


Article

Research on Knowledge Gap Identification Method in Innovative Organizations under the “Internet+” Environment

Lin Qi ^{1,2} , Xuejiao An ¹, Shuo Zhang ^{1,3} and Xiang Wang ^{4,*}

¹ School of Economics and Management, Beijing Information Science and Technology University, Beijing 100192, China; qilin@bistu.edu.cn (L.Q.); anxuejiao@bistu.edu.cn (X.A.); zhangshuo99@bistu.edu.cn (S.Z.)

² Beijing World Urban Circular Economy System (Industry) Collaborative Innovation Center, Beijing 100192, China

³ Laboratory of Bid Data Decision-Making for Green Development, Beijing 100192, China

⁴ College of Engineering, China Agricultural University, Beijing 100083, China

* Correspondence: wxzrj@cau.edu.cn; Tel.: +86-1861-832-8592

Received: 5 November 2020; Accepted: 2 December 2020; Published: 7 December 2020



Abstract: Under the “Internet+” environment, the R&D intensity of products and services has increased; hence, organizations need to improve their ability to integrate knowledge and technology resources. Knowledge gaps will arise when an organization’s knowledge reserves fail to meet the needs of innovation activities. This research established a network of complete knowledge topics under the “Internet+” environment based on the Word2Vec model. The word vectors and word frequencies of organizational reserve knowledge texts were analyzed to establish an organizational reserve knowledge topic network. The Term Frequency-Inverse Document Frequency algorithm was used to identify the demanded knowledge topic. The satisfaction capability of demanded knowledge in the reserve knowledge topic network was calculated via the eigenvector centrality and the fuzzy evaluation method. The corresponding strategies were then put forward to make up the knowledge gap. Finally, a case study was conducted and compared with SWOT (Strengths, Weaknesses, Opportunities, and Threats) and Venn diagram analysis on the economic and management college of a university in Beijing to verify the effectiveness of this method.

Keywords: Internet+; knowledge gap; Word2Vec; TF-IDF; eigenvector centrality; fuzzy evaluation

1. Introduction

The rapid development of the Internet presented by information technology has gradually integrated into the real economy and has become widely used in all walks of life. Driven by a new round of scientific and technological revolution, more and more countries and regions have launched Internet development strategies, which take the network as an important means to improve competitiveness and constantly promote the integration of the Internet with various fields of the economy and society [1]. Under the “Internet+” environment, fierce market competition has led to an accelerated pace of product and service research and development, forcing organizations to improve their ability to integrate large amounts of knowledge and technical resources.

When these innovations require more and more knowledge in the form of systems, it becomes increasingly impossible for a single organization to complete the rapid iteration of products and services only by relying on its capabilities due to facing rapid and fierce market competition. Knowledge gaps may arise when an organization’s knowledge reserves and existing technologies fail to meet the needs of innovation activities [2,3]. From a strategic perspective, it is quick and reasonable to choose the path

of differentiation and specialization, cooperate with other organizations or enterprises, and carry out collaborative innovation represented by intellectual property authorization or capability transactions for general organizations.

According to the realistic requirement of collaborative innovation, important means to improve the competitiveness of organizations under the Internet+ environment are the rapid, objective, and quantitative analysis of the organization's knowledge gaps, choosing effective remedies methods, and reducing the uncertainty of innovation activities. Identifying the knowledge gap and choosing the strategy is the key link to avoid the mutual influence between the knowledge gap and R&D investment, business model, development ability, and degree of opening to the outside world to reduce the risk of technological innovation [4–7].

The existing literature on knowledge gap identification can be divided into qualitative analysis and quantitative analysis. (1) Qualitative analysis. Zack has believed that the knowledge gap of an organization comes from the strategic gap; therefore, SWOT analysis of strategic research was adopted to transform the knowledge strategy of an organization into an organizational and technical framework to make up the knowledge gap [8]. This method is a tool for the identified and qualitative analysis of knowledge gaps, but the identified knowledge gaps cannot be quantified [9]. (2) Quantitative analysis. Chen et al. identified the knowledge gap and proposed the strategies for studying the different types of gaps by matching and comparing the knowledge demand set and the stock knowledge set based on Venn diagram analysis. This method quantifies the knowledge gap, but the mathematical form of the set ignores the correlation and hierarchy of knowledge elements; that is, it only studies the knowledge content and ignores the structural characteristics of knowledge. Qiu et al. established a structural model of organizational knowledge demand and knowledge reserve by using a tree structure and identified an organizational knowledge gap by using the tree-matching algorithm without further measures to fill the knowledge gap [9,10]. In addition, whether qualitative or quantitative methods are used, the analysis of knowledge demand and knowledge reserve depends on subjective experience and methods. On the one hand, conclusions are affected by people's preferences and abilities; on the other hand, the digital literature distributed on the Internet is not fully mined.

Therefore, this paper proposes a fast, objective, and quantifiable knowledge gap identification method for collaborative innovation required to iterate organizational products and services under Internet+. It provides a structured representation of the organization's stock knowledge and demanded knowledge and quantitative suggestions on the strategies to fill the knowledge gap, which will become the basis for organizations to choose reasonable competition and cooperation strategies in the future. This paper focuses on research on the knowledge gap identification method in innovative organizations under the "Internet+" environment and builds a network of complete knowledge topics under the "Internet+" environment based on the Word2Vec model. In the context of the complete network of knowledge topics, the reserve knowledge topic network is built by word vector and word frequency analysis of organizational reserve knowledge texts. Analyses of the degree of organization satisfaction with reserve knowledge based on the eigenvector centrality were conducted. Based on the TF-IDF model to identify the demanded knowledge topics [11–14], the fuzzy evaluation method is used to analyze the ability of the demanded knowledge topics in the reserve knowledge topic network, and the corresponding gap compensation method is proposed. Finally, an empirical study was conducted on the economics and management college of a university in Beijing to verify the effectiveness of this method.

The structure of this paper is as follows: the following section summarizes the definition of the knowledge gap and the research work of domestic and foreign scholars on identification and remedy methods. The third section presents the knowledge gap identification method under the "Internet+" environment. The case-based method is validated in the fourth section. The fifth section compares the proposed method with the existing method, and the sixth section draws conclusions and makes improvements.

2. Literature Review

2.1. The Definition and Identification of Knowledge Gaps

Knowledge gaps play an important role in organizational competition and innovation. There are two typical definitions of the knowledge gap. One definition is that the knowledge gap originates from the strategic gap, which is the gap between the knowledge needed by the organization to implement the strategy and the knowledge actually possessed [8]. Another definition for the knowledge gap is the knowledge that the organization lacks at any moment, but this knowledge is crucial for the survival and growth of the organization and must be filled [15]. In the theoretical research of knowledge gap analysis, significant research works on its basic concepts and recognition methods have been conducted by academic and industry scholars. By analyzing the application scenarios, Vos divided the classification of the knowledge gap into product R&D, manufacturing, marketing, and management and proposed a process paradigm for identifying the needed knowledge by SMEs to explore market opportunities [16]. Chen introduced a Venn diagram to analyze the knowledge gap of the enterprises and provided compensation strategies for different knowledge gaps according to the status of the enterprises' knowledge reserve [10]. Dang proposed a strategy for finding and replenishing technological innovation gaps under the network environment and conducted a strategic analysis of knowledge security [17].

2.2. The Application of Knowledge Gap Identification

In the application research of a knowledge gap, Lafuente-Ruiz-de-Sabando proposed that knowledge gaps for college image and reputation should be identified and compensated by stakeholders in the process of effective resource input in colleges and universities [18]. Based on the analysis of the causes of the knowledge gap in the manufacturing industry, Li built an evolutionary game model revealing the game behavior between manufacturing enterprises and customers and analyzed the model equilibrium points and their stability under different situations [19]. Malhotra et al. analyzed the knowledge gap caused by group diversity in an online platform strategy selection process and proposed four methods to reduce the risks [20]. Qiu et al. studied the construction method of organizations' knowledge structures based on text mining, designed a tree structure knowledge expression method, used a tree-matching algorithm to identify knowledge gaps, and empirically studied the method based on the patent literature of an organization [9,21]. Li et al. established an element matrix with members, knowledge, and goals and proposed a knowledge gap identification method for scientific research teams based on the decomposition of goals and knowledge [22]. In addition, there is research on the analysis of innovative knowledge, construction projects, and gap countermeasures to analyze the impact mechanism of different knowledge gaps [6,23–25].

3. The Identification Method of Knowledge Gap under “Internet+” Environment

3.1. The Construction of the Network of Complete Knowledge Topics under the “Internet+” Environment

The network of complete knowledge topics under the “Internet+” environment is an undirected weighted connected graph $G = (V, E)$. The nodes $V = \{v_1, v_2, \dots, v_m\}$ represent knowledge themes, and the edges $E = \{e_{ij} | v_i, v_j \in V\}$ represent associations between topics. The weight d_{ij} of the e_{ij} is the strength of the association between the topics, and the greater the value, the stronger the association. The process of constructing a network of complete knowledge topics under the “Internet+” environment is the process of determining the strength of association (d_{ij}) between node sets, associatio sets, and arbitrary nodes. Under the “Internet+” environment, the network of complete knowledge topics is the sum of all kinds of knowledge carriers in the form of texts on the Internet, and the content of quantity of knowledge carriers emerge as the topics of knowledge with semantics and co-occurrence similarity. Identifying knowledge topics from the knowledge carriers under the

“Internet+” environment and obtaining the associations between different topics is key to building a network of complete knowledge topics.

Based on the Word2Vec model, this study represents the networked knowledge carrier text as the semantic information of the word vector so that the knowledge topics are semantically vectorized and the semantic similarity is obtained. The Word2Vec model is a shallow neural network. By inputting words and context information, words can be mapped to embedded space vectors without supervised learning, and the curse of dimensionality could be avoided with the dense mapping of continuous dimensions to realize the semantic vectorization of words [26,27]. The Word2Vec model includes the continuous bag of words (CBOW) and the skip-gram model. The CBOW model is suitable for predicting input words for a given contextual semantics, and the skip-gram model is suitable for a given input word and predicting contextual semantics. In this study, the stock knowledge analysis of the innovative organization belongs to the process of semantic prediction of the given input words, therefore the skip-gram model is selected. In the skip-gram model, ω_t represents the current word, the c value represents the length of the context, and $p(\omega_i|\omega_t)(t-c \leq i \leq t+c)$ is the probability that the current word will appear together with a word in the window. The training goal of the model is to maximize the H value, where T is the length of the text.

$$H = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(\omega_{t+j}|\omega_t) \quad (1)$$

After each round of training, the Softmax classification function is used to calculate the loss and perform backpropagation. After the training, the vector representation $v_i = (S_{i1}, S_{i2}, \dots, S_{in})$ of the subject v_i can be extracted from the hidden layer of the neural network, where n is the vector dimension, and S_{in} is the value of the dimension of the vector. The Pearson correlation coefficient (PCC) is used to express the semantic relevance d_{ij} of the topic v_i and v_j .

$$d_{ij} = \frac{\sum_{k=1}^n (S_{ik} - \bar{S}_i)(S_{jk} - \bar{S}_j)}{\sqrt{\sum_{k=1}^n (S_{ik} - \bar{S}_i)^2} \sqrt{\sum_{k=1}^n (S_{jk} - \bar{S}_j)^2}} \quad (2)$$

where \bar{S}_i, \bar{S}_j is the mean of all dimensions in the vector representation of the topics, and the larger the value of P_{ij} , the stronger the semantic association between the topic v_i and v_j .

3.2. The Construction of Reserved Knowledge Topic Network

Under the “Internet+” environment, the establishment of the network of complete knowledge topics reflects the distribution and association of social knowledge topics. For an innovative organization, it focuses on the cluster of topics formed by one or more knowledge topics, and the relation of the topics is also different from the social average. Therefore, it is necessary to establish an organization’s reserve knowledge topic network and identify the reserve knowledge topics and topic associations.

The innovative organizations’ reserve knowledge topic network $G' = (V', E')$ is also an undirected weighted connected graph, where $V' \subseteq V, E' \subseteq E$. The topic relation of the reserve knowledge topic network takes co-occurrence associations and semantic associations into consideration at the same time. The semantic relevance is obtained by the Word2Vec model in Section 2.1. The co-occurrence correlation degree is obtained by the co-occurrence frequency analysis of the topic words and quantified by the Ochiai coefficient. The relationship is shown in Equation (3).

$$O_{ij} = \frac{tf_{ij}}{\sqrt{tf_i \times tf_j}} \quad (3)$$

tf_{ij} represents the total number of keywords v_i and v_j in the same document, tf_i is the total number of occurrences of the topic words v_i , tf_j is the total number of occurrences of the topic words v_j , and O_{ij} is the Ochiai coefficient between the topic words v_i and t_j . The larger the value of O_{ij} , the greater the co-occurrence relationship between the topics v_i and v_j [28,29]. Finally, the degree of association between the topics v_i and v_j in the knowledge reserve semantic network can be expressed as:

$$R_{ij} = 1 - (\beta O_{ij} + (1 - \beta) d_{ij}) \quad (4)$$

where $\beta \in (0, 1)$ is the weight coefficient [30]. The correlation between the knowledge topic in the reserve knowledge topic network can be used to amend the topic relevance of the network of complete knowledge topics. The specific formula is as follows:

$$d_{ij} = \begin{cases} R_{ij} & R_{ij} > d_{ij} \\ d_{ij} & R_{ij} \leq d_{ij} \end{cases} \quad (5)$$

3.3. The Required Knowledge Topic Identification

The TF-IDF algorithm is used to extract the topic words in the text of the demanded knowledge carrier [31,32] and realize the identification of the demanded knowledge topic. The TF-IDF algorithm considers both the word frequency and the reverse document frequency. From the perspective of the word frequency, the higher the frequency of a word in a single document, the more prominent the topic is represented by the word. From the perspective of the reverse document frequency, it is considered that a word appears in all documents. The frequency of occurrence is high, the general importance of the word is high, and the topic represented is less significant. The word frequency tf_{ij} is denoted as follows:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}} \quad (6)$$

where n_{ij} is the frequency of word v_i in document d_j , and $\sum_k n_{k,j}$ is the total frequency of all words in document d_j . The reverse document frequency idf_i is expressed as follows:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (7)$$

where $|D|$ is the total number of documents describing demanded knowledge, and $|\{j : t_i \in d_j\}|$ is the number of documents including v_i . At the same time, word frequency tf_{ij} and reverse document frequency idf_i are considered. The importance of a topic word is shown in Formula (8).

$$tfidf_i = tf_{ij} * idf_i \quad (8)$$

The importance threshold is set to α , and the demanded knowledge topic set is $V'' = \{v_i | v_i \in V \text{ and } tfidf_i > \alpha\}$. Calculate $tfidf_i'$ according to Formula (9):

$$tfidf_i' = \frac{tfidf_i}{\sum_{j=1}^n tfidf_j} \quad (9)$$

as the weight of v_i in demanded knowledge topic set. Thus, the weight coefficient matrix of the demanded knowledge topic set V'' can be written as follows:

$$A = (tfidf_1', tfidf_2', \dots, tfidf_n') \quad (10)$$

where n is the number of elements in V'' .

3.4. Knowledge Gap Identification and Filling

On the basis of constructing the knowledge collection network, the reserve knowledge topic network, and identifying the demanded knowledge topic in the previous sections, this section identifies the knowledge gaps in the demanded knowledge topic and proposes corresponding compensation methods. This study suggests that whether the demanded knowledge topics v_i can be met within the organization depends on whether the topic exists in the reserve knowledge topic network and whether the topic is important in the reserve knowledge topic network. This standard for measuring importance considers both the number of related neighbor topics and the importance of the neighbor topics. Therefore, the centrality of the eigenvector in the complex network model can be used to describe the importance of a knowledge topic in the organization's reserve knowledge topic network. The formula as follows:

$$E_i = \begin{cases} 0 & v_i \notin G' \\ \gamma \sum_{j=1}^n R_{ij} E_j & v_i \in G' \end{cases} \quad (11)$$

where γ is a proportional constant, R_{ij} is the degree of association of topics v_i and v_j in the reserve knowledge topic network, E_i is the centrality of the eigenvector of topic v_i in the reserve knowledge topic network, and E_j is the centrality of eigenvector of the neighbor topic v_j of topic v_i .

Among them, for $v_i \in V''$, E_i reflects the ability of the reserve knowledge topic network to meet the demanded knowledge topic v_i . However, due to the ambiguity of the organization's reserve knowledge topic, the demanded knowledge topic, and the knowledge gap, the ability to satisfy accurate calculations may not be optimal. Therefore, the ability to satisfy the knowledge topic needs to be blurred. In order to map the exact eigenvector centrality to the fuzzy domain of knowledge satisfaction, a fuzzy evaluation set needs to be established as follows:

$$U = \{u_1, u_2, u_3\} \quad (12)$$

where u_1, u_2, u_3 indicate that the knowledge topic satisfies ability a poor, general, and good degree of membership, respectively, and its linear degree of membership function is defined as follows:

$$u_1 = \begin{cases} 1 & E_i < a \\ \frac{b-E_i}{b-a} & a \leq E_i < b \\ 0 & E_i \geq b \end{cases} \quad (13)$$

$$u_2 = \begin{cases} \frac{E_i-a}{b-a} & a \leq E_i < b \\ 1 & b \leq E_i < c \\ \frac{d-E_i}{d-c} & c \leq E_i < d \\ 0 & E_i < a, E_i \geq d \end{cases} \quad (14)$$

$$u_3 = \begin{cases} 0 & E_i < c \\ \frac{E_i-c}{d-c} & c \leq E_i < d \\ 1 & E_i \geq d \end{cases} \quad (15)$$

where $0 \leq a < b < c < d \leq 1$. The fuzzy membership of the subject v_i in the demanded knowledge topic set V'' for each capability level can be calculated according to Equations (13)–(15). The fuzzy relation matrix can then be obtained by:

$$R = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & u_{n3} \end{pmatrix} \quad (16)$$

Through the compatibility of fuzzy relations, the fuzzy evaluation vector for knowledge satisfaction ability is obtained when the organization is oriented to the demanded knowledge topic set.

$$B = A \circ R \quad (17)$$

The organizational knowledge satisfaction ability is identified according to the membership degree distribution of each component in the vector B . If the evaluation of satisfaction ability is good, it means the demanded knowledge topic set V'' has corresponding themes in the reserve knowledge network. The topic also occupies the network center position with the neighbor topics, indicating that the organization's demanded knowledge can be fully met within the organization when the organizations are engaged in knowledge innovation activities on such topics for a long time. If the evaluation is poor, it means the knowledge topic involved in the demanded knowledge topic set V'' does not appear in the reserve knowledge network or the corresponding topic and neighbor topic are at the edge of the network and need to seek out-of-organization support. If the evaluation is general, it means the demanded knowledge set V'' has a corresponding topic in the reserve knowledge network; however, if the neighboring topics do not occupy the ideal network center position, the organization can gradually meet the demanded knowledge topic through special training and other means.

4. Case Study

In order to analyze the effectiveness of this method, an empirical study was conducted in the college of economics and management at a university in Beijing. The college has eight undergraduate majors, including economics, international trade, accounting, financial management, marketing, business administration, quality management, and human resource management, with three authorized disciplines of first-level Masters, namely management science and engineering, business administration, and applied economics. It has a research foundation in econometrics, knowledge management, science and technology management, quality management, human resources, asset evaluation, securities investment, corporate growth and mergers and acquisitions, financial accounting, business management teaching, and circular economy.

The Chinese Wikipedia corpus is the complete knowledge set under the "Internet+" environment, using the PyNLPIR (Chinese word segmentation system) provided by the Chinese Academy of Sciences for word segmentation and the Word2Vec model for semantic modeling [33]. In this study, the skip-gram model in Word2Vec is used for training. The dimension value of the word vector was set to 200. For the power-law distribution of the word frequency in the corpus, the low-frequency words with a word frequency of fewer than five times were filtered to reduce the size of the knowledge collection network.

In order to establish an organization's reserve knowledge network, the reserve knowledge text carrier was obtained through the China Knowledge Network full-text database. The college was used as the author unit to search for journal articles published in the range of 2014–2018, and 336 articles were obtained. The topics and abstract information of the above articles were exported, and the PyNLPIR thesaurus was used to segment the abstract information: (1) Semantic modeling was performed using the Word2Vec model. The skip-gram model was selected, and the dimension value of the word vector was set to 150. The frequency that to filter words was set to 5 times, and the semantic relevance d_{ij} between high-frequency topic words was obtained. (2) The TF-IDF algorithm was used to extract keywords and obtain the co-occurrence degree O_{ij} between high-frequency topic words. The topic words were extracted from the complete set network and obtained by the semantic modeling of the organization's reserve knowledge. The intersection of the keywords was obtained from the co-occurrence analysis of the reserve knowledge. As the topic set of the organization reserve knowledge network, $\beta = 0.50$ an organizational reserve knowledge network was built with 517 points, 5228 edges, an average degree of 20.22, an average aggregation coefficient of 0.43, an average path length of 2.87,

and a degree distribution that obeyed the organizational reserve knowledge network of the power-law distribution feature. The topology of the network is shown in Figure 1.

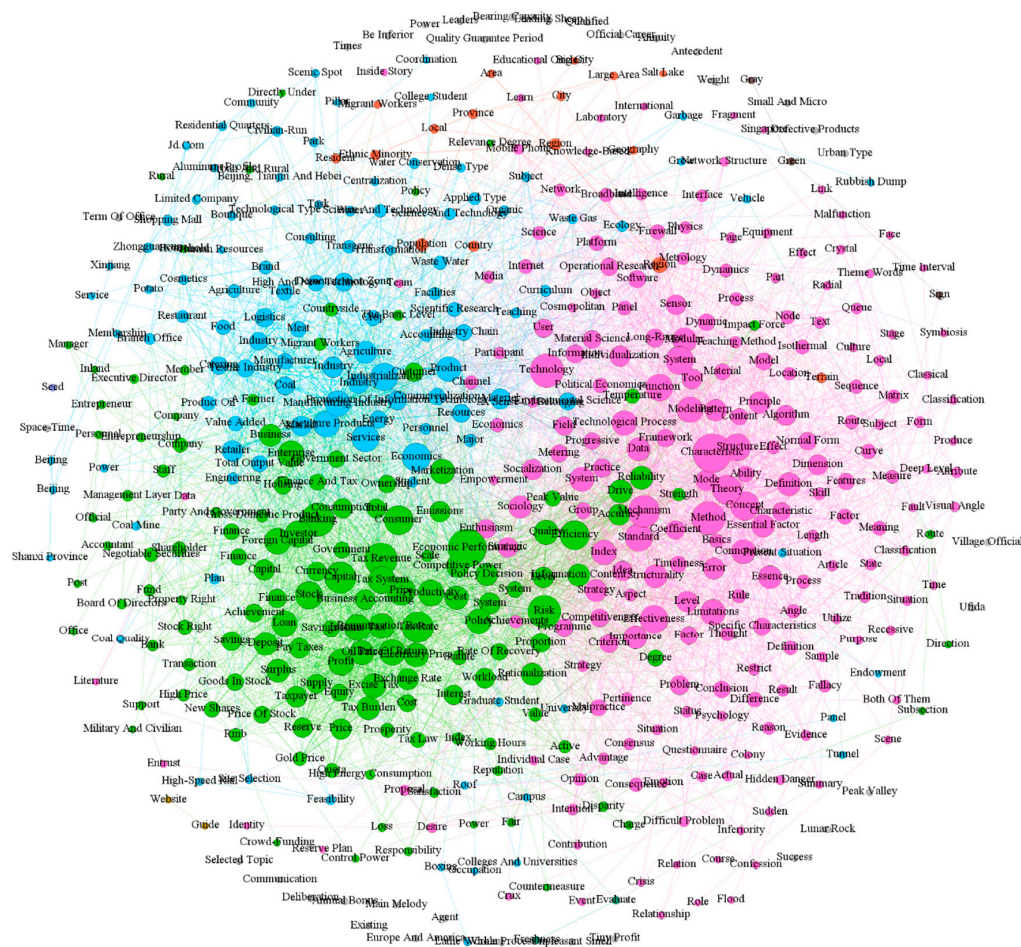


Figure 1. Reserved knowledge network of the economic and management college of a university.

It can be seen from Figure 1 that the reserve knowledge of the economic and management college has formed three clusters of topics with close associations located at the center of the network. The first is the financial management topic cluster, with “economic benefits” as the core and “banking,” “Capital,” “tax,” “investor,” “inventory,” “reward rate,” “return rate,” and so on as important topic terms. The second is the system evaluation topic cluster, with “features” as the core and “system,” “modeling,” “structure,” “function,” “efficiency,” and so on as important topic terms. It is worth noting that there is also a circular economy topic cluster with the main keywords of “population,” “region,” “cluster,” “area,” and “ecology”, shown in Figure 1. Compared with the above two topic clusters, the topic of the circular economy topic group is still smaller, the association between the topic words is still weak, and the network location is also far from the center, which is consistent with the status quo and trend of the discipline development of the college.

The topic of demanded knowledge is extracted from the demanded knowledge text carrier by applying the TF-IDF model. The demanded knowledge text carrier is the research content and research method of the project application of the Beijing Philosophy and Social Science Planning Office. Through this empirical analysis, we can understand whether this project can be completed independently in this college. In order to obtain the background corpus required for the calculation of the inverse document frequency, a wide range of materials were used to extract the corpus material of research content and the research method section, including 26 applications for the National 863 Program project since 2002, 14 applications for the 973 Program, 10 applications for the National

Science and Technology Research Project, 121 applications for the National Natural Science Foundation of China and general programs, 17 applications for youth and general programs of Beijing Natural Science Foundation, 18 applications for national social science fund projects, and 27 applications for social science fund projects in Beijing.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

At $\alpha = 0.1$ and after applying Formulas (6)–(8) to the above corpus materials to extract high-frequency topic words, 18 topic words were obtained. The weights $tfidf_i'$ of each topic word in the demanded knowledge were calculated according to Formula (9). At $\gamma = 1$, the eigenvector centrality E_i of each subject word could be calculated in the reserve knowledge network according to Formula (11). At $a = 0.10$, $b = 0.20$, $c = 0.30$, and $d = 0.40$, Formulas (13)–(15) could be used to calculate the degree of membership of the knowledge satisfaction content of each subject word, including poor, general, and good. The corresponding results are shown in Table 1. the degree of satisfaction of the organization on the reserve knowledge topic set was calculated according to Formula (17), and the fuzzy evaluation vector was $B = (0.51, 0.34, 0.15)$. It can be seen from the fuzzy evaluation vector that the membership degree belonging to the “poor” level is the largest, which is 0.51. According to the principle of maximum membership degree, the ability of the organization to meet the current knowledge needs is “poor.” It is recommended to seek external support from the organization, namely to complete the project through cooperation with other research units.

Table 1. The weight of the required knowledge subject and fuzzy membership degree.

No	Theme	$tfidf_i$	$tfidf_i'$	E_i	Fuzzy Membership		
					u_1	u_2	u_3
1	Park	0.48	0.14	0.04	1.00	0.00	0.00
2	Trend	0.33	0.09	0.30	0.00	1.00	0.00
3	Recourse	0.30	0.09	0.29	0.00	1.00	0.00
4	Circulation pattern	0.23	0.07	-	1.00	0.00	0.00
5	Network	0.22	0.06	0.09	1.00	0.00	0.00
6	Network structure	0.22	0.06	0.05	1.00	0.00	0.00
7	Environment	0.20	0.06	0.28	0.00	1.00	0.00
8	Eigenvector	0.19	0.06	-	1.00	0.00	0.00
9	Dynamics	0.19	0.05	0.16	0.40	0.60	0.00
10	System	0.17	0.05	0.51	0.00	0.00	1.00
11	Economics	0.15	0.04	0.55	0.00	0.00	1.00
12	Complexity	0.15	0.04	-	1.00	0.00	0.00
13	Network topology	0.11	0.03	-	1.00	0.00	0.00
14	Measurement	0.11	0.03	0.13	0.70	0.30	0.00
15	Modeling	0.11	0.03	0.46	0.00	0.00	1.00
16	Index	0.10	0.03	0.54	0.00	0.00	1.00
17	Substance	0.10	0.03	0.22	0.00	1.00	0.00
18	Information	0.10	0.03	0.28	0.00	1.00	0.00

5. Results and Discussion

In this section, the method of identifying and filling the knowledge gap of natural language processing proposed in this study is compared with SWOT analysis [8] and Venn diagram [10] to verify the effectiveness of the method.

(1) Analysis based on SWOT analysis. With a focus on industry trends, research directions of the organization, existing research capabilities, and current research needs, 12 professors and young teachers were invited to brainstorm and discuss in a conference room of the college on November 17, 2020. The analysis shows that the opportunities faced by the organization are “the contradiction of resources and environment facing economic development is prominent.” The threat is that “similar

colleges in the local region have relatively distinctive industry characteristics.” The strength is that the disciplines are relatively complete. The weakness is that the discipline of economics and financial management is relatively deep, and the advantage of management is not prominent. The strategy of organizational development should be “condensing the industry characteristics of circular economy and management,” and the knowledge gap is “management decision-making in the field of the circular economy, including system analysis, evaluation, and decision-making.” The knowledge gap obtained by the SWOT analysis method is shown in Figure 2.

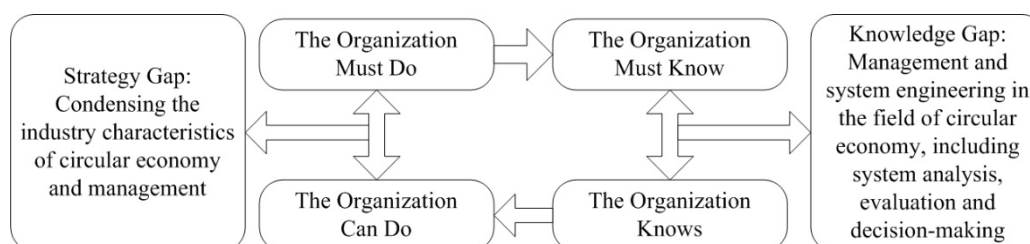


Figure 2. The knowledge gap obtained by SWOT analysis.

(2) Analysis based on the Venn diagram. On November 19, 2020, a total of 15 representatives (project team members, professors, and young teachers) were invited to a conference room of the college to conduct expert interviews based on the project research content to be completed in Section 4. According to the content of the interview meeting and combined with the Venn diagram method, the knowledge demand and knowledge reserve set needed to complete the project were sorted to obtain the knowledge gap. It can be seen that the set of the organization’s reserve knowledge includes “Finance,” “Accounting,” “Investment,” “Policy,” “Tax Revenue,” “Environment,” “Industry,” “Park,” and so on, whereas the set of demanded knowledge includes “Network,” “Environment,” “Industry,” “Park,” and so on. Among them, “Environment,” “Industry,” “Park,” and so on are the intersection of the reserve knowledge set and the demanded knowledge set, which are the knowledge needs that can be satisfied, and “Network” is the knowledge gap. In contrast with the methods proposed in this study, knowledge gaps such as “Network structures,” “Complexity,” “Network Topology,” and “Measurement” under the concept of “Network” were not identified. This is because, in expert interviews, the overall structure of the knowledge is a blind spot once the knowledge other than expert experience appears. As a result, the intensity of the gap is difficult to be quantified. Therefore, the analysis based on Venn diagram considers “Network” as “a knowledge gap with certain knowledge accumulation.” The knowledge gap obtained by the Venn diagram analysis method is shown in Figure 3. The comparison of SWOT, Venn diagram and the Method in this research on knowledge gap identification and fill is listed in Table 2.

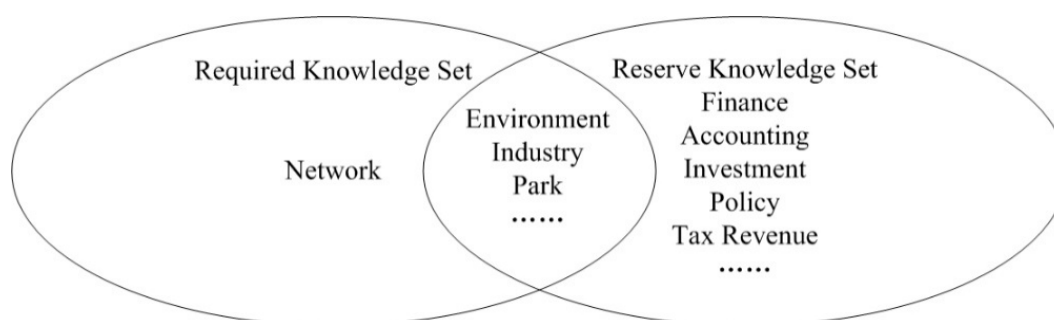


Figure 3. The knowledge gap obtained by Venn diagram analysis.

Table 2. The comparison of SWOT, Venn diagram and the Method in this research on knowledge gap identification and fill.

Methods	SWOT [8]	Venn Diagram [10]	Method in This Research
Set up the knowledge requirements set	Organization members adopt brainstorming and other methods to discuss and clarify the strategic intention of the organization and determine the knowledge needed to carry out its expected strategy.	Set up a knowledge demand set and draw a knowledge structure chart by means of brainstorming, interview, and investigation.	TF-IDF algorithm is used to extract the subject words in the text of the required knowledge carrier and construct the requirement knowledge network.
Create a knowledge store set	Perform a knowledge-based SWOT analysis to create a map of existing knowledge resources.	Establish a knowledge storage set, describe organizational status, and draw a knowledge distribution map.	Semantic vectorization is carried out based on the Word2Vec model, and the knowledge co-occurrence relationship and semantic association are considered to establish the subject network of reserve knowledge.
Identification of knowledge gaps	Identify knowledge gaps by matching organizational knowledge resources and capabilities to strategic opportunities and threats.	Manually compare knowledge structure diagrams and knowledge distribution diagrams to identify the knowledge gap.	Feature vector centrality is used to describe the importance of the required knowledge topic in the reserve knowledge topic network and identify organizational knowledge gaps.
Knowledge gap compensation method	Transform an organization's knowledge strategy into an organizational and technical architecture to support knowledge creation, management, and utilization processes to bridge these gaps	Proposed three kinds of knowledge gaps, knowledge gaps with knowledge accumulation, and knowledge gaps without knowledge accumulation.	Establish a fuzzy evaluation set to evaluate organizational knowledge satisfaction ability. If the ability evaluation is better, the knowledge required by the organization can be fully satisfied within the organization. Instead, seek support outside the organization or gradually meet the requirements of knowledge topics through special training and other means.

6. Conclusions and Discussion

Under the “Internet+” environment, on the one hand, the rate of organizational knowledge innovation has accelerated significantly. On the other hand, the amount of Internet distribution of knowledge text carriers has increased dramatically, presenting an information explosion. Therefore, the use of knowledge of the network text carrier, efficient and accurate identification of innovative organizational knowledge gaps, and providing corresponding gap compensation methods are the key to winning the knowledge innovation competition under the “Internet+” environment.

In view of the above problems, this study builds a network of complete knowledge topics under the “Internet+” environment based on the Word2Vec model. In the context of the complete network of knowledge topics, the word vector and frequency of an organization's reserve knowledge texts are analyzed, and the organization's reserve knowledge topic network was established based on characteristics the vector centrality analysis organizes the satisfaction degree of the reserve knowledge topics. The demanded knowledge topics are identified based on the TF-IDF model. The fuzzy evaluation method is used to identify the satisfaction ability of the demanded set for knowledge topics in the reserve knowledge topic network and propose corresponding compensation methods.

Through these methods, taking the college of economics and management of a university in Beijing as the research object, the reserve knowledge topic network has 517 points, 5228 edges, an average degree of 20.22, an average clustering coefficient of 0.43, an average path length of 2.87, and the degree distribution. They all obey the power ratio characteristic of a scale-free network. The network has already formed two thematic clusters of financial management and economic system evaluation, and the circular economy topic cluster is being formed. The demanded knowledge was obtained from

the research content of the project application of a Beijing Philosophy and Social Science Planning Office, and 18 topics were extracted. The fuzzy evaluation vector is obtained. The value is $B = (0.51, 0.34, 0.15)$, which indicates that the college's current ability to meet the above knowledge needs is "poor," and it is recommended to seek external support.

In the research of existing knowledge gap identification methods, SWOT analysis is a tool for strategic analysis, and the identified knowledge gap is difficult to be quantified. It is not suitable for rapid response identification. The Venn diagram is a quantitative method to study knowledge gaps, but it ignores the hierarchy and correlation between knowledge, and the strength of knowledge gaps obtained is affected by expert experience. This study adopts the quantitative analysis method based on the Word2Vec model to obtain the co-occurrence relationship and semantic relationship among knowledge, establishes the fuzzy evaluation set according to the method of the fuzzy comprehensive evaluation, and obtains different compensation methods according to different satisfaction degrees. This paper presents a fast, objective, and quantitative method for knowledge gap identification and filling for the demand for organizing rapid collaborative innovation under the Internet+ environment. Digital document resources of distributed network storage are mined, and the topic of knowledge reserve and its correlation are expressed in the form of a network graph, reflecting the background knowledge structure inside the organization. The knowledge gap of the organization was determined by matching the knowledge needs and background knowledge of the organization. The corresponding remedy strategies are given according to the degree of knowledge gap satisfaction.

The following further improvements about this study can be made: (1) The fuzzy evaluation results are affected by the values of the parameters a , b , c , and d of the linear membership function. In the future, the parameters can be fed back according to the ability of the organization to complete the knowledge innovation activities. (2) There is no corresponding topic in the reserve knowledge network due to the lack of professionalism of the Chinese Wikipedia corpus and the word segment library used in the establishment of the knowledge collection network. Corpus and word segmentation library for knowledge topic mining in innovation activities should be built in the future. (3) This research corresponds to the innovative organization level. Under the "Internet+" environment, this research method can be applied to the level of innovative talents to achieve more precise knowledge topic analysis and gap identification. (4) The method proposed in this study can be applied to the topic analysis and matching of technology, data, service, and content resources in addition to knowledge gap identification.

Author Contributions: Conceptualization, L.Q.; methodology, L.Q.; software, X.A. & S.Z.; validation, X.W.; writing—original draft preparation, L.Q.; writing—review and editing, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2017YFB1400400), Youth Talent Promotion Program of Beijing Association for Science and Technology (2020-2022-16), Social Science Research Program of Beijing Education Committee (SM202011232005), Program for Promoting the Connotative Development of Beijing Information Science & Technology University (521201090A, 5026010961).

Acknowledgments: The authors are grateful to the anonymous reviewers and the editor for their valuable comments and suggestions that have greatly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, P.; Guo, Y.H.; Li, Y. The Impact of Internet Development on the Transformation and Upgrading of Regional Industrial Struct—Research Based on Mediating Effect. *J. Ind. Technol. Econ.* **2020**, *39*, 135–144.
2. Zhang, Y.S.; Wang, D.M. Research on the Random Supervision of the Risk of Internet Financial Enterprises Excessive Innovation. *Econ. Rev.* **2017**, *7*, 100–105.
3. Dang, X.H.; Gong, Z.G. Impact of multidimensional proximities on cross region technology innovation cooperation: Experical analysis based on Chinese coinvent patent data. *Stud. Sci. Sci.* **2013**, *31*, 1590–1600.

4. Glisson, C. The role of organizational culture and climate in innovation and effectiveness. *Hum. Serv. Organi. Manag. Leadersh. Gov.* **2015**, *39*, 245–250. [[CrossRef](#)]
5. Grillitsch, M.; Asheim, B. Place-based innovation policy for industrial diversification in regions. *Eur. Plan. Stud.* **2018**, *26*, 1638–1662. [[CrossRef](#)]
6. Han, Y.; Gao, C.Y. Compensation of knowledge gaps among high-tech enterprises. *Stud. Sci. Sci.* **2009**, *27*, 1370–1375.
7. Mohd, S.S.A.; Prakoonwit, S.; Sahandi, R.; Khan, W.; Ramachandran, M. Big data analytics—A review of data-mining models for small and medium enterprises in the transportation sector. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1238. [[CrossRef](#)]
8. Zack, M.H. Developing a Knowledge Strategy. *Calif. Manag. Rev.* **1999**, *41*, 125–145. [[CrossRef](#)]
9. Qiu, J.N.; Wu, M.J.; Nian, C.L. Research on identifying organizational knowledge gap. *Sci. Res. Manag.* **2013**, *34*, 85–93.
10. Chen, J.H.; Sun, Q.X.; Zhu, Y.L. Study on the identification method and filling strategies of knowledge gap. *Stud. Sci. Sci.* **2007**, *25*, 750–755.
11. Liu, X.; Jia, W.; Wang, Y.; Guo, H.; Ren, Y.; Li, Z. Knowledge discovery and semantic learning in the framework of axiomatic fuzzy set theory. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1268. [[CrossRef](#)]
12. Rashid, J.; Shah, S.M.A.; Irtaza, A. Fuzzy topic modeling approach for text mining over short text. *Inf. Process. Manag.* **2019**, *56*, 102060. [[CrossRef](#)]
13. Zhu, Z.; Liang, J.; Li, D.; Yu, H.; Liu, G. Hot topic detection based on a refined TF-IDF algorithm. *IEEE Access* **2019**, *7*, 26996–27007. [[CrossRef](#)]
14. Tang, M.; Xia, Y.; Tang, B.; Zhou, Y.; Cao, B.; Hu, R. Mining Collaboration Patterns between APIs for Mashup Creation in Web of Things. *IEEE Access* **2019**, *7*, 14206–14215. [[CrossRef](#)]
15. Haider, S. Organizational knowledge gaps: Concept and implications. In Proceedings of the Druid Summer Conference, Copenhagen, Denmark, 12–14 June 2003.
16. Vos, J.P.; Keizer, J.A.; Halman, J.I.M. Diagnosing Constraints in Knowledge of SMEs. *Technol. Forecast. Soc. Chang.* **1998**, *58*, 227–239. [[CrossRef](#)]
17. Dang, X.H.; Ren, B.Q. Research on Knowledge Gaps and Compensation Strategies in Enterprises' Technological Innovation under Network Environment. *Sci. Res. Manag.* **2005**, *3*, 12–16.
18. Lafuente-Ruiz-de-Sabando, A.; Zorrilla, P.; Forcada, J. A review of higher education image and reputation literature: Knowledge gaps and a research agenda. *Eur. Res. Manag. Bus. Econ.* **2018**, *24*, 8–16. [[CrossRef](#)]
19. Li, G.H.; Chen, C.; Luo, J.Q. The Study on the Implements of Knowledge Gaps in Servitization of Manufacturing Based on Evolutionary Game Theory. *Ind. Eng. Manag.* **2014**, *19*, 40–46.
20. Malhotra, A.; Majchrzak, A.; Niemiec, R.M. Using Public Crowds for Open Strategy Formulation: Mitigating the Risks of Knowledge Gaps. *Long Rang Plan.* **2017**, *50*, 397–410. [[CrossRef](#)]
21. Qiu, J.N.; Nian, C.L. A construction method of organizational knowledge structure and its applications in the patent documents. *Sci. Res. Manag.* **2012**, *33*, 48–56.
22. Li, G.; Liu, X.H. Research on the selection of research team members based on knowledge gap. *Sci. Technol. Prog. Policy* **2015**, *32*, 139–143.
23. Zhang, L.Y.; Li, Y.N.; Gu, L.Z. Knowledge Risk Identification of Construction Projects from the Perspective of Knowledge Gap. *J. Engin. Manag.* **2015**, *29*, 89–94.
24. Li, C.; Yang, H.T. Research on the Underlying Mechanism of Different Knowledge Gap Affecting Organization Innovation and Countermeasures. *Sci. Technol. Prog. Policy* **2012**, *29*, 115–118.
25. Nalchigar, S.; Yu, E. Business-driven data analytics: A conceptual modeling framework. *Data Knowl. Eng.* **2018**, *117*, 359–372. [[CrossRef](#)]
26. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
27. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
28. Leydesdorff, L. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 77–85. [[CrossRef](#)]
29. Leung, X.Y.; Sun, J.; Bai, B. Bibliometrics of social media research: A co-citation and co-word analysis. *Int. J. Hosp. Manag.* **2017**, *66*, 35–45. [[CrossRef](#)]
30. Ba, Z.H. Research on the Domain Theme Evolution Analysis Based on Keywords Semantic Network. *Inf. Stud. Theory Appl.* **2016**, *3*, 14.

31. Zhang, J. A method of intelligence key words extraction based on improved TF-IDF. *J. Intell.* **2014**, *33*, 153–155.
32. Wang, J.Z.; Qiu, T.X. Focused topic Web crawler based on improved TF-IDF algorithm. *J. Comp. Appl.* **2015**, *35*, 2901–2904.
33. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).