Journal of Computer Science and Cybernetics, V.31, N.2 (2015), 149–158
DOI: 10.15625/1813-9663/31/2/3971

BUILDING ONTOLOGY BASED-ON HETEROGENEOUS DATA

TA DUY CONG CHIEN AND PHAN THI TUOI

Faculty of Computer Science and Engineering, HoChiMinh City University of Technology;
chientdc@cse.hcmut.edu.vn; tuoi@cse.hcmut.edu.vn

Abstract. Ontologies play an important role in the distinct areas, such as information retrieval, information extraction, question and answer. They help us in capturing and storing knowledge in a particular domain and can be used for distinct applications. In recent years, research relevant to ontology development has produced tangible results concerning semantic web, information extraction, etc. In this paper, a domain specific ontology called Information Technology Ontology (ITO) is proposed. This ontology is built basing on three distinct sources of Wikipedia, WordNet and ACM Digital Library. An information extraction system focusing on computing domain based on this ontology in the future will be built. In order to have an ontology with highest quality and performance as expected, the authors combine some algorithms between machine learning and natural language processing (NLP) for building ontology. Results generated by such experiments show that these algorithms outperform others, especially in semantic relations among entities of ontology.

Keywords. Domain ontology, information extraction, natural language processing.

1. INTRODUCTION

Building ontology is a necessary task for application domain relevant to artificial intelligent, semantic web, information extraction, etc. Ontologies are the structural framework for organizing information. They allow users to find and request complex data from distinct applications. Over the years, knowledge engineering research has been focusing on the development of theories, methods, algorithms, and software tools, which aid human to acquire knowledge in computer. They use scientific and mathematical approaches to discover the knowledge [1].

Ontology modeling in computer system, called computational ontology, is rather simpler than that in philosophy. It provides a symbolic representation of knowledge objects, classes of objects, properties of objects, and the relationships among objects to explicitly represent knowledge about an application domain [2]. Thereby, many ontologies have been built by research with different purposes. In recent years, researchers have trended to the use of ontologies for building applications relevant to information retrieval, information extraction and question answering systems. Tru H.Cao et al. [3] designed and constructed VN-KIM ontology, focusing on particular concepts of Vietnam in its politic, economic and social situations. M.A Salahli et al. [4] built domain-specific ontology basing on WorldNets database and consisting of Turkish and English terms on computer science and informatics.

However, the above mentioned research does not mention how to refer the synonym of these ontologys concepts and how to enrich ontologies. Furthermore, the research also does not regard how to integrate the available ontologies, such as WordNet, Wikipedia and ACM Digital Library. This paper introduces an approach combining Wikipedia [5], WordNet [6] and ACM Digital Library [7] in order to construct the Information Technology Ontology, which covers many different topics in this area. Besides, the authors propose several algorithms to find out synonyms, hyponyms, and hypernyms of concepts and extract sentences from documents with a focus on semantic relationships of concepts. These algorithms are composed of natural language processing, machine learning and statistic method.

Since the Information Technology Ontology (ITO) is an automatic integration of WordNet and Wikipedia, ITOs synsets may contain WordNet and Wikipedia entries, which have the same category. Moreover, in order to enrich the ontology the authors use the ACM Digital Library, which includes text files belonging to the information technology domain.

The paper is organized as follows: section 2 discusses the related work in building specific domain ontology; section 3 presents the details for building Information Technology Ontology (ITO); the evaluation and the performance results of ITO are given in section 4; and the concluding remarks in section 5.

2. RELATED WORK

Information retrieval, information extraction, and question and answer trend to the use of ontology as a knowledge base.

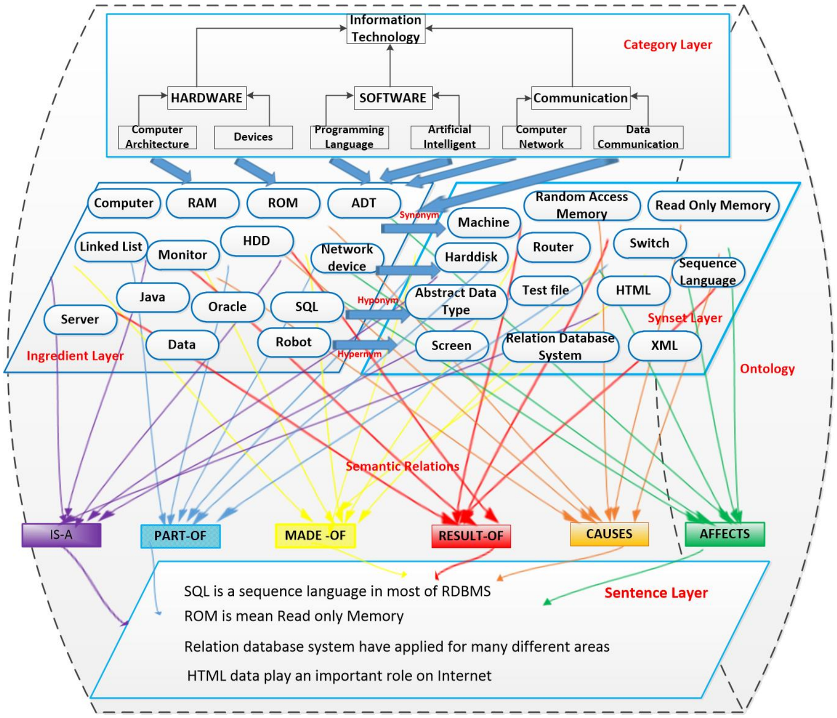
A.Pease et al. [8] has been proposed as a starter document for the SUO working group. It creates a hierarchy of top-level things as Entities, and subsumes Physical and Abstract. SUMO divides the ontology definition into three levels: the upper ontology (the SUMO itself), the mid-level ontology (MILO), and the bottom level domain ontology. Mid-level ontology serves as a bridge between the upper abstraction and the bottom-level rich details of domain ontologies. Beside the upper and mid-level ontology, SUMO also defines rich details of domain ontologies, including Communications, Countries and Regions, distributed computing, etc. W. Sun et al. [9] proposed some methods to build a domain ontology automatically. Based on the specific domain thesauri, he proposed a kind of way to reengineer the thesauri, in particular, on how to get and adjust the semantic relations automatically. Ultimately, he achieves the ontology automatically constructed. M.A.Shilahli et al. [4] built bilingual Turkish English ontology based on Wikipedia. His ontology focused on concepts of laptop devices. P. Q. Dung et al. [10] built domain specific ontology in order to sever in education area. He concentrated on personalized e-learning systems using both ontology technology and intelligent agents. This ontology describes the learning material that composes a course in terms of both learning resource and acquired knowledge, as well as the learners and their learning styles.

3. INFORMATION TECHNOLOGY ONTOLOGY (ITO)

3.1. Building ITO

In generic feature, a domain specific ontology life-cycle can be schematized by four main stages: the specification stage, the formalization stage, the maintenance stage, and finally the evaluation stage [1]. Based on ontology life-cycle, a model of Information Technology Ontology is given in Figure 1.

Since the ontology only focuses on information technology domain, it is called Information Technology Ontology (ITO). There are four layers in ITO, namely Category, Ingredient, Synset and Sentence layers. The terms of Synset layer are synonyms, hypernyms, hyponyms of the terms of Ingredient layer. Some of semantic relations, e.g., IS-A, PART-OF, will be derived from the hyponym and hypernym relations. A random sample of semantic relations for illustration is only picked, as shown in Figure 1.



SQL is a sequence language in most of RDBMS

ROM is mean Read only Memory

Relation database system have applied for many different areas

HTML data play an important role on Internet

Figure 1: Information Technology Ontology (ITO) Hierarchy

The first layer is known as Category layer. In order to build this layer, we extract items from ACM Category [11] are extracted. Over 170 different categories that belong to Information Technology are taken for building this layer. This layer is shown in Figure 2.

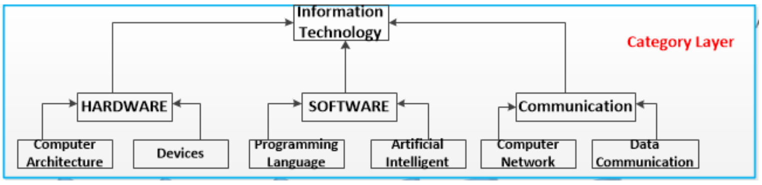


Figure 2: The hierarchy of Category Layer

The root of category tree is information technology. The left and the right sides of the tree are superclass and subclass that belong to information technology domain, such as hardware, software, computer, and devices. A superclass/subclass of this layer is converted into XML format as follows

```
<TOPICS><TOPIC>
      <NAME> Program </NAME>
      <ID> 0000021</ID>
      <LEVEL> 3 </LEVEL>
      <SUPER>Software</SUPER>
      <SUBOF> PAYROLL PROGRAM </SUBOF>
      <SUBOF> HRM PROGRAM </SUBOF>
</TOPIC></TOPICS>
```

Next layer is known as Ingredient layer. Firstly, let us define instances. Instances could be nouns or compound nouns, which are terms on information technology area, e.g. robot, Support vector machine, Local area network, wireless, UML, etc. In order to setup this layer, the authors start from an available ontology Wikipedia. Wikipedia is an ontology, which includes various fields and many different languages. However, the focus is only on English language and information technology domain. In order to extract items from Wikipedia with our target, Java-based Wikipedia Library (JWPL) [12] is used. NLP tools are also used, such as OpenNLP [13], Stanford Lexical Dependency Parser [14] for Parser, POS TAG and sentence detect. A processing model is proposed in Figure 3.

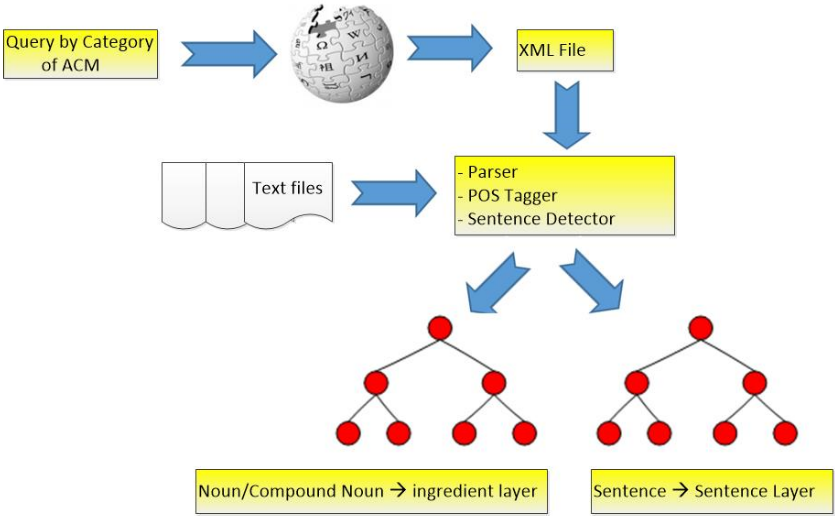


Figure 3: Model extraction from Wikipedia

An instance of category layer is converted into XML format as follows

```
<MEMBERS>
      [<MEMBER>
      <MEMBERID>101</MEMBERID>
      <VALUE>Robotics</VALUE>
<IG1>0.8743</IG1>
<IG2>0.703</IG2>
<CATEGORYIDs>102, 104</CATEGORYIDs>
      <HREF>Wikipedia</HREF>
</MEMBER>]
</MEMBERS>
```

In this model, some manually designed IT queries with JWPL [12] are used to get data relevant to 170 categories resulting in 170 XML files. These files will be parsed, POG tag, and sentence detecting to identify nouns, compound nouns. Furthermore, Information Gain (IG) is used to filter nouns/compound nouns, which are not related to the information technology domain before being put into information technology ontology.

Since one concept can belong to more than one category, so that one instance of ingredient layer can belong to one or many instances of category layer, e.g. robotics may belong to NLP and Machine Learning, thereby the value of categoryIDs tag can be greater than one. In order to decide which category the concept belongs to, the system will calculate an information gain value of this concept in each category. The concept will belong to category having the highest value. When extracting lexical terms from Wikipedia, a statistical method is used to evaluate these terms. Information Gain [15, 16] is applied to this case and calculated as follows

$$IG(A) = E(B - A) - E(A), \tag{1}$$

$$E(A) = \sum_{j=0}^{C-1}(Pj \log 2Pj), \tag{2}$$

where $E(a)$: entropy of attribute a in B, $E(Ba)$: entropy of all attributes in B after deleting a from B, Pj: probability distribution of attribute a in B.

To solve problem, an Information Gain (IG) formula (1) is proposed as follows:

$$IG(a|Ci) = E(X|Ci)E(a), \tag{3}$$

where $IG(a|Ci)$: Information Gain of a in category Ci, $E(X|Ci)$: entropy of all attributes in category Ci after deleting a from Ci.

After calculating IG for each instance of ingredient layer, a threshold T is used to evaluate them before putting them into ITO.

Threshold T is a real number that is chosen based on experience results. There are two cases that will occur in this papers context:

- $IG \geq T$: instance is attached to the respective category
- $IG < T$: instance is not putted into ITO and stored in other place for search support.

With $T = 0.6$, the precision of instances in this layer is roughly 95An algorithm is proposed as follows

```
Procedure Filter_Instance
T = 0.6
While (folder is not empty)
    Open(XML file in this folder)
            Remove_Tag(XML file)
    While(term in XML is not null)
        Calcultae_Term_Frequency(XML file)
        Calculate_TotalWorlds(XML file)
        Calculate_Entropy_Term(XML file)
        IG = Information_Gain(Term)
        If (IG > =T) then
        Put(Term into Ingredient Layer)
        End if
    End While
End While
End Pro
```

The third layer of ITO is known as Synset layer (Figure 1). To set up instances of this layer, the WordNet version 3.0 is used. Similar to Wikipedia ontology, WordNet is also an ontology that includes many distinct domains. However, focus is only on information technology domain. This layer includes a set of synsets. Synset is set of synonyms, hyponyms, and hypernyms of an instance of ingredient layer. They are extracted from WordNet version 3.0. An algorithm for collecting these words is proposed as follows.

```
Procedure Building_Synset_Layer ()
    While (instance of Ingredient layer is not null)
        Begin
            Synonym_List = WordNet_query (instance)
            Hyponym_List = WordNet_query (instance)
            Hypernym_List = WordNet_query (instance)
            If (Synonym_List is not null)
                Link (instance to Synonym_list)
            End if;
            If (Hyponym_List is not null)
                Link (instance to Hyponym_list)
            End if;
            If (Hypernym_List is not null)
                Link (instance to Hypernym_list)
            End if;
        End
    End While
End Pro
```

As these two ontologies of Wikipedia and WordNet are proposed, there are over 400,000 instances which belong to many different categories of information technology domain. That is an advantage of ontology for its applications in the future.

The last layer of ITO is known as Sentence layer (Figure 1). In this layer, the sentences are also extracted from Wikipedia, as shown in Figure 3. These sentences present the semantic relationship between words in sentences. Hence, most of the sentences in this layer are linked to one or many terms in ingredient layer. The finding of the semantic relationship between terms of ingredient layer and storing them in a sentence layer plays an important role as they can be re-used to build information extraction system in the future. This layer also includes many semantic relations between instances, such as, IS-A, PART-OF, MADE-OF, RESULT-OF, etc. Additionally, a random sample of semantic relations is selected for description in this case.

3.2. Enriching ITO

Since data is always updated, enriching ontology also plays an important role. Text documents of the ACM Digital Library are used to update terms of ingredient and sentence layers. These text files are annotated with keywords or a category based on ACMs standard. Firstly, the authors preprocess these files, like merge them by category, de-capitalize all words in documents, and remove unnecessary character. Then, NLP tools such as OpenNLP, Stanford Lexical Dependency Parser are used for extracting keywords from documents and putting them in ITO.

4. EXPERIMENT RESULTS

Information Technology Ontologys performance has been measured by using three factors: Precision, Recall and F-Measure [3]. These factors are calculated by each category in ITO as below:

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

$$R(C_i) = \frac{Correct(Ci)}{Correct\,(Ci) + Missing(Ci)}, \tag{5}$$

$$F - Measure(C_i) = 2\frac{Precision * Recall}{Precision + Recall}, \tag{6}$$

where Ci represents a category in ITO and correct, wrong, missing represent the number of correct, wrong, and missing of returned results from users queries, respectively.

Tables 1, 2, 3, and 4 respectively show the experiment results. Five categories are picked out at random among 170 categories for illustration.

Categories	Quantity	Precision	Recall	F-Measure
Artificial Intelligent (AI)	5714	97.03%	88.62%	93.00%
Logic Design (LD)	4644	96.41%	54.72%	70.00%
Operating System (OS)	6785	84.47%	81.37%	83.00%
Process Management (PM)	3056	96.72%	76.02%	86.00%
Software (Soft)	4249	96.52%	92.19%	95.00%

Table 1: The number of Instances extracted from Wikipedia and ACM Digital Library

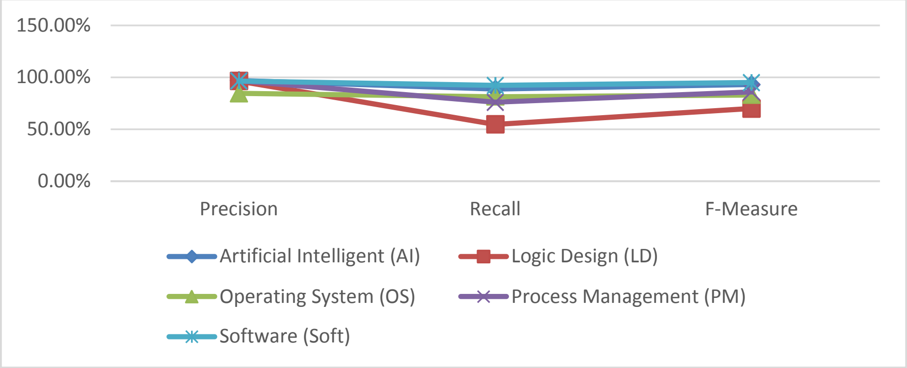


Figure 4: Evaluation the number of instances extracted from the ACM Digital Library

Categories	Quantity	Precision	Recall	F-Measure
Artificial Intelligent (AI)	689	94.41%	88.15%	91.18%
Logic Design (LD)	472	92.24%	84.27%	88.08%
Operating System (OS)	861	96.18%	91.58%	93.83%
Process Management (PM)	517	93.25%	86.16%	89.57%
Software (Soft) 583 94.26%	89.04%	91.58%		

Table 2: The number of synonyms extracted from WordNet

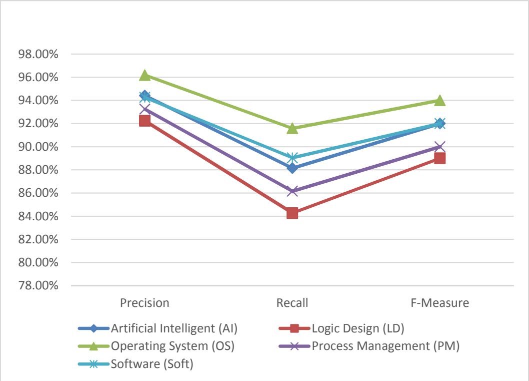


Figure 5: Evaluation the number of synonyms extracted from WordNet

Categories	Quantity	Precision	Recall	F-Measure
Artificial Intelligent (AI)	837	96.14%	88.29%	92.05%
Logic Design (LD)	718	87.54%	84.26%	85.87%
Operating System (OS)	972	96.82%	91.42%	94.05%
Process Management (PM)	728	88.31%	85.15%	86.71%
Software (Soft)	646	85.64%	81.04%	83.28%

Table 3: The number of hyponyms extracted from WordNet

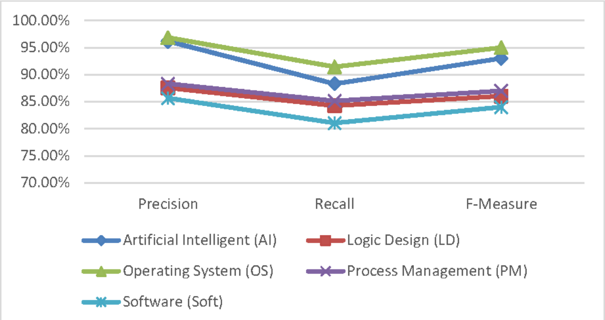


Figure 6: Evaluation the number of hyponyms extracted from WordNet

Categories	Quantity	Precision	Recall	F-Measure
Artificial Intelligent (AI)	1321	92.41%	91.17%	91.79%
Logic Design (LD)	954	84.62%	79.37%	81.92%
Operating System (OS)	1413	95.04%	96.81%	95.92%
Process Management (PM)	834	82.31%	84.55%	83.42%
Software (Soft)	893	85.48%	80.19%	82.76%

Table 4: The number of hypernyms extracted from WordNet

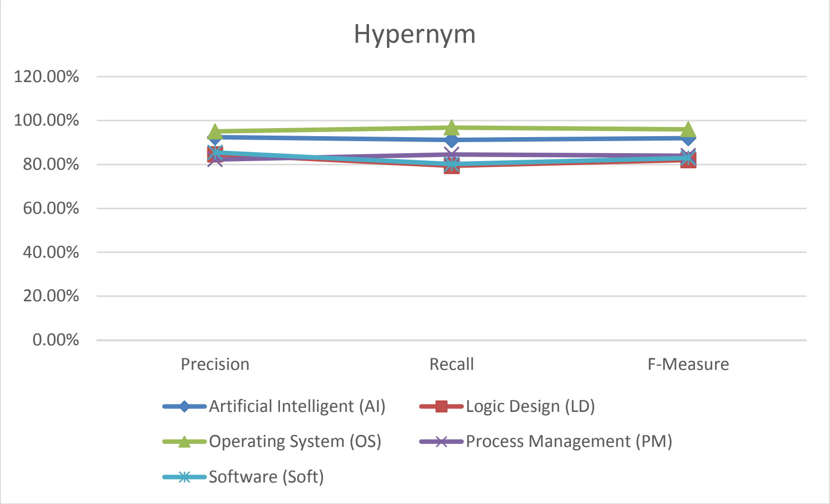


Figure 7: Evaluation the number of hypernyms extracted from WordNet

5. CONCLUSIONS

This paper presents the procedures to build a domain specific ontology on information technology based on corpora and two available ontologies of Wikipedia and WordNet. In order to build ontology, techniques from Machine Learning, NLP and Statistic are proposed. Overall evaluations are computed based on the factors, namely Precision, Recall and F-Measure. Efforts must also be invested in order to reduce the overall processing time of the system. Additionally, since data are collected from distinct sources, such as text files of the ACM Digital Library, Wikipedia and WordNet, there are approximate 1,000,000 distinct instances that belong to the information technology domain. That is an advantage of this ontology and its application in the future. Besides, it also ensures the semantic consistence of instances in this ontology.

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, the authors will focus particularly on automated ontology evaluations and how to detect automatically semantic relationships between concepts.

REFERENCES

[1] W. Wong, W. Liu, and M. Bennamoun, *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances.* USA: IGI Global, 2011.

[2] T. M. Nguyen, *Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications.* USA: IGI Global, 2009.

[3] T.-V. T. Nguyen and T. H. Cao, "VN-KIM IE: automatic extraction of Vietnamese named-entities on the web," *New Generation Computing*, vol. 25, no. 3, pp. 277–292, 2007.

[4] M. Salahli, T. Gasimzade, and A. Guliyev, "Domain specific ontology on computer science," in *Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009.* IEEE, 2009, pp. 1–3.

[5] Wikipedia. [Online]. Available: https://en.wikipedia.org/

[6] Princeton university. [Online]. Available: http://wordnet.princeton.edu/

[7] Association for computing machinery. [Online]. Available: http://www.acm.org/

[8] A. Pease, I. Niles, and J. Li. (2002) The association for the advancement of artificial intelligence. [Online]. Available: http://www.aaai.org/Papers/Workshops/2002/WS-02-11/WS02-11-011.pdf

[9] W. Sun, M. Jia, D. Zheng, H. Cao, B. Yang, and H. Yu, "Automatic domain ontology construction based on thesauri," in *Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09*, vol. 7. IEEE, 2009, pp. 415–418.

[10] P. Q. Dung and A. M. Florea, "An architecture and a domain ontology for personalized multi-agent e-learning systems," in *Third International Conference on Knowledge and Systems Engineering (KSE), 2011.* IEEE, 2011, pp. 181–185.

[11] Acm. [Online]. Available: http://www.acm.org/about/class/ccs98-html

[12] Google. [Online]. Available: https://code.google.com/p/jwpl/

[13] The apache software foundation. [Online]. Available: https://opennlp.apache.org/

[14] The stanford natural language processing group. [Online]. Available: http://nlp.stanford.edu/software/lex-parser.shtml

[15] C. D. C. Ta and T. P. Thi, "Improving the formal concept analysis algorithm to construct domain ontology," in *Fourth International Conference on Knowledge and Systems Engineering (KSE), 2012.* IEEE, 2012, pp. 74–78.

[16] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[17] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou, "Ontology change: Classification and survey," *The Knowledge Engineering Review*, vol. 23, no. 02, pp. 117–152, 2008.

Received on May 05 - 2014
Revised on March 21 - 2015