Impact Factor Value: 4.029 ISSN: 2349-7084
International Journal of Computer Engineering In Research Trends
Volume 4, Issue 11, November - 2017, pp. 539-542 www.ijcert.org

A Survey on Taxonomy learning using Graph-based Approach

Diksha R. Kamble*[1], Krishna S. Kadam [2]

[1] *PG Scholar, Dept. of Computer Science and Engineering, DKTE's TEI, Ichalkaranji (An Autonomous Institute), 416115, India.*
[2] *Professor, Dept. of Computer Science and Engineering, DKTE's TEI, Ichalkaranji (An Autonomous Institute), 416115, India.*
dikshakambale94@gmail.com[1], kadkrishna@gmail.com[2]

Abstract: - Taxonomy learning is an important task for developing successful applications as well as knowledge obtaining, sharing and classification. The manual construction of the domain taxonomies is a time-consuming task. To reduce the time and human effort will build a new taxonomy learning approach named as TaxoFinder. TaxoFinder takes three steps to automatically build the taxonomy. First, it identifies the concepts from a domain corpus. Second, it builds CGraphs where a node represents each of such concepts and an edge represents an association between nodes. Each edge has a weight indicating the associative strength between two nodes. Lastly TaxoFinder derives the taxonomy from the graph using analytic graph algorithm. The main aim of TaxoFinder is to develop the taxonomy in such a way that it covers the overall maximum associative strengths among the concepts in the graph to build the taxonomy. In this evaluation, compare TaxoFinder with existing subsumption method and show that TaxoFinder is an effective approach and give a better result than subsumption method.

Keywords: Taxonomy learning, ontology learning, TaxoFinder, concept taxonomy, concept graphs, similarity, associative strength.

1. Introduction

In the past, the documents are structured manually for of easy retrieval but it is a time consuming process, and it requires more knowledgeable person to structure the documents. It can be done by the concept of taxonomy and generate the structure by analyzing document corpus. Taxonomy learning keeps getting a more important process for knowledge sharing about a domain. It is also used for application development such as knowledge searching, information retrieval. The taxonomy can be build manually, but it is a very complicated process when the data are so large, and it also develops some errors while taxonomy construction. Various automatic taxonomy construction techniques are used to learn taxonomy based on keyword phrases, text corpus and from domain-specific concepts, etc. So it required to building the taxonomy with less human effort and with less error rate.

The most important goal of taxonomy learning is to build taxonomy from a text corpus which finds out the main characteristics of the given data. Hence it is more important to construct taxonomy for taxonomy learning. There are various techniques available for taxonomy learning. Some of the techniques are more accurate, and it clearly classifies a domain. Some of the techniques are a lexico-syntatic pattern, semi supervised methods, graph-based methods, etc. Basically, taxonomies are constructed from the collection of documents or websites or text corpus where the key phrases are extracted from the document and from the key phrases the concepts of the domain can be determined by using different algorithm and analysis the statistical and semantic relationship between the concepts to build the taxonomy. Likewise, various techniques are used to learn taxonomy. The main aim of all technique is to obtain enough data that covers the domain of interest thoroughly. There are various approaches and techniques among them TaxoFinder a graph-based approach for taxonomy learning to construct a good taxonomy.

TaxoFinder is an approach, which learns taxonomy based on graph representation. In this approach the concepts in text corpus were extracted, and the concepts were represented in graph representation to define the associative strength between the concepts. The associative strength determines how strongly the concepts are associated in the graph which is based on similarities and spatial distance between sentences. Yong-Bin Kang et al. proposed TaxoFinder he takes mainly three steps to automatically build taxonomy are as: First, from domain text corpus it identifies domain-specific concepts. Second, based on co-occurrences it builds a graph representation. Lastly, by using graph analytic algorithm TaxoFinder, induce taxonomy from the graph.

2. Literature Review

M. A. Hearst [1] described a method to automatic acquire the hyponymy lexical relation from unrestricted text. They motivate two main approaches first one is avoidance of the need for pre-encoded knowledge and the second one is applicability access a wide range of text. M.A.Hearst identifies easily recognizable set of lexico-syntactic patterns. This approach is low-cost automatic acquire of semantic lexical relations from unrestricted text.

F.M.Suchanek et al. [2] described the World Wide Web is an effective source of knowledge which is mostly in natural language. Data extract pairs of a given semantic relation from text documents automatically. Instead of surface text patterns, F.M.Suchanek et al. show that it's proposed approach profits significantly when deep linguistic structures are used. These structures are suitable for machine learning.

Wang Wei et al. [3] described for document modeling and topic extraction in information retrieval models are developed and utilized named is probabilistic topic models. In this approach topic models are used as efficient dimension reduction techniques, where they find out semantic relationships between word topic and topic document. They introduced two algorithms for learning terminological ontology using the principle of topic relationship and exploiting information theory with the probabilistic topic models learned. Compared the result of this method with two existing concepts of hierarchy learning methods on the same dataset, The result is shown this method gives better performance than another two existing systems regarding precision and recall measures.

For graph-based approaches builds a graph in which nodes represent concepts and edges are represents how to concepts are strongly connected to each other. Zornitsa Kozareva et al. [4] proposed a semi-supervised algorithm that uses a root concept. In this proposed method an algorithm is utilized to learn the different concepts of root concept, recursive surface level patterns and basic level concepts from the web hyponym-hypernym pairs subordinated to the root base. The learned hyponym-hypernym pairs are validated through a ranking mechanism in the web- based concept, and a graph algorithm is used to derive the combined taxonomy structure of all terms from scratch.

Clustering approach is created in the problem of taxonomy so solve this problem hierarchical clustering technique is suitable. E.A.Dietzet et al. [5] proposed TaxoLearn approach. In this approach combines existing approaches. But also they added one more new step to improving the quality of the resulted domain taxonomy. E.A.Dietz describes three main steps first one is word sense disambiguation to improving the quality (precision) of the taxonomy. The second step is used semantic-based hierarchical clustering for taxonomy learning. The third step describes the novel dynamic labeling procedure for clustering that used for large clustering are arranged properly. This approach is gives high precision and low recall because of many relations is hidden in the text semantics.

Another approach P. Velardi et al. [6] developed for definition sentences for each concept introduced term OntoLearn Reloaded. OntoLearn Reloaded method is used for automatic induction of taxonomy from numbers of documents and websites. In this approach learn the concepts and relations of documents to build taxonomy entirely from scratch. This concepts and relations are defined by automated terms extraction, automated definition extraction and hypernym extraction from this disconnected hypernym graph were obtained. Then the taxonomy is induced from novelweight policy and optimal branching.

K. Meijer et al. [7] presented a framework in which domain taxonomy from text corpora is automatically build. They named it Automatic Taxonomy Construction from Text (ATCT).ATCT is comprised in four steps. The first step, document corpus is extracted. Using filtering approach in second step most relevant term for the specific domain is selected. The third step, in which word sense disambiguation technique and concepts are generated. And finally, broader-narrower relations between concepts are determined. Using golden standard evaluation approach constructed taxonomy is compared with reference (benchmark) taxonomy. To retrieve quality of broader-narrower relations in the build taxonomy they use taxonomy precision and taxonomic recall. In generated ontology, K. Meijer et al. have additionally evaluated the effect of the disambiguation procedure. At the end to select most relevant in the domain of economics and management K. Meijer et al. constructed a taxonomy using a term filtering methods.

Y.B.Kang et al. [8] described in this paper CFinder method. Data extraction in domain corpus is a major step for ontology learning. The main aim of this is to build ontology by identifying relevant domain concepts and their semantic relationships from a text corpus. If the identified key concept is not closely related to the domain, then the constructed ontology will not be able to represent correctly. In this paper CFinder is used to extract key concept. They first extract noun phrases using their linguistic patterns based on part-of-speech (POS) tags as candidates for key concepts. CFinder combines their statistical knowledge indicating their relative importance within the domain for calculated the weights (or importance) of these candidates within the domain. The calculated weights are used later for inner structural pattern of the candidates. As per above discussion concluded that CFinder has a strong ability to improve the effectiveness of key concept extraction.

Yong-Bin-Kang et al. [9] mentioned in this paper a new taxonomy learning approach, which builds a high associative strength among the concepts called TaxoFinder. In this approach some concepts are given as input to the TaxoFinder to build taxonomies. Primarily there are three steps to construct taxonomy in that first step is identifies concepts from a domain corpus. Second is building Cgraph for extracted concepts. In the Cgraph node is represented concept and edge is a connection between those concepts. And the last step is, to calculate the associative strength of concepts and construct a good taxonomy. To calculated associative strength means how two concepts are strongly connected to each other.

3. Architecture of Proposed Work

The proposed system constructs taxonomy using graph-based unsupervised approach. The first step of taxonomy construction is extracting the concepts from given text corpus. Various approaches used to extract the concepts are machine learning approaches, Glossary-based approaches, Multiple-corpus based approach and hierarchical based approach. The second step is determining the optimal number of concepts and ranks those concepts. Then the third step is building a CGraph with an optimal number of concepts. This graph shows associative strength between the nodes and edges. Where nodes represent concepts and edges represent the relationship between those concepts. The main aim of constructing this CGraph is how strongly concepts are connecting with each other. Then finally taxonomies are constructing from the CGraph by maximizing the associative strength of all nodes in CGraph.

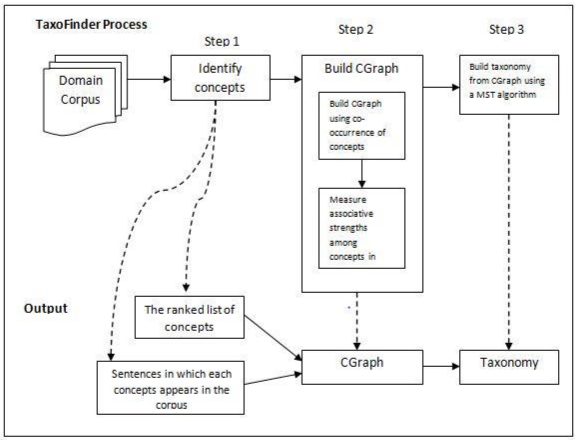
TaxoFinder Process
Step 1
Step 2
Step 3
Domain Corpus
Identify concepts
Build CGraph
Build CGraph using co-occurrence of concepts
Measure associative strengths among concepts in
Build taxonomy from CGraph using a MST algorithm
Output
The ranked list of concepts
Sentences in which each concepts appears in the corpus
CGraph
Taxonomy

Fig 1.System architecture diagram of TaxoFinder process [9]

4. Conclusion

In this survey of taxonomy learning studied different methods of taxonomy learning. Among them A TaxoFinder graph-based approach for taxonomy learning method generated good taxonomy. Because of some reason like, it measures the associative strength between concepts victimization the mix of the higher than 3 factors, unlike that determine classification relations victimization predefined lexico-syntactic patterns.

References

[1] M.A.Hearst, "Automatic acquisition of hyponyms from large text corpora," in Proc.14th Conf. Comput. Linguistics, 1992, vol. 2,pp. 539–545

[2] F.M.Suchanek, G.Ifrim, and G.Weikum, "Combining linguistic and statisticalanalysis to extract relations from web documents,"in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 712–717.

[3] E.-A. Dietz, D. Vandic, and F. Frasincar, "TaxoLearn: A semantic approach to domain taxonomy learning," in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol., 2012, pp. 58–65.

[4] W. Wang, P. Mamaani Barnaghi, and A. Bargiela,"Probabilistic topic models for learning terminological ontologies," IEEE

Trans.Knowl. Data Eng., vol. 22, no. 7, pp. 1028–1040, Jul. 2010.

[5] Z. Kozareva and E. Hovy, "A semi-supervised method to learn and construct taxonomies using the web," in Proc. Conf. Empirical Methods Natural Language Process., 2010, pp. 1110–1118.

[6] P. Velardi, S. Faralli, and R. Navigli, "OntoLearn Reloaded: A graph-based algorithm for taxonomy induction," Comput. Linguistics,vol. 39, no. 3, pp. 665–707, 2013.

[7] K. Meijer, F. Frasincar, and F. Hogenboom, "A semantic approachfor extracting domain taxonomies from text," Decision SupportSyst., vol. 62, pp. 78–93, 2014.

[8] Y.-B. Kang, P. D. Haghighi, and F. Burstein, "CFinder: An Intelligent Key Concept Finder from Text for Ontology Development,"Expert Syst. Appl., vol. 41, no. 9, pp. 4494–4504, 2014.

[9] Yong-Bin Kang, Pari Delir Haghigh, and Frada Burstein,"TaxoFinder: A graph-based approach for taxonomy learning." Vol.28, no 2,2016.

[10] Satish Kumar, Sujan Babu Vadde, " Typicality Based Content-Boosted Collaborative Filtering Recommendation Framework. "International Journal of Computer Engineering in Research Trends., vol.2, no.11, pp. 809-813, 2015.

[11] Y.Usha Sree,P.Ragha Vardhani." Pattern Finding in Large Datasets with Big Data Analytics Mechanism. "International Journal of Computer Engineering in Research Trends., vol.2, no.5, pp. 359-364, 2015.

Author Profile

Miss. Diksha R. Kamble pursed Bachelor of Engineering from DKTE Society's Textile & Engineering Institute, Ichalkaranji, India, in year 2016, She is currently pursuing Master of Technology from DKTE‘s TEI, (An Autonomous Institute), Ichalkaranji, India. Her research work focuses on text mining.

Prof .K. S. Kadam, Assistant Professor of Computer Science & Engineering, at DKTE Society's Textile & Engineering Institute, Ichalkaranji ,India. He is a member of the ISTE, CSI. His current research interests include Grid and Cloud Computing, Database Engineering, System Programming, Data Mining and Warehouse, Advanced Database and Compiler Construction, Big Data Analytics.