INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 14(1), 107-123, February 2019.

A Latent-Dirichlet-Allocation Based Extension for Domain Ontology of Enterprise's Technological Innovation

Q. Zhang, S. Liu, D. Gong, Q. Tu

Qianqian Zhang, Shifeng Liu,
Daqing Gong*, Qun Tu
School of Economics and Management
Beijing Jiaotong University, China
100044 No.3 Shangyuancun, Haidian, Beijing, China
15113121@bjtu.edu.cn, shfliu@bjtu.edu.cn
*Corresponding author:gongdq@bjtu.edu.cn
17113133@bjtu.edu.cn

Abstract: This paper proposed a method for building enterprise's technological innovation domain ontology automatically from plain text corpus based on Latent Dirichlet Allocation (LDA). The proposed method consisted of four modules: 1) introducing the seed ontology for domain of enterprise's technological innovation, 2) using Natural Language Processing (NLP) technique to preprocess the collected textual data, 3) mining domain specific terms from document collections based on LDA, 4) obtaining the relationship between the terms through the defined relevant rules. The experiments have been carried out to demonstrate the effectiveness of this method and the results indicated that many terms in domain of enterprise's technological innovation and the semantic relations between terms are discovered. The proposed method is a process of continuously cycles and iterations, that is the obtained objective ontology can be re-iterated as initial seed ontology. The constant knowledge acquisition in the domain of enterprise's technological innovation to update and perfect the initial seed ontology.
Keywords: Latent Dirichlet Allocation (LDA), ontology extension, enterprise's technological innovation, semantic web, text mining.

1 Introduction

With the pace of globalization of economy accelerated significantly, the market has stepped into the information age from the era of industrialization. As the market demand changes at a faster pace, the competition of the market has become extremely fierce. In this context, the technological innovation is increasingly becoming the inner motivation and main source of enterprise development. There is important significance to evaluate the capability of enterprise technological innovation scientifically and efficiently to set the technological innovation policy for government, revise technological innovation strategy reasonably for enterprises and improve the technological innovation ability. The evaluation of enterprise's technological innovation ability has drawn extensive attention of much scholars. Although much progress has been made on the theoretical research of enterprise's technological innovation [13, 33]. There still exist many problems such as evaluation mechanism and evaluation methodology, namely, the biggish subjectivity of evaluating indexes, strong dependence on declared data of evaluated enterprise, low evaluation accuracy, poor coincidence of evaluate results, etc. Enterprises produce large amounts of textual information in technological innovation process, including technological innovation activity report, meeting minutes, annual report and patent file. Hence, enterprises need to not only make use of these documents but to mine and discover valuable and hidden knowledge from large collections of data. It is also a pressing problem to transform massive textual data into knowledge that can serve and utilize for technological innovation of enterprise and provide decision-making

for technological innovation of enterprise. Therefore, it is the field which has not been involved by using techniques of text mining and machine learning to analyze massive textual information that generated by enterprise's technological innovation and further determined the enterprise's technological innovation ability from objective data. In this paper, we deal with three major problems as follows:

- Is it possible to discover the concepts from large amount of textual corpus of domain of enterprise's technological innovation?
- Is it possible to build rules for semantic relationship recognition to make the enterprise's technological innovation ontology subsumption hierarchy?
- Is it possible to make the enterprise's technological innovation domain ontology extension automatically?

To improve this situation, this paper presents an approach to extract core concepts from large textual data and proposes a new method of building rules for semantic relationship recognition based on LDA algorithm. The rest of paper is organized as follows, section 2 provides some background knowledge concerning concept and relative literature reviews. In section 3 explains the proposed methods, while section 4 presents the experimental results. Section 5 concludes the paper.

2 Background knowledge and related works

2.1 Technological innovation capability

Technological Innovation Capability (TIC) has become the key to improve productivity and maintain competitiveness in the constantly fluctuating environments for enterprises. However, the definition of TIC is hard to agree upon since the technological innovation involves numerous organizational functions and resources integration among various department [26]. The concept of innovation originally from the innovation theory proposed by Schumpeter. On the base of it, Burgelaman *et al.* [5] put forward that all TIC can be defined as a series of characteristics in an organization facilitating and supporting an innovation strategy. Based on differing perspectives, there are many scholars proposed various components of TICs of a firm [22, 30].Therefore, the measurement of TIC is difficult and complicated since the perceive objectives and criteria for TIC is different. Tsai *et al.* [24] established an evaluation model for the TIC of high-tech industries based on the AHP method. Wang and Chang [25] proposed a model for diagnose the value of TIC in enterprise and established an evaluation system by AHP method. Wang *et al.* [26] evaluated and analyzed TIC combined with fuzzy evaluation and non-additional fuzzy evaluation. Deng *et al.* [12] established a TIC evaluation system by factor analysis and the fuzzy synthetic assessment method is used to evaluate TIC. Guan *et al.* [13] developed an innovation measurement framework based on the traditional DEA method. By looking at literatures of the measurements of TIC [8, 32], few studies can avoid to involve the subjective judgement, previous experience and uncertain assessment by experts.

2.2 Ontologies construction and extension

In the last decade, many scholars have done a lot of researches on ontology definition, construction, extension and application aspects. Ontologies were defined as "an explicit specification of shared conceptualization" [14] provide the key to machine-processable data on Semantic Web, being fundamental components for sharing, reusing as well as reasoning over knowledge

domains [1]. Although there is a great progress in knowledge acquisition and ontology construction, the current ontology construction methods still rely heavily on manual parsing and existing knowledge bases. The process of ontology learning and extending is a costly, time-consuming and error-prone task when done manually. With the constant emergence of new domain knowledge, the domain ontology automatic updates are facing new challenge.

Many researchers have engaged into ontology construction and enriching automatically in recent years. In previous work, the machine learning and statistical analysis method has great advantages in accuracy and recall rate and has been proposed to solve this problem [9]. For instance, Jeroen *et al.* [10] proposed the subsumption method and a hierarchical clustering algorithm to arrange the domain terms hierarchically and compared the two methods performances. Researches [18] and [23] using the fuzzy mechanism to extract domain concept and generate the domain ontology through the fuzzy conceptual clustering. Khan and Luo [17] presented a modified self-organizing tree algorithm (SOTA) which is performs better than the hierarchical agglomerative clustering (HAZ) on ontology construction automatically. Gilles *et al.* [1] put forward a Mo'k workbench which is a framework using the agglomerative clustering techniques to generate concept hierarchies from parsed corpora. Cimiano and Völker [6] presented a Text2Onto which implementing variety algorithms and techniques for ontology learning. However, most of the existing methods require a certain scale of supervised training corpus as the learning object, and the result seldom consider semantic-aware which is difficult to recognize the relationship between terms in domain ontology.

Although the ontology construction and extending automatically has been achieved some progress, there are still some problems in this field. For example, the non-taxonomic relationship among terms were often omit in the ontology hierarchical relations construction. Besides, the parameter setting in the model and complex computing in the process cause the heavy computing burden and make the model overfitting which limits their application.

2.3 LDA topic model

The latent topic discovery researches have gained much attention to hierarchical relation learning in recent years. Latent topic discovery is invented to overcome the bottleneck of bag-of words processing model in information retrieval area, trying to advance the text processing technology from pattern to semantic calculation [23]. For the research in latent topic discovery, an earlier work in literatures is Latent semantic indexing (LSI), which is a retrieval technique to learn latent topic by performing a matrix decomposition (SVD) on the term-document matrix [31]. Through this technique, latent topics are revealed which are actually distributions over the words of the term space of the corpus [10]. For example, the work in [3] uses the technique of LSI to identify relationships among entities in large collections of text. The author in [4] also using the LSI for discovering new information relevant to a given topic in large textual databases. Although the LSI based on SVD having some early success on latent topic discovery and relationship identification, it lacks rigorous mathematical and statistical basis and the SVD decomposition is time-consuming. Probabilistic Latent Semantic Indexing (PLSI) was proposed to extend the LSI assuming which associates a latent context variable with each word occurrence and can deal with synonymy and polysemous words. The author in [16] proposed that PLSI has been considered as an unsupervised learning method used in the task of text learning. The work in [15] also using the PLSI to represent sentences and queries as probability distributions over latent topic to solve the multi-document summarization problem. Other than LSI and PLSI, the algorithm of Latent Dirichlet Allocation (LDA) is more advantageous since LDA model can avoid overfitting and large sets of parameters.

LDA model, proposed by David Blei *et al.* [2], is a statistical topic model and can analyzes

hidden topics in large-scale data. Ontology learning using LDA model is a relative new research approach. Elias *et al.* [34, 35] used the LDA model for discovery of topics that represent ontology concepts and comparing the high-probability terms in topics to arrange concepts in a subsumption hierarchy. However, it cannot infer subsumption relations in the case where a topic subsumes only one other topic. Yeh and Yang [29] developed an automatic domain ontology construction for historical documents. LDA model was used to extract latent topic from raw textual Chinese Recorder data and the basic cosine similarity with hierarchical agglomerative clustering is used to clustering the topic, but the relationship between the topic cannot be defined since the clustered latent topic is a hierarchical tree structure. Francesco *et al.* [7] present an automatic terminological ontological learning system which the common hypernyms between the aggregate root node and aggregate words are determined through the LDA model and then added the semantically similar root node to the ontology. however, the measurement in large set of data may cause heavy computing burden. Ni *et al.* [20] also used the LDA model to select the domain terms and through the word association analysis to discover the hierarchical relations among domain terms. Raghuveer [21] using the LDA model to obtain the topics from legal documents and clustering legal judgments by cosine similarity.

3 The proposed method

The paper has combined the ontology technique and LDA topic model, used the initial seed ontology guiding the LDA model to obtain the concept in the field of enterprise technological innovation. Adding the new concept to the initial domain ontology by defined rules to realize the iteratively updating and perfection of ontology. The framework of enterprise's technological innovation domain concept acquisition contains the following four modules:

- The module of seed ontology introducing. The paper needs to construct a seed ontology to guide the concept acquisition for enterprise's technological innovation domain. The basic concept and relationship of seed ontology in domain enterprise technological innovation mainly extracted from Chinese Classified Thesaurus. The protege 4.3 was used to visualize the construction of seed ontology. More details will be introduced in the next chapter.

- The module of text preprocessing. This is the process of converting a text into individual words or sequences of words which using the Natural Language Processing (NLP) technique including of word segmentation, Part-of-Speech (POS)tagging, stop-word filtering preprocessed the collected Chinese textual documents. Two words merging needs to satisfy adjacency and frequent co-occurrence both, the calculation method as follows. In order to guarantee the semantic accuracy after word segmentation, the method of entropy was adopted to merge the words [27, 28].

$$E(w_{m-1}, w_m) = \frac{p(w_{m-1}w_m)}{minp(w_{m-1}), p(w_m)} \tag{1}$$

  where $p(w_m)$ denotes the frequency of word w_m in documents and $p(w_{m-1}w_m)$ denotes the continuous frequency of word w_{m-1} and w_m in documents.

- The module of mining domain specific terms. LDA (Latent Dirichlet Allocation) is a three-level hierarchical Bayesian model which proposed by Blei [34]. It assumes that each document in corpus is represented as random mixtures over latent topic, where topic is characterized by a distribution over all the words. LDA is constructed for documents with "bag-of-words" which uses the statistical information of words to represent text in vector

space and explores the probabilistic relationships between words and text. In this paper, we use the LDA model was described below.

LDA taking the corpus D which after the preprocessing by module B as input and output the topic distributions and the distribution of words for each topic by training. The LDA generates the words in a two-stage process: words are generated from topics and topics are generated by documents. The graphical model of LDA is shown in Fig. 1. The terms of LDA was defined as follows:

A document is a sequence of N words denoted by $w = (w_1, w_2, \cdots, w_n)$ where w_n is the nth word in the sequence, and a corpus is a collection of M documents denoted by $D = d_1, d_2, \cdots, d_M$;

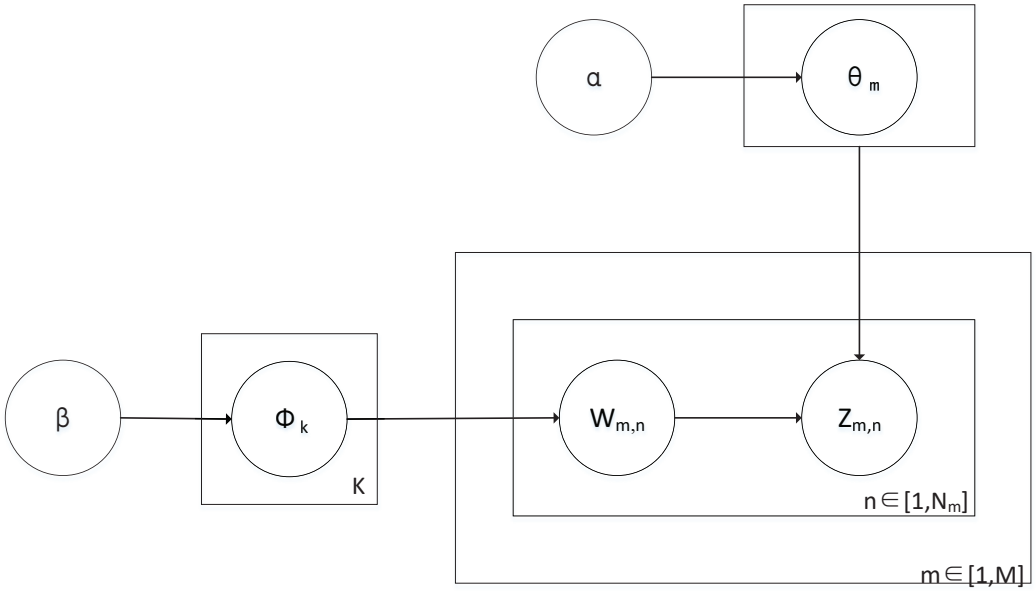

Figure 1: Graphical model representation of LDA

α and β are Dirichlet prior hyperparameters; All the words in document M will be clustered into Z topics, for each topic $Z \in 1, 2, \cdots, k$, sample a word distribution $\phi_k \sim Direchlet(\beta)$;

- Choose $N \sim Possion(\xi)$
- Choose a topic distribution $\theta_m \sim Direchlet(\alpha)$
- For each of word $w_{m,n}$ in mth document:
  - Choose a topic of the word $Z_{m,n} \sim Multinomial(\theta_m)$
  - Choose a word $w_{m,n} \sim Multinomial(\phi_{Z_{m,n}})$

Since the process to generate the topic for M documents are independent of one another, we can have M conjugated structures and the generative process of probabilistic of topics in corpus is as follows:

$$\begin{aligned} p(\vec{z}|\vec{\alpha}) &= \prod_{m=1}^{M} p(\vec{z_m}|\vec{\alpha}) \\ &= \prod_{m=1}^{M} \frac{\Delta(\vec{n_m} + \vec{\alpha})}{\Delta(\vec{\alpha})} \end{aligned} \tag{2}$$

The process to generate words for K topics are independent of one another, we can have K conjugated structures and the probabilistic of words in corpus is as follows:

$$\begin{aligned} p(\vec{w}|\vec{z},\vec{\beta}) &= \prod_{k=1}^{k} p(\vec{w_k}|\vec{z_k},\vec{\beta}) \\ &= \prod_{k=1}^{k} \frac{\Delta(n_k \vec{+} \vec{\beta})}{\Delta(\vec{\beta})} \end{aligned} \tag{3}$$

Thus, within a document, the probability distribution over words specified by the LDA model is given as follows:

$$\begin{aligned} p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) &= p(\vec{w},\vec{z}|\vec{\beta}) * p(\vec{z}|\vec{\alpha}) \\ &= \prod_{k=1}^{k} \frac{\Delta(\vec{n_k}+\vec{\beta})}{\Delta(\vec{\beta})} * \prod_{m=1}^{M} \frac{\Delta(\vec{n_m}+\vec{\alpha})}{\Delta(\vec{\alpha})} \end{aligned} \tag{4}$$

Thus, in this paper, the LDA topic model was used to train the term candidate set which obtained by the module of text preprocessing and to obtain the word probabilistic of domain concepts (topics) as shown in Fig.2.

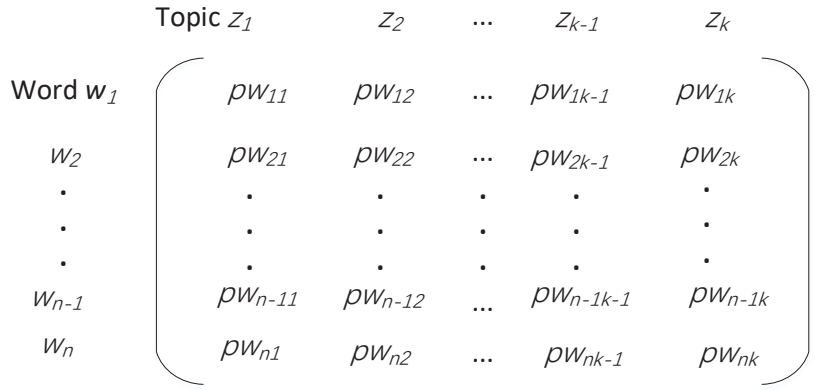

Figure 2: Words distribution probabilistic of topics

where pw_{nk} represents the probability of the word n in the topic k.

- The module of domain ontology updating. The module is the key point and difficulty of this paper. Take each concept in the initial enterprise technological innovation ontology as a document into the module (3) trained LDA model. We can get the topics probabilistic of documents as shown in Fig.3. Where, a corpus is a collection of M ontology concepts denoted by $C = (c_1, c_2, \cdots, c_{m-1}, c_m)$;

Where pz_{km} represents the probability of the topic k in concept(document) m.

According to the LDA algorithm, we can get the term probabilistic of documents, namely, the probabilistic of words in documents and concepts in initial ontology denoted as $p(w_n|c_m)$. Then by using the relevant rules to judge the relationship between topics generated by LDA model and concepts in initial domain ontology.

$$p(w_n|c_m) = \sum_{j=1}^{K} p(w_n|z=j) * p(z=j|c_m) \tag{5}$$

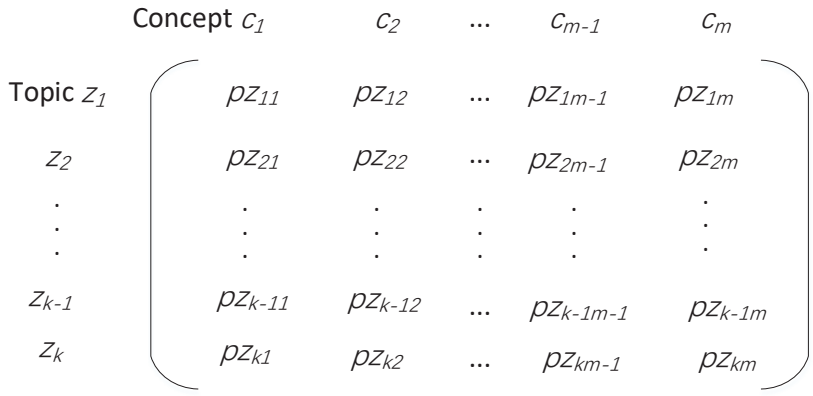

Figure 3: Topics distribution probabilistic of documents

When the $p(w_n|c_m)$ greater than the threshold value TH and the word n is not in the list of $C = (c_1, c_2, ..., c_{m-1}, c_m)$, therefore, the term w_n is an associated term of c_m.

$$p(w_n|c_m) > TH \tag{6}$$

Algorithm: Rules of the semantic relationship recognition were defined as follows:

$$\begin{aligned} W(W_n, C_m) &= \frac{p(W_n|C_m)}{p(z=j|C_m) + p(W_n|C_m)} \\ &= \frac{\sum_{j=1}^{K} p(w_n|z=j) * p(z=j|c_m)}{p(z=j|C_m) + \sum_{j=1}^{K} p(w_n|z=j) * p(z=j|c_m)} \end{aligned} \tag{7}$$

- Rule 1: Rules for synonymy relations recognition. If the $W(Wn, Cm) \geq 0.01$, the semantic relationship between word and concept is equivalent, namely, the related terms extracted by LDA is equal to the existed concept.

- Rule 2: Rules for hyponymy relations recognition. When the Rule 1 cannot be satisfied, if the $W(Wn, Cm) \geq 0.004$, the word includes the concept, namely, the related terms extracted by LDA is superclass of the existed concept, the relationship as "is-a" or "sub-class".

- Rule 3: Rules for correlation recognition. When the Rule 1 and Rule 2 are cannot be satisfied, the relationship between existed concept and related terms can be recognized as related or using people to identify the specific semantic relationship by external knowledge base.

Based on the above rules, the semantic relations between the existing concepts and their related terms are identified, add the obtained related terms and semantic relations to the original ontology O, the original ontology O was updated to Oi.

4 Experiments and result

4.1 Ontology acquisition

Enterprise ontology and TOVE (Toronto Virtual Enterprise Ontology) are the most popular ontology-based enterprise modeling methodologies. The two projects all point out the common key influencing factors in the process of enterprise ontology construction including of resources, organization, strategy, market and activity. In this paper, the five factors also considered as the first class of the enterprise's technological innovation ontology. The Chinese Classified Thesaurus has clear semantic structure which is more suitable for the extraction of concepts and relationship between concepts. Transforming thesaurus into ontology through further concepts analysis and semantic relationship adjustment of the words in F27 category of Enterprise Economy in Chinese Classified Thesaurus. There are 5 concepts extracted from the thesaurus including of Innovation resources, Marketing innovation, Strategic innovation, Organizational innovation and Innovation activities. The nested composite view provides a representation of the interrelation between the first classes in the entire ontology structure. It is convenient for considering whether the constructed domain ontology meets actual needs. The nested composite view of enterprise's technological innovation domain is shown as Fig.4. The relationship between domain ontology concepts includes the hyponymy relations and complex non-hierarchical relationship for specific application. The Fig.5 shows that the relationship between domain ontology concepts which takes the Strategic innovation as the center and reflects the complex relationship between concepts. The ontology of enterprise's technological innovation is a prototype, in which many concepts and relationships are still insufficient and need to continuously improved.

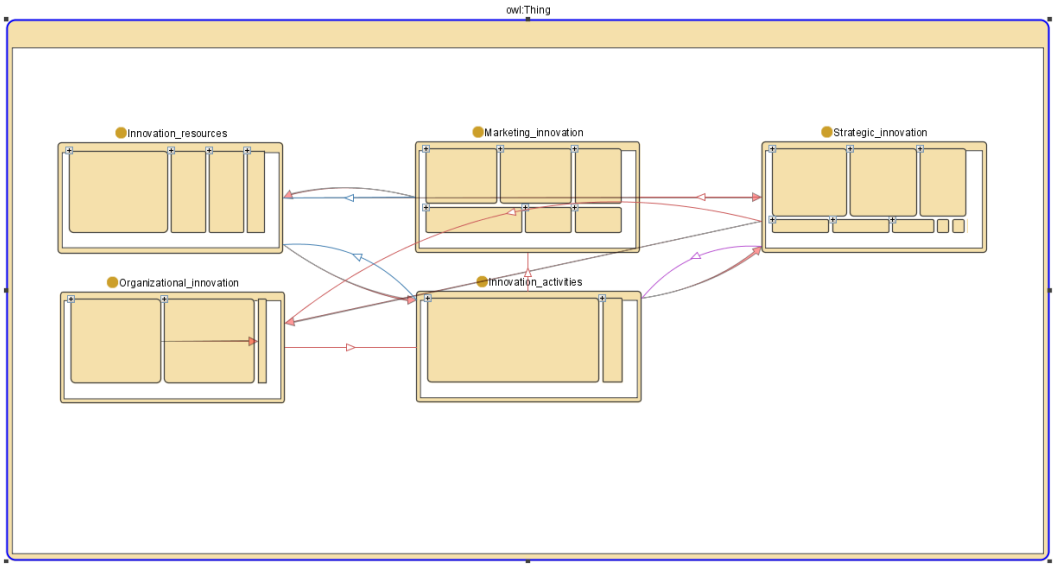

Figure 4: Nested composite view of enterprise technological innovation domain

4.2 Textual data collection

There are two aspects to collect the textual data of enterprise technological innovation, one is the internal information generated from daily production activities such as internal R&D, innovation activities, etc. The other type of collected data is generated when enterprise interacting with external customers and partners by social networks, mobile applications, etc. 863 sets of valid data are obtained which includes of 413 enterprise technology centers.

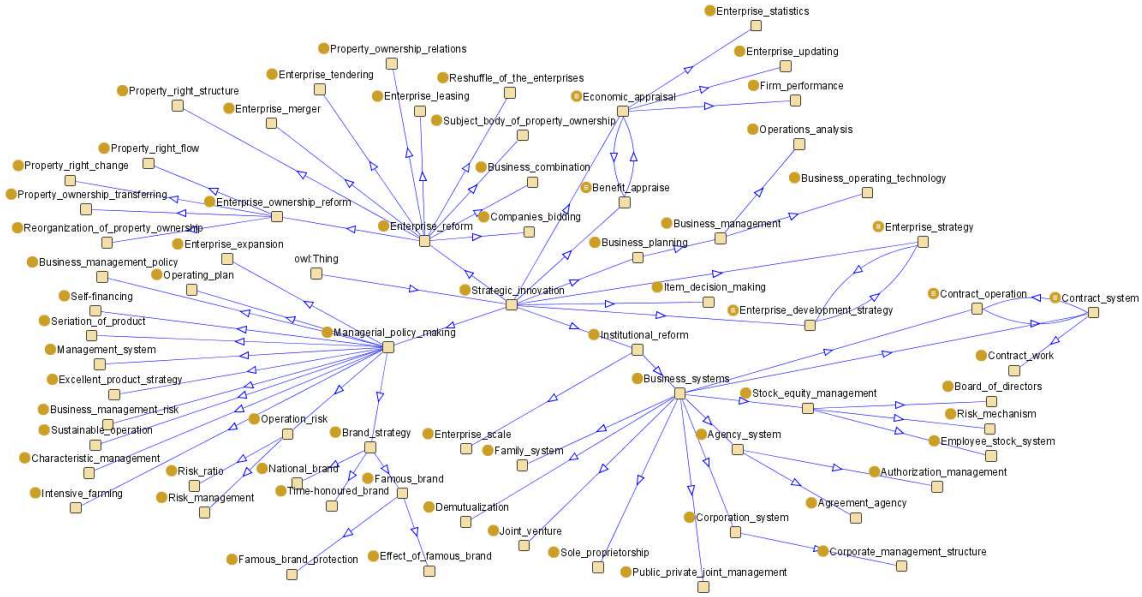

Figure 5: Visualization of relationship between concepts

4.3 Text preprocessing

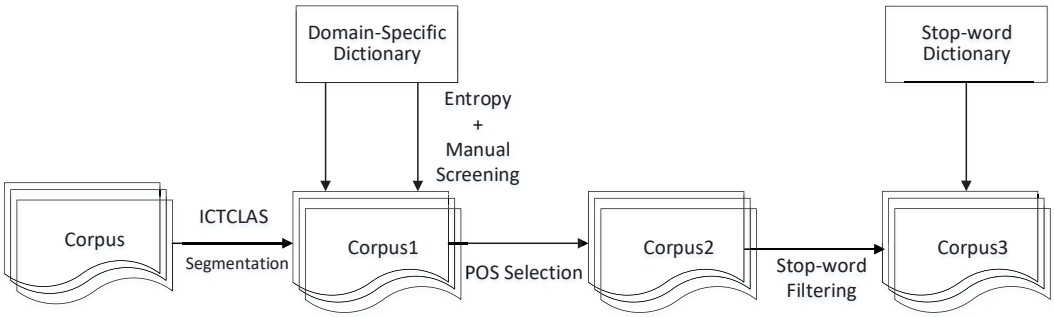

Figure 6: Process of text preprocessing

Chinese segmentation

Firstly, constructing the domain-specific dictionary for the field of enterprise's technological innovation by widely collected materials such as the cell thesaurus and imported the dictionary into the ICTCLAS segmentation system [32] which developed by the Chinese Academy of Sciences. Secondly, the result of segmentation will appear the problem due to a Chinese phrase was wrongly divided into many words. For example, the "enterprise's technological innovation" was divided into three small-grained words such as "enterprise", "technological" and "innovation". The method of entropy was adopted to merge the words which shown as equation (1). Combining two words that satisfy the conditions into a new phrase and adding to the domain-specific dictionary by manual screening. Then, segment the source document and iterate repeatedly.

POS selection

The documents of enterprise's technological innovation are the synthetic texts, in which, nouns are more representative important for semantic information in source documents. Hence, selecting the nouns and the word similar to nouns as the research object such as the verb with noun function, the adjective with noun function, etc.

Elimination of stop-words

Useless words selected from the domain of enterprise's technological innovation is used to build stop words dictionary. Filtering the stop words in documents which processed by the above two steps. It can reduce the size of the indexing structure considerably by elimination of stop words.

4.4 Mining domain terms from text corpus based on LDA

- Terms selection. According to the word frequency of terms in all corpora, the word frequency of [50, 1000] were selected as terms to represent each document in vector space model.

- Optimal number of topics. The perplexity index is adopted in optimal topic selection. Perplexity is an effective measurement to verify the model generalization ability. A lower perplexity indicates the better generalization performance. The perplexity is defined as follows:

$$perplexity(W_n|C_m) = e^{\frac{-\sum \log(p(W_n|C_m))}{N}} \tag{8}$$

  Where $p(W_n|C_m)$ is the probability of each word in candidate term set, N is the number of words.

  The perplexity of all documents generated under different topic numbers is shown as Fig.7 It looks like the 160-topic model has the lowest perplexity score. Hence, the optimal number of topic 160 (k=160) is selected for all corpus by perplexity analysis. The smoothing parameters α and β were fixed at 0.1 and 0.3. The threshold TH was set to 0.001.

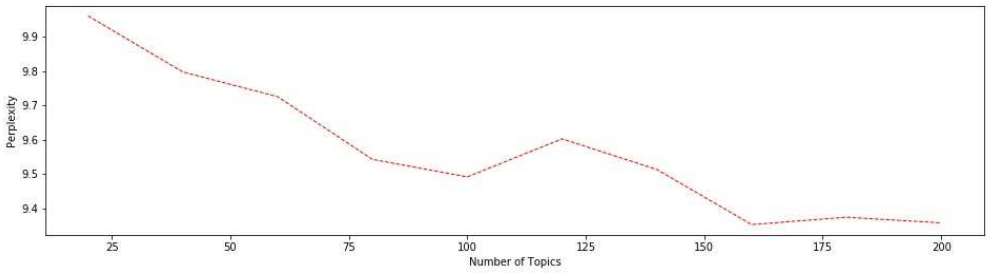

Figure 7: Perplexity result on enterprise technological innovation corpora for LDA model

- When the number of topics is 160, the LDA topic modelling is carried out to obtain the distribution of terms, that is each topic comprises of a series of related words. The order of the top terms of each topic is arranged by the probability and presented in the Table.1, in which only the first 10 topics with high probability of topic distribution were shown. The Fig.8 shows the resulting graph visualization LDA model for top terms of topics.

Table 1: The distribution probability of topics and words when topic K=160

Topic 5	$P(w_n\|k=160)$	**Topic 13**	$P(w_n\|k=160)$	**Topic 23**	$P(w_n\|k=160)$
Technology	0.001665396	Expert	0.0021303003	Patent	0.0095653171
Material	0.0014465271	Doctor	0.0009977445	Name	0.0047526103
Technique	0.0012552338	Senior engineer	0.0007695822	Number	0.0042560613
New Material	0.0007991531	Counselor	0.0006120452	Technology	0.0031178757
Product	0.0005905334	Master	0.0005237754	Information	0.0028379832
Technological innovation	0.0005886399	Bachelor	0.0005183982	Type	0.0027637654
High-performance	0.0005476442	Post-doctoral	0.0003545205	Invent	0.0038754337
Precision	0.0005436358	Degree	0.0002742724	Copyright	0.0016382793
New-technology	0.0004137853	Professor	0.0002726482	Authorized-patent	0.0007759799
New-product	0.0003711188	College	0.0002557352	Conservation	0.0006379388
Stability	0.0003021399	Associate-professor	0.0001320247	Authorization	0.0002794288
Practical	0.0002943092	Academic	0.0001297645	Intellectual-property	0.0002544165
Topic 27	$P(w_n\|k=160)$	**Topic 30**	$P(w_n\|k=160)$	**Topic 39**	$P(w_n\|k=160)$
Project	0.0008273301	Enterprise	0.0010061373	New-product	0.0008340621
Types	0.0008237544	Name	0.0009097208	New-techniques	0.0008340621
Invisible-asset	0.0004237544	Development Organization	0.0008218773	Name	0.0004272026
Fix asset	0.0004223029	Contact-telephone	0.0006368324	Market Occupancy	0.0004272025
Equipment	0.0004208453	Organization	0.0004940051	Profit	0.0004272025
Facility	0.0003637534	Company	0.0004912549	Period	0.0004272025
Total-amount	0.0003230094	Department	0.0004209615	Sale-quota	0.0004263008
Quantity	0.0003034324	Laboratory	0.0004209615	Sales-volume	0.0004262023
Cost	0.0002784593	Contact person	0.0004209615	Competitive	0.0004262010
Fund	0.0002764534	Research-institute	0.0004209615	Economic-benefit	0.0004260232
Amount	0.0002230895	Contact details	0.0004209615	Popularization	0.0003037646
Instrument	0.0002234943	Information	0.0003026468	Technical management	0.0003037564

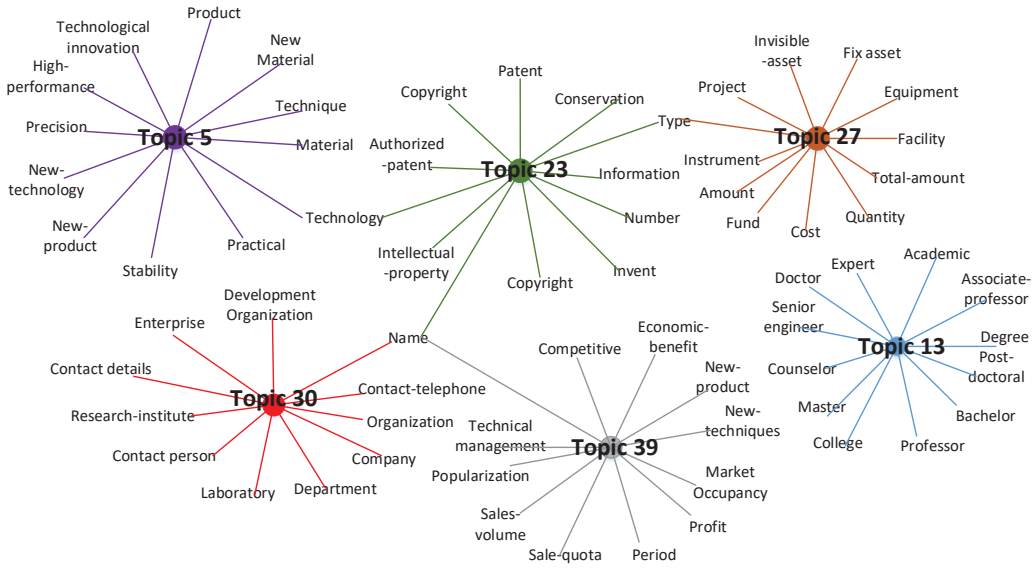

Figure 8: Graph of LDA for top terms of topics

4.5 Learning hierarchical relations among terms

Using the trained LDA model to infer each concept in the initial ontology and taking each concept (or word) as a document to calculate the topic probability of the document. Identify the semantic relations between existing concepts and its related terms, and add the related terms as the domain ontology concept to the appropriate position of the existing ontology to complete an update process of the domain ontology. Table.2 shows the results of the conceptual related terms extraction and relations recognition.

Table 2: Related terms extraction and relations recognition

Existing Concepts	**Topic**	$P(c_m\|z=j)$	**Related terms**	**Weights**	**Applicable rules**	**Semantic relations**
Profit Management	79	0.438621	Innovation Resources	0.00413586	(2)	Subclass
			Total Amount	0.00624005	(2)	Subclass
Technical Information	139	0.388462	Material	0.00651796	(2)	Subclass
			Painting	0.00331878	(3)	Related
			Alloy	0.00243573	(3)	Related
Visible Asset	27	0.236543	Fixed asset	0.01342533	(1)	Equivalent
			Equipment	0.00523433	(2)	Subclass
			Instrument	0.00243234	(2)	Subclass
Visible Asset	27	0.236543	Fixed asset	0.01342533	(1)	Equivalent
			Equipment	0.00523433	(2)	Subclass
			Instrument	0.00243234	(2)	Subclass
Technical-Quality	5	0.388462	Precision	0.00257653	(3)	Related
			New Technology	0.00323643	(3)	Related

Existing Concepts	Topic	$P(c_m\|z=j)$	Related terms	Weights	Applicable rules	Semantic relations
High-tech Product	136	0.446243	Strategy Innovation	0.00276406	(3)	Related
High-tech Product	136	0.446243	Technological Innovation	0.00143524	(3)	Related
Product Innovation	23	0.237643	Patent	0.00332763	(3)	Related
			Brand	0.00236232	(3)	Related
			Copyright	0.00323422	(3)	Related
			New Product	0.00332542	(3)	Related
Staff Management	13	0.376432	Expert	0.00335476	(3)	Related
			Doctor	0.003276543	(3)	Related
			Degree	0.002387432	(3)	Related
			Senior engineer	0.003723423	(3)	Related
Wage management	63	0.3412663	Wage	0.01472652	(1)	Equivalent
			Subsidy	0.00234653	(3)	Related
			Bonus	0.00334523	(3)	Related
			Insurence	0.00343263	(3)	Related

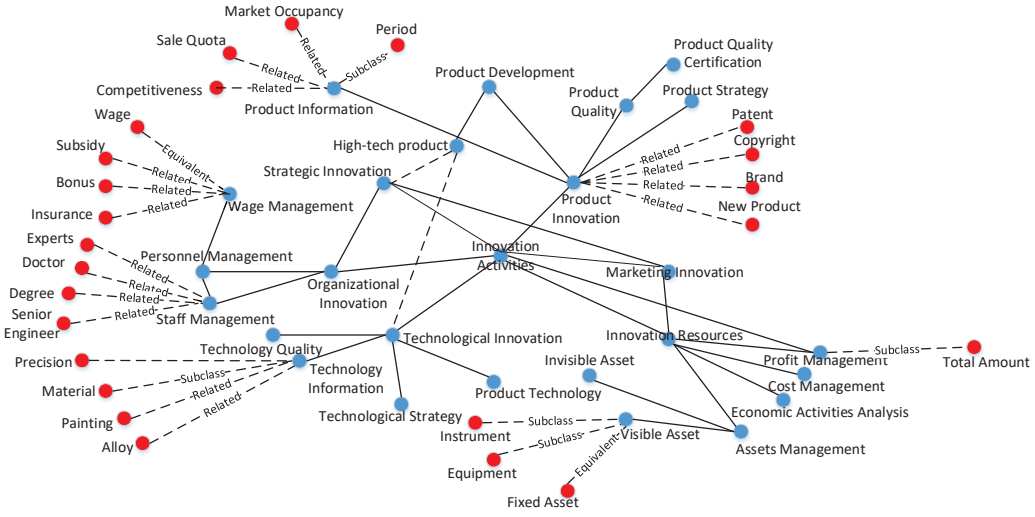

Figure 9: Parts of produced enterprise technological innovation domain ontology

The Fig.9 shows part of produced enterprise's technological innovation domain ontology. The blue dots represent the original terms of initial domain ontology and the red dots stand for the produced new terms. The original relations among entities in ontology are shown with solid lines, the dashed lines represent the new relations. The total amount of new terms in enterprise technological innovation domain ontology has updated about 163, the figure only shows parts of the result due to space limitation.

By looking at the literature of ontology evaluation, there are two approaches for measuring the ontology including of manual evaluation by human experts and gold standard-based approaches [11]. The first evaluation approach presents the learned ontology to one or more human experts and judge how far the extracted information is correct. The second method compare the learned

ontology with a previously created gold ontology which example for this kind of evaluation can be found in papers like [34]. The degree of matching between learned ontology and gold ontology determines the precision of learning ontology. The evaluation of ontologies when these ontologies are produced by an automated learning procedure is an open field of research. Since the enterprise's technological innovation is a new developing academic field which has not formed a generally acknowledged ontology yet. Therefore, the manual evaluation by human experts was the best way so far. The research chosen 5 groups and 20 terms and relations for each group in the updated enterprise's technological innovation domain ontology randomly. The assisted algorithm like following equation was defined as the ration between the right terms and relationships which evaluated by human experts and the total terms and relationships in ontology. According to the validation about the correct terms and relations with domain experts, the result of accuracy test is shown as Table 3.

$$precision = \frac{righttermsandrelationships}{totaltermsandrelationshipsontology} \tag{9}$$

Table 3: The accurate rate of the concepts in enterprise technological innovation domain

No.	Number of groups	Accurate number of groups	Precision
Group 1	20	19	95%
Group 2	20	19	95%
Group 3	20	18	90%
Group 4	20	17	85%
Group 5	20	19	95%

Compared with the traditional ontology construction methods such as OntoLearn and Text2-Onto, the proposed method has same precision which the average accuracy rate is 92%. The semantic content and relationship in the produced ontology is basically correct. The proposed automatic ontology extension method reduces the manual labor for ontology updating and solved the problem of automatic domain ontology acquisition and dynamic maintenance.

5 Conclusion and future work

This paper presented an automatic ontology extension method for the domain of enterprise's technological innovation. The main contributions of this paper present as follows: Firstly, this paper proposes an ontology-based LDA topic model for concept extraction and applies it to the realm of enterprise technological innovation, which not only discover the concepts from large amount of textual corpus, but also can provides data support for ontology construction. Secondly, this article takes a huge amount of enterprise technological innovation information in unstructured texts as the data source and proposes a method of building rules for semantic relationship recognition based on LDA topic probability distribution, and the process of automated domain ontology updating based on the LDA topic model is realized. Finally, the experiment results demonstrate the efficiency and validation of proposed method. The method focuses on discovering the domain terms via latent topics found by LDA algorithm from plain text corpus and recognizing the semantic relations among domain terms based on word association analysis. The proposed method is a process of continuously cycles and iterations, the domain ontology of enterprise's technological innovation will be updated and perfected automatically with the constant knowledge acquisition in the domain. The paper introduces the ontology on the basis of

the LDA topic model and the ontology is extended by the obtained related topics. The proposed method is an improvement for the single LDA algorithm.

The future work needs to solve several problems, firstly, improving the proposed method to achieve a better performance and continuing exploring automatic evaluation approaches on thesaurus constructing methods. Secondly, using the constructed enterprise technological innovation ontology and combined with the text mining methods to construct the mechanism of evaluation for enterprise's technological innovation.

Funding

This paper is supported by the Fundamental Research Funds for the Central Universities (2018YJS051,B18RC00070) and Beijing Social Science Funds (18JDGLA018).

Bibliography

[1] Bisson, G.; Nédellec, C. Canamero, D.(2000); Designing Clustering Methods for Ontology Building-The Mo'K Workbench, *ECAI workshop on ontology learning*, 31, 2000.

[2] Blei, D.M.; Ng, A.Y.; Jordan, M.I. (2003); Latent dirichlet allocation, *Journal of machine Learning research*, 3(Jan), 993–1022, 2003.

[3] Bradford, R.B. (2006); Relationship discovery in large text collections using latent semantic indexing, *Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism, and Security*, 2006.

[4] Bradford, R.B. (2005); Efficient discovery of new information in large text databases, *International Conference on Intelligence and Security Informatics*, 374–380, 2005.

[5] Burgelman, R.A.; Maidique, M.A.; Wheelwright, S.C. (1996); *Strategic Management of Technology and Innovation*, Chicago,IL:lrwin, 1996.

[6] Cimiano, P.; and Völker, J. (2005); text2onto, *International conference on application of natural language to information systems*, 227–238, 2005.

[7] Colace, F.; De Santo, M.; Greco, L.; Amato, F.; Moscato, V.; Picariello, A. (2014); Terminological ontology learning and population using latent dirichlet allocation, *Journal of Visual Languages & Computing*, 25(6), 818-826, 2014.

[8] Dai, Y.; Wu, W.; Zhou, H.B.; Zhang, J.; Ma, F.Y. (2018); Numerical simulation and optimization of oil jet lubrication for rotorcraft meshing gears, *International Journal of Simulation Modelling*, 17(2), 318–326, 2018.

[9] Dai, Y.; Zhu, X.; Zhou, H.; Mao, Z.; Wu, W.(2018); Trajectory tracking control for seafloor tracked vehicle by adaptive neural-fuzzy inference system algorithm, *International Journal of Computers, Communications & Control* 13(4), 465–476, 2018.

[10] De Knijff, J.; Frasincar, F.;Hogenboom, F. (2013); Domain taxonomy learning from text: The subsumption method versus hierarchical clustering *Data & Knowledge Engineering*, 83, 54-69, 2013.

[11] Dellschaft, K; Staab, S. (2008); Strategies for the evaluation of ontology learning, *Ontology Learning and Population*, 167, 253–272, 2008.

[12] Deng, L; Wang, X; Lin, Y; He, F.Z. (2005); Model of Multiple Fuzzy Synthetical Evaluation for Enterprise Technology Innovation, *Journal of Chongqing University (Natural Science Edition)*, 7, 004, 2005.

[13] Guan, J.C.; Yam, R.C.; Mok, C.K.; Ma, N. (2006); A study of the relationship between competitiveness and technological innovation capability based on DEA models, *European Journal of Operational Research*, 170(3), 971-986, 2006.

[14] Guarino, N.; Poli, R. (1993); Toward principles for the design of ontologies used for knowledge sharing, *In Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers, in press. Substantial revision of paper presented at the International Workshop on Formal Ontology*, 1993.

[15] Hennig, L. (2009); Topic-based multi-document summarization with probabilistic latent semantic analysis, *Proceedings of the International Conference RANLP-2009*, 144–149, 2009.

[16] Hofmann, T. (2001); Unsupervised learning by probabilistic latent semantic analysis, *Machine learning*, 42(1-2), 177–196, 2001.

[17] Khan, L.; Luo, F. (2002); Ontology construction for information selection, *Proceeding of Tools with Artificial Intelligence*, 122-127, 2002.

[18] Lee, C.S.; Kao, Y.F.; Kuo, Y.H.; Wang, M. H. (2007); Automated ontology construction for unstructured text documents, *Data & Knowledge Engineering*, 60(3), 547–566, 2007.

[19] Liu, Q.; Zhang, H.; Yu, H.; Cheng, X. (2004); Chinese lexical analysis using cascaded hidden markov model, *Journal of Computer Research and Development*, 41(8), 1421–1429, 2004.

[20] Ni, N.; Liu, K.; Li, Y. (2011); An automatic multi-domain thesauri construction method based on lda, *2011 10th International Conference on Machine Learning and Applications Workshops*, 235-240, 2011.

[21] Raghuveer, K. (2012); Legal documents clustering using latent dirichlet allocation, *International Journal of Applied Information Systems*, 2(1), 34-37, 2012.

[22] Saunila, M.; Ukko, J. (2012); A conceptual for the measurement of innovation capability and its effects, *Baltic Journal of Management*, 7(4), 355–375, 2012.

[23] Tho, Q.T.; Hui, S.C.; Fong, A.C.M.; Cao, T.H. (2006); Automatic fuzzy ontology generation for semantic web, *IEEE transactions on knowledge and data engineering*, 18(6), 842-856, 2006.

[24] Tsai, M.T; Chuang, S.S; Hsieh W.P. (2008); Using Analytic Hierarchy Process to Evaluate Organizational Innovativeness in High-Tech Industry, *Decision Sciences Institute 2008 Annual Meeting (DSI)*, 1231-1236, 2008.

[25] Wang, T. J; Chang, L. (2011); The development of the enterprise innovation value diagnosis system with the use of systems engineering, *System Science and Engineering (ICSSE), 2011 International Conference on IEEE*, 373–378, 2011.

[26] Wang, C; Lu, I; Chen, C. (2008); Evaluating firm technological innovation capability under uncertainty, *Technovation*, 28(6), 349–363, 2008.

[27] Wei, W.; Guo, C.; Chen, J.; Tang, L.; Sun, L. (2017); CCODM: conditional co-occurrence degree matrix document representation method, *Soft Computing*, 1-17, 2017.

[28] Wei, W.; Guo, C.; Chen, J.;Zhang, Z. (2017); Textual topic evolution analysis based on term co-occurrence: A case study on the government work report of the State Council (1954–2017), *Intelligent Systems and Knowledge Engineering*, 1-6, 2017.

[29] Yeh, J.H.; Yang, N. (2008); Ontology construction based on latent topic extraction in a digital library, *International Conference on Asian Digital Libraries*, 93–103, 2008.

[30] Yliherva, J. (2004); Management model of an organization's innovation capabilities; development of innovation capabilities as part of the management system, *dissertation, Department of Industrial Engineering and Management, University of Oulu.*

[31] Zhang, W.; Zhang, Z.; Chao, H.C.; Tseng, F.H. (2018); Kernel mixture model for probability density estimation in Bayesian classifiers. Data Mining and Knowledge Discovery, *Data Mining and Knowledge Discovery*, 32(3), 675–707, 2018.

[32] Zhang, W.; Zhang, Z.; Qi, D.; Liu, Y. (2014); Automatic crack detection and classification method for subway tunnel safety monitoring, *Sensors* , 14(10), 19307–19328, 2014.

[33] Zhao, W.; Zeng, Y. (2011); Construction and design of evaluation index system of innovative enterprises on innovative capacities, *Science and Technology Management Research*, 1, 005, 2011.

[34] Zavitsanos, E.; Paliouras, G.; Vouros, G.A.; Petridis, S. (2010); Learning subsumption hierarchies of ontology concepts from texts, *Web Intelligence and Agent Systems: An International Journal*, 8(1), 37-51, 2010.

[35] Zavitsanos, E.; Paliouras, G.; Vouros, G.A.; Petridis, S. (2010); Discovering subsumption hierarchies of ontology concepts from text corpora, *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 402–408, 2007.