AUTOMATIC ONTOLOGY ACQUISITION AND LEARNING

Sanju Mishra[1], Sarika Jain[2]

[1] Department of Computer Applications, Teerthankar Mahaveer University, Moradabad, U.P
[2]Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India

Abstract

It is widely accepted to consider ontology as a conceptualization of a domain of interest that can be used in several ways to model, analyse and reason upon the domain. It is an expensive and time consuming task to construct ontologies manually through domain experts and knowledge engineer. So there is a need of ontology learning from different sources. This paper provides a framework for ontology learning approaches and especially precise the term extraction using natural language processing.

Keywords— *Ontology Acquisition, Ontology Learning, Approaches, Term extraction, Natural Language Processing*

--***--

1. INTRODUCTION

An intelligent agent is a system that able to learn from its environment and interacts with it intelligently. They are human engineered systems to solve some problems related to the real world, also needs to be able to represent knowledge. Knowledge representation in a computer system constitutes an interconnection between a symbolic reasoning system and the outside world. The utilized formalisms and inference schemes never accomplished the computational power of a program where as this power can be acquired from the knowledge. To make a system intelligent, it is necessary to facilitate it with lots of high-quality, precise knowledge about some problem area of interest. Human experts observe and act in many thousands of cases to develop their skills. In the symbolic form, knowledge could be explicitly expressive, it must assume about the possibility of reducing all model of tacit knowledge (skills, intuition and the like) to explicit facts and rules [1].

Knowledge comes from different sources like structured, unstructured or semi structured text. Knowledge can be conquered in the form of both machines and human’s readable forms by using ontology. Ontologies capture the domain knowledge in an inclusive way and provide a shared agreed upon understanding of a domain. Ontology usually contains modeling primitives such as terms, concepts, generic relations between concepts, and axioms. Ontologies represent and share the knowledge within an application domain. Manual construction and population of ontologies, is a very time-consuming and labor-intensive task. Present research in the field of automatic and semi-automatic ontology acquisition and development provides methods and solution to solve this problem. Several methodologies for ontology learning and ontology population have been created in order to assist in building ontologies [19].

Ontology learning is the first step to build ontology which is concerned with knowledge acquisition and in the context of this paper more specifically with knowledge acquisition from text. There are two terms ontology population and ontology learning in ontology development, where ontology population define and observe the knowledgebase while ontology learning is (semi) automatic support in ontology development. Ontologies formalize the intensional aspects of a domain, whereas a knowledge base formalizes the extensional part that contains assertions about instances of concepts and relations which are defined by the ontology [6]. As an example, consider the EHCPRs System [1]. The globally spread knowledge treasure consists of the declarative knowledge (permanent knowledge structure + domain-specific KB + dynamic database) as well as the procedural knowledge. Ontology includes permanent knowledge structure and the procedural knowledge at the time of initiation of system on the server. Ontology is the soul, the KB is the mind and database is the personality or mode (a dynamic record of all actions and results) of the EHCPRs system.

Ontology learning systems can be categorised by data types, which they learned [11, 12]. Ontology learning system takes three types of input (i) unstructured data like any text file, books etc. (ii) Semi-structured data like HTML/XML files, (iii) structured data like databases. Ontology learning involves identifying ontology elements such as concepts, synonyms, relations, properties and axioms from textual sources. There is a number of Ontology learning approaches [3] like Machine Learning (ML), Statistical Based Method (SBM), Pattern Matching (PM), Logic Based Method (LBM) and the most common method is Linguistic Based Method (NLP). Information retrieval provides various algorithms for analysing associations between concepts in texts using vectors, matrices [9], and probabilistic theorems [14].

The complex task of ontology development can be sub divided into a layer stack, where lower layers represent the basic tasks upon which rely the more complex, higher layers [2, 3, 4]. In this view, terms are the most basic building blocks of ontology learning for unsupervised concept formation from text. These terms can be categorised as simple (i.e., single word) or complex (i.e., multi word), and are considered as lexical realizations of everything important and relevant to a domain. Layer 1 performs the linguistic analysis tasks required by the subsequent modules. Text pre-processing and term extraction are the main tasks with terms as the first layer of ontology learning. The pre-processing task brings the input texts in desired format. Term extraction task employed to identifying the set of terms which are characteristic for the domain and also able to form the ontology lexicon. The next layer of ontology learning is concept hierarchy which constitutes the "backbone" of the ontology. The aim of this task is to organise discovered concepts into a hierarchical structure or taxonomy, where each concept is related to its respective wider and narrower concepts. Attributes and relations are used to characterise the concepts to other concepts in the hierarchy. In the "relation" layer non-taxonomic relations are defined to upgrade the taxonomy with domain specific concept relationships. Concepts constructed the verb relating pair of concepts to depict the relation. [5].

The paper is organized as follows. Section II introduces the ontology definition used in this work. Section III presents an overview of the ontology learning approaches for term extraction. Section IV describes the Natural Language Processing for Ontology Learning. Section V describes the related work, section VI concludes the discussing results and future work.

2. DEFINING ONTOLOGY

The Greek word ontologia produces the term ontology and means "talking" (-logia) about "being" (on / onto-). Ontologies are important for the knowledge sharing and reuse. According to Gruber [7], Ontologies are the formal and explicit specifications of shared conceptualizations in the form of concepts and relations. Ontologies are basically semantic containers and capable to describe the set of terms, relationship between terms and axioms in a given domain or corpus. They have classes, relationships, constraints and axioms define a common vocabulary to share knowledge [16]. Ontology has number of definition in different domains. It is a study of existence, a theory of what is in the world or a hierarchy of concepts. Formally, ontology can be defined as the tuple:

According to Girardi [8], the Ontology O can be explained by using the formula as given below:

$$O = (C, H, I, R, P, A)$$

Where C represents the set of entities of the ontology. It is a concept that represents entities. H is the taxonomic relationship between concepts and denoted by 'kind_of' and 'is_a'. Ontology elements have their own instances also such as
"is_a(cuckoo, femalebird)",
"flying_birds(pigeon, cuckoo,parrot)"

are relationship between classes. I is an instance of concepts such as bird is an instance of animal class. R is the set of an ontology relationship that are neither "kind_of" nor "is_a". P is a set of properties of ontology entities and their data types like weight_of (bird, integer). A is the axioms and rules which checks the consistency of ontology.

According to shamsfard [11], Ontology O is described as O=(C, R, A, Top).

C represents non-empty set of concepts. *R* is the set of all hierarchical and non-hierarchical relations, in which two or more concepts are related to each other, *A* is the set of axioms and *Top* is the highest-level concept in the hierarchy.

Buitelar [13] presents the Ontology Learning Layer Cake in figure 1.

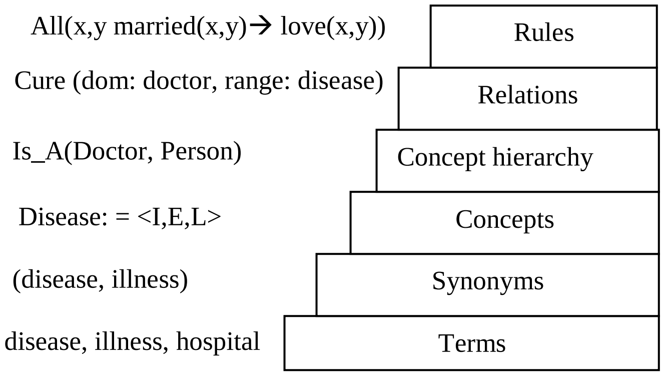


Fig 1 Ontology Learning Layer Cake

In this example, knowledge defined for the concept disease and related concepts, there can be number of terms in different languages to refer or associated with a single disease .There is a hierarchical relation between the concept doctor and person, while non-hierarchical relations between doctor and disease. A rule can be formed, defined over the person and disease concepts.

Combining all the efforts in ontology learning, ontology can be defined as:

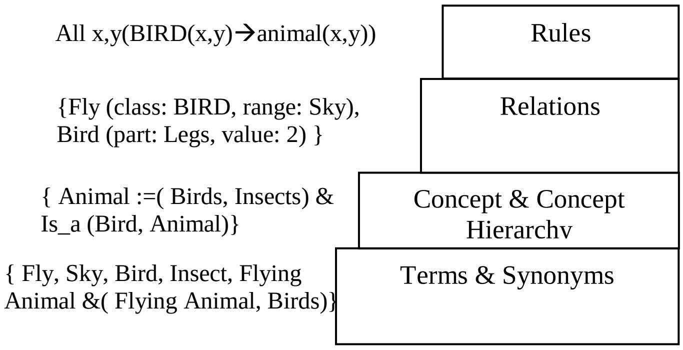

All x,y(BIRD(x,y)→animal(x,y))
Rules
{Fly (class: BIRD, range: Sky),
Bird (part: Legs, value: 2) }
Relations
{ Animal :=(Birds, Insects) &
Is_a (Bird, Animal)}
Concept & Concept Hierarchv
{ Fly, Sky, Bird, Insect, Flying
Animal &(Flying Animal, Birds)}
Terms & Synonyms

Fig 2 Ontology

O = {(T, S), (C, CH), R, A} Where O is Ontology.

T stands for Terms which is the basic building block of ontology learning. The main tasks associated with terms are texts pare-processing and term extraction. The pre-processing task brings the input texts in desired format. The term extraction task aims at identifying the set of terms which are characteristic for the domain.

S describes the synonyms between extracted terms. The synonym level addresses the term variants and term translations. For example, *flying* and *fly* both terms will belongs to same verb class and *flying animals with fur* and *birds* both represents the same concept *Animal.*

C stands for group of terms which belongs to a set and diagnosed as Concepts. These Concepts can be presented in the form of classes and sub-classes i.e. animal is a root class of bird where bird is a sub-class of animal class and parrot, cuckoo, pigeon, crow are the instances of class bird. The intensional definition provided by concept formation, a set of concept instances, i.e. its extension, and a set of linguistic realizations, i.e. (multilingual) terms for this concept.

CH describes the concept hierarchy. Once concepts are identified, there is a need to make hierarchy of concepts or taxonomy.

R describes the Relation between concepts which are not hierarchical. This relation is non-taxonomic. It can be a physical part of an entity or some property of an entity. Such as *flying (bird, sky).*In this statement there is relation, between *bird* and *sky* entity that is *flying.*

A stands for axioms, i.e., set of rules within a corpus. Axioms allow inquiring the consistency of ontology which represents a new knowledge using valid conditions. For example animals having wings and feathers indicates bird. So the rule is body_parts (wings, feathers)→ Bird.

3. APPROACHES FOR ONTOLOGY LEARNING

In Ontology learning there are four stages namely terms, concepts and concept hierarchy, relations (hierarchical relations, non-hierarchical relation) and axioms. To achieve these outputs shamsfard [11] presented some methods, tasks and approaches. In this paper we only discuss about the approaches of Ontology learning step by step. There are four approaches described in given figure 3 [11]. In ontology development we need to extract the concepts from given corpus. By using these approaches we can extract the concepts.

3.1 Statistical Based Approach

According to Wilson [15], In Statistic Based Methods information are mostly derived from Input and Data mining .There are confined words or batches of words, on which statistical analysis is operated.These words represented the frequency of the co-occurrences[16] in terms of words co-occurring with it. The occurrence of two or more words within a sentence or document is called a collocation [17].In statistical learning of ontological knowledge; Collocation and Co-occurrence are the most required method in statistical learning of ontological knowledge.

3.2 Logic Based Approach

There are number of logic based programming that can be used to extract the knowledge from given input such as Inductive Logic Programming (ILP), First Order Logic(FOL) based clustering, FOL rule learning (WEB→KB) and propositional learning[8].This approach is least used approach and is mostly used for complex task such as relation and axioms extraction.

3.3 Pattern Matching Approach

This approach is an important part of information extraction. To extract various ontology elements there are different types of templates such as syntactic or semantic and general or special purpose. To extract the relationship (hyponymy/ hypernymy) from text Hearst [19] introduced some lexico-syntactic patterns in the form of regular expressions. Lexico-syntactic patterns capture hypernymy relations using patterns. HASTI learns concepts, hierarchical and non-heirarchical conceptual relations, and axioms, to build ontologies upon the existing knowledge.

3.4 Linguistic Based Approach

To extract ontological knowledge from natural language text, which are language dependent, there is another a linguistic based approach. This approach usually performs the pre-processing on the input text to extract relevant knowledge for building ontologies from texts. According to Manning [20] Natural Language Processing (NLP) is a category of linguistics that consists of automatic generation and understanding of natural human language. Verb relating pairs, phrase structures and multi words are analyzed from the text by using NLP systems, according to their syntactic and semantic type. The process of information extraction and text mining are often dependent on NLP.

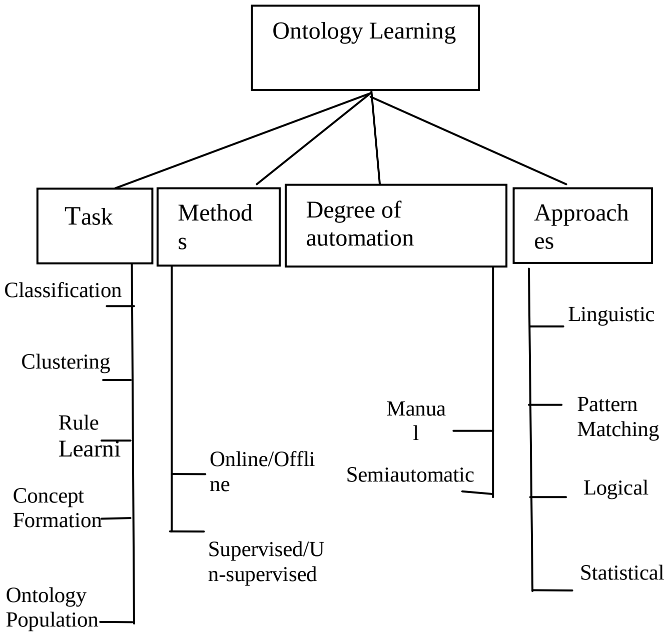


Fig: 3

4. NATURAL LANGUAGE PROCESSING FOR TERM EXTRACTION

According to Wilson [15], terms are the most basic building block for ontology learning. Terms can be simple or complex. So there is a need to preprocess text and extract the terms. These extracted terms make up the concepts for generating the hierarchy. In 2008 Maynard [31] describes a method by using linguistic and statistical techniques for term recognition and reviewed NLP techniques for term extraction. Term recognition and Term extraction, both required the combination of rule-based approaches and machine learning. Maynard also proposed that the core information extraction is carried out by linguistic pre-processing (Tokenization, POS tagging etc.), followed by a named entity recognition component. There are following NLP techniques [15, 21] used in term extraction from the corpus.

4.1 Sentence Splitting

This is an initial step of Natural Language Processing which is capable to split whole corpus into sentences. A corpus is a collection of paragraphs and sentences. These paragraphs and sentences are correlated with white spaces and punctuation marks and also line breaks. In this step, each sentence is splitted individually in a single line for the tokenization e.g.

“Birds are flying animals. Birds live in nests. Cuckoo is a bird. It flies in the high sky.”

The above example is of paragraph. In this paragraph, there are four sentences conjugated with full stops. After splitting in individual sentences, it will be like-
Birds are flying animals.
Birds live in the nests.
Cuckoo is a bird.
It flies in the high sky.

Now these sentences can easily be tokenized in the form of token in the next step.

4.2 Tokenization

After splitting the sentence the next process is to constituents the sentence in to tokens. A text corpus can be tokenized in the form of paragraphs, sentences, and words.

Tokenization is simplest form of the text which achieved after sentence splitting. It uses the spaces as the boundaries to pre-process the text as given in example-

Table 1 Tokenized Term

“Birds”	“Are”	“Flying”	“Animals”	“.”

There are 5 tokens in this sentence. Some tokens are useful to identity entities while others are not; like “are” and “.”.

In the next step it is required to apply Part-of-Speech tagging with each token.

4.3 POS Tagging

This step is an important part of the text pre-processing. It assigns each token to its corresponding syntactic word category (i.e. noun, verb, adjective etc). This is also called term annotation. All words of a sentence are annotated as noun, verb etc.

For example, in figure 4, consider the sentence, “The bird is a flying animal”.

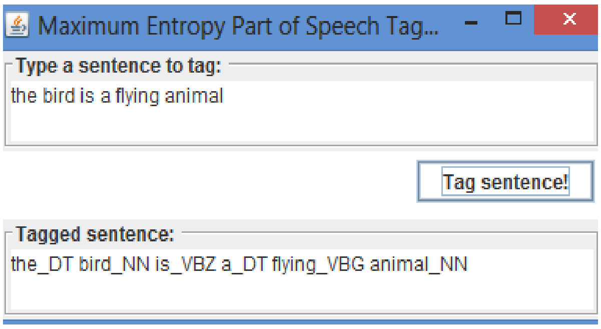
Table 2 Identification of Entities

The	**Bird**	Is	A	Flying	**Animal**
“DT”	**“NN”**	“VBG”	“DT”	“VB”	**“NN”**

There are some small boxes which indicate the terms with tokenization and tagging like “the” is a determiner, “bird” is a Noun and so on as below—

The /DT Bird/NN is/NN a/DT Flying /VBG animal/NN

There are two main entities *Bird* and *Animal in above table.*



Type a sentence to tag:
the bird is a flying animal

Tagged sentence:
the_DT bird_NN is_VBZ a_DT flying_VBG animal_NN

Fig 4. Extracted Tokens

4.4 Lemmatization or Morphological Analysis

English sentences are grammatically formed. Such as run can be Running, Polite, Politeness, Politely. In Pre-processing of text there is a need to reduce these words in their proper form like below-

Running □ *Run*
Politeness, Politely □ *Polite*
Organizing, Organizes, Organize□ *Organize*

A proper study for using the vocabulary and morphological analysis of words, refers to *Lemmatization*

In this step, there is extraction of text into morphological variants for ex, eaten becomes eat, running becomes run.

Consider the sentence, *“There are many mice in our houses.”*
After reducing the sentence:
There is a mouse in our house.

There is a term called *stemming which* chops off the ends of words by using a crude heuristic process. In stemming *Lifted* word can be converted into *lift* but word breed cannot be reduced as *bre.*

5. RELATED WORK

In 2005 Buitelaar[13] combined his research in two workshop on ontology learning and knowledge acquisition. The author organized ten papers included in their book into methodologies, evaluation methods, and application scenarios and also generalizes the concepts of “ontology learning layer cake” to describe the assorted layers ramified in ontology learning.

Alexander Maedche and Steffen Staab[22] propose the classification: ontology learning from dictionary, text, knowledge, relational schemata and from semi-structured schemata. Ontology learning approaches focus on the type of input used for learning.

HASTI [11] is an automatic ontology-building system, which learns concepts, hierarchical and non-hierarchical conceptual relations, and axioms, to build ontologies upon the existing system. This System builds dynamic ontologies from scratch. HASTI provides a fusion of linguistics, semantic analysis approach and a hybrid symbolic approach.

In 2003, a report presented by the OntoWeb Consortium [12]. In this survey there are 36 approaches has been listed for ontology learning from text.

At the same time in 2003 Shamsfard [11] performed the survey which studied about 50 approaches and focused on comparing ontology learning approaches

Text-to-Onto [23] is part of an ontology management infrastructure called KAON which is semi structured system. Text-to-Onto extracts plain text from various formats such as PDF, HTML, and XML.

Zhou [24] presented a concise survey looking at certain prominent challenges in this field and also proposed a learning-oriented model for the ontology development.

An ontology language should have an unambiguous well-understood meaning. The agents should be able to understand Ontologies, and also be able to use them properly. One such system, the EHCPRs System is an underlying methodology for representation, reasoning, learning, etc., of live multilingual thinking machine [25]. Jain & Jain [26] presents learning techniques in the EHCPRs based ontology. Jain & Mishra [27] presents various tools and languages for knowledge representation using Ontology.

6. CONCLUSION

The automatic acquisition of ontology and its learning proposed in this paper is based on natural language processing techniques and consists of four techniques: "sentence splitting", "Construction of Tokens uses tokenization", “POS tagging”, “Identification and extraction of Terms”. One of the main limitations identified in the approaches is that it is a domain dependent. Future work will also include the development of an integrated development environment for ontology learning and knowledge acquisition. The ontology layer cake consists of the various subtasks (increasing in complexity) involved in ontology learning.

REFERENCES

[1] N. K. Jain, S. Jain, “Live Multilingual Thinking Machine”, Journal of Experimental and Theoretical Artificial Intelligence, Taylor and Francis, vol. 25:4, pp. 575-587 2013.

[2] P. Buitelaar, P. Cimiano, B. Magnini, “Ontology Learning from Text: Methods, Evaluation and Applications”, IOS Press, Amsterdam, 2005.

[3] P. Cimiano, “Ontology Learning and Population from Text: Algorithms, Evaluation and Applications”, Springer, Heidelberg, 2006.

[4] J. Brank, M. Grobelnik, D. Mladenic, “A Survey of Ontology Evaluation Techniques”, In: Proc. of the Conf. on Data Mining and Data Warehouses, SiKDD 2005, Ljubljana, Slovenia, 2005.

[5] M. Kavalec A. Maedche, V. Svatek, “Discovery of Lexical Entries for Non taxonomic Relations sin Ontology Learning”, In: Van Emde Boas, P., Pokorn´y, J.,ielikov´a, M.,ˇ Stuller, J. (eds.) SOFSEM 2004. LNCS, vol. 2932, pp. 249–256. Springer, Heidelberg (2004).

[6] P Buitelaar, P Cimiano, B. Magnini, “Ontology learning from text: An Overview”, IOS Press 2003.

[7] T.R. Gruber, “A translation approach to portable ontology specification”, Knowledge Acquisition 5, pp. 199– 220, 1993.

[8] R. Girardi. “Analyzing the Problem and Main Approaches for Ontology Population”, Proceedings of 10th International Conference on Information Technology: New Generations, 2013.

[9] B. Fortuna, D. Mladenic and M. Grobelnik, “Semi-automatic construction of topic ontology”, In *Proceedings* of the Conference on Data Mining and Data Warehouses, SiKDD, 2005.

[10] Y. Yang and J. Calmet, “ OntoBayes: An ontology-driven uncertainty model”, In *Proceedings of the* International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC), 2005.

[11] M. Shamsfard and A. Barforoush. “The state of the art in ontology learning: A framework for comparison. The Knowledge Engineering Review”, Vol. 18 No.4, pp. 293-316, 2003.

[12] A. Gomez-Perez, D. Manzano-Macho, OntoWeb Deliverable 1.5: A Survey of Ontology Learning Methods and Techniques”, 2003.

[13] P. Buitelaar, P. Cimiano & B. Magnini, “ Ontology learning from texts: An overview. In P. Buitelaar, P. Cimiano, & B. Magnini (Eds.), *Ontology learning from* text: Methods, evaluation and applications”, Vol. 123, IOS Press 2005.

[14] E. Agirre, O. Ansa, E. Hovy, and D. Martínez, “Enriching very large ontologies using the WWW”, Proceedings of the ECAI 2000 Workshop on Ontology Learning, 2000.

[15] W. Wilson, W. Liu and M. Bennamoun, “Ontology learning from text: A look back and into the future”, ACM Comput. Surv. 44, 4, Article 20, 2012.

[16] D. Heyer, M. Läuter, U. Quasthoff, T. Wittig and C. Wolff, “Learning relations using collocations”, Proceedings of the IJCAI 2001 Workshop on Ontology Learning, 2001.

[17] A. Maedche and S. Staab, “Measuring similarity between ontologies”, Proceedings of EKAW, 251-263, 2002.

[18] AF. Bowers, C. Giraud-Carrier and JW. Lloyd. “Classification of individuals with complex structure”, Proceedings of the 17th International Conference on Machine Learning (ICML2000) 81–88, 2000.

[19] MA. Hearst, “Automatic acquisition of hyponyms from large text corpora”, Proceedings of the 14th International Conference on Computational Linguistics, 539–545, 1992.

[20] Christopher D Manning, H. Schutze, “Review: Foundations of statistical natural language processing”, 2000.

[21] http://www.stanford.edu/.

[22] S. Staab and A. Maedche. “Axioms are objects, to ontology engineering beyond the modeling of concepts and relations”, Technical Report 399, Institute AIFB, Univ. of Karlsruhe, 2000.

[23] P. Cimiano and S. Staab, “Learning concept hierarchies from text with a guided agglomerative clustering algorithm”, In Proceedings of the Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods, 2005.

[24] L. ZHOU, “Ontology learning: State of the art and open issues”, Info. Technol. Manage., pp. 241–252, 2002.

[25] S. Jain, N.K. Jain, “A generalized knowledge representation system for context sensitive reasoning: generalized HCPRs system”, Artif Intell Rev 30(1): pp. 39–52, 2009b.

[26] S. Jain, N.K. Jain, “Learning Techniques in Extended Hierarchical Censored Production Rules (EHCPRs) System”, Artificial Intelligence Review, Springer Netherlands, vol. 38:2, 2012.

[27] S. Jain, S. Mishra, "Knowledge Representation with Ontology Tools & Methodology", International Journal of Computer Applications® (IJCA) (0975 – 8887) International Conference on Advances in Computer Engineering & Applications (ICACEA-2014) at IMSEC, GZB 2014.

[28] D. Maynard, Li. Yaoyong and W. Peters, “NLP Techniques for Term, Extraction and Ontology Population”, Proceeding of the conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp 107-127 IOS Press Amsterdam, The Netherlands, 2008.