

# General ontology learning framework

Liu Baisong<sup>1,2</sup> Gao Ji<sup>1</sup>

(<sup>1</sup> College of Computer Science, Zhejiang University, Hangzhou 310027, China)

(<sup>2</sup> Network Center, Ningbo University, Ningbo 315211, China)

**Abstract:** In order to reduce the costs of the ontology construction, a general ontology learning framework (GOLF) is developed. The key technologies of the GOLF including domain concepts extraction and semantic relationships between concepts and taxonomy automatic construction are proposed. At the same time ontology evaluation methods are also discussed. The experimental results show that this method produces better performance and it is applicable across different domains. By integrating several machine learning algorithms, this method suffers less ambiguity and can identify domain concepts and relations more accurately. By using generalized corpus WordNet and HowNet, this method is applicable across different domains. In addition, by obtaining source documents from the web on demand, the GOLF can produce up-to-date ontologies.

**Key words:** ontology; ontology learning; ontology evaluation; semantic web

Since ontologies provide a shared understanding of a domain of interest, they have become a key technology for semantics-driven modeling, especially for the ever-increasing need for knowledge interchange and integration. In recent years, research related to ontology development has produced tangible results concerning the definition of language standards and increasingly powerful ontology editing and construction tools. Despite the availability of these tools, populating domain ontologies is a tedious and time-consuming process, preventing wide-scale production and usage of ontologies by industrial institutions. Automatic methods for ontology learning have been proposed in recent literature, such as TextToOnto<sup>[1]</sup>, EKAW2006<sup>[2]</sup>, EON2006<sup>[3]</sup>, OntoLT<sup>[4]</sup> or OntoLearn<sup>[5]</sup>. However, all these tools suffer from several shortcomings. First, they all depend either on very specific or proprietary ontology models which cannot always be translated to other formalisms in a straightforward way. This is certainly undesirable as ontology learning tools should be independent from a certain ontology model in order to be widely applied<sup>[6]</sup>. Secondly, the interaction with end-users, in contrast to linguists or machine learning specialists, has been largely neglected within such systems. As users are typically the ones who are most familiar with the domain, user interaction should be a

central part of the system architecture<sup>[7]</sup>. Thirdly, most of these tools lack a certain robustness with respect to changes made to the data set<sup>[8]</sup>.

In this paper, we propose an ontology learning method called GOLF (general ontology learning framework), which generates and publishes ontologies dynamically from corpus and web-pages.

## 1 GOLF's Architecture

To construct an ontology for a new domain, we need to collect domain keywords and find the relationships among them. An acquisition process, the GOLF, is designed that can construct a new ontology through domain corpus. Thus, with little human intervention, the GOLF can build a prototype of the domain ontology. Fig. 1 shows the GOLF's system architecture, which supports a three-phase process.

### Algorithm 1 GOLF ( $D, T$ )

```
/* construct an ontology for the terms in  $T$  on the basis of the web
documents in  $D$  */
Docs = preprocess(POS-tag( $D$ )) // pre-processing of corpus and web
documents;
Parses = parse(POS-tag( $D$ )) // keywords candidates generation;
DomainTerm = DR(Parses) and DC(Parses); // domain term extraction;
OntConcept = MI(DomainTerm); // ontology concept selection;
 $R_{can} = \{Tem_{hyp}, Tem_{par}, Tem_{syn}, Tem_{ins}\}$ ; // semantic relation learning;
 $T_c = AssociateRule(R_{can})$ 
DomainTree = getFormalContext( $T_c$ ) // create formal domain trees;
 $T_o = TaxMiner(DomainTree)$  // taxonomy construction;
 $O = Prune(T_o)$  // ontology pruning;
Evaluate( $O$ ) // result ontology evaluation;
Return  $O$ 
```

Received 2006-04-25.

**Foundation items:** The National Basic Research Program of China (973 Program) (No. 2003CB317000), the Natural Science Foundation of Zhejiang Province (No. Y105625).

**Biographies:** Liu Baisong (1971—), male, graduate, associate professor, lbs@nbu.edu.cn; Gao Ji (corresponding author), male, professor, gaoji@mail.hz.zj.cn.

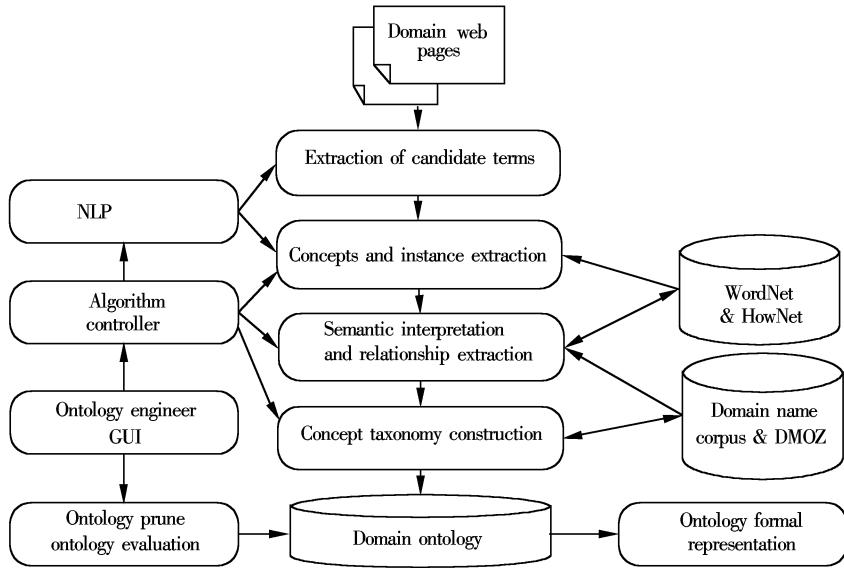


Fig. 1 Overall architecture of the GOLF

## 2 GOLF’s Method

### 2.1 Concept extraction

Terminology can be considered as the surface appearance of relevant domain concepts. Candidate terminological expressions are usually captured with shallow techniques that range from stochastic methods to more sophisticated syntactic approaches. To extract concepts, there are seven steps to follow, shown in algorithm 2.

**Algorithm 2** Ontology concepts extraction C-Extraction

#### Step 1 Web document features extraction

Web pages contain rich information, i. e., text, image, audio, video information, etc. But the textual information is one the most important. Analyze all web pages and extract the satisfying fragment, then, remove the stop words and word stemming.

#### Step 2 Candidates generation

Generate the candidates by statistical relevance information between glossary, domain feature information and the syntax information inside the phrase, also utilize shallow parser and heuristic rules, such as the hints of key sentences and paragraphs.

#### Step 3 Domain relevance calculation

A quantitative definition of the DR can be given according to the amount of information captured in the target corpus relative to the entire collection of corpora.

#### Step 4 Domain consensus calculation

A second filter operates on the principle that a term cannot be a clue for a domain  $D_k$  unless it appears in several documents; there must be some consensus on using that term in the domain  $D_k$ . The domain consen-

sus of term  $t$  in class  $D_k$  captures those terms that appear frequently across a given domain’s documents.

#### Step 5 Terminology order ascertainment

A linear combination of the above two filters obtains the terminology. Since the statistical significance can be influenced by the technicality of the language domain and by the dimension of the training corpus, we determine this threshold empirically.

#### Step 6 Extraction of domain concepts with inter-information

This step filters out irrelevant terms from the extracted. We employ a baseline pruning strategy which advocates that frequent terms in a corpus denote domain concepts while less frequent ones lead to concepts that can be safely eliminated. We consider the average frequency of the terms as a threshold value and prune all concepts that have a lower frequency than this value.

#### Step 7 Output: domain concepts list.

### 2.2 Semantic interpretation

Semantic interpretation is the process of first determining the right concept (sense) for each component of a complex term, this is known as semantic disambiguation.

At this stage, we solve the ambiguity problems with WordNet and HowNet. The main function of this module is to disambiguate domain terms and generate domain concepts. The input and output of this module are domain term sets and concept sets. The algorithm of the module is an improved SSI algorithm, which is described as follows.

**Algorithm 3** Semantic interpretation ( find-semantic-relationship)

Input: List(DC)  
 Output: Concepts list  $T_C$  (Including domain concepts and their relationships)  
 Initialize  $R_{can}, Tem_{hyp}, Tem_{par}, Tem_{syn}, Tem_{ins}$   
 For each concept in DC, extract every relationship template  $A$  or  $B$ :  
 $Tem_{hyp}^{A_i}, Tem_{hyp}^{B_j}; Tem_{par}^{A_i}, Tem_{par}^{B_j}; Tem_{syn}^{A_i}, Tem_{syn}^{B_j}; Tem_{ins}^{A_i}, Tem_{ins}^{B_j}$   
 Calculate  $R_{can} = \{Tem_{hyp}, Tem_{par}, Tem_{syn}, Tem_{ins}\}$ ,  
 Optimize A-type relationships using association rule  
 Optimize B-type relationships using association rule  
 Return: Concepts list  $T_C$

### 2.3 Taxonomy construction

The taxonomic relation learning is the task to build the concept hierarchy which constitutes the skeleton of the ontology<sup>[7]</sup>. The final result of the above outlined process is a flat list of terms. However, terms may be further structured in sub-trees, thus facilitating a subsequent linking of the sub-trees to the appropriate node of the domain ontology. We extract taxonomic relations starting from the syntactic head of multiword terms:

#### Algorithm 4 Taxonomy construction TaxMiner

Input: Concept list  $T_C$   
 Output: Taxonomy of  $T_C$   
 ① For each concept in  $T_C$ , run ② to ⑥ until  $T_C$  is empty;  
 ② Calculate neighborhood concept of each concept  $C$ ;  
 ③ Calculate hym/hypo of each concept;  
 ④ Calculate concept's part-whole relationship;  
 ⑤ Calculate hierarchy similarity  $T(a, b)$ , find sibling node for each concept;  
 ⑥ For  $H(C, Is-A)$ , find root node;  
 ⑦ Return

## 3 Experiments and Evaluation

We conduct the experiment on the web pages of education domain, and analyze 2 734 HTML documents, including more than 1 700 000 words and HTML tags. The results of the term extraction evaluation stage are shown in Tab. 1, and the results of ontology evaluation are shown in Tab. 2.

**Tab. 1** Term extraction results

$correct_{extracted}$	936
$all_{corpus}$	1 427
$all_{extracted}$	1 289
$T_{Recall}/\%$	65.59
$T_{Precision}/\%$	72.61
$F_{measure}/\%$	68.92

**Tab. 2** Ontology evaluation results

Correct	473
New	46
Fault	157
$O_{Precision}/\%$	76.77
$L_O/\%$	47
$O_1/\%$	126
TaxoPrecision/ $\%$	93.28

### 3.1 Term extraction

This evaluation stage is only concerned with the performance of the term extraction modules. We use term recall ( $T_{Recall}$ ) to quantify the ratio of relevant terms that are extracted from the analyzed corpus ( $correct_{extracted}$ ) over all terms to be extracted from the corpus ( $all_{corpus}$ ). Term precision ( $T_{Precision}$ ) denotes the ratio of correctly extracted terms over all extracted terms ( $all_{extracted}$ ). We also compute the  $F_{measure}$  by assigning an equal importance to both precision and recall.

$$T_{Recall} = \frac{correct_{extracted}}{all_{corpus}}, \quad T_{Precision} = \frac{correct_{extracted}}{all_{extracted}}$$

$$F_{measure} = \frac{2T_{Precision}T_{Recall}}{T_{Precision} + T_{Recall}}$$

### 3.2 Ontology evaluation

During the concept evaluation per concept analysis of the extracted ontologies, the domain experts rated concepts correct if they were useful for ontology building and were already included in the gold standard. Concepts that were relevant for the domain but not considered during manual ontology building were rated as new. Finally, irrelevant concepts, which could not be used, were marked as faulty. We express the ratio as ontology precision ( $O_{Precision}$ ):

$$O_{Precision} = \frac{correct + new}{correct + new + fault}$$

We evaluate domain coverage by comparing the extracted ontologies to the corresponding gold standard ontologies<sup>[5,9]</sup>. Our first metric is lexical overlap ( $L_O$ ). Let  $L_{O_1}$  be the set of all domain relevant extracted concepts and  $L_{O_2}$  the set of concepts of the gold standard ontology. The lexical overlap is equal to the ratio of the number of concepts shared by both ontologies (denoted as correct) and the number of all gold standard ontology concepts (denoted as all). Ontological improvement ( $O_1$ ) is the ratio between all domain relevant extracted concepts that are not in the gold standard ontology (denoted as new) and all the concepts of the gold standard ontology (denoted as all).

$$L_O(o_1, o_2) = \frac{|L_{O_1} \cap L_{O_2}|}{|L_{O_2}|} = \frac{correct}{all}$$

$$O_1(o_1, o_2) = \frac{|L_{O_1} \setminus L_{O_2}|}{|L_{O_2}|} = \frac{new}{all}$$

We also evaluate the quality of the taxonomic relations. For this we count the number of taxonomic relations discovered between domain relevant concepts. Then an expert assesses how many of these taxonomic relations express indeed an is-A relation ( $allRels_{Correct}$ ). The TaxoPrecision metric is the ratio of correctly identified is-A relations over all taxonomic relations between domain relevant concepts that are automatically

discovered.

$$\text{TaxoPrecision} = \frac{\text{allRels}_{\text{Correct}}}{\text{allRels}_{\text{Relevant}}}$$

## 4 Conclusion and Future Work

Comprehensive ontology construction and learning has been an active research field for the past few years. Several workshops and institutes have been dedicated to ontology learning and related issues<sup>[1, 5-6, 10]</sup>. In this paper the framework GOLF with the aim of learning ontologies from WWW is presented. The contributions of our work are as follows: ① The GOLF includes total process of ontology construction, so it has more integrated features; ② The GOLF is independent of the actual ontology model or knowledge representation language and ③ Our work integrates corpus-driven change discovery strategies, so it increases the efficiency of the system as well as the traceability of the learned ontology with respect to changes in the corpus, thus making the whole process more transparent and robust.

We apply our approach to many university websites in order to show its usefulness. As a result of our experiments, we confirm that our method can extract ontology concepts and their relationships in an efficient way. We also conduct ontology evaluation by comparing the result ontologies with gold standard ontologies. The future work focuses on more semantic relationships learning and ontology instance populations.

## References

- [1] Buitelaar P, Handschuh S, Magnini B. ECAI workshop on ontology learning and population: towards evaluation of text-based methods in the semantic web and knowledge discovery life cycle[A]. In: *The 16th European Conference on*

*Artificial Intelligence*[C]. Valencia, Spain, 2004. 1 – 6.

- [2] Navigli R, Velardi P. Ontology enrichment through automatic semantic annotation of on-line glossaries [A]. In: *Proc of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2006)*, LNAI [C]. Pödebrady, Czech Republic, Springer, 2006. 4248: 126 – 140.
- [3] Brank Janez, Mladenec Dunja, Grobelnik Marko. Golden standard based ontology evaluation using instance assignment [A]. In: *The 4th International EON Workshop* [C]. Edinburgh, United Kingdom, 2006. 47 – 54.
- [4] Buitelaar P, Olejnik D, Sintek M. OntoLT: a protégé plug-in for ontology extraction from text [A]. In: *Proceedings of the International Semantic Web Conference (ISWC)* [C]. Florida, USA, 2003. 17 – 22.
- [5] Velardi P, Navigli R, Cuchiarrelli A, et al. Evaluation of ontoLearn, a methodology for automatic learning of domain ontologies[A]. In: *Ontology Learning from Text: Methods, Evaluation and Applications* [C]. IOS Press, 2005. 1 – 32.
- [6] Sabou Marta. Learning web service ontologies: an automatic extraction method and its evaluation [A]. In: *ISWC2005* [C]. Osaka, Japan, 2005. 98 – 116.
- [7] Cimiano P, Staab S, Tane J. Automatic acquisition of taxonomies from text: FCA meets NLP[A]. In: *ECML/PKDD Workshop on Adaptive Text Extraction and Mining* [C]. Cavtat-Dubrovnik, Croatia, 2003. 10 – 17.
- [8] Thanh Tho Quan, Siu Cheung Hui, ACM Fong, et al. Automatic generation of ontology for scholarly semantic web [A]. In: *ISWC2004, LNCS* [C]. Hiroshima, Japan, 2004, 3298: 726 – 740.
- [9] Maedche A, Staab S. Measuring similarity between ontologies [A]. In: *Proceedings of European Knowledge Acquisition Workshop (EKAW2002)* [C]. Madrid, Spain, 2002. 251 – 263.
- [10] Maedche A, Staab S. Ontology learning for the semantic web [J]. *IEEE Intelligent Systems*, 2001, 16(2): 72 – 79.

# 通用本体学习框架研究

刘柏嵩<sup>1,2</sup> 高 济<sup>1</sup>

(<sup>1</sup>浙江大学计算机科学与技术学院, 杭州 310027)

(<sup>2</sup>宁波大学网络中心, 宁波 315211)

**摘要:**提出了一种通用本体学习框架 GOLF, 通过对网络上各专业领域 web 文档集进行挖掘来实现本体自动构建, 讨论了本体学习中本体概念的抽取、概念之间语义关系的抽取和分类体系的自动构建等关键技术, 通过实验对算法进行了测试, 并对本体评价方法进行了探讨. 由于集成了多种机器学习算法, 该方法在概念抽取和语义关系学习方面具有更高的准确性. 采用通用本体 WordNet 和 HowNet 作为语料库, 它可适用于不同的专业领域. 同时, 通过按需获取 web 文档, 该方法能实时生成本体.

**关键词:**本体; 本体学习; 本体评价; 语义 web

**中图分类号:** TP312. 1