Ontology Extraction and Integration from Semi-structured Data

Shaobo Wang[1], Yi Zeng[1], and Ning Zhong[1,2]

[1] International WIC Institute, Beijing University of Technology, P.R. China
victorwhy@emails.bjut.edu.cn, yizeng@bjut.edu.cn
[2] Department of Life Science and Informatics, Maebashi Institute of Technology, Japan
zhong@maebashi-it.ac.jp

Abstract. Domain ontologies are usually built by domain expert manually. They are accurate and professional from the perspective of domain dependent concepts, instances and relations among them, nevertheless, maintaining and creating new ontologies need too much manual work, especially when the ontology goes to large scale. Semi-structured data usually contain some semantic relations for concepts and instances, and there are many domain ontologies implicitly exist in these types of data sources. In this paper, we investigate automatic hierarchical domain ontology generation from semi-structured data, more specifically, from HTML and XML documents. The main process of our work includes domain terms extraction, pruning, union and hierarchical structure representation. We illustrate our study based on Artificial Intelligence related conference data represented in HTML and XML documents.

1 Introduction

Ontology plays a key role in Artificial Intelligence and the development of the Semantic Web [1]. A large number of ontologies are needed for describing the world wide knowledge in different domains and inferring new knowledge from them [2]. However, domain ontology constructions are usually carried out by domain experts manually, which does not scale well. On the other hand, there are a great many implicit ontologies embedded in the data sources on the Web. How to automatically extract and build ontology from existing information sources like Web pages has been an emerging field of study and an urgent task.

Semi-structured data is widely distributed on the Web, such as HTML Web pages and XML files. These kinds of data sources contain many concepts, instances and relations among them. Hence, semi-structured data has become an important source for automatic ontology learning. Kavalec uses machine learning to obtain the rule of elements mapping automatically [3]. By the pre-definition rule, Doan et al. find the relationship between DTD and concepts, and they build ontologies based on their findings [4]. Mitchell et al. argue the macro-reading of the Web by coupled semi-supervised learning algorithm to populate the ontologies on the Semantic Web [5].

In previous studies, the implicit structures (especially hierarchical relations) within the semi-structured document have not been well investigated for automatic ontology

N. Zhong et al. (Eds.): AMT 2011, LNCS 6890, pp. 39–48, 2011.

construction. In addition, relationships among concepts and instances are distributed in different semi-structured data sources, which need to be merged to build a holistic and more complete ontology. In this paper, based on our previous study, we focus on the hierarchical relationship in the semi-structured document, and we build domain specific hierarchical ontology based on these relationships distributed in different data sources. As an extension to our previous study introduced in [6-8], from methodology perspective, we provide deeper discussions on the construction details for domain ontology integration based on semi-structured data. From implementation perspective, we extend our previous preliminary work on domain ontology construction based on conference proceedings HTML pages and XML files to even larger scale, with involvement of workshop HTML pages to produce even more complete ontology. We choose Artificial Intelligence ontology as a specific one to build from these data sources. Detailed construction process and preliminary results are provided.

2 Domain Concept Extraction

Domain terms play central roles in domain ontology construction. In this section, we mainly discuss the details of domain terms extraction from semi-structured data sources. More specifically, we discuss how to extract domain concepts from conference and workshop proceedings HTML pages and XML files. Although these kinds of data sources are semi-structured, most of the structure related domain concepts are still marked with specific tags, so that we can use them wisely. In the following two sub-sections, two lines of examples are given to illustrate how to extract domain concepts.

2.1 Domain Concept Extraction from Conference Data Sources

In most cases, conference proceedings information is organized in files described by semi-structured markup languages. Conference names usually focus on certain general domains, while session and sub-session names are usually branch topics of general domains. Hence, they naturally contain domain ontologies and can be used as a source for domain concept extraction, and then build hierarchical ontology based on these extracted terms.

Since the DBLP dataset contains most of the conference information and it is publically available [9], we choose its XML version[1] for our investigation. We extract the branch information from "Artificial Intelligence" related conference proceedings. In the dblp_bht.xml file, the URLs of conference series on the topic of AI are assembled together with the label of "<h1>Artificial Intelligence</h1>". And this is where we get the domain name as the most general term for this domain. In the data segments that are corresponding to specific conference record, the label in the form of "<h2>*</h2>" are used to mark the relevant branch topics in the form of session names. Some conference proceedings information even contains sub-session names marked with "<h3>*</h3>". These sub-session names can be considered as even finer concepts for the specific domain. An example of such tags in DBLP dataset is given in Figure 1 [8].

[1] DBLP in XML (`http://dblp.uni-trier.de/xml/`)

```
<h1>18. <ref href="db/conf/ijcai/index.html">IJCAI</ref> 2003: Acapulco, Mexico</h1>
<h2>Learning</h2>
<h3>Clustering and Bayes Net Learning</h3>
```

Fig. 1. An illustrative example of the session branch tags in the DBLP dataset [8]

As shown in Figure 1, "Learning" and "Clustering and Bayes Net Learning" can be extracted as domain concepts to build the Artificial Intelligence ontology. According to the tags in the dblp_bht.xml file, we have extracted all the branch topics which belong to "Artificial intelligence" from the conference information lists. More than 400 session and sub-session names are obtained. We should notice that many terms extracted from conference session names cannot be considered as branch domain concepts, and they need to be filtered out. Our study has collected a list of filtered terms in the process of building the Artificial Intelligence ontology[2].

2.2 Domain Concept Extraction from Workshop Data Sources

Compared to conference names, workshop names usually focus on more specific topics, and the workshop names can be considered as branch fields for more general domains. Different from conference information in the semi-structured data source such as DBLP data, most workshop proceedings information does not contain session and sub-session names. While this kind of information is always available in workshop call for paper (CFP) pages on the Web. They are organized as "Topic of Interests" of workshops. They can be treated as more specific branch topics compared to the topic of the workshops.

In order to obtain the domain concepts related to the workshop, we need to download and analyze the workshop Web pages. Firstly, we can find workshop titles and links from co-located conference website. Here we choose two workshops co-located with IJCAI[3,4]. In most cases, workshop pages provide co-located conference names, and by using these kinds of information, the domain concepts embedded in the workshop title can find their super concepts that are coarser than them. Figure 2 presents two source code segments of the ITWP 09 workshop[4], and we can clearly find the conference and workshop titles as well as the domain terms embedded in them. Compared to the organization in the DBLP dataset, the phrase after "conference on" is the root node which can be tagged as <h1>, while the terminology after "workshop on" is the domain term that summarize the workshop focus, which can be tagged as <h2>, as shown in Figure 1.

After extracting domain terms from workshop titles, a step forward need to be taken for extracting finer concepts from the workshop page. Almost all the workshop pages contain sections titled "Topic of Interests". We observe that most topics of interests are tagged in the form of * and wrapped by ordered lists (tagged

[2] Filtered words for building the Artificial Intelligence ontology (http://www.wici-lab.org/wici/dblp-sse/Filterwords.txt)

[3] The 2003 Workshop on Information Integration on the Web (http://www.isi.edu/integration/workshops/ijcai03/iiweb.html)

[4] The 7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems (http://www.dcs.warwick.ac.uk/~ssanand/itwp09/)

with <ol>) or unordered lists (tagged with <ul>). Hence, domain concepts that are finer than the workshop topics can be extracted from the lines marked with <li>. Figure 3 presents an example on the source code level organization on the "Topics of Interests".

```
<p align=center style='text-align:center'><b>Held in conjunction with<br>
    <a href="http://ijcai-09.org/" target="_blank"><span style='color:windowtext'>The
    Twenty-first International Joint Conference on Artificial Intelligence
    (IJCAI-09)</span></a><o:p></o:p></b></p>
```

(a) Domain Concept Extraction from a Conference Title

```
<meta http-equiv="content-type" content="text/html; charset=ISO-8859-1">
<title>IJCAI Workshop on Intelligent Techniques for Web Personalization and
Recommender Systems - ITWP 2009</title>
```

(b) Domain Concept Extraction from a Workshop Title

Fig. 2. Title information in the source codes of the ITWP 2009 workshop Web page[4]

```
<UL>
  <LI>Information extraction</LI>
  <LI>Wrapper learning</LI>
  <LI>Automatic wrapper generation</LI>
  <LI>Information gathering</LI>
  ...
</UL>
```

Fig. 3. Topic information in the source codes of the ITWP 2009 workshop Web page[4]

We notice that except for the topics of interests, some other messages which are irrelevant for domain ontology construction are also marked with <li>, like submission deadline, author notification date, etc. Luckily, we find that these code segments always contain time and number related information, while branch topics seldom use them. Hence, we can filter out these information to avoid irrelevant concepts being included in the domain ontology.

The domain concepts extracted from workshop Web pages are also organized together as a level of domain concepts which are finer than the concept embedded in the workshop titles. They will be used to produce finer levels in the domain ontology construction phase, together with the domain concepts extracted from conference proceeding information in Section 2.1.

3 Domain Ontology Constructions and Optimization

After the domain concept extraction process, we need to build and represent hierarchical ontologies from single sources first, then these distributed ontologies are integrated into a holistic one. In this section, the ontology construction and representation process is investigated, then the ontology integration and optimization issues are discussed in detail.

3.1 Hierarchy Representation

Based on the discussion in Section 2, we have collected several hundreds of branch domain concepts of Artificial Intelligence from Web pages of both conferences and workshops. Because there is no explicit representation of ontology structures in the original semi-structured data sources, we need to make the structure explicit and describe these structures by knowledge representation languages. Relevant tags discussed in the previous section are the basis for the hierarchies.

We can get the hierarchical relationship among domain concepts according to the tags in the semi-structured data sources. Extracted domain concepts originally tagged with <h2> belong to their super class labeled with <h1>, while they are direct super class of the concepts originally marked with <h3>. For example, "Knowledge Representation and Reasoning" is a sub-concept of "Artificial Intelligence", while it also has "Description Logic" as its branch field. These three domain concepts constitute a partial knowledge structure of AI. In this paper, we use RDF and OWL to represent ontologies. The sub-class relation is described by using the predicate "rdfs:subClassOf". Figure 4 illustrates how the partial structure is represented [8].

```
<owl:Class rdf:about="http://www.wici-lab.org/wici/web-kr3/terminology/AI#DL">
   <dc:title>Description Logic</dc:title>
   <rdfs:subClassOf rdf:resource="http://www.wici-lab.org/wici/terminology/AI#KRR">
     <dc:title>Knowledge Representation and Reasoning</dc:title>
   </rdfs:subClassOf>
   <rdfs:subClassOf rdf:resource="http://www.wici-lab.org/wici/terminology/AI">
     <dc:title>Artificial Intelligence</dc:title>
   </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="http://www.wici-lab.org/wici/web-kr3/terminology/AI#KRR">
    <dc:title>Knowledge Representation and Reasoning</dc:title>
    <rdfs:subClassOf rdf:resource="http://www.wici-lab.org/wici/terminology/AI">
      <dc:title>Artificial Intelligence</dc:title>
    </rdfs:subClassOf>
</owl:Class>
```

Fig. 4. A Partial Hierarchical Knowledge Structure Representation from IJCAI 2001[5] [8]

For topics of interests tagged with <li> and extracted from workshops, they are organized as branch concepts finer than the domain concepts extracted from workshop titles. Hence, each workshop forms a hierarchy.

We connect workshop titles with the ontology structure generated from conferences in the following way: If the domain concept appeared in the existing structure, then the workshop sub-structure is connected to the conference ontology directly, with the matched concept as the connection point. If the domain concept does not appear in the conference ontology, then it is connected directly to the root node (in our example, the root node is "Artificial Intelligence"). Figure 5 presents an example of the partial structure generated from the IIWeb 2003 workshop[3].

[5] IJCAI 2011 Proceeding Information from DBLP: http://www.informatik.uni-trier.de/~ley/db/conf/ijcai/ijcai2001.html

```
<owl:Class rdf:about="http://www.wici-lab.org/wici/web-kr3/terminology/AI#SML">
   <dc:title>Source meta-data learning</dc:title>
   <rdfs:subClassOf rdf:resource="http://www.wici-lab.org/wici/terminology/AI#IIWeb">
     <dc:title>Information Integration on the Web</dc:title>
   </rdfs:subClassOf>
   <rdfs:subClassOf rdf:resource="http://www.wici-lab.org/wici/terminology/AI">
     <dc:title>Artificial Intelligence</dc:title>
   </rdfs:subClassOf>
</owl:Class>
```

Fig. 5. A Partial Hierarchical Knowledge Structure Representation from IIWeb 2003[3]

3.2 Ontology Pruning and Union

The previous section introduces hierarchical ontology generation based on single sources. After this step, integration of different sub-ontologies into a holistic structure is needed for generating relatively complete domain ontology. In this paper, we focus on the concept duplication and the level division conflict. Four types of concepts duplication are analyzed in level division conflicts and ontology pruning solutions are provided to each of them. Here the shadowed circle is used to denote the redundant keyword, while the blank ones to denote the normal nodes.

The first situation: The same concept *C* appears twice in the same structure and they have direct "sub-class of" relationship, they are distributed in the *n*th and the *n'*th level (*n' = n+1*), as shown in Figure 6. In this case, the concept in the *n'*th level is deleted, and relevant sub-concepts of *C* in the *(n'+1) th level (if any)* are assigned as direct sub-classes of the concept *C* in the *n*th level.

For example, in the IJCAI 2001 conference proceedings[6], the domain concept "Diagnosis" appears twice in the session organization hierarchy for some reason. In this case, we delete the "Diagnosis" in the third level, and we keep the one in the second level.

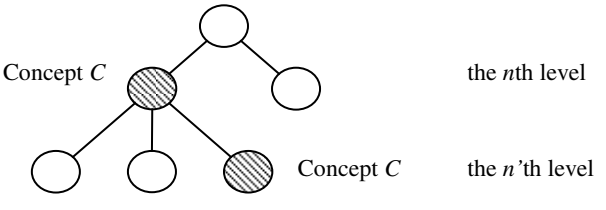


Fig. 6. Situation I: Concept duplication with direct hierarchical relation

The second situation: Two sub-structures that contain the same root node *C* and from two sources share the same super class, as shown in Figure 7. In this case, the root node *C* of these two sub-structures need to be combined together as one, and sub-concepts of the node *C* from the two sub-structures need to be combined together.

[6] Proceedings of IJCAI 2011 information:
(http://www.informatik.uni-trier.de/~ley/db/conf/ijcai/ijcai2001.html)

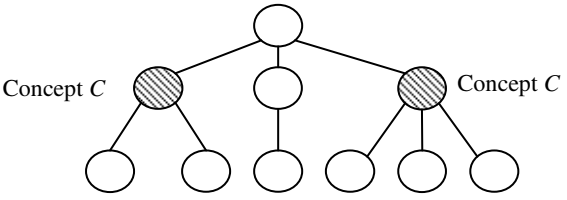


Fig. 7. Situation II: Concept duplication with shared super class

From the representation perspective, we use OWL to represent the union process of the sub-structures. "owl:unionOf" is used to denote that the new structure is based on the union of two existing sub-structures. Figure 8 presents an illustrative example. Two sub-structures of Artificial Intelligence are from IJCAI 2011 and IJCAI 2009. They are identified by different URLs. Each of them contains a sub-structure of AI. By using owl:unionOf, we combine these two sub-structures as a whole, and a new URL is used to identify the new and more complete structure.

```
<owl:Class rdf:about="http://www.wici-lab.org/wici/web-kr3/terminology/AI">
  <owl:unionOf rdf:parseType="Collection">
    <owl:Class rdf:about="http://www.wici-lab.org/wici/terminology/AI#AI-IJCAI2001">
      <owl:oneOf rdf:parseType="Collection">
        <owl:Thing rdf:about="http://www.wici-lab.org/wici/terminology/AI#MAS" />
        <owl:Thing rdf:about="http://www.wici-lab.org/wici/terminology/AI#KRR" />
        <owl:Thing rdf:about="http://www.wici-lab.org/wici/terminology/AI#CS" />
        <!-- many more -->
        </owl:oneOf>
    </owl:Class>
    <owl:Class rdf:about="http://www.wici-lab.org/wici/terminology/AI#AI-IJCAI2009">
      <owl:oneOf rdf:parseType="Collection">
        <owl:Thing rdf:about="http://www.wici-lab.org/wici/terminology/AI#KRR" />
        <owl:Thing rdf:about="http://www.wici-lab.org/wici/terminology/AI#CM" />
        <owl:Thing rdf:about="http://www.wici-lab.org/wici/terminology/AI#Ga" />
        <!-- many more -->
      </owl:oneOf>
    </owl:Class>
  </owl:unionOf>
</owl:Class>
```

Fig. 8. Representation on the union of sub-structures (MAS: Multi-agent Systems, CS: Constraint Satisfaction, KRR: Knowledge Representation and Reasoning, CM: Cognitive Modeling, Ga: Game)

The third situation: The concept *C* appears in the same level twice, but they have different direct super class *A* and *B*, as shown in Figure 9. In this case, we keep the concept *C* in both levels, as well as their sub-concepts (if any), since in ontology engineering, it is allowed that one concept may be direct subclass of different concepts [10].

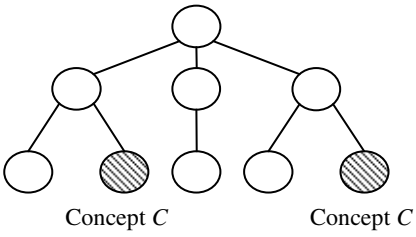


Fig. 9. Situation III: Concept duplication with different super class

The fourth situation: The redundancy occurs in not only different branches but also different levels, as shown in Figure 10.

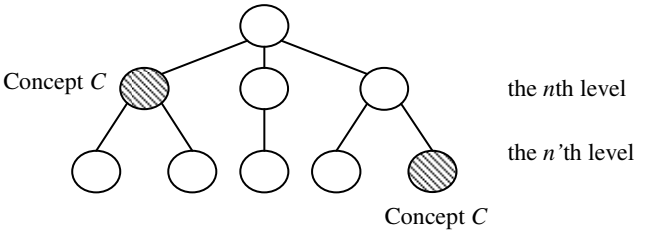
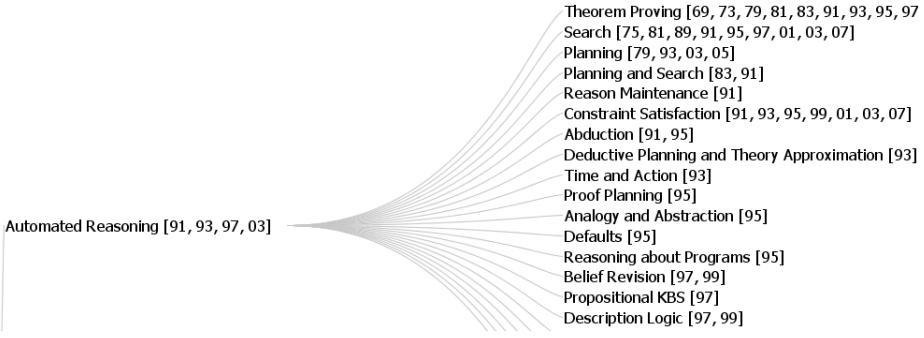


Fig.10. Situation IV: Concept duplication with different super class and located in different levels

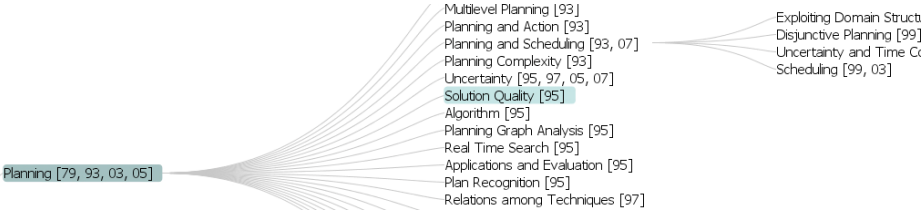
In this case, we make a statistical analysis of where the specified concept appears more times. Let *C* be a domain concept and *f(C,n)* denotes the times of *C* appears in the *n*th level, while *n* and *n'* be two arbitrary levels that contain the concept *C*. If *f(C,n)< f(C, n')*, then *C* will be in the *n'*th level and the one in the *n*th level will be deleted. In addition, sub-concepts and corresponding relation of *C* in the *n*th level need to be moved to and integrate with those of *C* in the *n'*th level. If *f(C,n)= f(C, n')*, and the concept in the *n'*th level is from workshop proceedings, and the concept in the *n*th level is from conference proceedings, then the concept *C*'s location from the *n*th level will be kept, and the one from workshop proceedings will be removed. If *f(C,n)= f(C, n')*, and the concepts in both levels are from the same type of sources (i.e. both of them are from conference proceedings or workshop proceedings), then a level is randomly selected from *n* and *n'*, and the concept in the selected level is kept, while the other one is deleted.

Following the method introduced in this section, we analyzed all proceedings related to "Artificial Intelligence" in the DBLP dataset, including 235 conferences in 14 AI related conference series [8], as well as a few workshops. We build a five-leveled hierarchical ontology on the topic of "Artificial Intelligence" based on the extracted domain concepts as well as their relationships. A visualized interactive ontology of Artificial Intelligence has been built based on the Prefuse toolkit[7], as shown in Figure 11.

[7] Prefuse: an interactive information visualization toolkit (`http://prefuse.org/`)



(a) A partial structure of levels 2 and 3 in the Artificial Intelligence Ontology



(b) A partial structure of levels 3, 4 and 5 in the Artificial Intelligence Ontology

Fig. 11. A partial example of the "Artificial Intelligence" visualized ontology

4 Conclusion and Future Work

In order to make the implicit ontologies on the Web explicit and integrate them together, in this paper, we provide an approach of building domain ontology hierarchy from semi-structured data such as XML and HTML files. We mainly divide the whole process into two steps, namely, domain Concept Extraction as well as ontology pruning and union. Implicit ontologies in the conference and workshop information are selected for investigation. More specifically, we select XML version of the DBLP data set and several workshop Web pages as our data sources. A domain ontology in the filed of Artificial Intelligence based on multiple distributed semi-structured data sources has been built based on the approach introduced in this paper.

In this paper, we have introduced concept duplication and relevant pruning methods in four different types of situations. Nevertheless, we did not investigate how to handle semantically similar concepts in ontology integration. In the future work, we are going to investigate on this direction.

Acknowledgments. This study is supported by China Postdoctoral Science Foundation (20110490255), Beijing Postdoctoral Research Foundation (2011ZZ-18), and the Large Knowledge Collider (LarKC) Project (FP7-215535) under the European Union 7th framework program.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 5, 34–43 (2001)
2. Kerrigan, M.: WSMOViz: An Ontology Visualization Approach for WSMO. In: Proceedings of the 10th International Conference on Information Visualisation, pp. 411–416 (2006)
3. Kavalec, M., Svátek, V.: A Study on Automated Relation Labelling in Ontology Learning. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2005)
4. Doan, A., Domingos, P., Levy, A.: Learning Source Descriptions for Data Integration. In: Proceedings of the Third International Workshop on the Web and Databases, pp. 81–86 (2000)
5. Mitchell, T.M., Betteridge, J., Carlson, A., Hruschka,Jr., E.R., Wang, R.C.: Populating the Semantic Web by Macro-reading Internet Text. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 998–1002. Springer, Heidelberg (2009)
6. Zeng, Y., Zhong, N.: On Granular Knowledge Structures. In: Proceedings of the 1st International Conference on Advanced Intelligence, pp. 28–33 (2008)
7. Zeng, Y., Zhong, N., Wang, Y., Qin, Y., Huang, Z., Zhou, H., Yao, Y., van Harmelen, F.: User-centric Query Refinement and Processing Using Granularity Based Strategies. Knowledge and Information Systems 27(3), 419–450 (2011)
8. Zeng, Y.: Unifying Knowledge Retrieval and Reasoning on Large Scale Scientific Literatures. PhD thesis, Beijing University of Technology (2010)
9. Ley, M.: DBLP - Some Lessons Learned. Proceedings of the VLDB Endowment 2(2), 1493–1500 (2009)
10. Antoniou, G., van Harmelen, F.: A Semantic Web Primer, 2nd edn. The MIT Press, Cambridge (2008)