Complex & Intelligent Systems (2023) 9:2459–2473
https://doi.org/10.1007/s40747-021-00388-9

ORIGINAL ARTICLE

Knowledge organization of node enterprises' technological innovation under supply chain environment

Qianqian Zhang[1] · Shifeng Liu[1] · Qun Tu[1]

Received: 20 November 2020 / Accepted: 24 April 2021 / Published online: 12 May 2021

Abstract
An improved text classification method based on domain ontology is proposed in this paper to organize the mass information that records node enterprises' innovation activities under the supply chain environment. This method can classify the documents of node enterprises under the supply chain without a training set. It achieves a precision of 80% for documents' classification, which outperforms the baseline method. Besides, the paper constructs a domain ontology of enterprises' technological innovation under the supply chain that effectively enhances the semantic relationship between words. Therefore, it can summarize and classify the textual information generated by node enterprises in product design, production, storage, logistics, and sales.

Keywords Supply chain · Technological innovation · Knowledge organization · Semantic classification · Domain ontolog

Introduction

With the development of technology and the economy, a new supply chain management model has been formed in the global business community. As a result, the enterprises' technology innovation model has been changed from a single enterprises' original independent innovation to a collaborative innovation model of upstream and downstream enterprises in the supply chain. The supply chain involves multiple entities such as suppliers, manufacturers, retailers, and customers. The innovation activities and processes of all entities form the enterprises' technological innovation in a supply chain. Therefore, the supply chain entities should innovate collaboratively to improve the entire supply chain's competitiveness. Figure 1 shows the main entities in a supply chain and the process of technological innovation. The node enterprises and upstream enterprises in the supply chain need to convey the market supply and cost information. Node enterprises in the supply chain need knowledge sharing and integration. Node enterprises and downstream enterprises or customers need to transfer the market demand and product information. Therefore, the information collection, classification, and knowledge system reconstruction of the entire supply chain is the key to promote an enterprise's technological innovation, which is conducive to enhance product competitiveness and even the whole supply chain.

Due to the complexity and the huge amount of information produced by innovation under the supply chain, a text classification method that can automatically process, organize and mine textual data is highly demanded. However, most of the existing studies focus on exploring influencing factors and cooperation modes of innovation under the supply chain by the empirical study [1–4]. A complete knowledge system should be built to search, organize, and analyze each node enterprise's knowledge to develop the text classification method. Furthermore, it makes information sharing, synchronize planning, and process coordination between members across different regions and industries come true.

Therefore, this paper constructs an ontology of the enterprise's technological innovation field under the supply chain environment and classifies and summarizes the textual information. This method can generate the influential factors of innovation under the supply chain dynamically, providing researchers or managers with influential factors of innovation under the supply chain and understanding the production knowledge dynamically in this field. Besides, the knowledge organization and sharing for node enterprises can realize enterprises' continuous innovation and enhance the

✉ Qun Tu
17113133@bjtu.edu.cn

[1] School of Economics and Management, Beijing Jiaotong University, No.3 Shangyuancun, Haidian, Beijing 100044, China

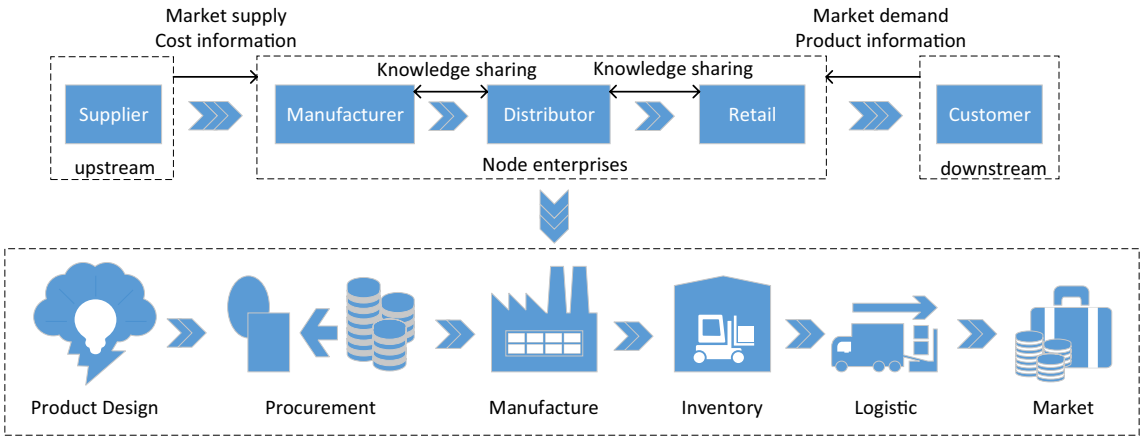

Fig. 1 The structure model of main entities in the supply chain

entire supply chain's competitiveness. The remainder of this paper is organized as follows. The second section presents the literature review on existing semantic text classification algorithms and semantic similarity methods based on ontology. The third section provides the implementation process of the proposed methodology. The fourth section presents the experiments, results analysis, and performance evaluation. The final section concludes the work and contribution of this paper and presents the limitation and future works.

Literature review

Supply chain management and enterprises' technological innovation

The supply chain is a functional network built around by the core enterprises [5, 6]. It revolves around the core enterprise and connects suppliers, manufacturers, distributors, retailers, and end-users through information flow, logistics, and capital flow. Supply chain management spans all activities from raw materials to final products. The synergy of demand and supply brings competitive advantages for enterprises in terms of value and cost. Technological innovation in supply chain management helps enterprises to reduce procurement costs and production costs.

Most researchers explored the association between efficient supply chain management and enterprises' innovation by empirical inquiry or survey methods [1–7]. An increasing number of scholars recently realized the importance of data analysis and text mining for supply chain management. Schniederjans et al. [8] enhanced the supply chain digital research paradigm through a large-scale literature review and a textual analysis of digitization technologies and topics. Kim et al. [9] explored sustainable supply chain management trends and firms' strategic positioning and execution based on news articles and sustainability reports with text-mining algorithms. Chu et al. [10] proposed a text-mining-based global supply chain risk management framework to identify region-specific supply chain risks. Chircu et al. [11] presented research examining the use of business analytics, big data, and business intelligence methods in operations and supply chain management by analyzed 625 published papers with text mining. Rozados et al. [12] concluded the trend and related research of big data analysis in supply chain management.

Semantic text classification algorithms

Text classification is a text-mining algorithm that automatically assigns the analyzed document to one or more predefined categories based on its content [14]. Traditional supervised text classification methods such as Support Vector Machines (SVM), Naïve Bayes, decision trees, and Latent Semantic Analysis (LSA) K-Nearest Neighbor (KNN) generally presented by the terms and their feature weights, also known as the "Bag of Word" (BOW) representation model. The number of words determines the word vector dimension in the vocabulary, which usually results in a very high and sparse dimensional document vector [15–21].

Ontology is a conceptual, structured, and standardized knowledge representation and organization method that can describe semantics and hidden knowledge from enormous amounts of information. Using domain ontology for knowledge representation can explore similar topics or events in the documents. Hence, it can construct a text representation model with the pre-defined semantic relationships between recognized entities and knowledge from the ontology and augment it with important background facts that are not directly present in the document. With this knowledge, the system can distinguish which terms or concepts

are more important and focus on categorizing more precise information.

1. Some researchers utilized the domain ontology to enrich the semantic feature vector representation and improve text classification accuracy. For example, Elhadad et al. [22] proposed building the feature vector for web text document classification based on the WordNet ontology. Abdollahi et al. [23] utilized the UMLS domain ontology to extract the key features and classify the medical text document.
2. Some researchers utilized the hierarchical taxonomy of domain ontology in the text classification task. For example, Cerri et al. [24] classified proteins in functions organized according to the Gene Ontology hierarchical taxonomy. Liu et al. [25] proposed the text classification method based on the ontology graph and structure.
3. Some researchers proposed a method based on the semantic similarity of concepts in the ontology for text classification. For example, Albitar et al. [26] proposed new text-to-text semantic similarity measures to replace classical similarity measures for text classification.

There is no research that utilizes the big data techniques for knowledge organization of enterprises' technological innovation under the supply chain environment from the above literature survey. The traditional text classification methods are usually represented by BOW, which ignores the semantic relationship between terms and usually requires a large number of labeled training texts, which increased manual annotation workload. Using the hierarchy of knowledge from domain ontology directly in the text classification process can obtain the semantic relations between terms and directly skip classifier construction training steps without any pre-categorized training sets.

Therefore, there are two research points in our paper. First, use the big data techniques to automatically process, organize, and mining the large amounts of textual data generated by the node enterprise's technological innovation and realize the knowledge service among node enterprises in the supply chain. Second, an improved text classification method does not require a large amount of training text to automatically organize and analyze the large amounts of textual data generated by the node enterprises' technological innovation to realize the knowledge classification of enterprises' technological innovation.

Methodology

Therefore, this paper utilizes the semantic concept model based on the domain ontology of enterprises' technological innovation under the supply chain to improve the text classification and proposes an enhanced text classification method based on the semantic similarity and relatedness between keywords and categories. This paper mapped the target categories and the keyword sets extracted from the collected textual documents to constructed domain ontology concepts. Then, the mapped target category-concept set and keywords-concept set are obtained. The domain ontology-based semantic similarity calculation and the concept distribution-based relatedness calculation are used to obtain the weight matrix of semantic similarity and relatedness between keywords and categories. Compared to the maximum weighted value of semantic similarity and relatedness between keywords and categories in the matrix's transverse space, the document categories can be obtained by the category corresponding to the keyword with the maximum value. The framework of the process on the improved text classification method based on the semantic conceptual model is shown in Fig. 2. According to the framework, there are mainly four steps in the improved methodology, and the detail is as follows. The main parameters used in the following equations are shown in Table 1.

Text preprocessing

The module of text preprocessing mainly includes word segmentation, part-of-speech tagging, and stop word removal. First, utilize the Python software Jieba to segment the collected textual documents. The result of Chinese word segmentation will lead to the problem that Chinese phrases are incorrectly divided into multiple words, such as the phrase "enterprises technological innovation," which were divided into three small-grained words "enterprises," "technology," and "innovation." Hence, the custom dictionary utilized to defined particular terms in the field, such as "enterprise technological innovation," "product innovation," and "mechanism innovation". Furthermore, tagged the text with part-of-speech (POS), where nouns are more representative and essential to the source document's semantic information. Therefore, nouns, gerundial phrases, adjective-noun collocation were selected as the research objects. Finally, the useless words were filtered through the stop word dictionary, such as "a, the, we, us, they" and other terms with high frequency without meanings. The index structure's size can be significantly reduced by stop word removal, and the keyword sets can be obtained. The general process of text preprocessing was shown in Fig. 3.

Domain ontology-based concept mapping

The key to constructing an improved semantic conceptual vector representation model based on domain ontology is the concept mapping from text keywords to ontology. The concepts of domain ontology are usually defined by attributes,

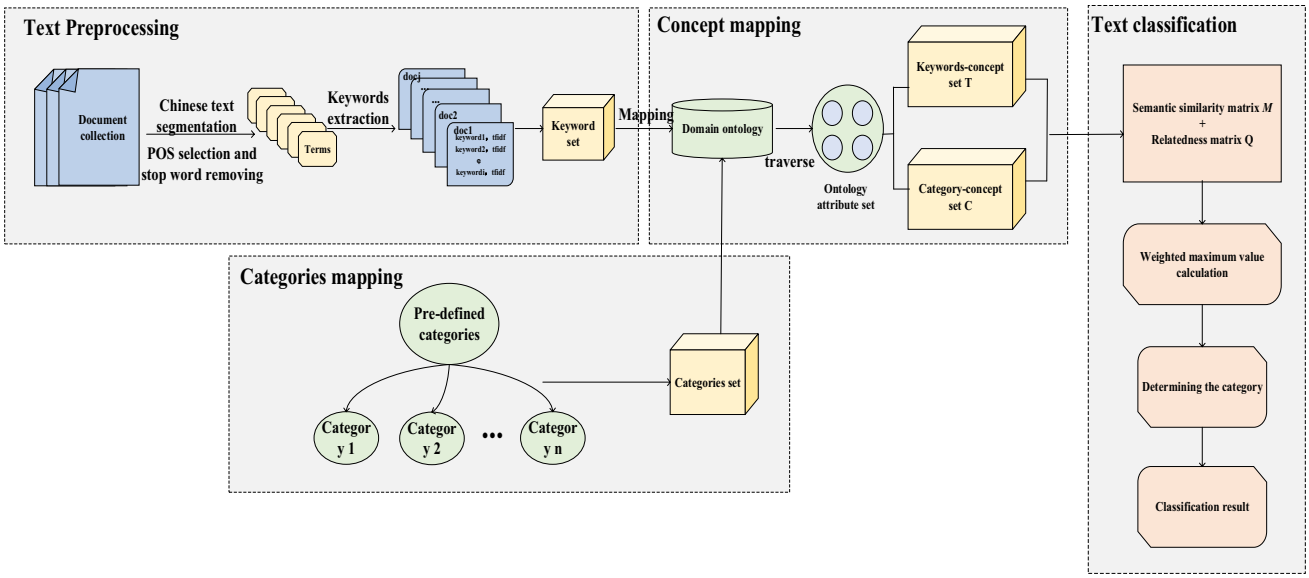

Fig. 2 The framework of improved text classification based on semantic conceptual mode

Table 1 The main parameters in equations' definition

Elements	Definition
t_j	Text keywords
c_i	Concepts in the domain ontology
nc_i	The number of multiple concepts c_i in ontology that matched with the keyword t_j
S_{tc}	The matching degree between the keyword and each concept attribute in the domain ontology
TF	The frequency of the keyword in data set
μ	The threshold value of keyword frequency
K	The rate at which the weight value decreases with the ontology hierarchy
depth(c_j)	The depth from root to concept c_j in ontology
w[sub(c_i, c_j)]	The value to the path of each node in ontology
Dist(c_i, c_j)	The semantic distance between concepts c_i and c_j in ontology
sim(c_i, c_j)	The semantic similarity between concepts c_i and c_j in ontology
λ	The influence factor of semantic distance on semantic similarity
E_{ij}	The keyword pair co-occurrence matrix
$f^k(c_i, c_j)$	The number of times that concept c_i and c_j appear simultaneously in the k words window at the entire corpus
$f^c(c_i)$	The frequency of the concept c_i at the entire corpus
rel(c_i, c_j)	Relatedness between concept c_i and c_j
Sim_Rel(c_i, c_j)	Semantic similarity and relatedness between concept of c_i and c_j
α	The weight of semantic similarity
H_j	The weighted sum of each transverse dimension vector in d_j

keywords, or synonyms in the texts. Hence, there are four situations when mapping text keywords to domain ontology as follows.

1. when the keywords in the dataset cannot be directly mapped with any concepts in the domain ontology, retained the keywords as the unregistered words while the frequency of the keyword is high. Calculate the frequency TF of the keyword, if $TF > \mu$, keep the keyword in the unregistered word set $w\{w_1, w_2, w_3, ..., w_l\}$, otherwise delete the keyword.
2. 1:1 mapping. When the keywords in the text can directly be matched with the attributes in the domain ontology,

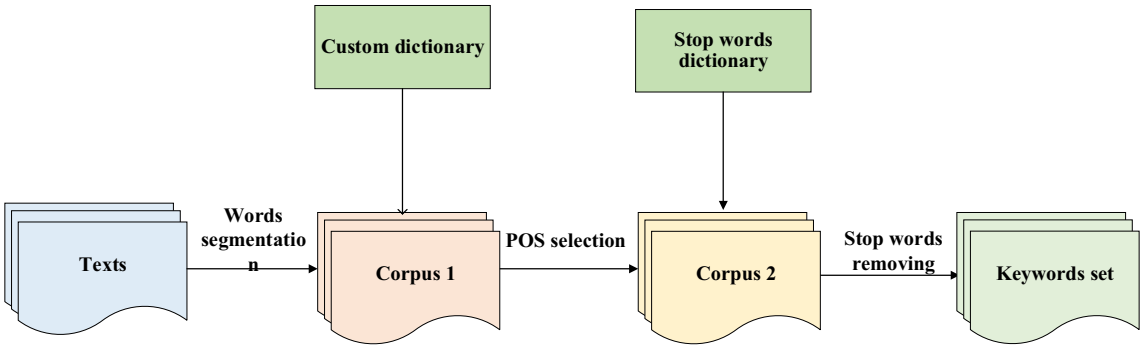

Fig. 3 The process of text preprocessing

the keywords can be directly replaced by the ontology concepts.

3. 1: *n* mapping. When the keyword t_j corresponds to multiple concept attributes c_i in the domain ontology, the mapping concept is determined by the matching degree between the keyword and each concept attribute in the domain ontology shown in formula (1). Selected the maximum value of the concept in S to replace the keyword t_j, where nc_i represents the number of multiple concepts c_i in the ontology that matched with the keyword t_j .

$$S_{tc} = \frac{\sum_i^n \sum_j^m s_{tc}(t_j, c_i)}{|nc_i|}. \tag{1}$$

4. The mapping relationship between keywords and concepts in *n*:1 and *n*:*m*, since the concepts in the domain ontology are usually composed of professional compound words. It is not easy to find concepts that directly and exactly matched the keywords. Therefore, utilize the maximum matching method to map multiple feature items to the same concept in mapping keywords to the domain ontology concepts. There are two situations for mapping keywords to multiple concepts. First, when one or more keywords are cross-mapped to multiple concepts, keep the multiple concepts from multiple keywords mapped to the domain ontology. For example, keywords t_1, t_2 mapped to concept c_1, while keywords t_1, t_2, t_3 mapped to concept c_2 and then kept the concepts c_1 and c_2. Second, when one or more keywords are mapped to multiple concepts without cross-over, keywords are unique in the text and retain the concepts directly.

Semantic similarity and relatedness calculation based on domain ontology

According to the previous literature review on ontology-based semantic similarity measures, this paper proposes a new calculation method that combines domain ontology-based semantic similarity and concept distribution-based relatedness. The proposed method obtained the semantic similarity matrix between concepts by calculating the semantic distance of concepts in the domain ontology, then calculating the relatedness matrix between concepts by co-occurrence frequency in the text, and fused the semantic similarity and the correlation matrix to obtain the final weight matrix.

First, assigned value to each node's path in the ontology and calculated the semantic distance between concepts with the following formula:

$$w[\mathrm{sub}(c_i, c_j)] = \frac{1}{2K^{\mathrm{depth}(c_j)}} + 1, \tag{2}$$

where K represents the rate at which the weight value decreases with the ontology hierarchy, $\mathrm{depth}(c_j)$ represents the depth from root to c_j in the ontology, and the $\mathrm{depth(root)} = 0$. Therefore, the semantic distance Dist of the two concepts can be defined by assigned the path weights between two concepts and shown as follows:

$$\mathrm{Dist}(c_i, c_j) = \begin{cases} 0, & c_i \equiv c_j; \\ w[\mathrm{sub}(c_i, c_j)], & c_i \to c_j; \\ \sum_{c \in s\mathrm{Path}(c_i, c_j)} wc[\mathrm{sub}(c_i, c_j)], & \text{others} \end{cases}, \tag{3}$$

where when the concept nodes c_i and c_j are the same concept, the semantic distance is 0; when there exists a direct path between the concept node c_i and c_j, the semantic distance is the path weight value between the two concepts; when there is an indirect path connected the two concept nodes c_i and c_j, the semantic distance is the sum of the path weights. The path weight assignment formula proposed above has the following properties.

1. The value of the semantic distance between concepts at the upper level in domain ontology is bigger than that at the lower level because that the more abstract concepts

in the ontology hierarchy have less similarity, and the more specific concepts have a greater similarity.
2. The semantic distance between concepts in the parent class and subclass is smaller than the value of the sibling concepts, which indicates that different types of concepts have different weights.
3. There is symmetry in the distribution of path weights between concepts.

The relationship between semantic distance and semantic similarity is inversely proportional. Hence, the semantic similarity $\mathrm{Sim}(c_i, c_j)$ can be calculated according to the semantic distance between concepts. The semantic similarity generally has the following properties.

- $0 \le \mathrm{sim}(c_i, c_j) \le 1$ defined the scope of the semantic similarity. When the c_i and c_j are the same concept, the semantic similarity is 1; when the concept c_i and c_j have nothing in common, the semantic similarity is 0.
- $\forall c_i : \mathrm{sim}(c_i, c_j) = 1$ defined the semantic similarity between c_i and itself as 1.
- $\forall c_i, c_j, c_k$: if $\mathrm{dist}(c_i, c_j) > \mathrm{dist}(c_i, c_k)$, then $\mathrm{sim}(c_i, c_j) < \mathrm{sim}(c_i, c_k)$ defined the relationship between conceptual semantic distance and semantic similarity. If the semantic distance between concepts c_i and c_j is greater than the semantic distance between concepts c_i and c_k, the semantic similarity between concepts c_i and c_j is less than that of concepts c_i and c_k. Therefore, the calculation of semantic similarity is shown as the following formula:

$$\mathrm{Sim}(c_i, c_j) = \frac{1}{1 + \lambda \mathrm{dist}(c_i, c_j)}, \tag{4}$$

where λ is the influence factor of semantic distance on semantic similarity, $0 < \lambda \le 1$.

After the preprocessing of the texts, selected the most representative keywords as the keywords set. To calculate the relatedness of a given keyword pair, the calculation formula of the $i \times j$ co-occurrence matrix E_{ij} generated for the terms in a certain window size k of the corpus is shown as the following formula:

$$E_{ij} = f^k(c_i, c_j), \tag{5}$$

where f^k represents the number of times that concept c_i and concept c_j appear simultaneously in a window containing k words at the entire corpus. The generated co-occurrence matrix E_{ij} was further processed by the mutual information method based on word distribution. The relatedness matrix of concept c_i and c_j was obtained, and the calculation formula is shown as the following formula:

$$\mathrm{rel}(c_i, c_j) = \begin{cases} 1, c_i \equiv c_j; \\ \log 2 \frac{f^k(c_i, c_j)}{f^c(c_i) \times f^c(c_j)}, \text{ others,} \end{cases} \tag{6}$$

where f^k represents the number of times that concept c_i and c_j appear simultaneously in the k words window at the entire corpus. $f^c(c_i)$ and $f^c(c_j)$ represent the frequency of the concepts c_i and c_j at the entire corpus.

The co-occurrence frequency information represents the strength of the content relatedness between concepts in the corpus. The similarity of concepts in the domain ontology represents the strength of the semantic relationship between concepts. Combined the semantic similarity and relatedness between concepts can represent documents more accurately. The following formula is used to normalize and fuse the similarity matrix and co-occurrence matrix of concepts which α represents the weight of semantic similarity:

Table 2 Matrix W generated by keyword t_i and category C_m in document d_j

	C_1	C_2	...	C_m
t_1	W_{11}	W_{12}	...	W_{1m}
t_2	W_{21}	W_{22}	...	W_{2m}
...	...	...	...	...
t_i	W_{i1}	W_{i2}	...	W_{im}

Table 3 The main element definition

Elements	Definition
D	A collection of all text documents
d_j	The ith document in set D
Dic	A collection of all keywords extracted from D
T_j	A collection of keywords in the text d_j
t_i	The ith keyword in T_j
C_m	Target category with total number of m
$\mathrm{sim}(t_i, c_m)$	Semantic similarity between the keyword t_i and the category c_m
$\mathrm{rel}(t_i, c_m)$	Relatedness between keyword t_i and category c_m
W_{im}	Semantic similarity and relatedness between keyword t_i and category c_m

$$\begin{aligned}\mathrm{Sim\_Rel}(c_i, c_j) &= \alpha \times \mathrm{sim}(c_i, c_j) + (1-\alpha) \times \mathrm{rel}(c_i, c_j) \\ &= \alpha \times \frac{1}{1+\lambda \mathrm{dist}(c_i, c_j)} + (1-\alpha) \times \log 2 \frac{f^k(c_i, c_j)}{f^c(c_i) \times f^c(c_j)}.\end{aligned} \tag{7}$$

Improved text classification algorithm

In this paper, the concept model based on the domain ontology proposed above was applied to text categorization. An improved text classification method based on the semantic similarity and relatedness between keywords and categories was proposed. The corpus D contains j documents and denotes as $D = \{d_1, d_2, ..., d_j\}$. First, constructed a vector space model for each text, extracted the keywords whose TF is greater than the threshold μ, sorted the keywords according to TF weight, selected the top 20 most representative keywords and the document d_j can be represented as $d_j = \{(t_1, tf_1), (t_2, tf_2), ..., (t_{20}, tf_{20})\}$. Obtained the keywords set $\mathrm{Dic} = \{t_1, t_2, t_3, ...t_{|\mathrm{Dic}|}\}$ by deleting the repeated words and the pre-defined categories denotes as $C = \{C_1, C_2, ..., C_m\}$. Mapped the keyword set $\mathrm{Dic} = \{t_1, t_2, t_3, ...t_{|\mathrm{Dic}|}\}$ with the target categories set $C = \{C_1, C_2, ..., C_m\}$ with the constructed domain ontology O. And then constructed the similarity and correlation matrix between each keyword t_i in the keyword set $T_j = \{t_1, t_2, ..., t_i\}$ of d_j and the category C_m, then t_i can be expressed as an m-dimensional vector $\{w_{i1}, w_{i2}, ..., w_{im}\}$. The semantic similarity matrix M between the keyword t_i and the category C_m is calculated by the formula (4). The relatedness matrix Q of the keyword t_i and the category C_m based on the word distribution is calculated by the formula (6). The matrix W is obtained by fusing the matrices M and Q through the formula (7). The element W_{im} in the matrix W represents the semantic similarity and relatedness of the keyword t_i to the category C_m the matrix W generated by the keyword t_i and the category C_m in the text d_j can be denoted as shown in Table 2.

Finally, the weighted sum of each transverse dimension vector in d_j is obtained by the formula (8), took the maximum value W_{im} corresponded category C_m as the text category. The improved text classification method based on semantic similarity and relatedness of keywords and categories is described as follows, and Table 3 defines the main elements:

$$H_j = \sum_{i=1}^{m} w_{im}. \tag{8}$$

Algorithm The semantic text classification

Input:

Pre-processed text sets $D = \{d_1, d_2, ..., d_j\}$;
Target categories $C = \{C_1, C_2, ..., C_m\}$;
Domain ontology O.

Begin:

1: load textual data $D = \{d_1, d_2, ..., d_j\}$ and categories $C = \{C_1, C_2, ..., C_m\}$ and extract keywords $Dic = \{t_1, t_2, t_3, ...t_{|Dic|}\}$, map to the concept in ontology O;

2: *For j=1*
Construct the vector of document $d_j = \{(t_1, tf_1), (t_2, tf_2), ..., (t_{20}, tf_{20})\}$, and obtained $T_j = \{t_1, t_2, ..., t_i\}$;
End For

3: *For* category C_m *do*
For text d_j includes t_i *do*
Calculate the semantic similarity by $sim(t_i, c_m) = \frac{1}{1+\lambda dist(t_i, c_m)}$ and obtained the similarity matrix M
End For
End For

4: *For* category C_m *do*
For text d_j includes t_i *do*
Calculate the relatedness by $rel(t_i, c_m) = \log_2 \frac{f^k(t_i, c_m)}{f^c(t_i) \times f^c(c_m)}$ and obtained the relatedness matrix Q
End For
End For

5: Confuse the matrix M and Q;
For W_{im} *in* W
$W_{im} = \alpha \times sim(t_i, c_m) + (1-\alpha) \times rel(t_i, c_m)$
End For

6: *For i in rows of W do*
$H_j = \sum_{i=1}^{m} w_{im}$, selected the *max* H_j and selected the *max* W_{im} of the keyword corresponding category
End For

7: output the category, end;

Output:

The category of text;

Compared with the traditional text classification method based on machine learning, the improved text classification method based on the semantic similarity of keywords and categories has the following advantages. First, the proposed improved text classification method does not require enormous amounts of labeled training text. The method is friendly to the textual data without the label. Second, this

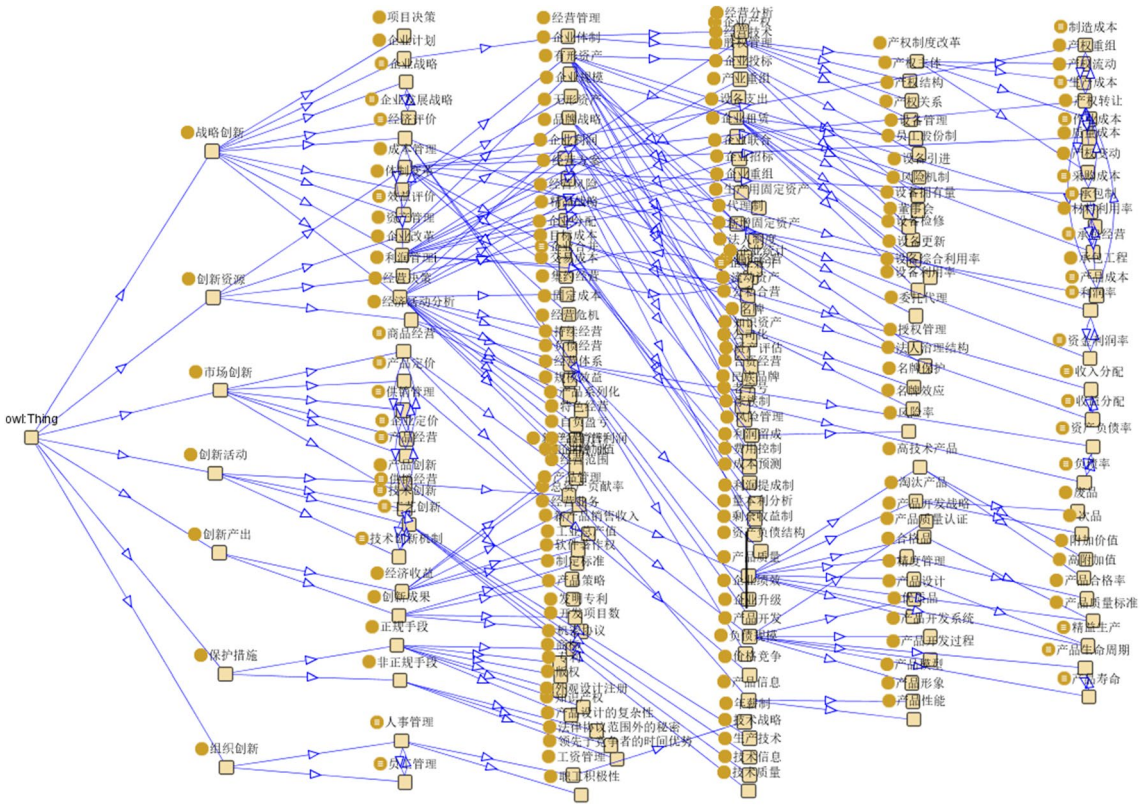

Fig. 4 The domain ontology of node enterprise's technological innovation under supply chain environment

Table 4 Summaries of data collection

Area of focus	Format	Number
Enterprise profiles	.doc	388
Enterprise technical and financial reports	.xlsx	212
Enterprises products	.txt	131
Enterprises rewards	.pdf	136

method uses the domain ontology to map concepts, convert text into low-dimensional space vectors, and reduce space complexity. Thirdly, this method calculates the semantic similarity and relatedness between keywords and categories through domain ontology and overcomes the defect of ignoring the semantic relationship between concepts in the traditional vector space representation method.

Experiment and analysis

This paper presents a methodology for text classification of enterprise's technological innovation under supply chain without the training set. However, to compare the performance with other text classification methods, labeled the collected textual data with pre-defined categories. The result analysis and performance evaluation are as follows. The structure of the enterprise's technology innovation domain ontology under the supply chain environment is shown in Fig. 4.

Dataset

This paper's experimental data mainly consist of enterprises' application form for technical center certification, provided by the Beijing Municipal Commission of Economic Informatization. The textual data consist of 400 enterprises in Beijing, and after data cleaning and selection, there are 867 valid texts, and the overall data size is about 20 M. Table 4 briefly shows the details of the data collection result. The experimental operating environment is Windows 10 system, 2.70 GHz core processor, 8.0 GB memory, and the Python 3.6.2 used for programming. There are seven pre-defined categories: manufacturing capability, innovation resource, mechanism innovation, innovation output, market innovation, protection measures, and innovation strategy. The labeled textual document set was divided into 70% training

Table 5 Labeled textual dataset of node enterprises

doc_id	Partial content	Type_id	Category
0001	Enterprise technical center, enterprise name, Airsys, Refrigeration Engineering, Technology, limited company, industry type, main business, development and production, self-produced product, innovation cooperation, technique fusion, strategic planning…	7	Innovation strategy
0002	Anton Oilfield Services, Technology, limited company, enterprise technical center, innovation system, construction situation, mechanism innovation, QHSE, Technological innovation activities, top talents, organization construction, management system, management system …	3	Mechanism innovation
0003	Tianlong Tungsten-Molybdenum Technology, refractory materials, manufacturing service, equipment, high-technique, product research, manufacture factory, nonferrous metals, high and new technology, emerging industry…	1	Manufacturing capability
0004	Antong construction, limited company, innovation trend, R&D team, core competitive advantage, university-industry cooperation, development tendency, engineering construction, resource integration…	7	Innovation strategy
0005	Airsys, technical center, data center, high availability, air conditioning equipment, refrigeration equipment, technical personnel, senior experts, total assets, equipment value…	2	Innovation resource
0006	Austar Hansen, Packaging Technology, competitive advantage, corporate culture, innovation-driven, innovation management, enterprise development…	5	market innovation
0007	Ankong, Technology Development, innovative product, solution, product seriation, industrial value chain, market demand, market research, core competitiveness, social benefit, brand influence…	4	Innovation output
0008	Orion Energy Technology Development, innovation trend, technological innovation system, operational condition, organizational construction, cooperation innovation, innovation project…	7	Innovation strategy
0009	Babcock Wilcox Beijing Company, innovation trend, competitive advantage, enterprise development, internal resources, innovation research, technical cooperation…	7	Innovation strategy
0010	Bestpower electrical technology, main business, manufacturing, technical process, quality control, product research, R&D expenditure…	1	Manufacturing capability

Table 6 The comparison of experimental performance results

No	Categories	Recall		Precision		F-measure	
		TF*IDF with KNN	Improved method	TF*IDF with KNN	Improved method	TF*IDF with KNN	Improved method
1	Manufacturing capability	0.6397	0.6523	0.8454	0.8633	0.7283	0.7431
2	Innovation resource	0.5762	0.6025	0.8195	0.8032	0.6766	0.6885
3	Mechanism innovation	0.6843	0.7181	0.9031	0.9121	0.7786	0.8035
4	Innovation output	0.5421	0.5717	0.7633	0.7781	0.6339	0.6591
5	Market innovation	0.5325	0.5626	0.7538	0.7643	0.6241	0.6481
6	Protection measures	0.5121	0.5531	0.7328	0.7439	0.6028	0.6344
7	Innovation strategy	0.6635	0.7261	0.8976	0.9021	0.7629	0.8045

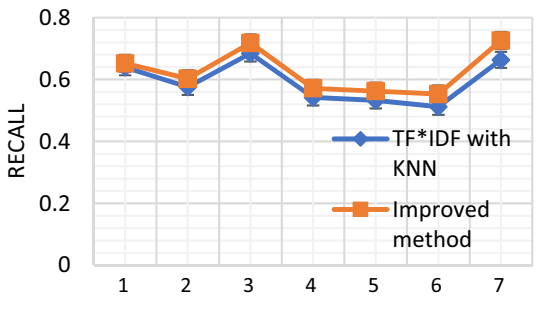

Fig. 5 The performance comparison of text classification based on recall rate

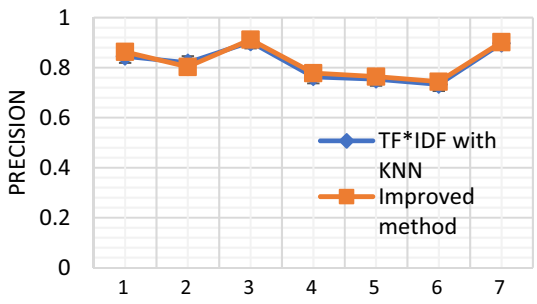

Fig. 6 The performance comparison of text classification based on precision rate

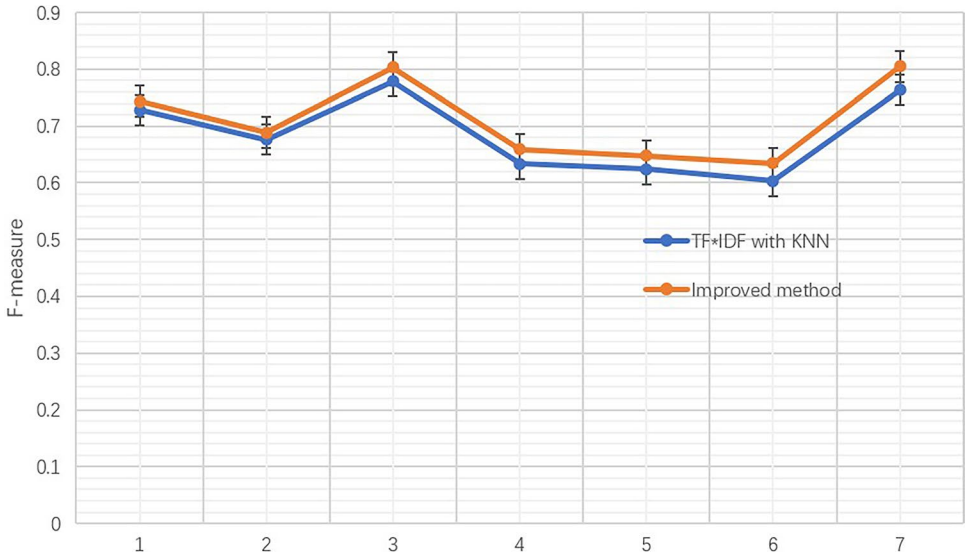

Fig. 7 The performance comparison of text classification based on F-measure

categories / keywords	Manufacturing innovation	Innovation Resources	Mechanism innovation	Innovation output	Market innovation	Protection measure	Innovation strategy
Incentive mechanism	0.409232421	0.627508	0.781177	0.518857	0.462342	0.603356	0.580323
Intellectual property	0.641641855	0.564063	0.683527	0.455306	0.463243	0.839392	0.568743
Patent invention	0.67301271	0.55274	0.486433	0.733763	0.512832	0.602892	0.428746
Organizational construction	0.69839865	0.668527	0.790104	0.616732	0.506782	0.402332	0.520222
Technical personnel	0.438324223	0.654337	0.561444	0.502384	0.432349	0.607395	0.576423
Product research	0.785231447	0.538483	0.454505	0.671037	0.571337	0.432324	0.625185
University-industry cooperation	0.520017564	0.468533	0.578246	0.498363	0.619147	0.320983	0.774316
Internal resources	0.52416122	0.66232	0.521515	0.573007	0.517614	0.356832	0.698937
Market research	0.576206028	0.49886	0.691827	0.503014	0.843423	0.402721	0.517614
Sales profit	0.613100946	0.689746	0.506281	0.793242	0.649051	0.317532	0.452393
Protection management	0.60043323	0.518345	0.532267	0.580918	0.565172	0.62926	0.600433
Technology import	0.416164637	0.804451	0.573007	0.545604	0.535277	0.362765	0.617735
Quality control	0.828452766	0.5989	0.706344	0.50087	0.630923	0.420392	0.563083

Fig. 8 The result of semantic similarity between part of concepts and categories

categories / keywords	Manufacturing innovation	Innovation Resources	Mechanism innovation	Innovation output	Market innovation	Protection measure	Innovation strategy
Incentive mechanism	2.638707275	0.607185	4.0111553	2.3209775	4.1109196	0.928166	1.2048906
Intellectual property	0.251433634	2.5459927	0.0593423	1.4602706	2.0850918	7.5892233	0.1128665
Patent invention	2.015428553	4.148717	0.2323221	7.1657431	0.9199524	4.0381608	0.4852805
Organizational construction	1.636869592	0.3776617	3.9677349	2.565018	0.7512595	1.973433	0.0687809
Technical personnel	4.523854552	3.897988	2.3987741	0.7997024	2.7704133	0.8165206	1.1224684
Product research	3.400643062	1.106084	4.7265101	2.1123473	3.4625604	0.2628881	0.038671
University-industry cooperation	0.238663364	2.0517401	1.8090781	0.2100064	0.0308283	0.4796633	1.7105993
Internal resources	2.418166292	1.7143676	1.9710924	0.0173069	1.746789	0.0273069	1.9582429
Market research	3.756618638	3.0045868	3.5485212	4.4258876	7.0674618	0.4701681	3.6276309
Sales profit	3.740094387	2.9655931	1.7888253	3.9494495	4.467494	0.0253212	0.6239273
Protection management	0.815889047	2.349235	1.928166	2.3554982	3.6494918	7.299783	0.8987964
Technology import	4.067820326	8.4688146	3.3697825	2.194562	0.112351	1.8749777	1.5092883
Quality control	5.670877648	0.1416427	0.5402546	0.6847937	1.4953687	0.0561352	0.0531394

Fig. 9 The result of relatedness between part of concepts and categories

Fig. 10 The result of semantic similarity and relatedness between part of concepts and categories

keywords \ categories	Manufacturing innovation	Innovation Resources	Mechanism innovation	Innovation output	Market innovation	Protection measure	Innovation strategy
Incentive mechanism	1.18954862	0.620395	1.911669	1.149599	1.739344	0.717039	0.798922
Intellectual property	0.505068978	1.257738	0.465062	0.807044	1.03089	3.201833	0.409186
Patent invention	1.142858255	1.811332	0.397494	2.984956	0.655324	1.805236	0.448533
Organizational construction	1.02686348	0.566724	1.902275	1.298632	0.592349	0.952218	0.362218
Technical personnel	1.868259838	1.789615	1.204509	0.606446	1.250672	0.680589	0.767539
Product research	1.700625512	0.737143	1.949707	1.175496	1.583265	0.373022	0.419905
University-industry cooperation	0.421543594	1.022655	1.009037	0.397438	0.413235	0.376521	1.102015
Internal resources	1.187062995	1.030537	1.028867	0.378512	0.947826	0.241498	1.139694
Market research	1.689350442	1.375864	1.69167	1.87602	3.021837	0.426328	1.60612
Sales profit	1.70754865	1.486292	0.955172	1.897915	1.985506	0.215258	0.51243
Protection management	0.675842766	1.159157	1.020832	1.202021	1.644684	2.963943	0.70486
Technology import	1.694244128	3.486978	1.551879	1.122739	0.387253	0.892039	0.929778
Quality control	2.523301475	0.43886	0.648213	0.565243	0.933479	0.292902	0.384603

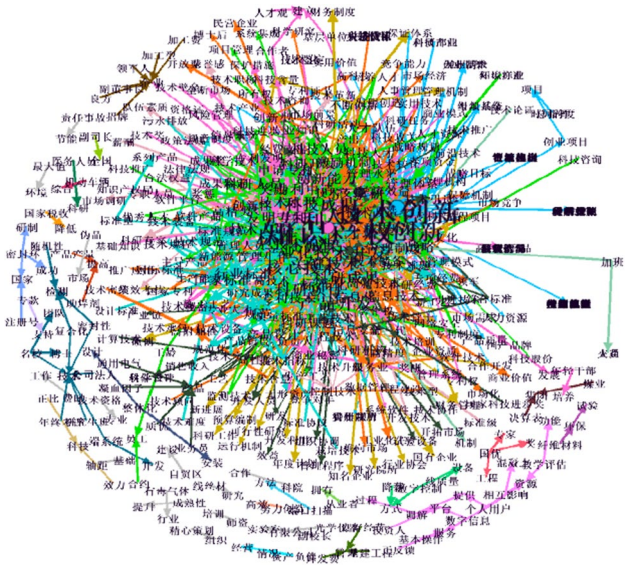

Fig. 11 The visualization of relatedness between concepts and categories

set and 30% test set. Part of the labeled textual dataset is shown in Table 5.

Performance comparison with KNN

According to the different application backgrounds, scholars have proposed various indicators for evaluating text classification systems' performance, including Accuracy, Precision, Recall, F-measure, and Macro-averaging, etc. The most commonly used indicators include precision rate, recall rate, and F-measure. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The recall is the ratio of correctly predicted positive observations to all observations in the actual class. The F-measure combines precision and recall, which is the harmonic mean of precision and recall. The following formulas represent the definition of the three methods.

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \tag{9}$$

$$\text{Recall} = \frac{\text{ture positive}}{\text{true positve} + \text{false negative}}, \tag{10}$$

$$F\text{-measure} = \frac{2 \times \text{Presicion} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{11}$$

This paper used Precision, Recall, and F-measure indicators to compare the proposed text classification method's performance based on the semantic similarity of keywords and categories and the KNN classification method based on TF*IDF.

The above performance comparison analysis shows that the value of Recall rate, Precision, and F-measure on the improved text classification based on semantic similarity and relatedness was higher than the KNN method based on TF*IDF (Table 6; Figs. 5, 6, 7). The mean value of precision of the improved text classification method proposed in this paper over 80%. Therefore, compared with the KNN classification method based on TF*IDF, the proposed text classification based on semantic similarity and relatedness between keywords and categories presented in this paper has better classification performance on texts related to enterprise technology innovation under the supply chain environment.

The improved semantic text classification method proposed in this paper used domain ontology concept sets instead of keywords as each element of the feature vector, enhancing the semantic relationship between words, highlighting the semantic expression, and improving the classification precision rate. The text representation based on domain ontology reduces the space vector's dimension

Table 7 The influencing factors system of node enterprise's technological innovation in supply chain

Type of impact	Meso-level concept	Linking terms
Manufacturing capability	Advanced equipment	Equipment level
		Equipment update
	Process design	Construction technique
		Process technology
		Technical process
		High-tech
	Quality management	Quality control
	Product innovation activities	Product research
		Information technology
		Industrialization
Innovation resources	Quality and quantity of R&D staff	Technical personnel
		Senior engineer
		Senior expert
		Employees number
		Bachelor degree or above
	Fund investment of R&D	R&D expenditure
		Total assets
		Expenditure on science and technology activities
		Main business proportion
	Equipment investment of R&D	R&D Equipment investment
		R&D laboratory construction
	Investment of non-R&D	Technology import
		Asset-liability ratio ownership structure
		Enterprise scale
Mechanism innovation	Staff incentive mechanism	Incentive mechanism
		Performance review
		Post-doctor
		Excellent talents
		Rewards system
	Organizational system management	Organizational construction
		Operating mechanism
		Organizational construction
		Management system
		Organizational structure
Innovation output	Technical output	Patent invention
		Industry standard
		Science and technology progress award
		Method number
		Number of patent applications
		Number of technology development projects
		Number of new product development projects
		Software copyright
		Utility model
		Design patent
	Innovation income	Sales profit
		Main business product sales revenue
		Industrial output
		Industrial added value

Table 7 (continued)

Type of impact	Meso-level concept	Linking terms
Market innovation	Product sale	Market research
		Market competitiveness
		Core competence
		Market-driven
		Market demand
		Development trend
		Products sale
	Economic benefit	Social benefit
		Brand influence
		Diversification
		Business area
		Market occupancy
		Market share
Protection measures	Intellectual property protection	Intellectual property
		Patent warning
		Confidential agreement
		Trade secrets
		Technology protection
		Protection method
	Intellectual property management	Intellectual property protection management
		Independent intellectual property
		Intellectual property application
		Property rights transformation
		Transfer of property rights
		Promote application
		Intellectual property layout
Innovation strategy	Joint innovation	University-industry cooperation
		R&D team
		Research institutes
		Cooperation
		Colleges and universities
	Resource allocation	Internal resources
		Resource Integration
	Technique fusion	Technical cooperation
		Technology exchange
		Technology fusion
	Leadership strategy	Core competence
		Strategic planning
		Overall planning
		Leader

and saves the calculation time. Furthermore, the improved method can also realize the text classification in node enterprise's technological innovation under the supply chain environment without a labeled training set and has a better classification effect. To some extent, this method solves the problem of text classification that lacks a training set due to the enormous workload of manual labeling in reality.

Result analysis

The semantic similarity and relatedness between keywords and categories are calculated based on the domain ontology of node enterprise's technological innovation under the supply chain environment. The result is shown from Figs. 8, 9, 10 and 11. The following shows part of the semantic similarity and relatedness matrix between concepts and categories due to space limited.

The improvement semantic text classification method proposed in this paper can effectively classify node enterprises' collected information in the supply chain and organize the concepts based on semantic similarity and relatedness of enterprise's technological innovation in the supply chain. The concepts here are the key influencing factor of the node enterprise's technological innovation within the supply chain. The classification system for key influencing factors can be obtained through the above experimental analysis of semantic similarity and relatedness between keywords and categories. There are seven types of influencing factors of node enterprise's technological innovation under the supply chain, including manufacturing capability, innovation resources, mechanism innovation, innovation output, market innovation, protection measures, and innovation strategy. According to the semantic text classification, the seven types of first-class factors can be divided into 20 kinds of second-class factors, shown in Table 7.

The influence of manufacturing capability on enterprises' technological innovation is mainly reflected in transforming the R&D results into manufacturing production. The word "quality control" has a high value of similarity and relatedness with the manufacturing capability, reflecting product quality management's content. Hence, the item belongs to the category of manufacturing capability. The innovation resources mainly refer to the enterprises' investment in technological innovation resources. For example, the investment in staff, funds, and equipment in R&D. The mechanism innovation is an innovation activity in various operating mechanisms to enhance the whole enterprise's competitiveness. The innovation output reflects the production of enterprises' innovation and the innovation benefits. Market innovation refers to that innovation in product sales and promotion made by enterprises to meet market demands. Protection measures reflect the content of protection measures of intellectual property. Protection measures reflect the protection measures of intellectual property. Technical knowledge protection can promote technology diffusion and attracting foreign capital and technology introduction. The innovation strategy refers to integrating and arranging the enterprise's internal and external innovation resources and technologies from the overall system with enterprise operation.

Conclusions

The knowledge of enterprises' technological innovation under the supply chain environment is the information source such as the database or documents collected from the supply chain's node enterprises. The knowledge organization is a process of classification and analysis of messy, complex, and huge information. This paper introduces domain ontology to make the knowledge organization system semantic and knowledgeable and constructs an ontology of the enterprises' technology innovation under the supply chain. It utilizes the relationship between the domain ontology concepts to describe the existing enterprises' knowledge management system's semantic information. An improved semantic text classification method was proposed in this paper, which can obtain a document's category by calculating the weighted maximum value semantic similarity and relatedness of the text's key feature words and categories. This method enhances the semantic relationship between words, reduces the space vector's dimension, and saves calculation time. Furthermore, this paper's improved method can classify the document based on the domain ontology hierarchy without a labeled training set—the mean value of precision of the improved text classification method is over 80%.

The contributions of this study are twofold. From an academic perspective, the improved text classification method proposed in this paper had a better performance than the KNN classification method based on TF*IDF. From a practical standpoint, this paper constructs a domain ontology for enterprises' technological innovation under the supply chain from a practical standpoint. It helps to summarize and classify the innovation information under the supply chain, providing researchers or managers with influential factors of innovation under the supply chain and understanding the production knowledge dynamically in this field.

However, there are still some limitations in this paper that future researches should solve. For example, first, the method proposed in this paper requires domain ontology to provide background knowledge and concept mapping. Future researchers may consider using the general ontology that can be applied to more fields. Second, future researchers can consider more influential factors of similarity and relatedness between concepts to increase the word association and improve text classification accuracy.

Acknowledgements This work was supported by Beijing Social Science Foundation under Grant 18JDGLA018, 19JDGLA002, MOE (Ministry of Education in China) Project of Humanities and Social Sciences under Grant 19YJC630043, and was partially supported by Beijing Logistics Informatics Research Base. We appreciate their support very much.

References

1. Modi SB, Mabert VA (2010) Exploring the relationship between efficient supply chain management and firm innovation: an archival search and analysis. J Supply Chain Manag 46(4):81–94
2. Li G, Li L, Choi TM, Sethi SP (2020) Green supply chain management in Chinese firms: innovative measures and the moderating role of quick response technology. J Oper Manag 66(7–8):958–988
3. Ju KJ, Park B, Kim T (2016) Causal relationship between supply chain dynamic capabilities, technological innovation, and operational performance. Manag Prod Eng Rev 7(4):6–15
4. Lee VH, Ooi KB, Chong AYL, Sohal A (2018) The effects of supply chain management on technological innovation: the mediating role of guanxi. Int J Prod Econ 205:15–29
5. Squire B, Burgess K, Singh PJ, Koroglu R (2006) Supply chain management: a structured literature review and implications for future research. Int J Oper Prod Manage 26(7):703–729
6. Mentzer JT, DeWitt W, Keebler JS, Min S, Nix NW, Smith CD, Zacharia ZG (2001) Defining supply chain management. Int J Oper Prod Manage 22(2):1–25
7. Saleem H, Li Y, Ali Z, Ayyoub M, Wang Y, Mehreen A (2020) Big data use and its outcomes in supply chain context: the roles of information sharing and technological innovation. J Enterp Inf Manag. https://doi.org/10.1108/JEIM-03-2020-0119
8. Schniederjans DG, Curado C, Khalajhedayati M (2020) Supply chain digitisation trends: an integration of knowledge management. Int J Prod Econ 220:107439
9. Kim D, Kim S (2017) Sustainable supply chain based on news articles and sustainability reports: text mining with Leximancer and DICTION. Sustainability 9(6):1008
10. Chu CY, Park K, Kremer GE (2020) A global supply chain risk management framework: an application of text-mining to identify region-specific supply chain risks. Adv Eng Inform 45:101053
11. Chircu A, Kononchuk N, Li G, Qi Y, Stavrulaki E (2016) Business analytics and supply chain and operations management–a text mining-based literature review. In: Proceedings for the northeast region decision sciences institute, NEDSI, pp 1–24
12. Rozados IV, Tjahjono B (2014) Big data analytics in supply chain management: trends and related research. In: 6th International conference on operations and supply chain management, OSCM, pp 10–13
13. Sathya S, Rajendran N (2015) A review on text mining techniques. Int J Comput Sci Trends Technol 3(5):274–284
14. Thangaraj M, Sivakami M (2018) Text classification techniques: a literature review. Interdiscip J Inf Knowl Manag 13:117–135
15. Kim HJ, Kim J, Kim J, Lim P (2018) Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning. Neurocomputing 315:128–134
16. Goudjil M, Koudil M, Bedda M, Ghoggali N (2018) A novel active learning method using SVM for text classification. Int J Autom Comput 15(3):290–298
17. Wang Z, Qu Z (2017) Research on Web text classification algorithm based on improved CNN and SVM. In: IEEE 17th International Conference on Communication Technology (ICCT). IEEE, pp 1958–1961
18. Azam M, Ahmed T, Sabah F, Hussain MI (2018) Feature extraction based text classification using *k*-nearest neighbor algorithm. Int J Comput Sci Netw Secur 18(12):95–101
19. Moldagulova A, Sulaiman RB (2017) Using KNN algorithm for classification of textual documents. In: 8th International conference on information technology (ICIT), pp 665–671
20. Thorleuchter D, Van den Poel D (2013) Technology classification with latent semantic indexing. Expert Syst Appl 40(5):1786–1795
21. Kou G, Peng Y (2015) An application of latent semantic analysis for text categorization. Int J Comput Commun Control 10(3):357–369
22. Elhadad MK, Badran KM, Salama GI (2018) A novel approach for ontology-based feature vector generation for web text document classification. Int J Softw Eng Appl 6(1):1–10
23. Abdollahi M, Gao X, Mei Y, Ghosh S, Li J (2019) An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimization. In: 2019 IEEE congress on evolutionary computation (CEC). IEEE, pp 119–126
24. Cerri R, Barros RC, de Carvalho AC (2015) Hierarchical classification of gene ontology-based protein functions with neural networks. In: 2015 international joint conference on neural networks (IJCNN). IEEE, pp 1–8
25. Liu JNK, He Y, Lim EHY, Wang XZ (2014) Domain ontology graph model and its application in Chinese text classification. Neural Comput Appl 24(3):779–798
26. Albitar S, Fournier S, Espinasse B (2014) An effective TF/IDF-based text-to-text semantic similarity measure for text classification. In: International conference on web information systems engineering, Springer, pp 105–114