International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Implementation of taxonomy classification using Graph-based Approach

D. R. Kamble [1*], K. S. Kadam [2]

[1]Dept. of Computer Science and Engineering, D.K.T.E'S TEI (An Autonomous Institute), Ichalkaranji, India
[2] Dept. of Information Technology, D.K.T.E'S TEI (An Autonomous Institute), Ichalkaranji, India

e-mail: dikshakambale94@gmail.com

Available online at: www.ijcseonline.org

Accepted: 23/Jul/2018, Published: 31/July/2018

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.

I. INTRODUCTION

In the past the documents are structured manually for the purpose of easy retrieval but it is 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. The taxonomy can be build manually but it is a complex process when the data are so large and it also produce 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 is required to build 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 are available for taxonomy learning. Some of the techniques are more specifically classifies a domain. Some of the techniques are lexico-syntatic pattern, semi supervised methods, graph based methods etc. Basically taxonomies are developed from the collection of websites or documents 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 construct 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 techniques and approaches among them TaxoFinder a graph based approach for taxonomy learning to develop a good taxonomy.

TaxoFinder is an approach that learns taxonomy based on graph representation. In this approach the concepts extraction in text corpus and the concepts were represented in graph representation were nodes represents concepts or sentences and edges represents 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 method he take mainly three steps to automatically build taxonomy are as: First, identifying concepts from a domain corpus. Second, based on co-occurrences it builds a graph representation. Lastly, by using graph representation developed a good taxonomy.

II. RELATED WORK

M.A.Hearst [1] describes 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 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] 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 its proposed approach profits significantly when deep linguistic structures are used. These structures are suitable for machine learning.

Used hierarchical Clustering approach for taxonomy learning because of clustering approach developed some problems. E.A.Dietzet.et al. [3] proposed TaxoLearn approach. In this approach combines existing approaches. But also they added one more new steps for improve the quality of the resulted domain taxonomy. E.A.Dietz describes three main steps first one is word sense disambiguation for improve the quality (precision) of the taxonomy. Second step is use semantic-based hierarchical clustering for the purpose of taxonomy learning. In third step describes the novel dynamic labeling procedure for clustering that used for large clustering are arranged properly. This approach is give high precision and low recall because of many relations is hidden in the text semantics.

Wang Wei et al. [4] 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, were 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 in terms of precision and recall measures.

For graph based approaches builds a graph in which nodes are represent concepts and edges are represents how to concepts are strongly connected to each other. Zornitsa Kozareva et al. [5] proposed semi-supervised algorithm that uses a root concepts. In this proposed method an algorithm is utilized to learn the different concepts like 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 the scratch.

Another approach P. Velardi et al. [6] developed for definition sentences for each concepts introduced term OntoLearn Reloaded. OntoLearn Reloaded method is used to 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 the scratch. This concepts and relations are defined by automated terms extraction, automated definition extraction and hypernym extraction from this disconnected hypernym graph was obtained. Then the taxonomy is induced from novelweight policy and optimal branching.

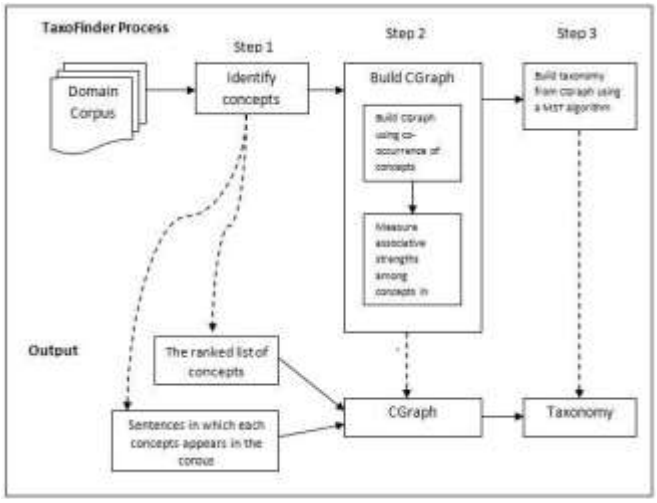
K. Meijer et al. [7] present a framework in which domain taxonomy from text corpora is automatically build. They named it Automatic Taxonomy Construction from Text (ATCT).ATCT is comprises in four steps. In first step, document corpus is extracted. Using filtering approach in second step most relevant term for specific domain is selected. 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.

III. METHODOLOGY

The TaxoFinder system is to classify basically in three steps. Using this steps create a taxonomy. The figure shows the system architecture of TaxoFinder process, the propose system construct taxonomy for finance domain using graph based unsupervised approach. 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.



System architecture diagram of TaxoFinder process[9]

Here, we used machine learning approach. Second step is determining the optimal number of concepts and rank those concepts. Then third step is building a CGraph with optimal number of concepts. This graph shows associative strength between the nodes and edges. Where nodes represent concepts and edges represents 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.

The TaxoFinder system has the following modules,

1. Identifying concepts from a domain corpus
2. Building a CGraph using extracted concepts
3. Deriving taxonomy from a graph
4. View result graphically

5.2.1. Identifying concepts from a domain corpus

Given a domain corpus, concept extraction is the first step for taxonomy learning. If extracted concepts are irrelevant, taxonomy may not correctly represent domain knowledge as such irrelevant concepts can also lead to generating irrelevant taxonomic relations. Most of existing approaches used to concept extraction like machine learning, multiple corpus-based, glossary-based and heuristic-based. In our system used, machine learning approaches that identify concepts from domain corpus using NLP techniques and then learn a classifier to identify which candidates are most likely to be concepts. The output of concept extraction step used in next step which is building a CGraph using extracted concepts.

5.2.2. Building a CGraph using extracted concepts.

In next step using extracted concepts and the sentences build a CGraph where a node represents each of such concepts and an edge represents an association between nodes. CGraph is built from concepts joined by undirected edges. This undirected graph called as bipartite graph (Bigraph). Bigraph consist of two types of nod8es one that represents concepts in C and other represents the collection of sequential sentences in S that contain the concepts. In Bigraph concepts are not connected to any other concepts, and sentences are also not connected with any other sentences. In Bigraph edge between the concept and a set of sequential sentences are represents concept appears in the set. After that using Bigraph construct CGraph in which two concepts are connected if they are appear together in the same set of sequential sentences. In CGraph construction associative strength is the most important estimation. The associative strength between two concepts con1 and con2 are defined as with respect to the document corpus is:

$$w\,(c1,\,c2)=\frac{1}{k}\sum_{j=1}^{k} w_j(c1,c2)|$$

Where k is the number of documents in D (i.e k=|D|), and (c1 c2) represents the associative strength between two concepts.

5.2.3. Deriving taxonomy from a graph

Once build a C-Graph, the third step is to derive taxonomy from it. Our eventual goal is to build taxonomy in such a way

that it maximizes the overall associative strengths among all concepts in C-Graph to find a good taxonomy.

5.2.4. View result graphically

Finally, display relation of different concepts graphically

I. Generative process of TaxoFinder system

All dataset are given as input to the TaxoFinder process. Using the following TaxoFinder method generates the taxonomy.

Input: Domain corpus

Output: Taxonomy

- **Generative process of LDA**

1. TaxoFinder processes are created in the following way:
2. Identify concepts from domain corpus.
3. Extract data from manually entered dataset.
4. Building a CGraph using co-occurrence of concepts.
5. Measuring associative strength of concepts

$$w(c_1, c_2) = \frac{1}{k}\sum_{j=1}^{k} w_j(c_1, c_2),$$

6. Deriving a taxonomy using CGraph
7. View result graphically.

IV. RESULTS AND DISCUSSION

A. *Result analysis using gold standard evaluation*

Evaluating a learned taxonomy is a crucial task for assessing products and branch analysis for representing of a domain. Our study represents the categories and sub-categories for products and their overall effect in an individual branch economy. For this evaluation 100% dataset was created by us. Our eventual goal is to build taxonomy in such a way that it maximizes the overall associative strength among the concepts in CGraph to find a good taxonomy.

In any evaluation Precision and recall are both extremely useful in understanding what set of documents or information was presented and how many of those documents are actually useful to the question being asked.
For example:

8. Domain	9. No of records
10. Learned taxonomy	11. 33000(Total Dataset)
12. Gold standard taxonomy	13. 20000(derived from Cgraph)

A user searches for Kurtis of first category out of 20000 possible records 5000 are kurtis. Let's say out of these 5000 records 2000 are results, to the relevant query links to query. So precision of query is 2000/5000 is equals to 0.04 is our precision value. And for recall 20000/33000 is equals to 0.060.

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

[3] [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] [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] [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] [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] [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] [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] [9] Yong-Bin Kang, Pari Delir Haghigh, and Frada Burstein,"TaxoFinder: A graph-based approach for taxonomy learning." Vol.28, no 2,2016.

Authors 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.

