TRANSACTIONS OF THE CHINESE SOCIETY FOR AGRICULTURAL MACHINERY
Vol. 47, No. 9, 2016

doi:10.6041/j.issn.1000-1298.2016.09.038

# Research on Ontology Non-taxonomic Relations Extraction in Plant Domain Knowledge Graph Construction

Zhao Ming　Du Yaru　Du Huifang　Zhang Jiajun　Wang Hongshuo　Chen Ying
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

**Abstract**: In order to provide more specific knowledge and technology of plant field, the main task of KG (knowledge graph) is to extract a wealth of concepts and relationships. Due to the relation extraction is the most difficult in KG construction, this paper makes use of ontology learning, and proposes a non-taxonomic relation learning method to obtain representative concepts and their relations from unstructured and semi-structured texts of Baidu Encyclopedia entry content by using lexicon-syntactic patterns based on dependency grammar analysis. Moreover, the methods of adding constraint models and words filtering were adopted to build heavy weight ontology automatically based on a lightweight ontology and greatly improved the precision of the relation extraction. The approach established a concept structure from the plant domain corpus, ameliorated the discovery of the most representative non-taxonomic relation, and formalized them in the standardized OWL 2.0. A set of experiments was performed using the approach implemented in the plant domain. The results indicated that extraction by patterns should be performed directly after natural language processing, which has a comparatively high accuracy compared to the former algorithms, and this approach can extract non-taxonomic relations with high effectiveness, which lays the foundation for KG construction of plant field.

**Key words**: plant domain ontology; knowledge graph; non-taxonomic relation; ontology learning; Baidu Encyclopedia

## 0 Introduction

"Knowledge Graph"[1], which is defined as the knowledge representation method, contains a large number of entities, taxonomy relationships and non-taxonomy relationships between entities, and it is a semantic knowledge base. Knowledge graph supports comprehensive information retrieval, intelligent question-answering, intelligent decision-making and so on. The main task of knowledge graph construction is to extract rich entities and relationships, in which extracting non-taxonomy relationships is the most difficult problem. Extracting non-taxonomy relationships though ontology learning not only enhances the completeness and complexity of knowledge representation, but enriches and extends the knowledge graph to a great extent[2].

Ontology[3], used for describing or representing the entities and relationships in a certain field, is a basic knowledge system and the formal explicit specification of shared conceptual model. The applications and solution based on ontology is implemented on the basis of ontology construction. However, traditional ontology construction tools (e.g., Protégé, KAON) support the process of manually building an ontology, which is often time-consuming and difficult to dynamically update in a timely way, thus causing the so-called "knowledge acquisition bottleneck" problem. How to build ontology automatically or semi-automatically, namely, ontology learning, has long been a problem worthy of study[4-5]. Currently, we summarized the methods in three categories: methods based on the dictionary, methods based on pattern matching and methods based on association rules.

Methods based on the dictionary can abstract relations that exist in dictionaries, e.g., WordNet; only synonymy, antonymy and part-whole relations can be drawn from the WordNet. This method has a great limitation. In the researches of the methods based on pattern matching, researchers set various parameters such as correlation threshold of entities[6], similarity weight[7], log likelihood[8-9] and granular computing

Received date: 2016-03-09　Accepted date: 2016-05-15
Supported by National Natural Science Foundation of China (61503386)
Corresponding author: Zhao Ming, Professor. E-mail: zhaoming@cau.edu.cn. Tel: +86-138-81027032.

model[10] to extract non-taxonomy relationships. Since the mistakes of part-on-speech tagging have a great influence on the accuracy, we have to explore more effective methods.

Methods based on association rules[11] summarize the language patterns used frequently to identify the appropriate semantic relationships though studying the related texts. Researchers adopt semantic dependency method (part-of-speech tagging, role labeling, semantic analysis) to gain the verb frameworks with semantic dependency, build corpus of verb vector[12], and calculation of sentence similarity. Another part of researchers have proposed a heuristic classification of non - relational learning framework, which integrated semantic patterns and statistics[13]. The method has a high extraction accuracy, but some headwords may impact on the selection of non-taxonomy relationships. In addition, non-taxonomy relationships contain a mount of attribute relationships such as origin, plant diseases and insect pests, instead of the relationships between entities. And these attribute relationships are more complex, so the extraction effect is poor. Due to the entity ambiguity has influence on non-taxonomy relationship category, the precision rate and recall rate of the methods based on association rules can be further optimized in ontology construction.

The methods of non-taxonomy relationships extraction cover agriculture[14], medical diagnosis and treatment[15], website[16], aviation management[17] and so on. In addition, some researchers focus on the unsupervised learning method[18] to extract non-taxonomy relationships form web text, and make evaluation of the method[19]. Chinese ontology construction automatically, especially researches of non-taxonomy relationships learning in agricultural field just getting started. Hence, aiming at limitation of non-taxonomy relationships extraction based on pattern matching method, this paper selected plant terms in Baidu Encyclopedia as corpus, adopted lexicon-syntactic patterns to extract non-taxonomy relationships, improved it by stoplist filtering and method of increasing the restriction for the the pattern, and made discussion on the relationship between non-taxonomy relationships categories to carry out the research.

## 1 Non-taxonomic relation learning methods

The specific steps of the non-taxonomic relation extraction method are as follows:

(1) Grab the web terms content as corpus, preprocess the corpus and uselexicon-syntactic pattern matching method to extract relation.

(2) According to characteristics and deficiencies of this method, put forward a type of guidance mode to improve the accuracy of the extraction.

(3) Based on relation triples of the above extracted methods, use OWL language to formalize the relations.

The overall process is shown in Fig. 1.

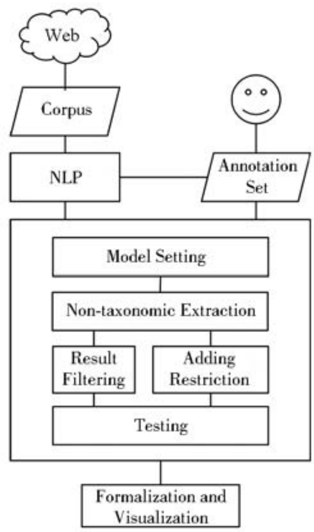


Fig. 1 Flow chart of technological process

### 1.1 Non-taxonomy relationships extraction based on lexicon-syntactic patterns

#### 1.1.1 Data acquisition and preprocessing

This paper grabbed 9623 plant entries and their respective contents as corpus from the Baidu Encyclopedia using a corpus collection script tool. The corpus are stored in GBK format text files. The preprocessing process uses an open source tool named LTP (Language technology platform)[20], which mainly uses three modules, i.e., word segmentation, POS tagging and interdependence syntactic analysis, to preprocess the corpus. Preprocessed results are saved in XML format.

#### 1.1.2 Modes acquisition

To acquire lexical-grammatical patterns, we selected a small group of high-quality and typical entries of plant materials from the Baidu Encyclopedia (97 entries and 19 types) to find the typical statement that expressed the non-taxonomic relation. In this paper, we focus on non-taxonomic relations with agricultural value, such as the geographical distribution, fit environment, diseases and insect pests, economic value of plants. Tab. 1 lists some of these statements. At present, we identify these statements by means of artificial reading of the entry content. This approach is only suitable for the condition in which there are relatively few selected entries and they do not have a scope of scalability. In the future, we can automatically look for two or more agricultural concepts that appear at the same time by using a program.

Using interdependence syntactic analysis to address these statements and by combining manual summary and statistical analysis, we summarize the lexicon-syntactic patterns expressing a non-taxonomic relation. First, we conduct interdependence syntactic analysis on these statements and determine concepts and non-taxonomic relations in these sentences, denoted as $R_i$ ($X_i$, $Y_i$). Then, we count the number of interdependent relation sequence $X_i$, $R_i$, $Y_i$ occurrences in other annotations, i. e., $X_j$, $R_j$, $Y_j$ ($i \neq j$). After artificially eliminating the ungrammatical sequences, we take the remaining multi-occurrence sequences as the lexicon-syntactic patterns. Moreover, we summarize some patterns as a supplement. A set of lexicon-syntactic patterns are shown in Tab. 2 (see interdependence syntactic tagging definition in LTP official document).

**Tab. 1 Some of statements listed**

| Category | Entry name | Sentence |
|---|---|---|
| Flowers | Dendrobium | 虫害有介壳虫危害,用40%氧化乐果乳油2 000倍液喷杀 |
| Tea | Tie Guayin | 铁观音原产于福建安溪县西坪 |
| Root herbs | Polygonatum Odoratum | 宜温暖湿润气候,喜阴湿环境,较耐寒,在山区和平坝都可栽培 |
| Crops | Solanum Tuberdsm | 它除了具有普通红薯的营养成分外,还富含硒元素和花青素 |

**Tab. 2 Set of lexicon-syntactic patterns**

| Patterns | Examples |
|---|---|
| SBV (Y, X), HED (Root, Y), VOB (Y, Z)→Y (X, Z) | 主要虫害有介壳虫和斑蛾。→have (Insect peat, Scale insect) |
| SBV (Y, X), HED (Root, Y), CMP (Y, Z), POB (Z, W)→ Y_Z (X, W) | 铁观音原产于福建安溪县西坪。→born in (Tie Guanyin, Xiping) |
| VOB (X, Y), DE (Z, X), ATT (W, Z)→X (W, Y) | 松毛虫是危害马尾松的常见害虫。→harm (peats, pinus massoniana) |
| SBV (Y, X), HED (Root, Y), COO (W, Z)→Y (X, Z) | 主要虫害有介壳虫和斑蛾。→have (insect peat, burnet moths) |

Patterns are expressed in the following form: Dependency (head node, dependent node) → non-taxonomic relation (Agent concept, patient concept), which can be formalized as follows:

$$\bigwedge_{i=1}^{n} D_i(s_i, t_i) \Rightarrow L(A, B) \tag{1}$$

Formula (1) says that $L(A, B)$ is determined when all dependencies $D_i$ are met, among which $L$, $A$ and $B$ are the specified elements of $s_i$ and $t_i$.

### 1.1.3 Adding rules for pattern

For the lexicon-syntactic patterns extracted, a question worth attention is that, some patterns contaion parallel relationship (COO) and the serial verb contruction (VV). If the dependencies of a node is COO, the nodes equated with dependencies node pointed to, and the property is transitive, i. e., in the sentence"主要虫害有介壳虫和斑蛾。", the "介壳虫" and "斑蛾" will be considered equal. If the dependencies of a node is VV, the nodes and the dependencies node pointed to will be considered as shared subject, that is VV(X, Y), SBV(X, Z)→SBV(Y, Z), and the property is transitive too, i. e., in the sentence"黄芪产于内蒙古等地,为国家三级保护植物。", "产"and "为" are VV, "为" share the same subject.

Another noteworthy problem is, the patterns in Tab. 2 not using attribute relationship (ATT) and adverbial relationship (ADV), which will lead to that the entity and relationship of non-taxonomy relationships extracted are all the central word in the statement. For example, in the sentence"铁观音原产于福建安溪县西坪。", we use the pattern SBV (Y, X), HED (Root, Y), CMP (Y, Z), POB (Z, W)→Y_Z (X, W) to extract the pattern"产_于(铁观音,西坪)", the adverbial "原" of relationship is missing, which affect the accuracy of relationship name. Considering the need to raise the recall in summarizing the lexicon-syntactic patterns, we used the weak qualified pattern, and post-processing method to solve the conditions of lack of semantic extraction result. If there is a restrictive modifiers before entity and relationship names in non-taxonomy relationships extracted, it will be make up.

In addition, due to the text of Baidu Encyclopedia describe surrounds the terms, so many subject of the sentence is the default entity name. It results that the non-taxonomy relationship extracted lacks agent concept. In this paper, we treat the entry name as a default agent concept to solute the situation above.

### 1.1.4 Non-taxonomic relation extraction

The detailed process of extracting non-taxonomic relations on the basis of lexicon-syntactic patterns is as follows: first, LTP is used to preprocess the original document. Then, we convert the pattern matching issue to the problem of searching for a sub-tree in a D-tree (dependency tree). If the sub-tree nodes are in the tree and meet every constraint of their former parts, the pattern match is a success, and these nodes can be translated into non-taxonomic relations according to the latter part. Take the second line in Tab. 1 as an example, i. e., "铁观音原产于福建省安溪县西坪。" The sentence D-trees are shown in Fig. 2, where "铁观音" and "省" have a subject-predicate relation (SBV), the virtual node Root and "产于" have an HED relation, "产" and "于" have a verb-complement structure (CMP), and "于" and "西坪" have a preposition-object relation (POB). To meet the former part of the model, according to latter part of the model, we translate these nodes into non-taxonomic relations, e. g., product_ in (铁观音,西坪).

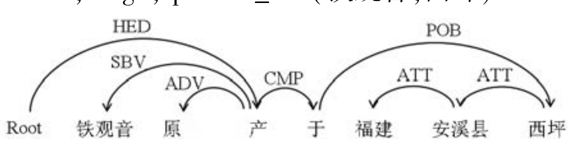


Fig. 2 An example of a D-tree

### 1.1.5 Add constraints for lexical-grammatical patterns

Only adopting an interdependent relation as the constraint condition of a model can lead to the model being set too broadly and some common semantic relations being extracted outside the plant field. To avoid this problem, this paper adds more constraints to the lexicon-syntactic patterns to improve the extracting accuracy.

(1) Filtering initial results

This paper adopts a word filter method, saving a relation only if the agency and patient concepts of the relation are all in the word list. This approach can greatly improve relation extraction accuracy.

(2) Adding constraints when manually setting patterns

The flow chart of the improved method is shown in Fig. 3.

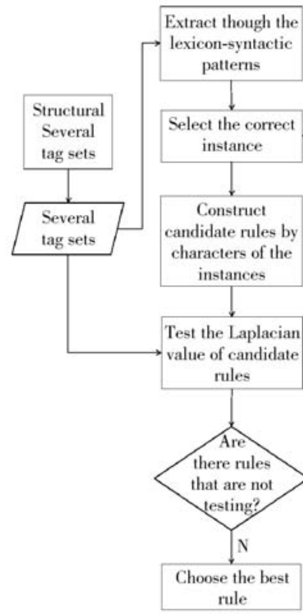


Fig. 3 Flow chart of the improved method

A single constraint is shown in Fig. 4. Several restriction combinations can be represented by "Constraint Combination" objects.

| Para ID | Sentence ID | Standard ID | Offset Related to Stardard | Confined Value |
|---|---|---|---|---|

Fig. 4 Data structure of a single constraint

By first using basic patterns such as SBV – HED – VOB to extract tagging sets preliminarily, for a successful extraction example, this paper transforms the agency concept, relation name, patient concept and their part of speech to restrictions, as in Fig. 4, and then calculates the Laplacian value of the restrictions of all subsets under the tagging sets. The Laplacian value is calculated as

$$\text{Laplacian} = \frac{e+1}{n+1} \tag{2}$$

where $e$ is the number of extraction errors and $n$ is the total number of extractions. The Laplacian is used to estimate the constraint; a low Laplacian value means that the restriction combination performs well on the tagging sets extraction.

## 1.2 Non-taxonomic relation extraction from semi-structured text

Apart from the unstructured free text, there is a large amount of semi-structured text in the Baidu Encyclopedia. The text is the knowledge that has been summarized and refined. Compared with the NLP (Natural Language Processing), our method reused the knowledge, though collecting the semi-structured text has the advantages of simplicity and high accuracy; thus, it is also an important approach to ontology learning. The Baidu Encyclopedia consists of entry names, cards, paragraph titles, entry text, body pictures, entry links, among other text. Among these models, the semi-structured information of the entries that can be used for non-taxonomic relation extraction exists in the tables.

Most of the entry body text is free text with a semi-structured pattern, but some of it will also express knowledge with the table at the same time. Tab. 3 summarizes the nutrients of Solanum tuberdsm. The entry name "Solanum tuberdsm" demonstrates a non-taxonomic relation of "nutrient" with "anthocyanin", "selenium", "moisture", and "inorganic salt", among others.

**Tab. 3　Nutrients of *Solanum tuberdsm* in Baidu encyclopedia (per 100 g)**

| Composition | Content |
|---|---|
| Anthocyanin/g | 0. 01 |
| Selenium/mg | 0. 02 |
| Moisture/g | 9. 9 |
| Axunge/g | 0. 2 |
| Inorganic Salt/g | 0. 68 |
| Protein/g | 4. 768 |
| Vc/mg | 28. 4 |
| Carbohydrate/g | 82. 5 |
| Cellulose/g | 2. 7 |

## 1. 3　Formalizing extraction results

This paper expresses the extraction triples in OWL (Ontology Web Language) and formalizes the extraction results with the help of a Protégé visualization plug-in OWL Prop Viz. Part of the visual results is shown in Fig. 5.

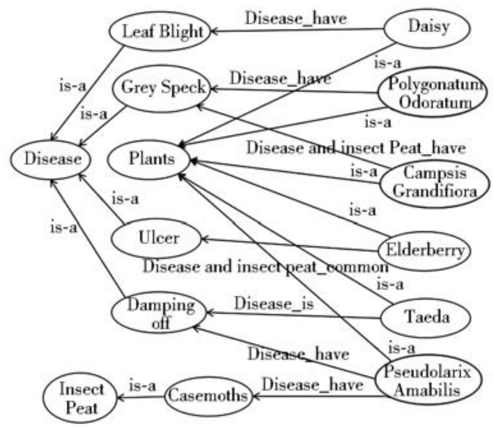


Fig. 5　Partial results of visualization

# 2　Discussion

In this research, we select 70 sentences containing plant distribution relations and 31 sentences containing plant diseases and pests from the corpuses. Then, we test the performance of the non-taxonomic relation extraction methods. As a baseline for the extraction methods, the results based on manually setting patterns are shown in Tab. 4. The SBV－HED－VOB pattern is mainly used to extract plant diseases and pests relations, whereas the SBV－HED－CMP－POB pattern is primarily used to extract distribution relations.

**Tab. 4　Evaluation results based on manually setting patterns**

| Models | Total extraction | Correct number | Accuracy/% | Extraction apart from the LTP errors | Accuracy apart from the LTP errors/% |
|---|---|---|---|---|---|
| SBV－HED－VOB | 24 | 18 | 75. 0 | 20 | 90. 0 |
| SBV－HED－CMP－POB | 22 | 15 | 138. 2 | 16 | 93. 8 |
| SBV－HED－VOB (Table Filter) | 18 | 18 | 100 | N/A | N/A |

Tab. 4 shows that based on LTP preprocessing, the extraction accuracy using the lexicon-syntactic patterns method can reach 70%, which indicates much room for improvement. This work adopts the word filtering method as an improvement.

Another method to improve the model accuracy is to add constraints to the lexicon-syntactic patterns. The 31 sentences that contain plant diseases and pests relations can be divided into two parts; the first 15 sentences are used to calculate the Laplacian value of constraints and then select constraints with the smallest values from these sentences; patterns and constraints are combined together to extract the relations from the other 16 sentences. We attempt to choose constraints with the lowest Laplacian value because the 5 constraints have the same Laplacian. Then, according to the heuristic strategy, we should choose as few restrictions as possible, and the constraints should be split into different objects. Finally, the results obtained for testing accuracy are presented in Tab. 5.

**Tab. 5 Extraction after adding restriction**

| Constraints with the Lowest Laplacian Value | Laplacian | Total Extraction | Correct number | Accuracy/% |
|---|---|---|---|---|
| 2,0,2,0,0,Disease | 0.125 | 11 | 10 | 90.9 |
| 2,0,0,0,1,n<br>2,0,2,0,0,Disease | 0.125 | 11 | 10 | 90.9 |
| 2,0,0,0,0,Harm<br>2,0,2,0,0,Disease | 0.125 | 9 | 9 | 100 |

As shown in Tab. 5, because the method makes full use of the language features in addition to the D-tree by adding constraints to the models, the results obtained by using the adding constraints method are much better than those obtained when using the manual pattern-setting method.

Over the past several years, there have been some studies on non-taxonomic relation learning at home and abroad, such as the OntoCmaps tools by Canada ZOUAQ[21] and Chinese ontology non-taxonomic relation extraction methods developed by GU[22]. To a certain extent, their studies represent the current level of non-taxonomic relation learning, the performances of which are shown in Tab. 6.

**Tab. 6 Related research extraction status**

| Correlational research | Average accuracy/% | Corpus |
|---|---|---|
| ZOUAQ[21] | 28.5 ~ 80.8 | SCORM, AI |
| GU[22] | 64.47 | Corpus of text classification, FDU (Fudan University) |

In OntoCmaps, the manually lexical-grammatical patterns that ZOUAQ took had an undulating performance in different corpuses; the average accuracies of the maximum and minimum values are shown in Tab. 6. In a word, the accuracies of our experiments and the manually setting model are the same overall. ZOUAQ also noted that the process of filtering the results after extracting non-taxonomic relations by a model was necessary. GU Linglan used the semantic role labeling function of LTP in his extraction method. Nevertheless, because of the LTP performance issues and the lack of filter performance of the results, the extraction accuracy is not sufficiently high.

Synthesizing the experiment's results, this paper manually sets lexicon-syntactic patterns as the basic method for non-taxonomic relation extraction and has achieved approximately equal accuracy with the same type of method. Thus, using the methods based on word list filtering and adding restrictions to patterns greatly improve the accuracy of relation extraction.

## 3 Conclusions

(1) The non-taxonomic relation extraction in this paper greatly retains its original state; relations with the same meaning can be clustered as one relation by using a semantic similarity clustering algorithm in the next work.

(2) Considering the applicable scope of the LTP Chinese natural language processing tools, their performance on Chinese Wiki text has decreased. This result suggests that the current technology in the relatively deep natural language processing, such as interdependence syntactic analysis, still needs development before its complete practical use. Making full use of structured and half-structured data and reusing the purified knowledge constitute a more practical and feasible method of ontology learning.

**References**

[1] WANG Haofen. Large Scale Knowledge Graph Technology. Communications of the CCF, 2014, 10(3): 64-68.

[2] DESHPANDE O, LAMBA D S, TOURN T, et al. Building, maintaining, and using knowledge bases: a report from the trenches[C]//2013 SIGMOD'13, 2013: 1209-1220.

[3] CHENG Tongling, LI Juanzi. Entity relationship discovery and semantic annotation based on the wikipedia encyclopedia of knowledge resources[J]. Electronic Technology and Software Engineering, 2015(18): 170-173.

[4] MAEDCHE A, STAAB S. Ontology learning for the semantic web[J]. IEEE, Intelligent Systems, 2001, 16(2): 72-79.

[5] WONG W, LIU W, BENNAMOUN M. Ontology learning from text: a look back and into the future[J]. Acm Computing Surveys, 2012, 44(4): 1-36.

[6] LIAO Fuyan. Research on domain ontology concept and relation acquisition[D]. Xi'an: Xi'an University of

Architecture and Technology,2011. (in Chinese)

[7] GU Jun, YAN Ming, WANG Hao. Research on ontology relation extraction based on improved association rule [J]. Information Studies, 2011, 34(12):121 - 125. (in Chinese)

[8] SHU Wanli. Research on concept and relation extraction of Chinese domain ontology[D]. Chongqing: Chongqing University, 2012. (in Chinese)

[9] HU Yunfei. Research on relations acquisition of ontology learning[D]. Xi'an: Xi'an University of Architecture and Technology, 2012. (in Chinese)

[10] QIU T R, HUANG H Q, DUAN W Y, et al. Research on granular computing model for non-taxonomic relations learning[J]. Journal of Nanchang University, 2012,34(3):273 -278. (in Chinese)

[11] LIANG Jizhen. Research on non-taxonomic relationships learning based on domain concept knowledge [D]. Changchun: Jilin University, 2012. (in Chinese)

[12] WEICHSELBRAUN A, WOHLGENANNT G, SCHARL A. Refining non-taxonomic relation labels with external structured data to support ontology learning[J]. Data & Knowledge Engineering, 2010, 69(8):763 -778.

[13] XIANG Yang, ZHANG Bo, HAN Jie. Agent driven intelligent construction of Chinese ontology [J]. Computer Engineering and Appfication, 2009, 45(10): 133 -137. (in Chinese)

[14] YE Qiong. Research on cloudization method of agricultural ontology knowledge [D]. Hefei: Anhui Agricultural University, 2012. (in Chinese)

[15] DENG Ziping. Research and development of a ontology automatic generation system oriented medical diagnosis [D]. Guangzhou: Guangdong University of Technology, 2011. (in Chinese)

[16] MA Li, CHEN Zhixin. Domain ontology learning mehtod based on structure of the site [J]. Computer CD Software and Applications, 2014 (16): 83, 85. (in Chinese)

[17] WANG Hong, GAO Siting, PAN Zhenjie, et al. Application and research of non-taxonimic relation extraction method based on NNV association rule[J]. Application Research of Computers, 2012, 29(10): 3665 -3668. (in Chinese)

[18] SÁNCHEZ D, MORENO A. Learning non-taxonomic relationships from web documents for domain ontology construction [J]. Data & Knowledge Engineering, 2008, 63(3):600 -623.

[19] SERRA I, GIRARDI R, NOVAIS P. Evaluating techniques for learning non-taxonomic relationships of ontologies from text [J]. Expert Systems with Applications, 2014, 41(11):5201 -5211.

[20] CHE W, LI Z, LIU T. LTP: a Chinese language technology platform [C] // Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, 2010:13 -16.

[21] ZOUAQ A, GASEVIC D, HATALA M. Linguistic patterns for information extraction in OntoCmaps[C] // Proceedings of the 3rd Workshop on Ontology Patterns, 2012:1 -12.

[22] GU Linglan, SUN Suyun. Approach to Chinese ontology non-taxonomic relation extraction based on semantic dependency[J]. Computer Engineering and Design, 2012, 33(4):1676 -1680. (in Chinese)

doi:10.6041/j.issn.1000-1298.2016.09.038

# 植物领域知识图谱构建中本体非分类关系提取方法

赵　明　杜亚茹　杜会芳　张家军　王红说　陈　瑛

（中国农业大学信息与电气工程学院，北京 100083）

**摘要：** 采用本体学习的方法，以百度百科植物类词条内容的非结构和半结构化中文文本信息作为语料进行处理。使用一种有指导的基于依存句法分析的词汇-语法模式来获取植物领域的概念、分类和非分类关系，并分别利用基于词表过滤的方法和给模式添加限制的方法，较大程度地提高了关系抽取的精确度，完成在轻量级本体的基础上自动构建重量级本体。该方法建立了一个特定领域语料的概念层次，提高了最具代表性的分类和非分类关系的发现，并使用 OWL 语言形式化表达抽取结果。实验表明，该方法在非分类关系抽取上取得了较好的结果，为该领域知识图谱构建奠定了基础。

**关键词：** 植物领域本体；知识图谱；非分类关系；本体学习；百度百科

**中图分类号：** TP391　　**文献标识码：** A　　**文章编号：** 1000-1298(2016)09-0278-07

# Research on Ontology Non-taxonomic Relations Extraction in Plant Domain Knowledge Graph Construction

Zhao Ming　Du Yaru　Du Huifang　Zhang Jiajun　Wang Hongshuo　Chen Ying

(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

**Abstract:** In order to provide more specific knowledge and technology of plant field, the main task of KG (knowledge graph) is to extract a wealth of concepts and relationships. Due to the relation extraction is the most difficult in KG construction, this paper makes use of ontology learning, and proposes a non-taxonomic relation learning method to obtain representative concepts and their relations from unstructured and semi-structured texts of Baidu Encyclopedia entry content by using lexicon-syntactic patterns based on dependency grammar analysis. Moreover, the methods of adding constraint models and words filtering were adopted to build heavy weight ontology automatically based on a lightweight ontology and greatly improved the precision of the relation extraction. The approach established a concept structure from the plant domain corpus, ameliorated the discovery of the most representative non-taxonomic relation, and formalized them in the standardized OWL 2.0. A set of experiments was performed using the approach implemented in the plant domain. The results indicated that extraction by patterns should be performed directly after natural language processing, which has a comparatively high accuracy compared to the former algorithms, and this approach can extract non-taxonomic relations with high effectiveness, which lays the foundation for KG construction of plant field.

**Key words:** plant domain ontology; knowledge graph; non-taxonomic relation; ontology learning; Baidu Encyclopedia

## 引言

"知识图谱"[1]作为一种知识表示方法，包含了大量概念（实体）以及概念间的分类和非分类关系，使其成为具有语义性的知识库。它支撑综合性知识检索、智能问答、智能决策等方面的广泛应用。知识

收稿日期：2016-03-09　修回日期：2016-05-15

基金项目：国家自然科学基金项目（61503386）

作者简介：赵明（1963—），男，副教授，主要从事知识工程与计算机图形学研究，E-mail：zhaoming@cau.edu.cn

通信作者：陈瑛（1978—），女，副教授，主要从事自然语言处理研究，E-mail：chenying@cau.edu.cn

图谱构建的主要任务是抽取丰富的概念和关系，其中概念间非分类关系抽取是构建知识图谱的难点问题。而本体（Ontology）作为构建知识图谱的概念模型和逻辑基础，尤其是概念之间的非分类关系，不仅增加了知识表达的完备性和复杂性，还在很大程度上对知识图谱进行了丰富和扩展[2]。

本体[3]是用于描述或表达某一领域中术语、概念以及之间关系的一个基本知识体系，是共享概念模型的形式化的明确说明。基于本体的应用和解决方案是在本体构建的基础上实现的。然而，现有的各种本体开发工具（如 Protégé、KAON 等）支持的是手工构建本体的方式，存在费时、费力且不易及时动态更新等问题。因此如何快速、低成本、自动或半自动构建大规模领域本体——即所谓的本体学习（Ontology learning）就成为一个很有意义的研究方向[4-5]。目前，在本体学习领域抽取非分类关系的方法主要有：基于词典的方法、基于关联规则的方法和基于模式匹配的方法。

基于词典的方法抽取出的关系必须是 WordNet 中已经存在的关系，并仅能从 WordNet 中抽取同义、反义和部分/整体这几种关系，局限性比较大。基于关联规则的方法研究中，研究者分别设置不同的参数，诸如概念之间的关联度阈值[6]、术语之间相似度权重[7]、对数似然比[8-9]以及粒计算模型[10]等进行非分类关系抽取。但是该类方法中，由于词性标注的错误对准确率有较大的影响，因此还要探索更为有效的方法。

基于模式匹配的方法[11]是通过分析领域相关文本，归纳出频繁使用的语言模式，来识别相应的语义关系。研究者均采用语义依存[1]的本体非分类关系抽取方法，经过词性标注、角色标注[3]和语义分析得到具有语义依存的动词框架，构建动词向量语料库[12]，再进行句子相似度的计算。还有一部分研究者提出一种综合语义模式和统计学的启发式非分类关系学习框架[11,13]。该类方法中，抽取精度较高，但是有些中心词的选取会对非分类关系的选取造成影响。另外，非分类关系复杂多样，不仅仅包含术语和术语之间的关系，还包含一些诸如产地、病虫害等属性关系，此类关系更为复杂且丰富，以上方法没有针对性，因此提取效果较差。以上方法由于中文概念的多义性对非分类关系种类的影响等原因，该方法的准确率和召回率在领域本体构建中还可以进一步优化。

以上非分类关系抽取的研究方法所涉及的领域包括农业[14]、医学诊疗[15]、网站[16]、航空管理等[17]。还有一部分研究者利用非监督的本体学习方法[18]从 Web 文档提取非分类关系，并对该方法进行评测[19]。但是中文本体的自动构建，尤其是农业领域非分类关系的学习研究才刚刚起步。因此本文针对基于模式匹配方法提取非分类关系的局限，以百度百科植物词条内容为语料，采用词汇-语法模式进行非分类关系的抽取，利用停用词表过滤和为模式增加限制的方法对该方法进行改进，并且对非分类关系的类别进行讨论，开展基于 Web 的中文植物本体非分类关系学习方法的研究。

## 1　基于词汇-语法模式的非分类关系抽取

针对基于百度百科植物领域本体学习中非分类关系的学习方法，采用的主要技术和步骤如下：

（1）抓取相关词条的网页内容，以此为语料，对语料进行预处理，使用改进的词汇-语法模式进行非分类关系的自动抽取，提升抽取的准确度。

（2）抽取百度百科半结构化文本中的非分类关系。

（3）在抽取出关系三元组的基础上，使用 OWL 语言将其形式化。

总体流程如图 1 所示。

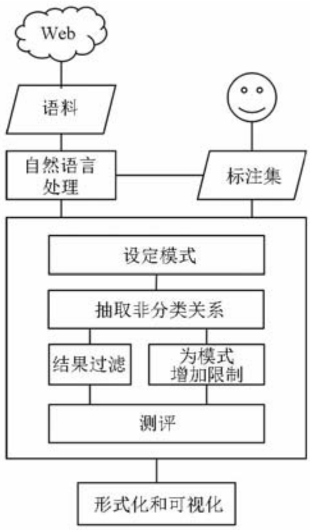


图 1　技术流程图

Fig. 1　Flow chart of technological process

### 1.1　基于词汇-语法模式的非分类关系抽取

#### 1.1.1　数据的获取与预处理

利用 Python 编写的语料采集工具从百度百科的植物分类下抓取了 9 623 个词条作为语料。这些语料用 GBK 编码的文本文件形式存储。为解决百度百科词条正文非结构化文本难以直接利用的问题，借助自然语言处理的开源工具 LTP（Language technology platform）[20]，主要使用分词、词性标注和依存句法分析这 3 个模块对语料进行预处理，得到的结果以 XML 的形式保存。

### 1.1.2 模式获取

从百度百科植物语料中选取一小批高质量并具有代表性的词条(在 19 个分类下共选取了 97 个词条),自动找出典型的表达非分类关系的语句(共 339 句),该类语句均含有农业概念,且主要集中在具有农业价值的非分类关系上,如植物的地理分布、适生环境、病虫害、经济价值等,表 1 列举了其中部分语句。

**表 1 部分语句**

**Tab.1 Some of statements**

| 所属分类 | 词条名称 | 语句 |
|---|---|---|
| 蔬菜 | 茄子 | 利用大棚茄子和晚水荸荠进行水旱轮作,可以解决连作障碍问题 |
| 茶叶 | 铁观音 | 铁观音原产于福建安溪县西坪 |
| 花卉 | 石斛兰 | 虫害有介壳虫危害,用 40% 氧化乐果乳油 2 000 倍液喷杀 |
| 农作物 | 紫薯 | 除了具有普通红薯的营养成分外,还富含硒元素和花青素 |

模式提取的任务实际上也是一种序列标注问题,解决该任务一般采用有监督学习方式。令 $\{(x_i, y_i)\}(i=1,2,\cdots,N)$ 为一个大小为 $N$ 的训练样本集。每个序列对 $(x_i, y_i)$ 为一个样本,其中 $x_i = <x_i,1,x_i,2,\cdots,x_i,n>$,$y_i = <y_i,1,y_i,2,\cdots,y_i,n>$。序列标注的目标就是寻找一个模型 $M$,可以在给定输入序列 $x$ 的基础上预测标注序列 $y$。

另外,考虑到植物领域词汇的特殊性,其概念词汇出现的上下文中往往包含特定的词汇,例如:又名、又叫、轮作、套种、有、富含、含有、栽培、病害、虫害等。因此在植物词汇出现的语句中,分别在浅层和深层依存句法分析结果进行标注,即找出并标明句子中存在非分类关系的概念,以及非分类关系的名称,记为 $R_i(X_i, Y_i)$。然后统计 $X_i$、$R_i$、$Y_i$之间的依存关系序列在其他标注 $X_j$、$R_j$、$Y_j$$(i\neq j)$中出现的次数。其中,浅层句法分析通过正则表达式匹配的方法获取非分类关系,深层句法分析通过句法依赖分析匹配模式的方法获得关系。再借助相似度计算等机器学习方法统计训练语料中出现频率较高的词语构成高频词表,进行多次迭代不断泛化来获取表达非分类关系的词汇-语法模式,剔除不符合语法的序列后,得到出现多次的依存关系序列即为表达非分类关系的词汇-语法模式,如表 2 所示(依存句法标注含义见 LTP 官方文档)。

**表 2 词汇-语法模式集**

**Tab.2 Set of lexicon - syntactic patterns**

| 模式 | 实例 |
|---|---|
| SBV(Y, X), HED(Root, Y), VOB(Y, Z)→Y(X, Z) | 主要虫害有介壳虫和斑蛾。→有(虫害,介壳虫) |
| SBV(Y, X), HED(Root, Y), COO(W, Z)→Y(X, Z) | 主要虫害有介壳虫和斑蛾。→有(虫害,斑蛾) |
| SBV(Y, X), HED(Root, Y), CMP(Y, Z), POB(Z, W)→Y_Z(X, W) | 铁观音原产于福建安溪县西坪。→产_于(铁观音,西坪) |
| VOB(X, Y), DE(Z, X), ATT(W, Z)→X(W, Y) | 松毛虫是危害马尾松的常见害虫。→危害(害虫,马尾松) |

其中,模式采用依存关系(头节点,依存节点)→非分类关系名称(施事概念,受事概念)的形式表示。可以形式化为

$$\bigwedge_{i=1}^{n} D_i(s_i, t_i) \Rightarrow L(A,B) \qquad (1)$$

其中 $L$、$A$、$B$ 为 $s_i$和 $t_i$中指定的元素,式(1)表示当所有依存关系 $D_i$都满足时,则 $L(A,B)$关系成立。

### 1.1.3 为模式增加规则

对于抽取的词汇-语法模式,一个值得注意的问题是有些模式含有并列关系(COO)和连动结构(VV):对于并列关系,如果某个节点的依存关系为 COO 的话,则该节点视为等同于依存关系所指向的节点,并且该性质具有传递性。例如,在句子“主要虫害有介壳虫和斑蛾。”中,将“介壳虫”与“斑蛾”视为等同;对于连动结构,如果某个节点的依存关系为 VV 的话,则该节点依存关系所指向的节点可视为与该节点共享主语,即 VV(X, Y), SBV(X, Z)→SBV(Y, Z),并且该性质具有传递性。例如,在句子“黄芪产于内蒙古等地,为国家三级保护植物。”中,“产”与“为”是连动结构,“为”共享“产”的主语“黄芪”。

另一个值得注意的问题是,表 2 中的模式没有利用定中关系(ATT)和状中结构(ADV),这样会导致抽取出来的非分类关系的概念和关系名称都是语句中相应成分的中心词,从而丢失了部分语义,例如,“铁观音原产于福建安溪县西坪。”这句话,利用模式 SBV(Y, X), HED(Root, Y), CMP(Y, Z), POB(Z, W)→Y_Z(X, W)会抽取出“产_于(铁观音,西坪)”,关系名称丢失了“原产于”中的状语“原”,也影响了关系名称的精确性。本文在归纳总结词汇-语法模式时,考虑到要尽可能提高模式的召回率,采用了这种弱限定的模式,对于上述抽取结果语义缺失的情况,本文采用后处理的方法,如果抽取出的非分类关系的概念和关系名称在语句中前面有限定性修饰成分的话,则会补上。

另外，由于百度百科词条的文本围绕着词条进行阐述说明，故许多句子的主语默认就是词条名称，从而被缺省。这种情况对非分类关系抽取的影响是导致抽取出来的非分类关系缺少施事概念，本文的处理方法就是将词条名称作为默认的施事概念。

#### 1.1.4　非分类关系抽取

在所获词汇-语法模式的基础上，可以进行非分类关系的提取。具体过程为：对待提取的文档利用 LTP 进行分段、分句、分词、依存句法分析；然后将模式匹配归结为在依存树中寻找子树的问题，即对句子的依存树进行检测，如果树中存在这些节点，其满足模式前件中的每项约束，则模式匹配成功；并将这些节点按照模式的后件转换为非分类关系。以表 2 中第 3 行为例，“铁观音原产于福建安溪县西坪。”这句话的依存树如图 2 所示，因为“铁观音”与“产”为主谓关系（SBV），虚拟节点 Root 与“产”为 HED 关系，“产”与“于”为动补结构（CMP），“于”与“西坪”为介宾关系（POB），所以满足了该模式的前件，根据模式的后件，将这些节点转为非分类关系：产_于（铁观音，西坪）。

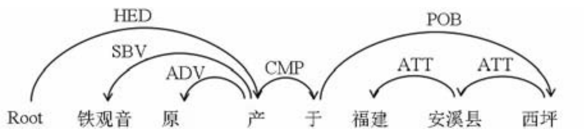


图 2　“铁观音原产于福建安溪县西坪”的依存树

Fig. 2　An example of a D-tree

另外，当利用 SBV - HED - VOB 模式对句子进行模式匹配时，将句子中依存关系名称为“SBV”、“HED”、“VOB”的词组成三元组，然后再去判断这些三元组内部的依存关系是否和模式中的完全一致，当一致时，则匹配成功，否则失败。

#### 1.1.5　对词汇-语法模式添加约束

（1）对结果进行过滤

针对模式过于宽泛而影响准确率的问题，为避免抽取出一些非农业领域的通用语义关系，采用词表过滤的方式，当关系的施事概念、受事概念都属于植物领域相关概念时，保留该关系，其余关系则为错误关系。这种方式可以大幅提高关系抽取的准确度。

（2）对词汇-语法模式增加限制

对上述词汇-语法模式进行改进，添加更多的约束，以提高模式的准确度。改进方法的流程如图 3 所示。

单个限制用如图 4 所示的方式表示。可以用 Constraint Combination 对象表示多个限制的组合。

首先用基本模式，如 SBV - HED - VOB 在标注集上进行初步抽取，对于抽取成功的实例，本文将施

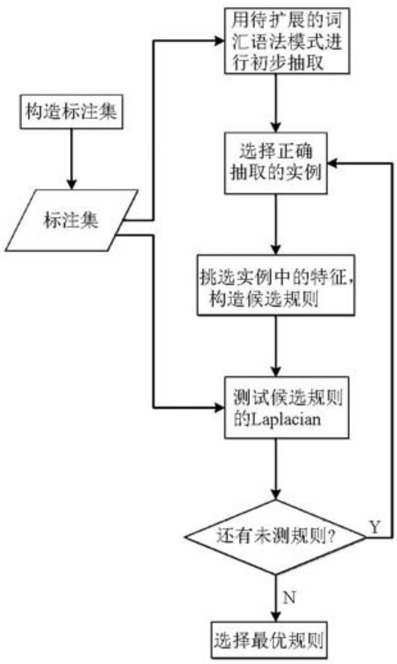


图 3　改进方法的流程图

Fig. 3　Flow chart of the improved method

| 段落id | 句子id | 基准id | 相对于基准的偏移量 | 限制的值 |
|---|---|---|---|---|

图 4　单个限制的数据结构

Fig. 4　Data structure of a single constraint

事概念、关系名称、受事概念的词性和词中的字用图 4 中的方式表示为限制（限制值均为 XML 处理过程中对文本的标记。其中，段落 id 和句子 id 分别为实例所在段落和句子的位置；基准 id 为依存父节点在句子中的位置；相对于基准的偏移量为抽取实例与依存父节点位置差值；限制的值为模式抽取的正确实例所含的特征），并计算这些限制的所有子集在标注集上的 Laplacian 值，Laplacian 因子计算公式为

$$V_{\text{Laplacian}} = \frac{e+1}{n+1} \tag{2}$$

式中　$e$——抽取的错误数

$n$——抽取的总数

Laplacian 用来估计所添加的约束的性能好坏，Laplacian 值低的表示该限制组合在标注集上抽取表现良好。

### 1.2　基于百度百科半结构化文本的非分类关系抽取

由于对模式增加限制会对抽取非分类关系的召回率造成影响，本文除了从非结构化文本信息中进行关系提取，还有效利用百度百科中大量的半结构化文本，这些半结构化文本通常是经过总结和提纯过的知识，相较于自然语言处理的方式，这种通过收割半结构化文本的知识重用方式不仅可以抽取非分类关系作为补充来提高召回率，还具有简单易行、准

确度高的优点,因而也是本体学习的一个重要途径。

百度百科的词条结构包括:词条名称、百科名片、段落标题和目录、词条正文等。可用于抽取非分类关系的半结构化信息存在于百科名片和词条正文的表格中。

百科名片中包含了大量的植物别名、产地等属性信息,该类信息也为植物领域非分类关系的一部分,且清晰易得。百度百科词条正文中的大部分都是非结构化的自由文本,但是有的词条正文中也会同时采用表格的方式表示知识。

本文利用百度百科词条半结构化信息的方法为:对网页的 DOM 树进行解析处理,查找出其中表示百科名片中词条基本信息栏表格和正文表格的 HTML 标签,获取表格中"分布区域"等表示非分类关系的表项,然后将其转换为对应的非分类关系三元组,其中,关系名称为表格中表项的名称,施事概念为词条名称,受事概念为表项的内容。因为表达受事概念的表项内容通常为一些简单的自然语言语句,因此需先对其进行 LTP 分词、命名实体识别等处理。例如表 3 为百度百科词条"紫薯"的正文中的一张表格,其中含有较多的表示营养成分的概念,因此触发了非分类关系的生成,这些概念均处于表示营养成分主题的列表中,故非分类关系的名称取"营养成分"。

表 3 百度百科中紫薯的营养成分

Tab. 3 Nutrients of *Solanum tuberdsm* in Baidu encyclopedia

| 成分 | 含量 |
|---|---|
| 花青素 | 0.01 g |
| 硒元素 | 0.02 mg |
| 水分 | 9.9 g |
| 脂肪 | 0.2 g |
| 无机盐 | 0.68 g |
| 蛋白质 | 4.768 g |
| 维生素 C | 28.4 mg |
| 碳水化合物 | 82.5 g |
| 纤维素 | 2.7 g |

注:各成分含量为每 100 g 中的含量。

### 1.3 抽取结果的形式化

将抽取的关系三元组用 OWL(Web ontology language)语言进行形式化,并借助 Protégé 的可视化插件 OWLPropViz 进行抽取结果的显示。对应的可视化结果如图 5 所示。

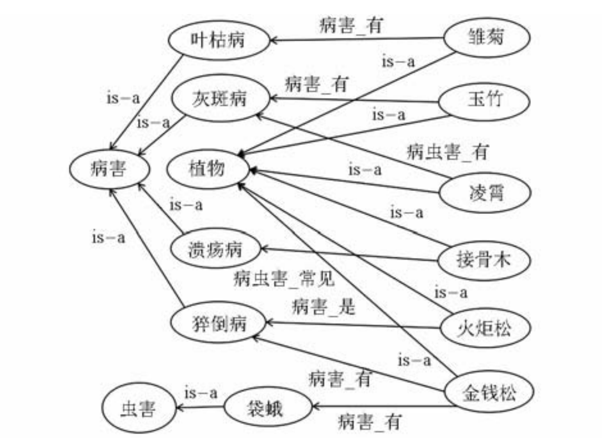


图 5 部分结果的可视化结果

Fig. 5 Partial results of visualization

## 2 实验与结果分析

本研究从收集的语料中挑选了 70 个含有植物分布区域关系的句子和 31 个含有植物病虫害关系的句子,测试上文提出的非分类关系抽取方法的性能。作为抽取方法的 baseline,基本模式的抽取情况如表 4 所示。其中 SBV - HED - VOB 模式主要用于病虫害关系的抽取,SBV - HED - CMP - POB 模式主要用于分布区域关系的抽取。

表 4 基本模式的抽取情况

Tab. 4 Extraction of basic pattern

| 模式 | 抽取总数 | 正确数 | 准确率/% | 除去 LTP 错误后抽取数 | 除去 LTP 错误后准确率/% |
|---|---|---|---|---|---|
| SBV - HED - VOB | 24 | 18 | 75.0 | 20 | 90.0 |
| SBV - HED - CMP - POB | 22 | 15 | 68.2 | 16 | 93.8 |
| SBV - HED - VOB(词表过滤) | 18 | 18 | 100 | N/A | N/A |

注:表中 N/A 表示不可用。

从表 4 中可以看出,在利用 LTP 对文本进行自然语言处理结果的基础上,直接用模式进行抽取,准确率在 70% 左右,还有较大的提升空间。一方面,LTP 对于长难句和语言风格偏科研文献句子的处理结果都会有一些问题。因此,本文把 LTP 处理有错误的句子进行过滤,模式抽取准确率大有提升。另一方面,针对模式过于宽泛而影响准确率的问题,本文采用词表过滤的方法作为对模式的一种改进。如表 4 所示,过滤后的结果具有较高的准确率。

另一种提高模式准确率的改进方法是为其增加限制。将 31 个含有植物病虫害关系的句子分为 2 部分,前 15 个句子用来计算限制的 Laplacian 因子,

并据此挑选在其上表现最优的限制；然后将模式和限制合在一起，用来抽取后 16 个句子中的病虫害关系。这时根据启发式的策略选择拥有最低 Laplacian 值的限制，测试语句上的准确率如表 5 所示。

**表 5　加上限制后模式的抽取情况**

**Tab. 5　Extraction after adding restriction**

| 有最低 Laplacian 值的限制 | Laplacian 值 | 抽取总数 | 正确数 | 准确率/% |
|---|---|---|---|---|
| 2,0,2,0,病 | 0.125 | 11 | 10 | 90.9 |
| 2,0,2,0,病 | 0.125 | 11 | 10 | 90.9 |
| 2,0,0,0,害 | 0.125 | 9 | 9 | 100 |

可以看出，通过对模式增加限制，可更充分地利用除依存关系外的其他语言特征，从而提高了模式的抽取准确度。

近几年内，国内外本体学习界也有一些非分类关系学习的研究，例如加拿大 ZOUAQ 等[21] 研究的 OntoCmaps 工具，以及国内的古凌岚等[22] 提出的中文本体非分类关系抽取方法，这些研究在一定程度上代表了当前本体非分类关系学习的水平，非分类关系抽取方法的性能如表 6 所示。

**表 6　相关研究的抽取情况**

**Tab. 6　Extraction condition of related study**

| 文献序号 | 平均准确率/% | 采用语料 |
|---|---|---|
| [21] | 28.5 ~ 80.8 | SCORM、AI |
| [22] | 64.47 | 复旦大学文本分类语料库 |

ZOUAQ 等在 OntoCmaps 中采用的是手工总结的词汇-语法模式，其在论文中的测评结果显示，该方法在不同语料上的性能波动较大，故表 6 中给出了其平均准确率的最大和最小值，总的来说，该结果和本文手工设定的模式准确率基本一致，但二者均由于 LTP 性能问题以及其方法缺少对结果的过滤功能，导致抽取的准确率偏低。

综合实验结果来看，本文用改进的词汇-语法模式作为非分类关系提取的基本方法，已经能够达到与同类方法相当的准确度。在此基础上，分别利用基于词表过滤的方法和给模式添加限制的方法，较大程度地提高了关系抽取的准确度。

## 3　结论

（1）以网络百科为知识来源，在信息提取、自然语言处理等领域的一些关系抽取方法基础上，提出了一种改进的基于词汇-语法模式的中文非分类关系抽取办法，实验表明其有效地提高了模式的准确度，取得了预期的效果。

（2）本文的研究属于尝试性的探索工作，测试的语句集还较少，未来还有一些需要改进的地方和可能的研究方向：鉴于 LTP 等中文自然语言处理工具在百度百科的文本上性能有所下降，说明目前的技术在依存句法分析等相对深层的自然语言处理中与完全实用尚有一定的距离。本体是知识图谱表示的概念模型和逻辑基础，在本体非分类关系抽取的基础上，进行实体和关系的映射，构建完整的植物领域知识图谱还需要进一步完善。

## 参　考　文　献


1　王昊奋．大规模知识图谱技术[EB/OL]．(2014－06－12)http://www.China-cloud.com/zhongyunxy/20140612_38070.html.

2　DESHPANDE O, LAMBA D S, TOURN T, et al. Building, maintaining, and using knowledge bases: a report from the trenches[C]//2013 SIGMOD'13, 2013:1209－1220.

3　程童凌，李娟子．基于维基类百科知识资源的实体关系发现和语标注[J]．电子技术与软件工程，2015(18):170－173.

4　MAEDCHE A, STAAB S. Ontology learning for the semantic web[J]. IEEE, Intelligent Systems, 2001, 16(2):72－79.

5　WONG W, LIU W, BENNAMOUN M. Ontology learning from text: a look back and into the future[J]. Acm Computing Surveys, 2012, 44(4):1－36.

6　廖福燕．本体构建中概念和关系获取方法研究[D]．西安：西安建筑科技大学，2011.
LIAO Fuyan. Research on domain ontology concept and relation acquisition[D]. Xi'an: Xi'an University of Architecture and Technology, 2011. (in Chinese)

7　谷俊，严明，王昊．基于改进关联规则的本体关系获取研究[J]．情报理论与实践，2011，34(12):121－125.
GU Jun, YAN Ming, WANG Hao. Research on ontology relation extraction based on improved association rule[J]. Information Studies, 2011, 34(12):121－125. (in Chinese)

8　舒万里．中文领域本体学习中概念和关系抽取的研究[D]．重庆：重庆大学，2012.
SHU Wanli. Research on concept and relation extraction of Chinese domain ontology[D]. Chongqing: Chongqing University, 2012. (in Chinese)

9　胡云飞．本体学习中关系获取的研究[D]．西安：西安建筑科技大学，2012.
HU Yunfei. Research on relations acquisition of ontology learning[D]. Xi'an: Xi'an University of Architecture and Technology, 2012. (in Chinese)

10　邱桃荣，黄海泉，段文影，等．非分类关系学习的粒计算模型研究[J]．南昌大学学报：工科版，2012，34(3):273－278.
QIU T R, HUANG H Q, DUAN W Y, et al. Research on granular computing model for non-taxonomic relations learning[J].

Journal of Nanchang University, 2012,34(3):273 – 278. (in Chinese)
11 梁吉震. 基于领域概念知识的非分类关系学习研究[D]. 长春:吉林大学, 2012.
LIANG Jizhen. Research on non-taxonomic relationships learning based on domain concept knowledge[D]. Changchun: Jilin University, 2012. (in Chinese)
12 WEICHSELBRAUN A, WOHLGENANNT G, SCHARL A. Refining non-taxonomic relation labels with external structured data to support ontology learning[J]. Data & Knowledge Engineering, 2010, 69(8):763 – 778.
13 向阳, 张波, 韩婕. Agent 驱动的中文本体智能构建研究[J]. 计算机工程与应用, 2009, 45(10):133 – 137.
XIANG Yang, ZHANG Bo, HAN Jie. Agent driven intelligent construction of Chinese ontology[J]. Computer Engineering and Appfication, 2009, 45(10):133 – 137. (in Chinese)
14 叶琼. 农业领域本体知识云化方法研究[D]. 合肥:安徽农业大学, 2012.
YE Qiong. Research on cloudization method of agricultural ontology knowledge[D]. Hefei: Anhui Agricultural University, 2012. (in Chinese)
15 邓子平. 面向医学诊疗的本体自动生成系统的研究与开发[D]. 广州:广东工业大学, 2011.
DENG Ziping. Research and development of a ontology automatic generation system oriented medical diagnosis[D]. Guangzhou: Guangdong University of Technology,2011. (in Chinese)
16 马莉,陈志新. 基于网站结构的领域本体学习方法[J]. 计算机光盘软件与应用,2014(16):83,85.
MA Li, CHEN Zhixin. Domain ontology learning mehtod based on structure of the site[J]. Computer CD Software and Applications,2014(16):83,85. (in Chinese)
17 王红,高斯婷,潘振杰,等. 基于 NNV 关联规则的非分类关系提取方法及其应用研究[J]. 计算机应用研究,2012,29(10):3665 – 3668.
WANG Hong, GAO Siting, PAN Zhenjie, et al. Application and research of non-taxonimic relation extraction method based on NNV association rule[J]. Application Research of Computers, 2012, 29(10): 3665 – 3668. (in Chinese)
18 SÁNCHEZ D, MORENO A. Learning non-taxonomic relationships from web documents for domain ontology construction[J]. Data & Knowledge Engineering, 2008, 63(3):600 – 623.
19 SERRA I, GIRARDI R, NOVAIS P. Evaluating techniques for learning non-taxonomic relationships of ontologies from text[J]. Expert Systems with Applications, 2014, 41(11):5201 – 5211.
20 CHE W, LI Z, LIU T. LTP: a Chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, 2010:13 – 16.
21 ZOUAQ A, GASEVIC D, HATALA M. Linguistic patterns for information extraction in OntoCmaps[C]//Proceedings of the 3rd Workshop on Ontology Patterns, 2012:1 – 12.
22 古凌岚, 孙素云. 基于语义依存的中文本体非分类关系抽取方法[J]. 计算机工程与设计, 2012, 33(4):1676 – 1680.
GU Linglan, SUN Suyun. Approach to Chinese ontology non-taxonomic relation extraction based on semantic dependency[J]. Computer Engineering and Design, 2012, 33(4):1676 – 1680. (in Chinese)