IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Recent Survey on Automatic Ontology Learning

R. Manimala[1*], G. MuthuLakshmi[2]

[1, 2]Dept. of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-12, Tamilnadu, India

Corresponding Author: rm.spkcit@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si8.143147 | Available online at: www.ijcseonline.org

***Abstract*—** Semantic Web allows machine to understand the data, for that machine-readable semantic metadata is needed.Intelligence is necessary for the creation and processing of semantic metadata. Ontologies play an important role to implement the idea of the semantic web. Ontology is about the exact description of things and their relationships to represent the knowledge. Nowadays Automatic annotation based on artificial intelligence is required for gathering such knowledge. Manual ontology construction is labour-intensive, error-prone process, inflexible, expensive, time consuming and complex task. Ontology Generation or ontology learning includes the automatic extraction of domain's terms and the relationships between the concepts from a corpus of text, and encoding them with an ontology language for easy information retrieval. Automatic Ontology generation and sharing it through the web make the web content more accessible to machine. Ontologies can be automatically extracted using various techniques. This paper describes the survey about automatic ontology extraction techniques and various methods used to extract the ontology.

***Keywords*—** Ontology, Resource description Framework, ontology learning, Ontology Acquisition

I. INTRODUCTION

Ontology is a formal and structural way of representing the concepts and its relationship for a specific domain. It can also be defined as concepts, relations, attributes and hierarchies present in the domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. Every field creates ontologies to organize information into knowledge. Ontology defines a common vocabulary for researchers to share information in a domain. It helps the computer to understand the things like a human. Ontologies have gained recognition in the semantic web because of their extensive use in web-based applications and artificial smart systems. It mainly improves the accuracy of Web search.

Ontology's also serve as semantic glue between heterogeneous information sources, e.g., sources using different terminologies or languages. Ontology learning is all about automatically extracting the ontologies from the whole documents. RDFS (RDF Schema), OWL (Web Ontology Language), RDF are used to define ontologies. The Resource Description Framework data model is similar to class diagrams which describes the concept and its relationship. It is based on the idea of making statements about resources in the form of subject–predicate–object, known as triples. The subject denotes the resource, and the predicate denotes behaviour of the resource and expresses a relationship between the subject and the object. These triples form the semantic metadata. Automatic ontology learning process generally consists of some common tasks, which are not applied to all ontology learning system. They are:

1) Domain terminology extraction or keyword extraction

A corpus of texts are collected and pre-processed in a domain to extract the domain-specific terms or keywords, which are used to derive concepts.

2) Concept & concept hierarchy extraction

In the concept discovery step, terms are grouped to meaning bearing units which is derived from domain-specific terms and their synonyms. The extracted concepts are arranged in a taxonomic structure. This is most probably achieved by hierarchical clustering methods

3) Relation extraction

In this step, semantic relationships between concepts are used to enable automated reasoning about data. It is used to attach richer semantic metadata to the documents [22].

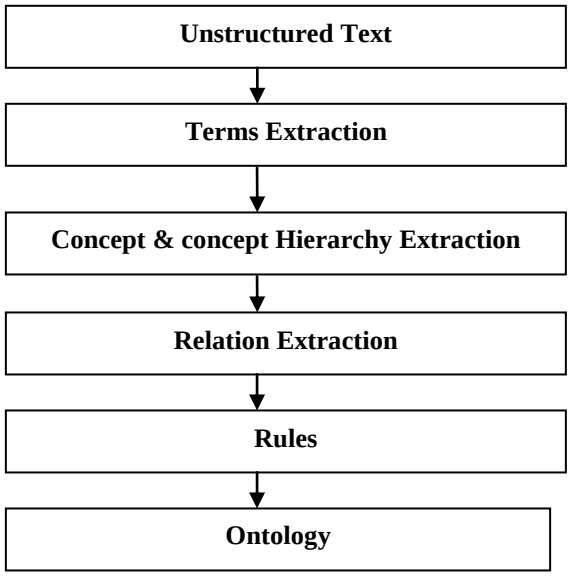

Figure 1. Common Steps to Extract Ontology

4) Rule extraction

Axioms (formal description of concepts) are generated for the extracted concepts. This can be achieved by analyzing the syntactic structure of a natural language definition and the application of transformation rules on the resulting dependency tree. The result of this process is a list of axioms, which is afterwards comprehended to a concept description. Logical statements are used for specifying knowledge about the classes and relationships. These axioms specify additional constraints on the ontology and can be used in ontology consistency checking and for inferring new knowledge through some inference mechanism.

5) Ontology Representation

In this phase, Individual classes are created based on properties of the class and domain. Ontology is represented using XML and OWL ontology description language.

II. LITERATURE SURVEY

Trent Apted & Judy Kay (2002) [23] describes a system that automatically constructs an ontology from the Free On-Line Dictionary of Computing (FOLDOC). The dictionary is parsed and a graph is generated with nodes representing concepts and weighted, directed edges represent the relationships in the ontology. It is used as a basis for making inferences for student models and other reasoning within a teaching system. This paper describes the possible machine learning applications of ontology.

Alani, Harith et al(2003)[1] presented the tool to automate an ontology-based knowledge acquisition process and maintain a Knowledge Base which is used to generate customised biographies of artists. The Web document is divided into paragraphs, and further into sentences. Each paragraph is analysed syntactically and semantically to identify any relevant knowledge to extract. The Apple Pie Parser6 is used for grouping grammatically related phrases as the result of syntactical analysis. Semantic examination then locates the main components of a given sentence (i.e. 'subject', 'verb', 'object'), and identify the entities using GATE and WordNet. Artequakt submits a query to the ontology server to obtain relationship between entities. In addition, three lexical chains (synonyms, hypernyms, and hyponyms) from WordNet are used in order to reduce the problem of linguistic variation between relations defined in the ontology and the extracted text. The output is an XML representation of the facts, paragraphs, sentences and keywords identified in the selected documents. The extraction process terminates by sending the new XML files to the ontology server to be inserted into the Knowledge Base.

Sidi Benslimane, Mimoun Malki, Mustapha Rahmouni, and Adellatif Rahmoun (2008) [21] proposed the domain ontology automatic generation, by applying reverse engineering technique which is defined as a process of analyzing a "legacy" system to identify all the system's components and the relationships between them. This approach can be used for migrating HTML pages to the ontology-based Semantic Web. The main aim of this migration is to make the relational database information that is available on the Web machine-processable, and reduce the time consuming task of ontology creation.

Mithun Balakrishna, Munirathnam Srikanth (2008) [14] presented the semi-automatic ontology library for the National Intelligence Priorities Framework. Jaguar-KAT tool is used for knowledge acquisition and domain understanding to create NIPF ontologies loaded with rich semantic content.NLP tools are used for preprocessing. Concept discovery module identifies noun concepts which are related to the NIPF topic target words or seeds. Well-formed noun concepts are extracted based on a set of syntactic patterns and rules. Classification module determines the hierarchical structure from the extracted NIPF topic noun concepts and semantic relations. Conflict resolution techniques are used for handling the conflict induced in the ontology to generate a merged ontology.

Roberto navigli and Paola velardia (2008) [19] focuses the ontology learning approaches to extract concepts and relation instances from the domain glossary, Art and Architecture Thesaurus (AAT) instead of web documents, i.e. from unstructured texts. They presented a method, based on the use of regular expressions, to automatically annotate the glosses of a thesaurus, the AAT, with the properties (conceptual relations) of core ontology, the CIDOC-CRM[10]. The annotated glosses are converted into OWL

concept descriptions and used to enrich the CIDOC. Here each input gloss is preprocessed with a part-of-speech tagger and Tree Tagger. Sentences are annotated with CIDOC properties and finally formalize the glosses.

Mohammad Syafrullah, Naomie Salim (2009)[7] proposed framework for ontology learning from textual data which contains three phases: First, is the terms extraction phase. Second, is the synonym, concepts and concept hierarchies extraction phase, and finally, is the relations extraction phase. Terms extraction phases developed using hybrid method i.e. the combinations of linguistic, NLP and statistical method. FCA[16] and clustering techniques with the combination of Fuzzy PSO algorithm is used to extracts synonym, concept and concept hierarchies. NLP techniques are used to extract verbs (relations) and their argument structure.

Jone Correia, Rosario Girardi, Carla 8Faria (2011)[9] proposes a approach for the automatic extraction of ontology taxonomic relationships from English texts using natural language processing techniques. The method consists of four steps: "Tagging", "Extraction of Candidate Classes", "Identification of Hyponyms and Synonyms" and "Identification and Representation of Taxonomic Relationships". "Tagging" step is to process the corpus by tokenization, splitting of sentences, Lemmatization and lexical analysis using NLP techniques. "Extraction of Candidate Classes" phase is responsible for selecting candidate classes from concrete and abstract nouns. "Identification of Hyponyms and Synonyms" identifies the synonyms and hyponyms from the obtained candidate classes with the help of wordnet. "Identification and Representation of Taxonomic Relations" phase is to identify the taxonomic relations using heuristic patterns and their representation in an ontology specification language. Using this approach T-NLPDumper prototype tool has been developed to automate the relation extraction in ontology learning.

J. I. Toledo-Alvarado, A. Guzmán-Arenas, G. L. Martinez-Luna (2012)[8] build an ontology automatically ,from a corpus of text documents without using dictionaries or thesauri. This is accomplished with data mining association Rule Mining algorithm such as Apriori algorithm[17] which extracts the multi-word concepts. Instead of looking for multiwords in all the vocabulary, they look up only the most frequent terms to reduce the computational effort. Relation between the concepts is established by creating pairs of concepts and looking for every pair in all the documents of the corpus. They proved the possibilities to establish unknown relations between new concepts of the domain in an unsupervised way. The set of relations and concepts form an undirected graph that represents the ontology. Here the preprocessing is achieved using NLP techniques such as stop word removal and stemming.

Andreia Dal Ponte Novelli, José Maria Parente de Oliveira (2012) [4] presented a method for ontology extraction from the documents using latent semantic, clustering and Wordnet. The latent semantic analysis examines the relation between terms and documents to build a vector space, which allows analysis between documents. The frequent terms in the documents are obtained. The first level (ground level) of the ontology hierarchy is obtained from its own index terms from the term-document matrix. The next level of the hierarchy is formed of concepts obtained from the term analysis of matrix, which provides the relation between terms and concepts. TF-IDF (Term Frequency Inverse Document Frequency)[5] is used to retrieve the concepts. The resultant matrix from the application of SVD provides the relation between concepts. Semantic relations are obtained using Wordnet. The process of the semantic relations obtainment is simplified by performing the analysis by level. The concept and semantic relations are organized into ontologies using OWL language.

Amel grissa touzi1, Hela ben massoud and Alaya ayadi (2013) [3] proposed a new approach for automatic ontology generation is called Fuzzy Ontology of Data Mining (FODM), with the fusion of conceptual clustering, fuzzy logic, and Formal Concept Analysis(FCA). Here the database records are organized into homogeneous clusters which have common properties. In this approach, they define ontology between clusters resulting from a preliminary classification on the data. The clusters are modeled by Fuzzy Cluster Lattice. From the lattice, the fuzzy ontology is generated which is automatically converted into the corresponding semantic representation using Fuzzy OWL2.They prove that this approach optimize ontology learning in case of memory space and execution time.

Alexandra Moreira, Jugurta Lisboa Filho, and Alcione de Paiva Oliveira (2016)[2] proposed a system for automatic generation of ontologies called AutOnGen (AUTomatic ONtology GENnerator which was developed in python .This tool was designed for the sector of production and supply of electric power. To extract the terms, Exterm information extraction tool is used which is based on Natural Language Processing techniques. Translator is used to translate the natural language. Then the term is analyzed by the Part of Speech (POS) tagger. The wordnet was used to obtain the hypernym and SUMO[15] ontology was used to access the higher hierarchy starting from the hypernym. The top level ontology was expressed in OWL language.

Mazen Alobaidi, Khalid Mahmood Malik and Susan Sabra. (2018) [12] Proposed fully automated ontology generation framework, Linked Open Data approach for Automatic Biomedical Ontology Generation (LOD-ABOG) which is empowered by biomedical Linked Open Data. It integrates natural language processing, syntactic pattern, graph

algorithms, semantic ranking algorithms, semantic enrichment, and RDF triples mining to make automatic large-scale machine processing and improve the accuracy of ontology generation. The NLP module of LOD-ABOG framework performs preprocessing tasks such as tokenization, segmentation, stemming, stop words removal, and part-of-speech tagging (POS). Entity Discovery identifies biomedical concepts from free-form text by using UMLS and LOD. Semantic Entity Enrichment module enriches all the discovered concepts. RDF Triple Extraction module identifies the well-defined triplet in LOD that represents a relation between two concepts within the input biomedical text.URI Ranking algorithm is used for ranking the URIs of each concept based on their semantic relatedness. Syntactic Patterns module performs pattern recognition to find a relation between two concepts. Ontology factory module automates the process of encoding the semantic enrichment information and triples candidates to ontology using an ontology language such as RDF, RDFS, OWL, and SKOS.

III. ONTOLOGY EXTRACTION METHODOLOGIES

There are various Ontology Extraction methods[11] such as Machine Learning (ML) Method, Statistical Based Method (SBM), Pattern Matching (PM), Logic Based Method (LBM) and the most common method is Linguistic Based Method (NLP).

Linguistic Based Method [20] is mainly used for preprocessing the input text to extract domain related terms. Natural Language Processing (NLP) is a kind 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 input text by using NLP systems, according to their syntactic and semantic type. NLP pre-processing techniques such as Tokenization, POS tagging, Lemmatization/stemming are mainly used for term extraction. Tokenization split the text corpus into tokens in the form of words, sentences and paragraph. Part of Speech tagging assigns each token to its corresponding syntactic word category (i.e. noun, verb, adjective etc). Stemming/Lemmatization reduce the inflectional forms of each word into its base or root word.

By using statistical approach [24], the occurrence of two or more words within a sentence or document i.e. collocation [6] and frequency of Co-occurrences in terms of words are find out for concept extraction.

Pattern Matching Approach [13] is used to extract the relationship (hyponymy/ hypernymy) from text. Lexico-syntactic patterns capture hypernymy relations using patterns which are in the form of regular expressions.

Logic Based Approach is used for relation and axioms extraction. There are various logic based programming methods such as Inductive Logic Programming (ILP), First Order Logic (FOL) based clustering, FOL rule and propositional learning [18].

Machine Learning (ML) approach is used for tokenization, PoS tagging, Syntax Analysis, Entity extraction, Sentiment Analysis. Machine learning-based method uses various supervised and unsupervised methods for automating ontology extraction.

IV. CONCLUSION

This paper summarizes ontology learning methods. We observed that a hybrid approach such as linguistic and statistical techniques produces better ontologies. The performance of ontology learning techniques is dependent on efficient preprocessing of data such as finding correct base words by using Lemmatization techniques instead of stemming technique based on the specific domain. After analyzing the literature of ontology learning, many researchers prefer to use statistical techniques for term and concept extraction. Nowadays, Linked open Data are used for refining the ontology construction.

REFERENCES

[1] Alani, Harith; Kim, Sanghee; Millard, David E.; Weal, Mark J.; Hall, Wendy; Lewis, Paul H. and N.R. Shadbolt, "Automatic ontology-based knowledge extraction from web documents". IEEE Intelligent Systems, 18(1) pp. 14–21. Jan-Feb 2003.

[2] Alexandra Moreira, Jugurta Lisboa Filho, and Alcione de Paiva Oliveira ,"Automatic creation of ontology using a lexical database: an application for the energy sector", International Conference on Applications of Natural Language to Information Systems NLDB 2016: Natural Language Processing and Information Systems pp 415-420, 17 June 2016.

[3] Amel grissa touzi1, hela ben massoud and alaya ayadi,"Automatic ontology generation for data mining using FCA and clustering", 7 Nov 2013.

[4]Andreia Dal Ponte Novelli,José Maria Parente de Oliveira, "Simple Method for Ontology Automatic Extraction from Documents" , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 12, 2012.

[5] Caden Howell," Machine Learning Methods of Mapping Semantic Web Ontologies",Published in November 22, 2008.

[6] 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.

[7] Mohammad Syafrullah , Naomie Salim ," A Framework for Ontology Learning from Textual Data ",published in 2009.

[8] J. I. Toledo-Alvarado, A. Guzmán-Arenas, G. L. Martínez-Luna," Automatic Building of an Ontology from a Corpus of Text Documents Using Data Mining Tools", Journal of Applied Research and Technology, Vol. 10 No.3, June 2012.

[9] Jone Correia, Rosario Girardi, Carla Faria," Extracting Ontology Hierarchies From Text", Conference: Proceedings of the 23rd International Conference on

Software Engineering & Knowledge Engineering (SEKE'2011), Eden Roc Renaissance, Miami Beach, USA, January 2011.
[10] M. Doerr, The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata, AI Magazine,24(3) (2003).
[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]Mazen Alobaidi, Khalid Mahmood Malik and Susan Sabra. Alobaidi, "Linked open data-based framework for automatic biomedical ontology generation ", BMC Bioinformatics (2018) 19:319.
[13] MA. Hearst, "Automatic acquisition of hyponyms from large text corpora", Proceedings of the 14th International Conference on Computational Linguistics, 539–545, 1992.
[14]Mithun Balakrishna, Munirathnam Srikanth, "Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)", Published in OIC 2008.
[15] Pease, A., Niles, I., and Li, J.,"The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications", in Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada, July 28-August 1, (2002).
[16]Philipp Cimiano, Andreas Hotho, Steffen Staab," "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis", Journal of Artificial Intelligence Research (JAIR) pages 305–339, 2005.
[17]Rakesh Agrawal and Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules in Large Databases, In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), 1994, pp 487-499, San Francisco, CA, USA.
[18]R.Girardi. "Analyzing the Problem and Main Approaches for Ontology Population", Proceedings of 10th International Conference on Information Technology: New Generations, 2013.
[19]Roberto navigli and Paola velardia,"From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions", Published in Ontology Learning and Population 2008.
[20]Sanju Mishra, Sarika Jain," Automatic Ontology Acquisition and Learning", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308, Volume: 03 Special Issue: 14 | Nov-2014 | SMART-2014.
[21] Sidi Benslimane, Mimoun Malki, Mustapha Rahmouni2, and Adellatif Rahmoun3 ," Towards Ontology Extraction from Data-Intensive Web Sites: An HTML Forms-Based Reverse Engineering Approach" , The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008 .
[22] Ting Wang12, Yaoyong Li1 , Kalina Bontcheva1 , Hamish Cunningham1 , and Ji Wang, "Automatic Extraction of Hierarchical Relations from Text", Proceedings of the 3rd European conference on The Semantic Web: research and applications Pages 215-229, June 11 - 14, 2006.
[23] Trent Apted, Judy Kay," Automatic Construction of Learning Ontologies", Published in ICCE 2002.
[24] 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.

AUTHORS PROFILE

R.Manimala is currently pursuing her part-time research in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. She completed her M.Phil degree in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli and Master Degree in Computer Applications from SCAD college of Engineering and Technology, Cheranmahadevi. She is working as an Assistant Professor at Sri Paramakalyani College, Alwarkurichi, and Tamilnadu. Her research interests include Data mining and Big data.

Dr.G.Muthulakshmi is an Assistant Professor in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. She received her B.E degree in Computer Science and Engineering at PSR Engineering College, Viruthunagar and M.E degree in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli. She received her doctorate in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli. She has 11 years of teaching experience and 9 years of research experience. Her areas of interest are Digital Image Processing, Data Mining, and Neural Networks.

