Research and Application on Domain Ontology Learning Method Based on LDA

Wang Hong, Zhang Hao*, and Shi Jinchuan
School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China

* Corresponding author. Tel.: +86 13820811445; email: zhanghao1205@gmail.com
Manuscript submitted February 10, 2017; accepted April 12, 2017.
doi: 10.17706/jsw.12.4.265-273

Abstract. Considering the problem of multi-source heterogeneous cross-media text information in the field of aviation safety is difficult to share, the paper proposed a domain ontology learning method for civil aviation emergency management. The use of adaptive the NLPIR word segmentation and filtering methods to obtain the candidate term dataset. LDA topic model of domain ontology was designed, through the LDA model training of Gibbs sampling and topic inference to realize the related terms of domain ontology concept core extraction. The construction method of basic semantic relation recognition rules was studied based on the LDA topic probability distribution. The recognition and implementation of the concept and its related term basic semantic relations were presented. Experimental results show that the method can effectively solves the problem of automatic updating of concepts and relations in large-scale domain ontology, and it provided a good data support for sharing and reasoning of civil aviation emergency cross-media information under the environment of big data.

Keywords: Cross media; text information; ontology learning; LDA; civil aviation emergency.

1. Introduction

Ontology [1] is an important tool for information and knowledge representation, and scholars at home and abroad have done a lot of research on ontology construction method and its application [2]-[5]. Ontology can realize the expression, sharing and reuse of domain knowledge effectively. In the field of civil aviation, in order to enhance the ability of civil aviation to deal with emergencies, the domain ontology has been established based on the domain dictionary [6]. With the advent of the era of cross-media information [7], [8], new domain knowledge is emerging, automatic update of domain ontology face new challenges.

Ontology learning [9]-[11] is an automatic or semi-automatic extraction of concepts, relationships and other elements of ontologies from data sources, and uses these newly extracted elements to reconstruct or extend the technology of updating existing ontologies. In the previous research, the statistical-based machine learning method [12], [13] has great advantages in both accuracy and recall, and now, this method has become the mainstream technology of ontology learning. However, most of the existing methods require a certain amount of training corpus as learning objects, and don't consider the semantic factors. Facing the massive unstructured text information in the cross-media information, these methods are difficult to carry out ontology learning effectively.

LDA (Latent Dirichlet Allocation) [14] is a three-layer Bayesian probability model, which can identify the semantic topic information in large-scale document sets or corpus. It is an unsupervised machine learning technique, and it's suitable for a large number of natural texts processing [15], [16].

In this paper, LDA model is introduced into the automatic updating of the domain of civil aviation emergencies, so we propose an ontology learning method based on LDA, and the method not only enriches and improves the ontology knowledge representation content, but also provides better data support for event emergency management and collaborative decision making.

2. Research Ideas

Domain ontology O is usually composed of four elements, namely $O=(C, R, A, I)$, where C is a set of domain-related concepts, R is a set of relations between domain concepts, A is a set of axiom and rules, and I is an instance set of domain concepts. On the basis of existing domain ontology O, the domain ontology learning method based on LDA update the concept and relationship of domain ontology by using cross-media text information as data source, and the updated domain ontology is denoted as O_N.

Namely: $O_N = (C \cup C_N, R \cup R_N, A, I)$.

The method of research route was shown in Fig. 1.

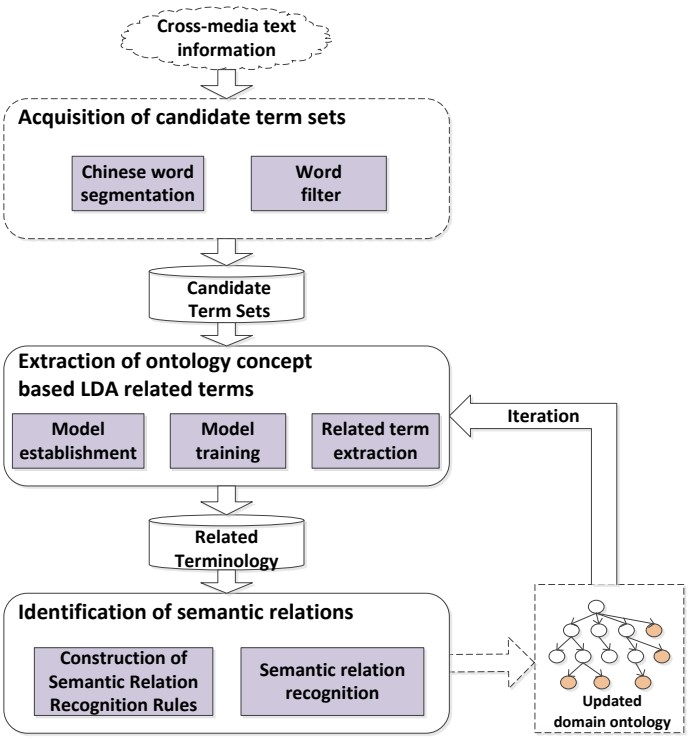

Fig. 1. Framework of the domain ontology learning system.

- Acquisition of candidate term sets: For the unstructured cross-media text information, the text data are pre-processed by using adaptive word segmentation in NLPIR (Natural Language Processing & Information Retrieval) and word-filtering to form candidate set of domain ontology learning.
- LDA-based ontology-related terminology extraction: A LDA model is built on the ontology domain of civil aviation emergencies. The model is trained by candidate term sets and the semantic related terms are deduced from the core concepts of domain ontology.

- Semantic relation identification: According to the probability distribution relationship defined by the LDA model, a set of rules for recognizing the domain ontology and semantic relations of the related terms are designed and constructed. The rules are used to identify the semantic relationship between domain ontology concepts and its related terms, and then we realize the automatic updating of concepts and relations.

The above process can be carried out periodically, in each cycle, the new cross-media text information is retrieved as a data source, the candidate concept set is acquired again, and the updated domain ontology is re-entered into the system as the initial domain ontology. The extraction of concept-related terms and the recognition of semantic relations are carried out, and the automatic learning of domain ontology is realized.

3. Acquisition of Candidate Term Sets

The process of acquiring the candidate term set is to segment and filter the text data of the cross-media information. Domain Collection S_{doc} is a collection of domain documents collected from massive cross-media network resources by crawling and parsing, and the collection as a data source for domain ontology learning. In this unstructured Chinese text data, the boundaries between words and vague, the computer is often difficult to identify, so it's necessary to segment Chinese word. In this paper, we use the open-source NLPIR Chinese word segmentation system, at first we invoke the NLPIR_NWI function, identify the domain vocabulary automatically, and add the domain vocabulary to the word segmentation dictionary, and then traverse the field S_{doc} collection of each document in the word segmentation operation, to achieve adaptive word segmentation, at last the formation of domain terms set S_{term} is formed.In order to improve the training efficiency of LDA subject model and enhance the quality of ontology learning, we need to stop word filtering, low frequency word filtering and part-of-speech filtering in terms of domain term set S_{term}.

After the above process, a candidate term set S_{cand} of domain ontology learning is formed, which provides the training sample data for LDA subject model.

4. Extraction of Ontology Concept Based LDA Related Terms

4.1. The Establishment of the LDA Theme Model

For domain ontology learning, each domain term in a domain document selects a domain topic with a certain probability, and selects a domain term from a certain probability from the domain topic. The LDA three-layer Bayesian Probabilistic model contains domain documents, domain topics, domain terms, three-tier structure.

The LDA model of domain ontology learning is a typical directed probability graph model, which is determined by the hyper-parameters α and β. The α reflects the relative strength between implied topics in the domain document set, and the β depicts the probability distributions of all implicit subjects themselves, as shown in Fig. 2.

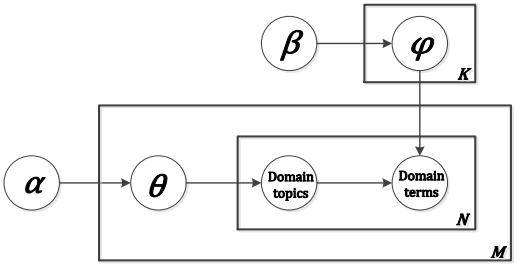

Fig. 2. Represents probabilistic graphical models.

where θ is the probability distribution of the topic documents in the domain, φ is the probability distribution of the domain terms under the specific topic, M is the domain document number in the domain document set, K is the number of domain document sets, and is the number of domain terms contained in each domain document.

The probability of a topic *topic* appears in the domain document *doc* and the probability of a domain term *term* appearing in the topic *topic* , and the probability of the domain term *term* appears in the domain document *doc* is calculated as follows:

$$p(term|doc) = p(term|topic)*p(topic|doc) \quad (1)$$

Where *doc* represents a domain document, *topic* represents a domain topic, and *term* represents a domain term. $p(topic|doc)$ calculated using θ, and $p(term|topic)$ calculated using φ.

4.2. The Extraction of Terms Related to Ontology Concept

The concept of domain ontology is an important part of domain ontology learning. How to automatically obtain semantically rich and accurate domain ontology concepts from a large number of candidate domain terms is a core problem of domain ontology learning. Using LDA thematic inference, we can derive the degree of correlation between domain terms, input domain ontology to deduce the topic, and get the domain terms related to the concept. The domain topic with the highest subject probability and the domain term under this topic And the input concept is most closely linked. The algorithm design is shown in Figure 3.

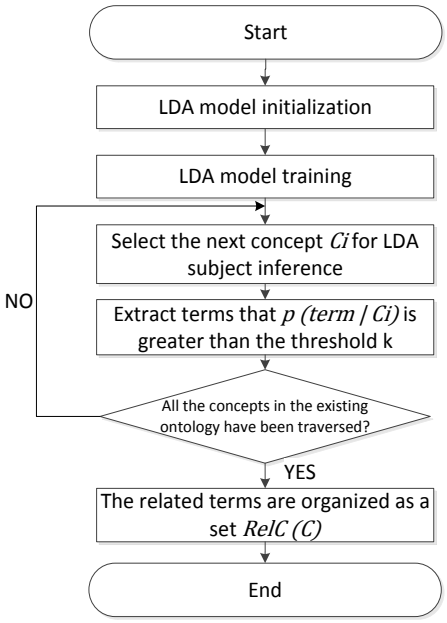

Fig. 3. Conceptual relevance term extraction algorithm based on LDA.

STEP 1. We construct the topic model of LDA and use the candidate term set obtained in section 3 as the training corpus. We use the Gibbs sampling method [18] to train the LDA model, and get the topic-term probability distribution $D(term|topic)$;

STEP 2. It traverses every concept C_i (assuming there are n concepts, i=1,2,…,n) of the existing domain ontology O:

a) With the current concept of C_i as input concept, the use of trained LDA model concept C_i thematic inference to give the concept-topic probability distribution $D(topic|C_i)$, by equation (1) can be calculated concept - term probability distribution $D(term|C_i)$ of concept C_i;

b) The threshold k is set, selection concept-term probability distribution $D(term|C_i)$ with probability values $p(term|C_i)$ greater than k and are not domain terms in the concept of existing ontologies, as a related term C_{ij} for the concept C_i (assuming there are m related terms, j =1, 2,…,m);

STEP 3. The b) process to get the relevant terms collated, and statistical probability distribution, the relevant term set *RelC(C)*.

$RelC(C)$={$RelC(C_1)$,$RelC(C_2)$,…,$RelC(C_n)$} $RelC(C) = \{RelC(C_1), RelC(C_2), \cdots, RelC(C_n)\}$, the $RelC(C_i)$ is the corresponding subset of the n concepts C_i(i=1,2,…,n) in the existing ontology.$RelC(C_i)$={C_{i1},C_{i2},C_{i3},…,C_{im}},C_{ij}(j=1,2,…,m) denotes m related terms related to concept C_i.

5. The Identification of Basic Semantic Relations Introduction

How to effectively identify and define the semantic relationship between these terms and concepts becomes another key problem of ontology updating after acquiring the ontology concept-related term set $RelC(C)$. Therefore, on the basis of the in-depth analysis of the probability distribution of $RelC(C)$, a set of semantic relation recognition rules based on LDA topic probability distribution is defined to realize the recognition of the relationship between the concepts of domain ontology.

The concept-term probability corresponding to the relate terms C_{ij} of the ontology core concept C_i is $p(C_{ij}|C_i)$ (i=1,2,…,n;j=1,2,…,m), can be derived from the concept-term probability distribution $D(term|C_i)$ in the LDA inference result. The degree of correlation between the respective term C_{ij} and the core concept C_i is defined by defining the weight $W(C_{ij},C_i)$ of the relevant term, the calculation formula is as follows:

$$W(C_{ij}, C_i) = p(C_{ij}|C_i)/\sum^{m}_{z=0} p(C_{iz}|C_i) \quad (2)$$

The rules for identifying basic semantic relationships are defined as follows:

1) Synonymous relationship recognition rules: If the weight W(Cij, Ci) of the related term Cij of the existing concept of ontology Ci is greater than or equal to 0.5 and the semantic relation between Cij and Ci is the same, and then the related term Cij has the same relationship with the existing concept Ci.
2) Upper and lower relationship recognition rules: When the condition of rule (1) is not satisfied, if the related term Cij of the existing concept of ontology Ci is maximal in its corresponding topic T and greater than 0.1, and considering that Cij have a general effect on the topic T, the related term Cij and the existing concept Ci exist upper and lower relationship, Cij as Ci's the upper word.
3) Related relationship recognition rules: When both the rule 1 and the rule 2 are not satisfied, the existing concept Ci is considered to be related to the related term Cij, or the specific semantic relationship can be identified manually using the external civil aviation professional knowledge base.

According to the above rules, the semantic relations between the existing concepts and their related terms are identified, and the related terms C_n and semantic relationships R_n which is obtained are added to the domain ontology as appropriate position by the Jena [19] tool, the existing ontology O is updated to O_{ij}, which enriches the concept and relation of domain ontology and completes the automatic learning of domain ontology.

6. Experiments

6.1. Experimental Data and Parameters

The experiment was done in the Microsoft Visual Studio 2015 environment using the C # language. Ten thousand , we collected 50,000 civil aviation emergencies related text information in 2015 form social platform(such as Sina micro-blogs , Tencent micro-blog and Wechat public platform)as experimental data, the average length of the text is 160 words. In addition, the civil aviation emergencies domain ontology which has been established is chosen as the initial ontology for domain ontology learning.

In the LDA model, we use the Bayesian standard method [20] to select the smallest number of subjects as the experimental parameters T, set T=50, as shown in Figure 4. Hyper parameter α and β are set according to general experience, where α=50/T=1, β=0.01. In addition, during the related term filtering process, the learning efficiency of the ontology is optimal when the threshold k=0.01 is set.

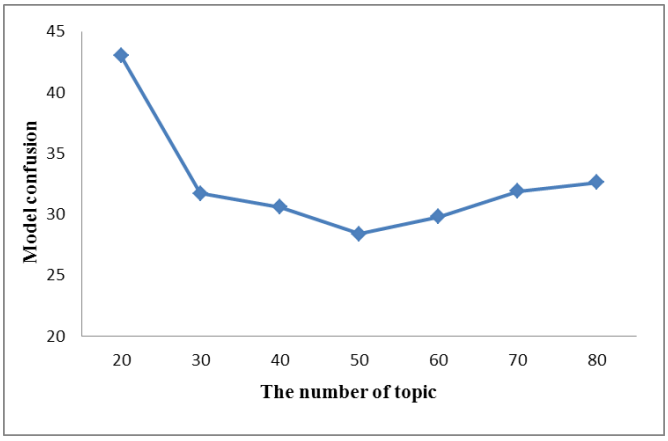

Fig. 4. LDA model distribution of confusion.

6.2. Experimental Process

Using the Web crawler and API interface technology to crawl a large number of cross-media text information which is related to civil aviation emergencies, and then we use word segmentation and filtering technology for cross-media text information pre-processing to get a candidate set of terms.

After the LDA model was established, the Gibbs sampling training method was used to train the candidate term set as the training corpus's input of LDA subject model. Then, the LDA model was used to deduce the concept of the existing ontology the related term set and the LDA topic probability are obtained after screening and sorting. The core code is as follows:

```
Text2Bag( "CandidateTerms.data","LDA_train.data")
LDA_Init(alpha,beta,topick_num,"LDA_train.data");
//Training process
LDA_RandomSampling(&Statistics);
While(converged >1E-6)
{   LDA_GibbsSampling(&Statistics);
    LDA_Calcu_param(Statistics,&phi , &theta); }
LDA_save_model(&phi , &theta , &Statistics);
ReadOntology("CAEDO.owl");
//LDA inference, related term extraction process
```

```
foreach(Concept in Concepts)
{  LDA_RandomSampling(&Infer_Statistics);
   While(infer_converged > 1E-6)
   {  LDA_Infer_GibbsSampling(&Statistics);
      LDA_Calcu_Infer_param(Statistics , &theta);}
   LDA_Infer_Result(&theta,&phi, "RelTerms.data");
}
```

The basic semantic relation between the existing domain ontology concept and its related terms is identified by using the basic semantic relation recognition rules based on LDA topic probability distribution. Finally, we invoke Jena interface and add the related terms as domain ontology Concept to the existing ontology proper position to realize ontology learning.

6.3. Experimental Effect Analysis

The results of concept-related term extraction and semantic recognition are shown in Table 1. And the "semantic relation" of "relevance relation" is obtained by external civil aviation professional knowledge base, and the base which is artificial identified.

Table 1. Conceptual Terms Extraction and Semantic Relation Recognition Results（English Comments）

Existing concepts C_{ji}	Related terms C_{ij}	Topic - term probability $p(C_{ij}\|T)p(C_{ji}\|T)$	Weights $W(C_{ij}, C_{ji})W(C_{ji}, C_i)$	Applicable Rules	Semantic relations (Basic relations)
民航突发事件(Civil aviation emergency)	事故灾害 (Accident disaster)	0.1526	0.4260	2	Subclass (upper & lower)
	民航紧急事件 (Civil aviation urgent event)	0.1163	0.5741	1	Equivalent (synonymous)
航空器突发事件 (Aircraft emergency)	失事(Crash)	0.0823	0.3527	3	Result (Related)
	故障(Malfunction)	0.0775	0.2880	3	Is-a (Related)
	干扰(Interference)	0.0518	0.1833	3	Is-a (Related)
	相撞(Collide)	0.0302	0.1051	3	Reason (Related)
	恐怖袭击(Terrorist attacks)	0.0920	0.3264	3	Is-a (Related)
非航空器突发事件(Non - aircraft emergency)	管理不当(Improperly managed)	0.0736	0.2183	3	Reason (Related)
	矛盾冲突 (Contradictory conflict)	0.0849	0.2011	3	Result (Related)
	自然灾害(Natural disaster)	0.0544	0.1362	3	Reason (Related)
	群体性事件(Mass incidents)	0.0503	0.1185	3	Is-a (Related)
航班延误(Flight delay)	天气(Weather)	0.0823	0.4078	3	Reason (Related)
	航班取消(Flight canceled)	0.0775	0.3234	3	Result (Related)
	准点率(Punctuality rate)	0.0518	0.2588	3	Affected (Related)
...	...	...	...	...	...

Figure 5 shows the civil aviation emergencies after the study of the local ontology effect map. The solid line ellipse represents the original concept of the domain ontology, and the dotted line ellipse represents the

updated concept of learning. The solid arrow represents the original relation of the domain ontology, and the dotted arrow represents the updated relationship of learning.

This field of ontology update a total of 247 groups of concepts and relationships, but because of limiting by space, this paper shows only part of the results. We selecte 50 groups of concepts and relationships randomly to test the accuracy, and we test 3 times according to the knowledge of the external civil aviation knowledge base. The test results shows that the accuracy of ontology learning is more than 90%.

The experimental results show that applying the method described in this paper greatly reduces the workload of updating the ontology, and the semantic contents and relations are basically accurate, and the output domain ontology is reasonable, which meets the application requirements of civil aviation airport, For the civil aviation airport safety management level of upgrading and service quality protection provides strong support.

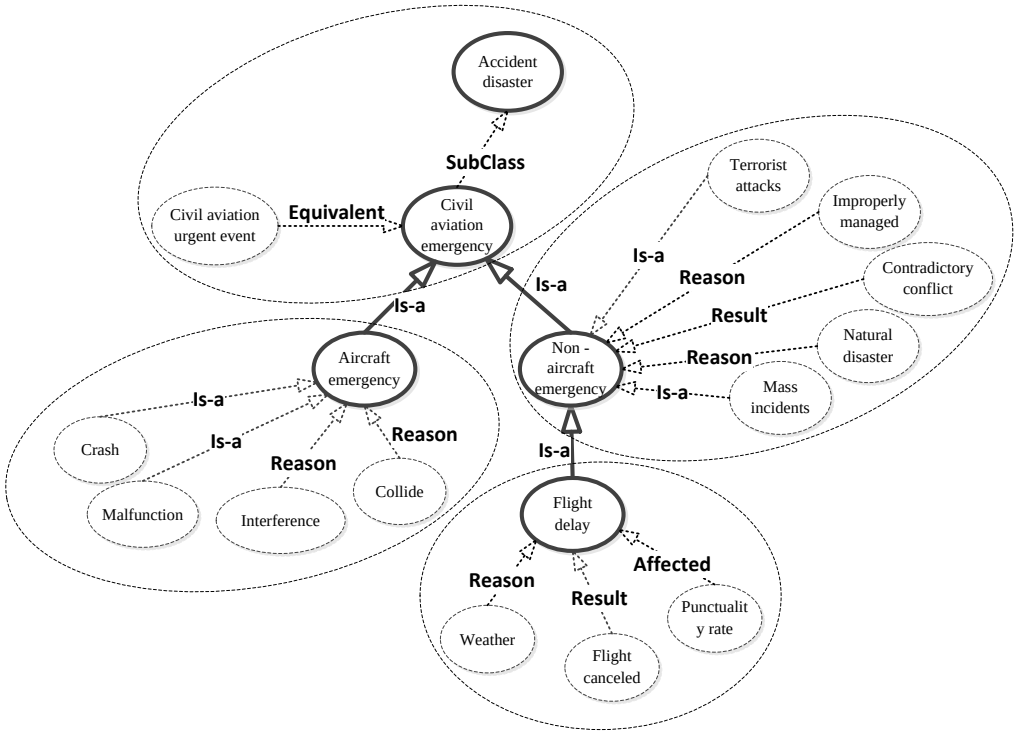

Fig. 5. The result of civil aviation emergencies ontology learning.

7. Conclusions

Based on the LDA ontology learning method and the construction and training of the LDA model, this paper realizes the automation of the related terms and semantic relations of domain-oriented concepts in cross-media text information. This provides a good method for ontology-oriented learning and domain ontology application. Due to the complexity of the concept of domain ontology and its semantic relations, there are still many problems in the construction of rules based on LDA probability distribution and the automatic learning of ontology instances, which are worthy of further study.

Acknowledgment

This work is partially supported by the Joint Funds of the National Natural Science Foundation of China and the Civil Aviation Administration of China (U1633110).

References

[1] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5(2)*,199-220.

[2] Li, S. P., Yin, Q., Hu, Y. J. *et al*. (2014). Overview of researches on ontology. *Journal Computer Research and Development*. *41(7)*,1041-1052.

[3] Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: methods, evaluation and applications . Amsterdam: IOS Press. 3-14.

[4] Zhang, K. L., Li, W. G., & Wang, H. L. (2015). Ontology-based question answering system for aviation domain. *Journal of Chinese Information Processing*, *29(4)*,192-198.

[5] Soylu, A., Giese, M. *et al*., Experiencing OptiqueVQS: A multi-paradigm and ontology-based visual query system for end users . *Universal Access in the Information Society*, *15(1)*, 129-152.

[6] Wang, H., & Yang, X. (2010). Research and implementation of the civil aviation emergency management domain ontology construction method. Civil Aviation University of China.

[7] Zhao, Y., Wei, S. K., Wang, S. H. *et al*. (2014). Knowledge representation in the era of cross media: perception, association and consistent representation. *Communications of the CCF*.

[8] Guo, Q., Jia, J., Shen, G. *et al*. (2016). Learning robust uniform features for cross-media social data by using cross auto encoders. *Knowledge-Based Systems*.

[9] Liu, W., Weichselbraun, A., Scharl, A. *et al*. (2005). Semi-automatic ontology extension using spreading activation . *Journal of Universal Knowledge Management*.

[10] Craven, M., Dipasquo, D., Freitag, D. *et al.* (2015). Learning to extract symbolic knowledge from the world wide web. *Coastal Management*, *31(2)*, 121-126.

[11] Du, X. Y., Li, M., & Wang, S. (2006). A survey on ontology learning research. *Journal of Software*, *17(9)*, 1837−1847.

[12] Wei, X. L., Sun, Y., & Zhang, S. K. (2009). Ontological concept extraction method based on maximum entropy model. *Computer Engineering*. *35(24)*.

[13] Yang, Q., Cai, K. M., Sun, J. L. *et el*. (2010). Design analysis and implementation for ontology learning model.

[14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*.

[15] Xu, J. J., Yang, Y., Yao, T. F. *et al*. (2016). LDA based hot topic detection and tracking for the forum. *Journal of Chinese Information Processing*.

[16] T. Lian, J. Ma, S. Q. Wang, *et al.* (2014). LDA-CF: A mixture model for collaborative filtering. *Journal of Chinese Information Processing*.

[17] NLPIR. Retrieved, from http://ictclas.nlpir.org/

[18] Griffiths, T. L., & Steyvers, S. M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*.

[19] JENA. Retrieved, from http://jena.apache.org/.

[20] Steyvers, M. & Griffiths, T. L. (2007). Probabilistic topic models. *Psychological Review*.

Wang Hong was born in 1963, is a M. S., a professor, a CCF member. Her research interests include ontology technology, data mining and intelligent information processing.