Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 1 (2012) 287 – 292

AASRI Procedia

www.elsevier.com/locate/procedia

2012 AASRI Conference on Computational Intelligence and Bioinformatics

Ontology Learning from Online Chinese Encyclopedias

Yingna Li[a], Tong Ruan[a,]*, Fanghuai Hu[a]

[a]*Department of Computer Science and Engineering, East China University of Science and Technology, No.130, Meilong Road, Shanghai, 200237, China*

Abstract

An ontology is very important in describing and sharing knowledge of a domain. This paper proposes a method to automatically generate domain ontologies from Chinese encyclopedias on the web. First, we use the terms appears in category systems of encyclopedias as concepts and resolute synonyms, then derive an original taxonomy from Chinese-Wikipedia and Hudong-Baike; for other concepts not in the original taxonomy, we use a set-theory like method to form a directed graph and generate a tree from the graph via maximum-spanning-tree algorithm, and merge the tree into the taxonomy. Then, we use titles of normal articles as instances and populate them via category labels in them. The attributes of concepts and instances are generated from special structures such as InfoBox modules. We learn a plant ontology successfully and the later experiments show that the learnt ontology has well precision and high coverage.

Keywords: ontology learning; encyclopedia; concept; taxonomy; Instance

1. Introduction

An ontology describes formally a domain of discourse, it typically consists a finite list of concepts, the relationships (both taxonomic and non-taxonomic) between these concepts and instances of them. Ontologies are useful in knowledge management, information retrieval, intelligent recommendation, and so on. However,

* Corresponding author. Tel.: +86-137-6431-7832.
E-mail address: ruantong@ecust.edu.cn.

doi:10.1016/j.aasri.2012.06.044

to build ontologies manually is very expensive, it needs many domain experts and programmers work together for a long time. Moreover, the manual process is tedious and error-prone, and the built ontologies suffer from low coverage and fast aging. There are only a few manually built ontologies; the most famous three are WordNet [Fellbaum, 1998], Cyc [Lenat, 1995] and Gene Ontology [GOC, 2000]. In Chinese, HowNet [Dong and Dong, 2006] provides a knowledge database, but it only has less than 100,000 words and is not free; to the best of our knowledge, there are no large Chinese ontologies published for free yet.

Therefore, in the last several years, people have engaged themselves in building ontologies (semi-)automatically, which is called ontology learning. According to the coverage of the ontology, ontology learning can be divided into general ontology learning and domain ontology learning. When distinguishing by the source from which the ontologies are learnt, there are ontology learning based on structured data, semi-structured data and unstructured data; more concretely, semi-structured data including h. This paper focuses on learning ontologies from Chinese encyclopedias. With the development of Web 2.0, there are more and more collaborative semi-structured data published on the Internet, the most famous one is Wikipedia [Wales, 2005], which contains more than 3.8 million English articles and 400 thousand Chinese articles. In Chinese, there are other two larger online encyclopedias, namely Baidu-Baike and Hudong-Baike, both of which contain more than 3 million articles.

The paper is organized as follows. Section 2 describes some related works and Section 3 defines the ontology learning problem; then, the learning process is shown in detail in Section 4; after that, the experimental result is listed in Section 5; Finally, Section 6 gives conclusions and future works.

2. Related work

There are many works about ontology learning, here we divide these works by which source they use.

Text based learning, which learns ontology from unstructured text directly. Text-to-Onto [Maedche and Staab, 2000] is an earlier ontology learning system which learns ontologies from text. TShamsfard and Barforoush [TShamsfard and Barforoush, 2004] proposed an automatic ontology building approach which started from a small ontology kernel and constructed the ontology through text understanding automatically. Lee et al. [Lee et al., 2006] presented a novel episode-based ontology construction mechanism to extract domain ontology from unstructured text document. In general, these methods relied heavily on NLP techniques, structured knowledge such as dictionaries, and sometimes human interaction.

HTML page based learning. Sánchez and Moreno [Sánchez and Moreno, 2004] developed an approach to automatically construct ontology from the Web which started from an initial keyword, they first used search engines to get web pages related to the keyword and then mined concepts from these web pages. Tian et al. [Tian et al., 2009] used an iterative learning approach to learn ontologies with the help of search engine and Protégé-OWL API. Shinzato and Torisawa [Shinzato and Torisawa, 2004] provided a method to acquire hyponymy relations from HTML documents, they used itemization and listing elements in documents. These methods could get shallow information only as lacking of useful structured data.

Encyclopedia based learning. Over 2 million concepts with mappings to over 3 million unique terms were mined from Wikipedia by Gregorowicz and Kramer. [Gregorowicz and Kramer, 2006]. Ponzetto and Strube [Ponzetto and Strube, 2007] derived a large scale taxonomy from Wikipedia by labeling the semantic relations between categories. KOG [Wu and Weld, 2008] is an autonomous system for refining Wikipedia's InfoBox-class ontology, SVMs and Markov Logic Networks were used to solve the ontology refinement machine learning problem. YAGO [Suchanek et al., 2007] is a high coverage and quality ontology which contains more than 1 million entities and 5 million facts, the ontology was automatically extracted from Wikipedia and unified with WordNet. As the English Wikipedia is really abundant, good ontologies can be learnt. In contrast,

Chinese-Wikipedia only has 386,137 articles, most of which contains short descriptions. Therefore, it is almost impossible to derive a good ontology from Chinese-Wikipedia with these methods.

Other source based learning. There are ontology learning method based on other source such as search log [Yin and Shah, 2010; Paşca, 2007; Paşca, 2008] and dictionaries [Suchanek et al., 2007].

3. Problem setting

Formally, an ontology can be defined as a structure of $O := (C, \leq_C, R, \sigma_R, \leq_R, A, \sigma_A, I)$, including a concept set C, the taxonomy of C, a relation set R, the taxonomy and the range of R, an attribute set A, the range of A, and a instance set I. In this article, we only learn the core of an ontology which contains concepts, the taxonomy of concepts, instances, and attributes of concepts and instances.

We will use the three largest online encyclopedias, Baidu-Baike, Hudong-Baike and Chinese-Wikipedia, as corpus for ontology construction. The useful structures are article title, redirection, category system, taxonomy, and InfoBox module, some of which are shown in Fig 1.

Fig. 1. Useful structure in encyclopedias for ontology construction

4. Ontology learning

In order to use the three encyclopedias, we should download them first. The Chinese-Wikipedia provides a repository for end users to download it directly; however, we have to write crawlers by ourselves for the other two encyclopedias as they only available via HTML pages. Totally, we get 4,009,755 articles from Baidu-Baike, 2,780,675 articles from Hudong-Baike and 377,635 articles from Chinese-Wikipedia.

4.1. Concept

Concepts are directly deduced from category systems of the three encyclopedias. As encyclopedias are freely edited by arbitrary end users, there may be some problems. Therefore, we use the following rule to filter the appeared categories: if a category is isolated and has no child categories and instance pages, we discard it. For the remaining categories, we have to merge synonyms. The redirection module in all the three encyclopedias is good clue for synonym resolution. Moreover, we find the fields named “alias” in Baidu-Baike and “alias in Chinese” in Hudong-Baike contain many synonyms of the current article.

4.2. Instance

The titles of articles can be seen as instances, the same problem is to merge synonyms, the same method as concept synonym resolution is adopted. Almost all articles have category labels, which can help us to populate them; but the category labels may be inappropriate when mapping them into ontologies directly.

Therefore, for all the categories one article belongs to, if there are inclusion relations, we populate the article to the sub-category. For example, the article titled with "lotus" has category labels "plant" and "Magnoliidae"; obviously, we should keep the category "Magnoliidae" and ignore the category "plant".

4.3. Taxonomy

Both Hudong-Baike and Chinese-Wikipedia have good category hierarchies, thus an original taxonomy can be deduced from them. However, there are scattered categories. Moreover, all the categories in Baidu-Baike are not well organized. In order to merge these categories into the original taxonomy, we first use the set-theory like method to calculate the belonging degree of them and thus form a directed graph $G = (V, E)$, with weight function $\omega: \mathbb{E} \rightarrow \mathbb{R}$, V stands for the set of categories. The weight function is defined as:

$$\omega_{A \supset B} = \frac{\left| I_A \cap I_B \right|}{\left| I_B \right|} \tag{1}$$

Where A and B are two concepts, IA means the set of instances of concept A, and $| \, |$ is the operator to calculate the size of a set. Then we use the maximum-spanning-tree algorithm [Chu and Liu, 1965; Edmonds, 1967] to generate a tree.

4.4. Attribute

The InfoBox modules are good resource for attribute extraction; however, only about ten percent of articles have InfoBox modules. Fortunately, later experiment shows that about eighty percent of plant articles have InfoBox modules. The InfoBox modules are usually defined by special HTML tags and are easily to parse.

5. Experiment

In this section, we first learn a plant ontology use the techniques specified above. Then, evaluation about the learnt ontology is described. Usually, there are four ontology evaluation methods, namely "golden standard" based evaluation, expert based evaluation, application based evaluation and domain-data scale based evaluation. As there is no "golden" ontology in Chinese, and we have no application based on our learnt ontology yet, we use the domain-data scale based method and expert based method. We will evaluate the ontology in two aspects, its scale and its precision.

5.1. Ontology learning information

We start with the category labeled with "plant" and extract concept, taxonomy, instance, and attribute step by step. In concept extraction, we extract 3,122, 4,431 and 4,507 concepts from Chinese-Wikipedia, Hudong-Baike and Baidu-Baike respectively; and the numbers of instance from them are 24,482, 54,977 and 56,128 respectively. After doing synonym resolution, the learnt plant ontology contains 4,788 concepts and 59,975 instances in total. Almost all the concepts and 49,023 instances have attributes. Fig 2 shows the information of the learnt ontology in detail, and Fig 3 shows a snapshot of the inner structure of the ontology (we only list part of the concepts and instances).

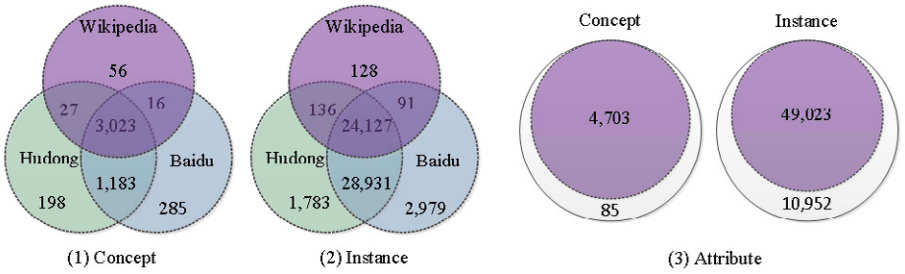

Fig. 2. The detail information of the learnt ontology

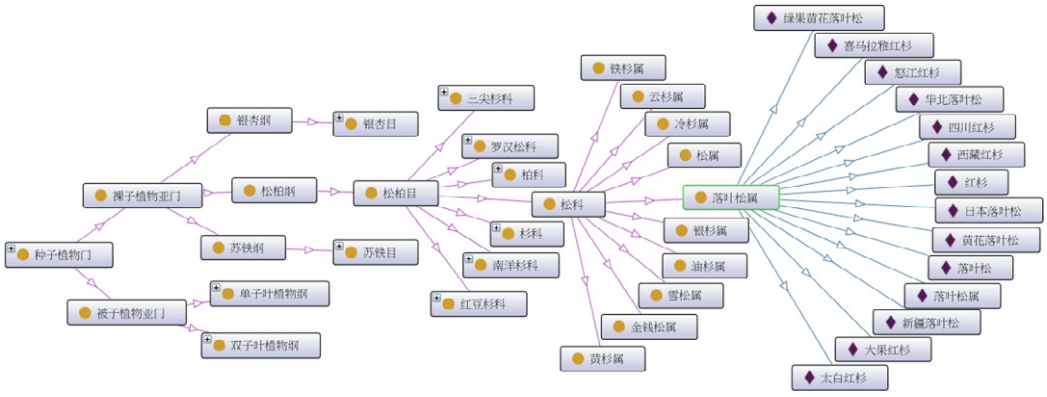

Fig. 3. Sample structure of the plant ontology

5.2. Ontology evaluation

According to Wikipedia, there are about 350,000 kinds of plants and 287,655 of which have been discovered until 2004. Our learnt plant ontology contains about 6 million instances, covering about 21 percent, which certainly contains most frequently-used plant. Obviously, the scale of the ontology is limited by the corpus.

In order to evaluate the precision, we chose all the decedents of the concept “Coniferales” for manual evaluation, including 52 concepts, 493 instances in which 484 having attributes. The result is shown in Table 1. From the table, we can conclude that the concept and attribute are all correct because their sources are reliable; Taxonomy error comes from the set-theory like assumption and the maximum-spanning-tree algorithm; There are 2 instances are not plant, but they are labelled as plant in the encyclopaedias; Instances are wrongly populated because the category system in the encyclopaedias are really too loose.

Table 1. Precision of the learnt ontology estimated from the sample data

	Concept	Taxonomy	Instance(is plant)	Instance(populate)	Attribute
Total	52	52	493	489	484
Correct	52	49	491	452	484
Precision	1	0.9423	0.9959	0.9243	1

6. Conclusions and future works

This paper shows how to learn ontologies from online semi-structured Chinese corpus and successfully learns a plant ontology with high coverage and good quality from the three largest Chinese online encyclopaedias, Baidu-Baike, Hudong-Baike and Chinese-Wikipedia. Concept, taxonomy, instances and attribute are derived from different kind of structured modules of encyclopaedia pages.

Future works will focus on enriching ontologies by using other corpus such as normal web pages, machine-readable dictionaries, and so on. Moreover, we will leverage machine learning methods in ontology learning.

Acknowledgements

This research is supported by the National Science and Technology Pillar Program of China under grant number 2009BAH46B03.

References

[1] Fellbaum C. WordNet: An electronic lexical database. Boston: The MIT Press; 1998.
[2] Lenat DB. CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM. 1995; 38 (11):33-41.
[3] GOC. Gene ontology: tool for the unification of biology. Nature Genetics 2000; 25:25-29.
[4] Dong Z, Dong Q. Hownet And the Computation of Meaning. World Scientific publishing Co.; 2006.
[5] Wales J. Wikipedia in the free culture revolution. Companion to the 20th annual ACM SIGPLAN 2005.
[6] Maedche A, Staab S. The Text-To-Onto ontology learning environment. Software demonstration at ICCS 2000; 14-18.
[7] Shamsfard M, Barforoush AA. Learning ontologies from natural language texts. Int. J. Hum-Comput. St. 2004; 60:17-63.
[8] Lee CS, Kao YF, Kuo YH, Wang MH. Automated ontology construction for unstructured text documents. Data Knol. Eng. 2006; 60:547-566.
[9] Sánchez D, Moreno A. Creating ontologies from web documents. CCIA 2004.
[10] Tian F, Jiang P, Ren F. A practical system of domain ontology learning using the web for Chinese. ICIW 2009; 58-63.
[11] Shinzato K, Torisawa K. Acquiring hyponymy relations from web documents. HLT-NAACL 2004.
[12] Gregorowicz A, Kramer MA. Mining a large-scale term-concept network from wikipedia. Technical report #06-1028. The MITRE Corp; 2006.
[13] Ponzetto SP, Strube M. Deriving a Large Scale Taxonomy from Wikipedia. AAAI 2007; 1440-1445.
[14] Wu F, Weld DS. Automatically Refining the Wikipedia Infobox Ontology. WWW 2008; 635-644.
[15] Suchanek FM, Kasneci G, Weikum G. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW 2007; 697-706.
[16] Yin X, Shah S. Building Taxonomy of Web Search Intents for Name Entity Queries. WWW 2010.
[17] Paşca M. Weakly-supervised discovery of named entities using web search queries. CIKM 2007.
[18] Pasca M, Turning web text and search queries into factual knowledge: hierarchical class attribute extraction. AAAI 2008: 1225-1230.
[19] Suchanek FM, Kasneci, G, Weikum, G. YAGO - A Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW 2007.
[20] Chu YJ, Liu TH. On the shortest arborescence of a directed graph. Science sinica 1965; 14:1396-1400.
[21] Edmonds J. Optimum branching. J Research of the National Bureau of Standards 1967; 71(B):233-240.