Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Novel Approach for Information Extraction Using Ontology

[1]Varsha Manwade, [2]Ritesh Shah
[1]Mtech Research Scholar Sanghvi Institute of Management and Science, Indore, India
[2]Assistant Professor Sanghvi Institute of Management and Science Indore, India

Abstract: Finding structured information from unstructured or semi-structured text is called information extraction. It's very necessary in point of view of text mining; it has a wide range of applications in domains such as biomedical literature mining and business intelligence. The main function of information extraction is to extract specific information from text. Information extraction technology includes the information technology based on the dictionary, rule-based extraction technology. This paper represents some relevant research based on ontology based information extraction.

Keywords: Information extraction, ontology, semantic web

I. INTRODUCTION

The general goal of information extraction is to discover structured information from unstructured or semi-structured text.

As an example for an information extraction system, we can describe a system that processes a set of web pages and extracts information regarding countries and their political, economic and social indicators. Some kind of model that specifies what to look for (e.g., country name, population, capital, main cities, etc.) is needed to guide this process. The system will attempt to retrieve information matching this model and ignore other types of data. Information extraction is extracting the structured data in accordance with specific rules from the natural language text or semi-structured text automatically.

Information extraction technology is a technology that grows out of the rapid growth of the internet and aims at extracting factual information to help people overcome the problem of information overloading. [13]

II. DEFINING ONTOLOGY

Originally, ontology is a philosophical term which is defined as "the description of the objective existence of the world, namely, the existence". Ontology is formalized clearly to the shared conceptual model. Ontology gives the basic terms of the vocabulary and the relationships, to capture the relevant domain knowledge, and propose a common understanding of the field to identify the common vocabulary, and give a clear formal definition [3].

Ontology is a system of concepts in which all concepts defined and interpreted in a declarative way [2].

Semantic web [15] is an extension of World Wide Web that aims to enable computers to discover, search, infer and collect Web's information without human effort. Semantic web allows efficient way of representing data in the World Wide Web.

III. DEFINING ONTOLOGY-BASED INFORMATION EXTRACTION

Key characteristics of ontology-based information extraction system:-

• Process unstructured or semi-structured natural language text: Since OBIE is a subfield of information extraction, which is generally seen as a subfield natural language processing, it is reasonable to limit the inputs to natural language text. They can be either unstructured (e.g., text files) or semi-structured (e.g., web pages using a particular template such as pages from Wikipedia).In this Approach system takes input in the form of image, audio and videos but it cannot be categorized as OBIE systems.

• Present the output using ontologies: Li et al. [4] identify the use of a formal ontology as one of the system inputs and as the target output as an important characteristic that distinguishes OBIE systems from IE systems. There are some OBIE systems that construct the ontology to be used through the information extraction process itself instead of treating it as an input (e.g., The Kylin system [5]).

•Use an information extraction process guided by ontology: We believe that "guide" is a suitable word to describe the interaction between the ontology and the information extraction process in an OBIE system: in all OBIE systems, we can extract things like that- classes, property, object, individual etc. by using ontology in information extraction system.

IV. SYSTEM ARCHITECTURE

Ontology-based Information extraction system has four parts. They are Knowledge base, text pre-processing module, ontology parsing module and semantic extraction module. With the help of this four parts system can easily extract classes, properties, individuals from text.

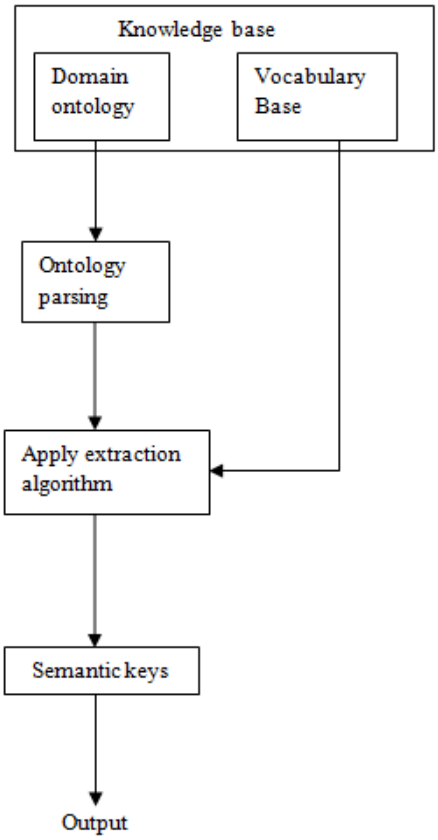

Figure1:- Ontology-based Information Extraction

4.1 Knowledge Base: - It consists of domain ontology and vocabulary base. Only domain expert can build the domain ontology.

4.2 Pre-Processing Module: - In this module text will divided into paragraphs, and paragraph is split into sentences and sentences will split into series of single words by some existing software.

4.3 Parsing Module: - Its main functions are to parse the inputted domain ontology and get all information about classes, properties and individuals from it.

4.4 Extracting Module: - This module will be used proposed algorithms to extract information. And Output will be relevant as compare to old approach.

V. CLASSIFICATION OF CURRENT OBIE SYSTEMS

The following methods are employed by the OBIE systems we have studied [1].

1. Linguistic Rules Represented by Regular Expressions
2. Gazetteer Lists
3. Classification Techniques
4. Construction of Partial Parse Trees
5. Analyzing HTML/XML Tags
6. Web-Based Search

VI. LITERATURE SURVEY

6.1 Text information extraction based on OWL ontologies

This paper [Hongsheng wang, Lu Yuan, Hong Shao] presented in IEEE 2008 an OWL ontology-based text information extraction system and has made a clear description of every module contained in the system and the cooperative relationships among them. This use two algorithm for implementation. One is semantic information extraction and other is semantic information re-recognizing. Experiment results show that, the two algorithms are effective and accurate, especially when the domain ontologies are well-defined. System has good portability.

6.2 Web information extraction based on news domain ontology theory

This paper [Junfang Shi, LiLiu] proposed in IEEE 2010 a web information extraction method based on news ontology. Accurate and Interested information was identified by using news domain ontology. With the help of page pre-processing, XPath and page conversion technology. Testing from news site shows that the approach proposed doesn't rely on the page structure and it increased the recall and precision of information extraction.

6.3 Ontology-based information extraction system in E-commerce websites

This paper [Yang Xiudan, IEEE 2011] used the concept of ontology to analyze the structure and content of the website, in order to extract the information based on ontology from the e-commerce website for the users with the help of ontology model.. The paper makes an experiment test of the test tool GATE to extract from website and evaluate the results objectively. This Increase the performance of the system.

6.4 Ontology-based information extraction from twitter

This paper proposed an approach for ontology-based information extraction from Twitter. This system provides an integrated disambiguation module based on popularity score and syntax-based similarity. Result shows that, the system performed significantly better using disambiguation process.

6.5 Abstraction based domain ontology extraction for Idea Creation

This paper [Delin Jing, Hongji Yang, Yingchun Tian] Proposed in IEEE 2013, an abstraction method to support one of the essential parts in creative idea creation- domain ontology extraction. Abstraction techniques are classified, selected and integrated while elements of domain ontology are defined including concepts and relations.

Survey Paper	Methods and algorithm	Performance
Text information extraction based on OWL ontology	Semantic information extraction and semantic information re-recognizing algorithm	System has good portability.
Web information extraction based on news domain ontology theory	Rule-based web information extraction method	Improved the performance of the system.
Ontology-based information extraction system based on E-commerce websites	GATE tool used to extract ontology based information.	Relevant performance of the system.
Ontology-based information extraction from twitter	Rule-based method used	Performance measured through BDM (balance distance metric).
Abstraction based domain ontology extraction for Idea Creation	Abstraction Based Method Used	Improved a performance.

VII. PROPOSED ALGORITHM

1. In this Paper we will implement the ontology-based information extraction search based on properties, keywords, Ontology search. Here we will use domain ontology and extract information from this ontology. And we will also compare this extracted information to old approach.

2. Semantic Web search: - We will also introduce the concept of semantic web search and we will compare it and then recommendation graph will be generated.

VIII. CONCLUSION

In this paper, we have reviewed the new field of ontology-based information extraction. In future there are several directions with OBIE like improving the efficiency of IE system to improve the precision and recall. OBIE system can be used for identified semantic content for semantic web and implemented the ontology using web services.

REFERENCES

[1] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 2010, 36(3): 306

[2] Delin Jing, Hongji Yang, Ying Chun Tian "Abstraction based domain ontology extracting for Idea creation" 2013 IEEE 13th International Conference on Quality Software.

[3] Yang Xiudan "Ontology-based information extraction system E-commerce website" 2011 IEEE Control, Automation and Systems Engineering(CASE).

[4] Y. Li and K. Bontcheva, Hierarchical, perceptron-like learning for ontology-based information extraction. In: Proceedings of the 16th International Conference on World Wide Web, (ACM, New York, 2007).

[5] F. Wu and D. S. Weld, Autonomously semantifying wikipedia. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, (ACM, New York, 2007).

[6] D.E. Appelt, J.R. Hobbs, J. Bear, D.J. Israel and M. Tyson, FASTUS: A Finite-state Processor for Information Extraction from Real-world Text. In: Ruzena Bajcsy (ed.), Proceedings of the 13th International Joint Conference on Artificial Intelligence, (Morgan Kaufmann, Chambéry, France, 1993).

[7] H. Cunningham, K. Bontcheva, V. Tablan, and D. Maynard, General Architecture for Text Engineering (GATE) (2003). Available at: http://www.gate.ac.uk (accessed 25 June 2009).

[8] H. Saggion, A. Funk, D. Maynard, and K. Bontcheva, Ontology-based information extraction for business intelligence. In: Proceedings of the 6th International and 2nd Asian Semantic Web Conference, (Springer, Berlin, 2007).

[9] P. Buitelaar and M. Siegel, Ontology-based Information Extraction with SOBA. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, (European Language Resources Association, Genoa, Italy, 2006).

[10] M. Moens, Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series) (Springer-Verlag, Secaucus, NJ, 2003).

[11] Y. Li, K. Bontcheva, and H. Cunningham, Using uneven margins SVM and perceptron for information extraction. In: Proceedings of the 9th Conference on Computational Natural Language Learning, (Association for Computational Linguistics, Morristown, NJ, 2005).

[12] J. R. Hobbs, M. Stickel, P. Martin, and D. Edwards, Interpretation as abduction. In: Proceedings of the 26th Annual Meeting on Association for Computational Linguistics, (Association for Computational Linguistics, Morristown, NJ, 1998).

[13] Hongsheng Wang, Lu Yuan, Hong Shao "Text information extraction based on OWL Ontologies" 2008 IEEE Fifth International Conference on Fuzzy Systems and Knowledge Discovery.

[14] YI Wei-Guo, LIU Ya-Qing YAN Ling-Wei, Liu Zhi "An ontology-based web information extraction algorithm" 2010 IEEE 2nd International Conference on future Computer and Communication.

[15] Berners-Lee T, HendlervJ, Lassila O. The Semantic Web. Sci Am, 2001, 284: 34-43.

