An Information Extraction Approach for Building Vocabulary and Domain Specific Ontology in Information Technology

Ta Cong Duy Chien, Phan Thi Tuoi

Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology
268 Ly Thuong Kiet, Ho Chi Minh City, Vietnam

Email: tuoi@cse.hcmut.edu.vn

***Abstract* - This paper introduces the HETEONTO system that extracts concepts from text files and existing ontologies such as Wikipedia, ACM, and WordNet in order to build the vocabulary and domain specific ontology focusing on Information Technology domain. It also describes how to import concepts into domain ontology. Data sources and techniques deployed in HETEONTO for ontology learning from texts, Wikipedia, WordNet are briefly presented herein. This paper then focuses on evaluating the generated domain ontology by selected methods. One of these methods that we introduce here, is comparative. Comparative evaluation performed in this study use the same corpus to contrast result from HETEONTO. Results generated by such experiments show that HETEONTO yields superior performance, especially regarding semantic relation.**

Keywords—Domain ontology, Knowledge based system, Ontology Evaluation, Information Technology Ontology, Information Extraction

I. INTRODUCTION

Ontologies are applied in many applications in recent years, especially in the fields Sematic Web, Information Retrieval, Information Extraction, or Question and Answer [HYPERLINK \l "RPo10" 1]. It is believed that, ontologies will play an important role in the future of Artificial Intelligent and Natural Language Processing systems. Ontologies can act as a common and reusable knowledge base in applications for multiple purposes. They make such applications adhere to the domain knowledge view expressed in the ontology.

The entities, which represent concepts in the domain ontology, are always updated 2] }. These entities should be evolved with new discovery and usages. In order to updating ontologies with such advances, automatic methods should be used since manual methods are not suitable for dynamic nature of ontologies. It takes time and effort for conceiving and building ontology by manual methods [HYPERLINK \l "MAS09" 3], 4]}.

In order to reduce and prevent these drawbacks, automatic methods for domain ontology building must be adopted. There are many resources such as text files, images; available ontologies can be used as sources of knowledge base. Since the ontology is built from text documents and available ontologies, it is possible to explain the presence of a particular concept, property, instance or attribute. Hence, the updating the instances of ontologies from texts will help retain somewhat of a semantic validity by providing the means to refer to the original texts [HYPERLINK \l "AZo09" 5].

However, automatic or semi-automatic methods for domain ontology building are still restricted 6] }. Automatic knowledge extraction techniques can only provide domain ontology skeletons and more complex building steps still require the intervention of human [HYPERLINK \l "MFM06" 7]. Another issue within the ontology community is relevant to the lack of methodologies to evaluate ontologies 5]}

This paper presents steps to build an ontology for Information technology domain, called Information Technology Ontology (ITO) and evaluates it. This domain ontology provides a solution for the aforementioned issues by generating semi-automatic domain ontologies from texts and available ontologies such as ACM Digital Library, Wikipedia and WordNet. We choose ACM Digital Library since our purpose focus on information technology domain. We also use Wikipedia since it is free Internet encyclopedia, multilingual and very popular in the recent years. Besides, this paper also recommends a

clear evaluation methodology based on traditional and comparative methods. It is organized as follows: Section 2 describes related work; Section 3 details the steps to build and enrich automatically domain ontology; Section 4 focuses on a methodology to evaluate the generated ontology. In this section, we introduce a comparative measure with one of the most advanced systems in building domain ontology: TerMine. Finally, Section 5 discusses the results of the evaluation methodology.

II. RELATED WORK

Several approaches have been proposed for building and evaluating domain-specific ontology. Trucao et al [HYPERLINK \l "Tru07" 8] designed and constructed of VN-KIM ontology, which was mainly on particular concepts of Vietnam regarding to its politic, economic and social situations M.A Salahli et al 3] } built domain-specific ontology based on WorldNet's database and consisted of Turkish and English terms on computer science and informatics. Besides, several approaches focused on to solve semantic relations between concepts of ontology. A.Weichselbraun et al [HYPERLINK \l "AWe11" 9] used unstructured documents, social evidences and structured evidences for ontology learning regarding semantic relations between ontology's concepts. W. Sun et al 10]} pointed to how to obtain the correct semantic relations of ontology's concepts. A.G. Valarakos et al [HYPERLINK \l "AGV" 2] use a novel name matching method based on machine learning regarding semantic relations between instances.

However, the above researches did not mention how to refer the synonym of these ontology's concepts and how to enrich ontologies by available resources. Otherwise, those researches did not also present how to combine the existing ontologies such as WordNet, Wikipedia and ACM. This paper introduces an approach combining Wikipedia, WordNet and ACM to construct domain ontology on Information Technology, which cover many different topics in this area. Besides, the authors propose several algorithms to find out synonyms of concepts, to extract sentences regarding semantic relation of concepts. These algorithms combine Natural Processing Language with Machine Learning and Statistic method.

III. BUILDING AND ENRICHING INFORMATION TECHNOLOGY ONTOLOGY (ITO)

A. How to Build ITO?

Some concepts should be defined first.

Definition 1: Entity relation is a relation between two or more entities that is detected and the relation possibly is typed with a semantic or syntactic role. Entities can be nouns or compound nouns.

Definition 2: An instance represents an entity. An entity represents a concept, which belongs to Information technology domain. In this ITO, instances can belong to topic layer, ingredient layer and synonym layer.

Building up a domain specific ontology usually requires overcoming many obstacles 1] } [HYPERLINK \l "MAS09" 3]because the data has been gotten from different resources. Therefore, how to consolidate data is a big problem 11]}. In this paper, we present a method of building and enriching ITO. We propose a structure of ITO, as shown in Fig. 1.

In this figure, ITO includes four layers: topic layer, ingredient layer, synonym layer, and sentence layer. We will explain in further detail how to build each layer in ITO.

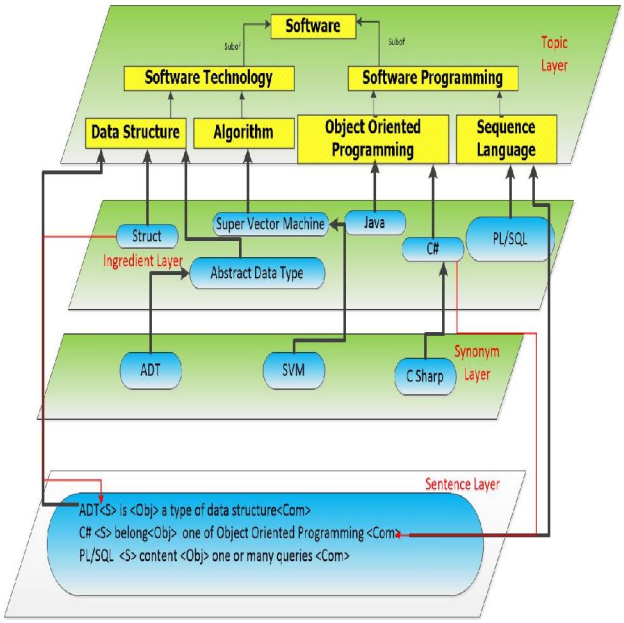


Figure 1. Domain Specific Ontology on Information.

The first layer is known as the topic layer. It is named for topic layer since it includes many instances that represent different categories of Information Technology domain. These instances are extracted from ACM Categories [HYPERLINK \l "ACM" 12]. An instance in this layer is representative one topic. After extracting from ACM, we have over 245

different topics. A small part of the hierarchical tree presents for these instances, as shown in Fig. 2.

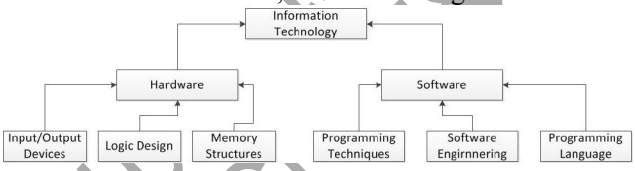


Figure 2. The hierarchical tree of topics in Topic layer.

The mapping from an instance in this layer to knowledge structure is represented by XML as shown below

<TOPICS>
<TOPIC>
<NAME> Information technology </NAME>
<ID> 0000001</ID>
<LEVEL> 1 </LEVEL
<SUBOF> 0 </SUBOF>
</TOPIC>

We define a relation "SUBOF' between the instances in low and high level as below

SUBOF (Hardware, Information technology)
SUBOF (Input/output Devices, Hardware)
SUBOF (Mainboard, Devices)

We have a hierarchical tree of topics in this layer based on this relation. It allows specifying far more about the properties of instances. For example, it can indicate that "if instance A belong to topic B and SUBOF(C, B) then this implies A belong also to topic C".

Next layer is known as the ingredient layer. It is named for ingredient layer since it includes the vocabulary about Information Technology domain, e.g., "robot", "Super Vector Machine", "Local Area network", "wireless", "UML", etc. The vocabulary is extracted from an available ontology, namely Wikipedia. Wikipedia is an ontology, which includes various areas and many different languages. However, we focus only on English language and Information Technology domain. In order to extract terms from Wikipedia with our target, we use Java-based Wikipedia Library (JWPL). We propose a processing model, as shown in Fig. 3.

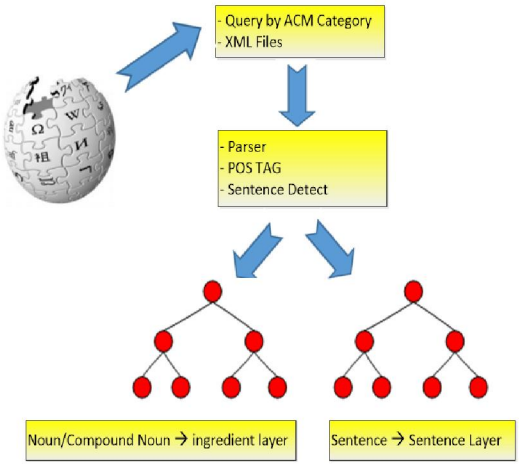


Figure 3. Model extracting terms from Wikipedia to build ITO

The mapping of an instance in ingredient layer to knowledge structure is presented by XML as below

<Member_list>
[<member>
<memberID>101</memberID>
<value>Robot</value>
<IGs>0.8743, 0.671, 0.242</IGs>
<categoryIDs>102,103, 104
</categoryIDs>
<href>Wikipedia</href>
</member>…]
</Member_list>

Since one instance of ingredient layer can have one or many topic, so value of "categoryIDs" and <IGs>tag can be greater than one.

When extracting lexical terms from Wikipedia, in order to ensure that the quantity of this ontology is suitable with target above, we use a statistical method to evaluate these items. Information Gain 13] } [HYPERLINK \l "ALB96" 14] is used in this case. It is calculated as below.

$$IG(a) = E(B - a) - E(a) \quad (1)$$

$$E(a) = -\sum_{j=0}^{C-1} (P_j \log_2 P) \quad (2)$$

Where

- E (*a*): Entropy of attribute "*a*" in B.
- E (*B - a*): Entropy of all attributes in B after "*a*" is deleted from B.
- Pj: Probability Distribution of attribute "*a*" in B.

To solve our problem, we propose an Information Gain (IG) formula (1) as follow:

$$IG(a|C_i) = E(X|C_i) - E(a) \quad (3)$$

Where

• $IG(a|C_i)$: Information Gain of "a" in category C_i

• $E(X|C_i)$: Entropy of all attributes in category C_i after "a" is deleted from C_i

With IG = 7.6, the precision of instances in this layer is approximately 85%.

The third layer of ITO is known as the synonym layer. To setup instances of this layer, we use the second ontology available: WordNet. WordNet is a lexical ontology that includes many fields. To find out synonyms corresponding with one instance of the ingredient layer, we propose an algorithm as below

```
Procedure Find_out_Sysnonym()
  While (instance of Ingredient layer is not null)
  Begin
        Synonym_List = WordNet_query (instance)
        If (Sysnonym_List is not null)
          Link (instance to Sysnonym_list)
        End if
  End
End While
End Pro
```

The last layer of ITO is known as the sentence layer. Instances of this layer are sentences that are extracted from Wikipedia while extracting terms of ingredient layer (Fig. 3).

These sentences present the semantic relationships between words in sentences. Hence, most of sentences in this layer are linked to one or many items in ingredient layer. Finding out the semantic relationships between items of the ingredient layer and storing them in the sentence layer plays an important role since we reuse this to build Information Extraction System in the future. However, to extract easier sentences in Wikipedia, we apply a number of English patterns, as shown in Table 1.

Table 1: English Pattern to extract sentences

No	English Pattern	Example
1	S + V	Computer is broken
2	S + V + Object	He was mowing the lawn
3	S + V + Adjective	The girl was tall
4	S + V + Indirect Object	The woman went to the house
5	S + V + Direct Object	The man hit the ball
6	S + V + Complement	Some students in the class are engineers
7	S + V + Prepositional Phrase	The cat waited for its owner yesterday
8	S +V+ Indirect Object + Direct Object	Granny left Gary all of her money
9	S + V + Direct Object + Adjective	The Jury found the defendant guilty
10	S + V + Direct Object + NP	The Jury found the defendant guilty
11	S + V + Direct Object + Complement	The class picked Susieclass representative.

We proposed an algorithm to extract sentences based on these patterns.

```
Procedure Extract_Sentence()
    ReadXMLFile(categoryID);
    Detection(noun/compound Noun)
    SentenceList = Detection(Sentence)
    While (Senetence: SentenceList is not
null)
        If (match(sentence,pattern))
            Sentence = Extract sentence;
            Put sentence into Sentence
layer;
        End if;
  End while;
End Pro;
```

B. How to Enrich ITO?

Since entities of ontology should be updated, hence enriching ontology plays an important role. We comprise a hybrid methodology including Natural Language Processing (NLP) and statistic for enriching ITO. We use also 20,000 papers from ACM Digital

Library for this task. Most of layers of ITO such as Ingredient, synonym, and sentence layer are updated. We propose a model to enrich ITO from text document, as shown in Fig. 4.

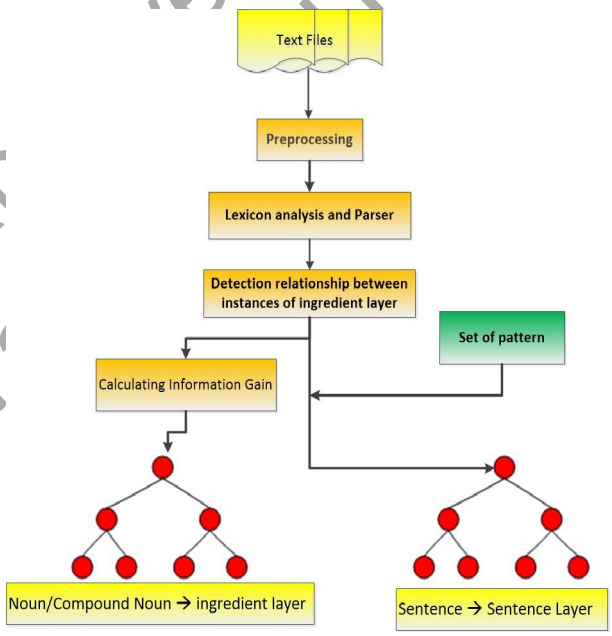


Figure 4. Model of enriching ITO from ACM Digital Library

Keep in mind that we reuse Information Gain factor to drop out the bad instances while putting them into ingredient layer.

IV. Evaluating Domain Ontology on Information Technology

The building domain specific ontologies based on heterogeneous data sources will face a number of challenges, such as consistency, disambiguation, and semasiology. The quantity of domain ontology plays an important role since they are used in the overall system and they can be reused in many different applications. This session will address the evaluation of the domain ontology generated through the models and algorithms as mentioning above. We propose two methods:

- The first is based on three measures: Precision, Recall and F-measure.
- The second uses the comparative method [15].

A. The Evaluation based on Three Measures

HETEONTO's performance has been measured using three measures: Precision, Recall and F-measure [8]. They are calculated by each category in ITO as below:

$$P(C_i) = \frac{Correct(Ci)}{Correct(Ci)+Wrong(Ci)} \quad (4)$$

$$RC_i = \frac{Correct(Ci)}{Correct(Ci)+Missing(Ci)} \quad (5)$$

$$F-Mesure = 2\frac{Precision*Recall}{Precision+Recall} \quad (6)$$

Where C_i represents a category in ITO and correct, wrong, missing represent the number of correct, wrong, missing, respectively.

B. Comparative Method

HETEONTO's performance is evaluated again with other method. We use TerMine[16], that is a term extract tool from text files, PDF files or URL In order to compare, three following testing corpora are used:

One corpus with 150 papers from ACM digital library. They belong to Operating System category.

One corpus with 200 papers from ACM digital library. They belong to Network Architecture and Designcategory.

One corpus with 200 papers from ACM digital library. They belong to Artificial Intelligent category.

C. Experimental Result

Table 2, Table 3 show the testing results from two methods above, respectively. We pick three categories at random at the topic layer for illustration.

Table 2: Testing over Wikipedia and ACM Digital Library

Category	Precision (P)	Recall (R)	F-Measure (F)
Operating System (OS)	76.21%	71.42%	73.73%
Network Architecture and Design (NA&D)	78.1%	72.74%	75.52%
Artificial Intelligent (AI)	84.03%	76.62%	80.15%

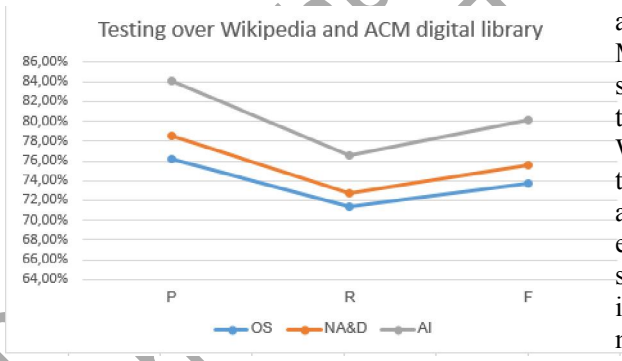
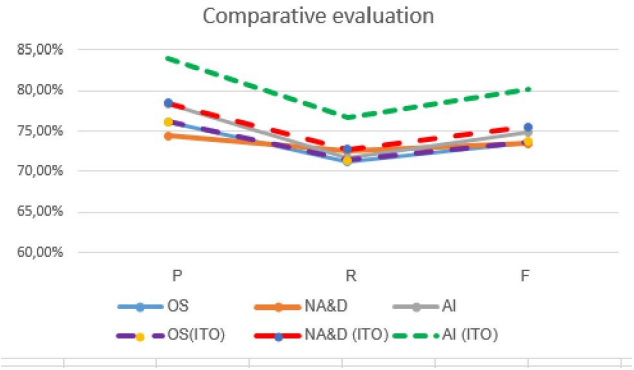


Table 3: Testing over 3 corpora with TerMine

CATEGORY	PRECISION (P)	RECALL (R)	F-MEASURE (F)
Operating System (OS)	76.14%	71.21%	73.59%
Network Architecture and Design (NA&D)	74.42%	72.53%	73.46%
Artificial Intelligent (AI)	78.37%	71.61%	74.83%



The scores reported in table 2 and table 3 reveal that our approach for extracting terms from Wikipedia and ACM Digital Library in order to build domain ontology outperforms the TerMinetool.

V. CONCLUSIONS

This paper presents the procedures to build a domain specific ontology on Information Technology generated through HETEONTO, and ontology evaluations. The ontology is built based on ACM Digital Library, Wikipedia and WordNet. This paper also proposes a composed methodology including Machine Learning, NLP and statistics. Additionally, since data are collected from distinct sources, such as text files of ACM Digital Library, Wikipedia and WordNet, we have over 700,000 distinct instance data that belong to information technology domain. That is an advantage of our IE system. Besides, we also ensure the semantic consistency of instances in this system. The overall evaluation of IE system implemented is based on Precision and Recall measures. A comparative evaluation using TerMine tool was also proposed. The experimental results show that our approach is outperformed.

There is no single best or preferred approach to ontology evaluations: the choice of a suitable approach must reflect the purpose of the evaluation, the application in which the ontology will be used [17]. In the future work, we will focus particularly on automated ontology evaluations and how to detect automatically semantic relationship between concepts, the necessary perquisites for the fruitful development of automated ontology processing techniques in order to solve a number of problems, such as ontology learning and matching.

REFERENCES

[1] R.Poli et al, *Theory and Applications of Ontology*, Michael Healy, and Achilles Kameas Roberto Poli, Ed.: Springer, 2010.

[2] A.G. Valarakos et al. Institute of Informatics and Telecommunications. [Online]. http://www.iit.demokritos.gr/

[3] M.A.Shalalhi et al, "Domain Specific Ontology on Computer Science," in *Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009. Fifth International Conference.*, Famagusta, 2009, pp. 1 - 3.

[4] N.R.Milton, *Knowlegde Acquisition in Practice.*: Springer, 2007.

[5] A. Zouaq et al, "Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project," *IEEE Transactions on Knowledge and Data Engirneering*, vol. 21, no. 11, pp. 1559 - 1572, Nov. 2009.

[6] H. Chu-Ren, *Ontology and the Lexicon*, 1st ed., HUANG Chu-Ren and Calzolari Nicoletta, Eds.: Cambridge University, 2010.

[7] M F Moens, *Information Extraction: Algorithms and Prospects in a Retrieval Context*, W.Bruce Croft, Ed. Belgium: Springer, 2006.

[8] Tru H. Cao et al, "VN-KIM IE: Automatic Extraction of Vietnamese Named Entities on the Web," *New Generation Computing*, vol. 25, no. 3, pp. 277-292, january 2007.

[9] A.Weichselbraun et al, "Evidence Sources, Methods and Use Cases for Learning Lightweight Domain Ontologies," in *Ontology learning and knowledge discovery using the Web: challenges and recent advances*, W.Wong et al, Ed.: Information Science Reference , 2011, ch. 1, pp. 1 - 15.

[10] W. Sun et al, "The 6th International Conference on Fuzzy Systems and Knowledge Discovery," in *Fuzzy Systems and Knowledge Discovery, 2009. FSKD* , Chongqing, China, 2009, pp. 415 - 418.

[11] Randy Goebel, "Trends In Applied Intelligent Systems," , Cordoba, Spain, 2010, pp. 535 - 575.

[12] ACM. [Online]. http://www.acm.org/about/class/ccs98-html

[13] Chien D C Ta, Tuoi Phan Thi, "Improving the Formal Concept Analysis Algorithm to Construct Domain Ontology," in *The 2012 Fourth International Conference on Knowledge and System Engirneering - KSE 2012*, Da Nang, VietNamese, 2012, pp. 74 - 79.

[14] A L. Berger et al., "Maximum Entropy Approach to Natural Language Processing," *Association for Computational Linguistics*, vol. 22, 1996.

[15] F. Gargouri et al, *Ontology Theory, Management and Design: Advanced Tools and Model*, Jamie Snavely, Ed.: IGI Global, 2010.

[16] The National Centre for Text Mining. [Online]. http://www.nactem.ac.uk/software/termine/

[17] G. Flouris et al, "Ontology change: classification and survey," *The Knowledge Engineering*, vol. 00, no. 0, pp. 1 -29, 2007.

AUTHORS' BIOGRAPHIES

Ta Cong Duy Chien is Ph.D. Student at the Faculty of Computer Science and Engineering, HCMC University of Technology, Vietnam. His research interests include Natural Language Processing and its applications, Information Extraction

Phan Thi Tuoi is a Professor at the Faculty of Computer Science and Engineering, HCMC University of Technology, Vietnam. She obtained her Ph.D. in Computer Science from Charles University, Czech Republic, in 1985. Her research interests are compiler, information retrieval, natural language processing. She has been the Chief Investigator of national key projects and published many papers in international journals and conference proceedings in those areas