An Ontology Learning Method for Product Configuration Domain

Shao Wei-ping, Hao Yong-ping, Zeng Peng-fei
R&D Center of CAD/CAM Technology
Shenyang Ligong University
Shenyang, China
E-mail: shaoweiping3008@sohu.com

Abstract-**Ontology learning, which seeks to discover ontological knowledge from various forms of data automatically or semi-automatically, can overcome the bottleneck of ontology acquisition in ontology development. Now most of existed ontology construction tools only support manual building ontology and it results in lowly efficiencies, highly workloads and many mistakes etc. At the same time, it's very difficult for knowledge updating and maintenance in time. In order to resolve these problems, a framework system of domain ontology automatic construction is proposed in this paper. Through ontology learning by using existed enterprise database and huge interrelated domain knowledge in Website, product configuration domain ontology is established. Key technologies of ontology learning, such as domain concepts extraction and semantic relationships between concepts extraction in different data sources structures, are discussed.**

Keywords-product configuration; ontology learning;concept ; relation

I. Introduction

An ontology is an explicit specification of a conceptualization[1]. It contains four implications: conceptualization, explicit, formal and share[2]. Ontology is used to describe the common and shared concepts and their relations in a specific-domain. It makes these concepts and their relations have common approbatory, explicit and exclusive definition in the shared area. As sharing conceptualization of knowledge presentation, ontology has been used widely in many domains, such as knowledge engineering, knowledge management, intelligent information integration, information retrieval, semantic WEB and digital library etc.

Product configuration technology is a valid strategy to fit for rapid response innovation design and variant design. Product configuration knowledge description method which based on ontology is becoming hotspot method gradually, for resolving semantic different meanings between concepts in domain. And ontology has been used more and more far-flung in configuration domain. Alexander has described virtual supply network configuration problem using ontology [3]. Wu[4] has built a Web Service Ontology and has put forward two calculation approaches of semantic similarity. Most of these methods mainly research on how to express the product configuration ontology. But not for how to built ontology automatically and build tools are almost few.

At recent years, some ontology construction tools are coming forth gradually. The typical tools are OntoEdit[5], KAON[6], OntoSaurus[7], WebOnto[8], Ontolingua[9] etc. These tools have provided friendly graphical interfaces and consistency check-up strategies for users. But most of them only support manual construction ontology and it results in lowly efficiencies, highly workloads and many mistakes etc. At the same time, it's very difficult for knowledge updating and maintenance in time. It makes the ontology building work as an arduous and fussy task. Therefore, automatic or semi-automatic ontology building methods have been proposed by researchers, that is ontology learning. Ontology learning, which seeks to discover ontological knowledge from various forms of data automatically or semi-automatically, can overcome the bottleneck of ontology acquisition in ontology development. Despite the significant progress in ontology learning research over the past decade, there remain a number of open problems in this field. Now, ontology learning needs to resolve some key problems as follow: ontology concepts extraction, semantic relations between concepts extraction, domain taxonomic system construction and ontology evaluation etc. We put the ontology learning methods into product configuration domain ontology construction automatically, for improving efficiency, reducing mistakes and updating information conveniently.

II. Product Configuration Domain Ontology Descriptions

Product configuration domain ontology defines domain-specific concepts, terms and glossaries in given configuration domain and describes their relations accurately. At the same time, domain ontology also describes the main theories and basic principles of product configuration domain. With the development and update continually of enterprise products, enterprise's knowledge repository and information are increasing quickly; new knowledge is coming forth constantly and dynamic renewal timely. Some old knowledge or information may be washed out, so concepts and their relations also change continually. How to search interrelated knowledge quickly and accurately from vast knowledge repository and extract ontology concepts and their relations, is the key to build the product configuration ontology automatically. This

Supported by the Natural Science Foundation of Liaoning Province, P. R. China (Grant No. 20082047)

978-1-4244-4639-1/09/$25.00 ©2009 IEEE

paper proposes an approach to describe knowledge ontology of product configuration domain according to water bump product.

Usually domain ontology can be expressed as a hierarchy concept tree. Product configuration domain ontology can be described as follow:

PCDO=(C,R,A,f,H^C,H^R)

C---a set of concepts in *PCDO*. They are nodes in concept tree. Such as part, function, port, resource, relation and kinds of constraints are all specific concepts in configuration domain. For bump product, its specific domain ontology contains follow concepts such as power machine, reducer, executive machine, modulator, piston, etc. Each concept is expressed using some attribute variables. The set $C^a \bigcup C^b$ consists of two subsets: $C^a = \{c : \exists ins \tan ce(c)\}$- a set of non-primitive class concepts i.e. concepts which can have instances and $C^b = \{c : \neg\exists ins \tan ce(c)\}$ -a set of primitive class concepts i.e. concepts which cannot have instances.

R---all kinds of relations existed between concepts. They are sides in the tree. *R* and *C* are two disjoint sets. There are many kind relations in concepts such as "is a" or "subclass of", "is part of/has part", "equivalence", "attribute of", "instance of", "hyper-hyponymy" and "domain/range" etc.

A---a set of concepts (or instances) attributes. Each concept has an attributes set and these attributes can be expressed as variants. Each attribute variant has its domain.

f---function *f*: *R*→*C*+ called signature.

H^C--- a partial order H^C on *C* called concept hierarchy or taxonomy.

H^R---a partial order H^R on *R* called relation hierarchy, where $r_1 \prec_R \ r_2$ implies $|f(r_1)| = |f(r_2)|$ and $\pi_i(f(r_1)) \prec_C \ \pi_i(f(r_2))$ for each $1 \le i \le |f(r_1)|$.

The concept hierarchy tree and the relations between concepts are very complicated. In order to simplify the problem of building configuration ontology, we make a rational hypothesis as follow:

Concept tree has a root and its any subtree has a root too. The same instance cannot be synchronously assigned to two concepts which has no paternity relation.

The aim of the approach presented in this paper is now to automatically acquire the concepts, relations and the partial order H^C and H^R.

III. Ontology Learning Framework of Configuration Domain

Using knowledge discovery and artificial intelligence technology, the basic ontology learning framework of configuration domain is proposed, as shown in Fig.1. It mainly contains data sources preprocessing, key words extraction, concepts and their semantic relations extraction, ontology classification system building, ontology building and pruning, ontology evaluation and editing.

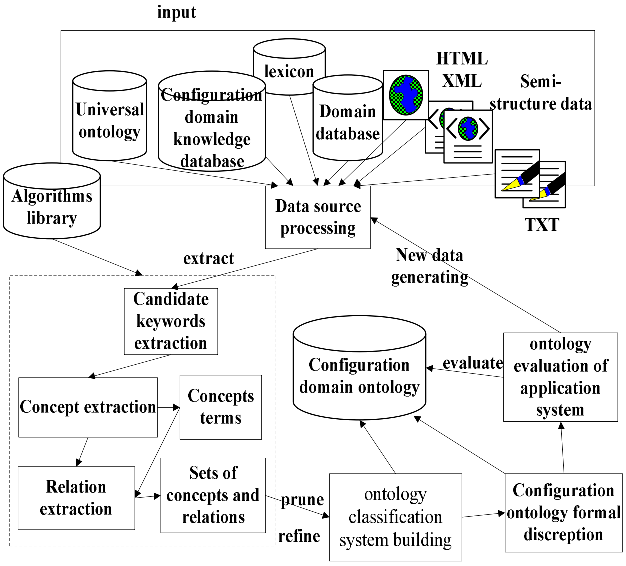
input
Universal ontology
~~Configuration~~ domain knowledge database
lexicon
Domain database
HTML
XML
Semi-structure data
TXT
Algorithms library
Data source processing
extract
New data generating
Candidate keywords extraction
Concept extraction
Concepts terms
Relation extraction
Sets of concepts and relations
prune
refine
Configuration domain ontology
evaluate
ontology evaluation of application system
ontology classification system building
Configuration ontology formal discreption

Figure 1. Ontology learning framework of configuration domain

A. *Configuration domain data sources input and preprocessing*

It can be seen from Fig.1, input data can be various data source and different structures. It includes structural data such as data in RDB or OODB, non-structural data such as vast plain text files in Web, semi-structural data such as vast XML or HTML format Web pages and XML schema or DTD etc. Different data source has different preprocessing method.

B. *Concepts and relations extraction*

In this phase, using inputted ontology and many kinds of ontology learning algorithms in algorithms library to extract new concepts and relations from different data sources. Concept extraction include: candidate keywords extraction, domain terms extraction and concept define. First, extract concept candidate from preprocessed result set. Then, use halt using word list and filtering rule to leach the non-concepts. Select proper algorithms to find the domain terms and concepts definitions finally.

C. *Ontology pruning and refining*

According to the semantic relations between concepts, concept classification hierarchy relations can be established. The extracted concepts and relations must be pruned and refined. So some concepts and relations which are not interrelated to product configuration domain-specific must be removed.

D. *Domain ontology formalization*

Acquired domain ontology, concepts set and relations set by ontology learning need to formalize so that can be understood and processed by computer. Many ontology description languages such as OWL, RDF/S, DAML+OIL and OML are used to express the acquired configuration ontology.

E. Domain ontology application evaluation

The acquired configuration ontology is used to product configuration design application system and its validity is evaluated by practicing in engineering. The evaluation criterions are precision, recall and F factor etc. Through evaluating and affirming, objective ontology will be added to ontology library.

IV. CONCEPTS AND RELATIONS EXTRACTION

A. Concepts extraction

There are multiform data structures in configuration domain. According to the different kind data source structure, different concept extraction method is used.

- Concepts extraction from non-structure data

Traditional ontology learning systems for concept extraction from non-structure data source were based on words. First, keywords were identified from the text. These words are typically single-word terms and will be seen as the concepts. Then, possible multiword terms were formed by combining these keywords. As a result, the multiword terms generated were not natural and most concepts extracted were only single-word terms. Therefore, many important multiword concepts are lost. When using the NLP component to process documents, we found most noun terms in the text are multiword terms. Traditional systems focusing on single-word term extraction will thus miss many concepts. We adopt a different strategy for concept extraction. First, multiword terms are induced from text directly. Then, single-word terms are extracted if they appear frequently in the multiword terms or they are found related to the multiword terms through certain semantic relations. This strategy reduces the chance of missing important concepts. This method identifies domain concepts mainly according to its different statistical feature between domain concepts and general words. Usually calculate the domain relativity and domain generality.

The concept learning procedure is described as follow. Extract all multiword terms using the predefined regular expression rules. As concepts are nouns or noun phrase in texts, only word patterns with the NP tag are collected. At the same time, remove articles and descriptive adjectives such as “a”, “many” and “several” etc. from the terms extracted. Generate all possible sets of two or more words in each extracted term as candidate terms. For each term *t*, compute a liner combination: $T-D(t)=\alpha T(t)+(1-\alpha)D(t)$, where α∈[0,1] is a weighting parameter. The *T* and *D* scores[10], are the statistical measures for evaluating terms. Terms with high *T-D* values are selected to form an initial concept list *T*. Let *SW* be the set of single-word terms appearing in the T as the syntactic head of a term t. Then compute for each single-word term in *SW* the occurrence frequency in *T*. Those with frequency above a threshold δ are added to the list *T*.

- Concepts extraction from structure data

In enterprise, vast product data of configuration domain are saved in database. Structural data mainly includes the data in RDB or OODB. Therefore, it is important to build configuration ontology using the profuse data in database. At present, most enterprises use RDB to save product information. RDB adopts relation model .Its structure is very simple so that the two-dimensional relation tabulation can be easily understood. Entities and its relations are both expressed in sheet in relation model. So it must be distinguished that which sheets are used to describe the entities and which sheets are used to describe the relations of entities. Then, entities are mapped into concepts of configuration domain ontology and relations of entities are mapped into relations of concepts in ontology. In this paper, we adopt the ontology learning method of literature [11]. By analyzing the sheet, attributes, primary key, foreign key and contain relations in RDB, a group of mapping rules from relation model to configuration domain ontology are extracted firstly. Then, candidate ontology is obtained based on these rules, and it is evaluated, pruned and refined.

- Concepts extraction from semi-structure data

Semi-structure data is that has connotative structure but no rigorous or fixed structure data, such as vast web pages of XML, RDP and HTML format. Due to this kind of data is intervenient between structure data and non-structure data, the concepts can be extracted using hereinbefore two methods.

B. Semantic relations between concepts extraction

Semantic relations between concepts are important contents of ontology learning. We mainly thank about taxonomic relations and non-taxonomic relations learning. Taxonomic relations are those in different logic levers relations between concepts, such as “*is a*” or “*instance of*” relation. While non-taxonomic relations are those relations such as “*part/whole*”, “*synonymous*” relations. Take water bump as an example, the relations are shown in Fig. 2.

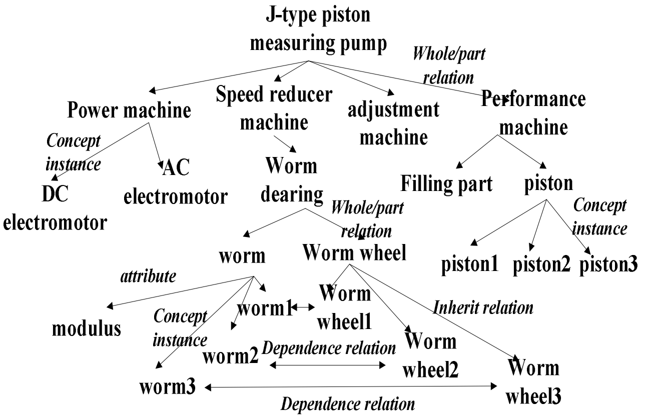

Fig. 2 Concepts and their relations of J-type measuring pump

Semantic relations extraction approaches mainly include: templet-based, association rule-based, concepts clustering-based, lexico-based etc. We adopt concept clustering-based and association rule-based methods learning methods according to the characteristics of machine product configuration. Calculate the semantic distance between concepts. The same kind concept has the approximate semantic relation, and the result of concept hierarchy clustering is taxonomic relations between concepts.

For learning non-taxonomic relations, use shallow text processing method to identify related pairs of concepts and calculate the *support* and *confidence* of concepts pairs. Our learning mechanism is based on the algorithm for discoverying generalized association rules proposed by Srikant[12]. The basic association rule algorithm is, for example, provided with a set of reducer $RS{:}=\{rs_i|i=1,2,...n\}$, where each reducer rs_i consist of a set of items $rs_i{:}=\{a_{ij}|j=1,2,...m_i,a_{ij}\in C\}$ and each item a_{ij} is from a set of concepts C. The algorithm computes association rules $X_k \Rightarrow Y_k (X_k, Y_k \subset C, X_k \cap Y_k = \{\ \})$ such that measures for *support* and *confidence* exceed user-defined thresholds. Thereby, *support* of a rule $X_k \Rightarrow Y_k$ is the percentage of reducers that contain $X_k \cup Y_k$ as a subset, and *confidence* for $X_k \Rightarrow Y_k$ is defined as the percentage of reducers that Y_k is seen when X_k appears in a reducer, *viz.*

$$\text{support}(X_k \Rightarrow Y_k) = \frac{|\{rs_i | X_k \cup Y_k \subseteq rs_i\}|}{n}$$

$$\text{confidence}(X_k \Rightarrow Y_k) = \frac{|\{rs_i | X_k \cup Y_k \subseteq rs_i\}|}{|\{rs_i | X_k \subseteq rs_i\}|}$$

Then, determine associations at the right of *taxonomy*, formally given by a taxonomic relation $H \subset C \times C$. For this purpose, we extend each reducer rs_i to also include each ancestor of a particular item a_{ij} , i.e. $rs_i' := rs_i \cup \{a_{il} | (a_{ij}, a_{il}) \in H\}$. Now, compute support and confidence for all possible association rules $X_k \Rightarrow Y_k$ where Y_k does not contain an ancestor of X_k as this would be a trivially valid association. Finally, prune all these association rules $X_k \Rightarrow Y_k$ that are subsumed by an “ancestral” rule $\hat{X}_k \Rightarrow \hat{Y}_k$, the item sets $\hat{X}_k$, $\hat{Y}_k$ of which only contain ancestors or identical items of their corresponding item set in $X_k \Rightarrow Y_k$.

V. CONCLUSION

Configuration domain ontology automatic (or semiautomatic) building is the basis of product configuration quickly. We have presented a new approach to automatically acquire concepts and their relations from different data sources. First, the ontology learning methods of different data sources in enterprise is discussed based on the theories and approaches of ontology building automatically. Then, the key technologies and algorithms of concepts and relations extraction in configuration domain are discussed respectively. Finally, the basic ontology learning framework and its functional modules of configuration domain is proposed. It overcomes the disadvantages of manual-building ontology and gives a new and effective approach for ontology building of product configuration.

REFERENCES

[1] T. R. Gruber. A translation approach to portable ontologies[J]. Knowledge Acquisition, 5(2):199-220, 1993.

[2] Fensel D. Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer. 2001.

[3] Alexander V S, Leonid B S, Nikolai C. Ontology-driven approach to constraint-based VSN configuration[A], Second World Conference on POM and 15th Annual POM, Conference, Cancun, Mexico,April 30 - May 3,2004.

[4] Wu J, Wu Z H, Li Y.etc. Web service discovery based on ontology and similarity of words[J]. Chinese Journal of Computer. 28(4): 595-602, 2005.

[5] Sure Y, Angele J, Erdmann M etc. OntoEdit: Collaborative ontology engineering for the semantic Web. In: Horrocks I, Hendler JA, eds. Proc. of the ISWC 2002. Heidelbeg: Springer-Verlag, 221-235,2002.

[6] Bozsak E, Ehrig M, Handschuh S etc. KAON-Towards a large scale semantic web. In: Bauknecht K, Mintjoa A, Quirchmayr G, eds. Proc. of the 3rd conf. on E-Commerce and Web Technologies. H eidelbeg: Springer-Verlag, 304-313,2002.

[7] Swartout B, Ramesh P, Knight K etc. Towards distributed use of large-scale ontologies. In: Proc. of the AAAI Symp. On Ontological Engineering.http://ksi.cpsc.ucalgary.Ca/KAW/KAW96 /swartout/ Banff_96_final_2.html.

[8] Duineveld AJ, Stoter R, Weiden MR. Wonder tools? A comparative study of ontological engineering tools. Journal of Human-Computer Studies, 52(6): 1111-1133, 2000.

[9] Farquhar A, Fikes R, Rice J. The ontolingua server: A tool for collaborative ontology construction. Journal of Human-Computer Studies, 46(6): 707-727,1997.

[10] T.Berners-Lee, J.Hendler, O.Lassila. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific Americal,2001.

[11] Stojanovic L, Stojanovic N, Volz R. Migrating data-intensive web sites into the semantic Web. In: Proc. Of the 17th ACM Symp. On Applied Computing. New York: ACM Press,1100-1107 ,2002.