of the 2013 NAACL HLT student research 
workshop. Atlanta, pp.16-23.
[2] Zhou Y., Lan M., Wu Y. (2018) Effective Semantic Relationship Classification of Context-Free 
Chinese Words with Simple Surface and Embedding Features. In: Huang X., Jiang J., Zhao 
D., Feng Y., Hong Y. (eds) Natural Language Processing and Chinese Computing. NLPlCC 
2017. Lecture Notes in Computer Science, vol 10619. Springer, Cham. 
https://doi.org/10.1007/978-3-319-73618-1_38
[3] Minaee S, Kalchbrenner N, Cambria E, et al. (2021) Deep Learning--based Text Classification: A 
Comprehensive Review. ACM Computing Surveys (CSUR), 54:1-40.
[4] Joulin A, Grave E, Bojanowski P, et al. (2016) Bag of tricks for efficient text classification. arXiv 
preprint arXiv:1607.01759, 2016.
[5] Kim Y. (2014) Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 
2014 Conference on Empirical Methods in Natural Processing (EMNLP).Doha, pp.1746-1751.
[6] Mikolov T, Chen K, Corrado G, et al. (2013) Efficient estimation of word representations in vector 
space. Computer Science,.
[7] Devlin J, Chang M W, Lee K, et al. (2019) BERT: Pre-training of deep bidirectional transformers 
for language understanding. In: Proceedings of the 2019 Conference of the North American 
Chapter of the Association for Computational Linguistics: Human Language Technologies. 
Minnesota. arXiv:1607.01759.
[8] Vaswani A, Shazeer N, Parmar N. et al. (2017) Attention is all you need. In: Proc of the 31st Conf 
on Neural Information Processing Systems. New York. pp.5998-6008.