ISSN (online) : 2277-5390
ISSN (Print) : 2277-5382

A Survey on Various Approaches for Taxonomy Construction

S. Sritha[*1], B. Mathumathi [2]

[1]*Student,* [2]*Assistant Professor, Dept of Computer Science, Sree Narayana Guru College, Coimbatore-641105, Tamil Nadu, India*
*srithamphil@gmail.com, madhumathisankar@gmail.com

Abstract

Objectives: To analysis different approaches for taxonomy construction to improve the knowledge classification, information retrieval and other data mining process.
Findings: Taxonomies learning keep getting 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 complex process when the data are so large and it also produce some errors while taxonomy construction. There is 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. This paper provides detailed information about those techniques.
Methods: The methods such as lexico-syntatic pattern, semi supervised methods, graph based methods, ontoplus, TaxoLearn, Bayesian approach, two-step method, ontolearn and Automatic Taxonomy Construction from Text are analyzed in this paper.
Application/Improvements: The findings of this work prove that the TaxoFinder approach provides better result than other approaches.

Keywords: Taxonomy learning, knowledge searching, TaxoFinder, keyword phrases.

1. Introduction

Taxonomy is the outcome of a classification process where categories are arranged in a hierarchical [1] subclass structure. In modern days the extraction and the implementation of domain specific taxonomies has become gradually more related. This is because of two main facts. One is, it is a tedious process in the field of information science and another one is domain taxonomy construction manually it takes more time for constructing taxonomies for a domain and it has been done by experts of a exacting domain. 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 accurate and it clearly classifies a domain. Some of the techniques are 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 taxonomy. Like as 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.

In [2] proposed a probabilistic method called lexico-syntactic pattern probabilistic method which is used to learn taxonomies. In this method two probabilistic models are defined are direct probabilistic model and induced probabilistic model. In the first model direct probabilistic model from the observations of text collection taxonomies are directly estimated. In induced probabilistic model induced the probabilities of derived events based on transitivity over direct probabilistic. In this model while estimating direct probabilities singular value decomposition is used as unsupervised method for feature selection.

In [3] presented a semi supervised method for taxonomy construction. In this proposed method an algorithm is utilized to learn the different concepts like root concept, recursive surface level patterns and a basic level concept 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.

In [4] proposed a methodology for learning taxonomic relations. In this the documents are collected where each document explained different concepts and define the relationship between the concepts by using three different feature extraction schemes. One of the schemes is based on statistical key phrase extraction which is language independent approach and another one scheme is combination of fuzzy logic-based feature weighting and selection and rule based stemming and one more schemes is based on rule based stemming with traditional tf-idf weighting scheme and hierarchical clustering is used. This approach is easily convenient and automatic to other domains and languages.

In [5] presented two new algorithms to learn terminological ontologies with the help of topic relationship and exploiting information theory by the probabilistic topic models learned standards. The dimension of the input data is reduced using an improved dimension reduction methods thus it can confine the semantic relationship among word-topic and topic-document take to mean in terms of probability distributions.

In [6] presented a methodology called ontoplus for the purpose of semi-automatic ontology extension and it fully depends upon text mining methods. It can be processed through a ranked list of relationships and potentially relevant concepts that provide a new concept have to be included in the ontology. Thus it provided an efficient extension of huge ontologies. Measures for ranking are depends on co-occurrence information, incorporating ontology content, and structure.

In [7] proposed a method to construct taxonomy from the system categories in Wikipedia. In this method category system is taken as a conceptual network and labeled the semantics relationship between categories. It is manually the quality of taxonomy and this method automatically compares the convergence of taxonomies with biggest manually ontology created and also the lexical database WordNet called ResearchCyc. Finally found semantic similarity between words for extrinsic evaluation.

In [8] proposed a new approach called TaxoLearn which automatically construct domain taxonomy. Initially detected concepts in text by using word sense disambiguation then learn the taxonomies by a semantics-based hierarchical clustering. Finally to cluster the concepts a novel dynamic labeling procedure is used. It used hierarchical clustering to construct domain cluster.

In [9] presented a new approach for automatic lexical taxonomy induction from text documents. A graph is used for taxonomy induction whose nodes define taxonomic terms and edges of the graph represent the degree of relationship. This graph is given as input and fits taxonomy to the graph by the combination of maximum likelihood approach with a Monte Carlo Sampling algorithm called Hierarchical Random Graph model (HRG).

In [10] developed a Bayesian approach to build taxonomy and described the problems involved in the taxonomy construction from keyword phrases instead of from text corpus in a document. The keywords category a domain more accurately but it does not explicit does not contain explicit relationships from which taxonomy can be constructed. In order to overcome problems in taxonomy construction from set of keywords knowledge along with context is proposed. With the help of Bayesian approach for taxonomy construction from set of keywords reduces time complexity of clustering approaches.

In [11] proposed a method to construct task specific taxonomies to maintain browsing in subjective document collections. This method was developed in two sub parts. One is handling path consistency and including specifications from users. With the help of supervised distance learning algorithm described a pair wise semantic distance thus it create a browsing taxonomies. The utilized supervised distance learning algorithm found out proximity between concepts and to know about the metric function. It permits the users to determine the way to arrange the concepts and it also found most excellent hierarchical structure as the browsing taxonomy.

In [12] proposed a new framework to develop taxonomies from collection of text corpus. Initially it utilized part-of-speech parser to taken out terms from the input text corpus. Then the extracted terms are filtered using domain consensus, structural relevance, domain pertinence, and lexical cohesion. The enduring terms represents the idea in the taxonomy. The subsumption method or hierarchical clustering algorithm is used to arrange the concepts in a hierarchy. In the subsumption method which determines the parent of concept for concept ancestors. Whereas in hierarchical clustering algorithm which utilized text based window and document scopes for idea co-occurrences for arrangement of concepts.

In [13] explained a method for constructing custom taxonomies. The taxonomies were constructed from the document collection. The construction of custom taxonomy involves five steps are initialization, extraction, connection, identification and selection. In initialization step, the documents are converted to text then in extraction stage used a Natural Language Processing (NLP) tools to extract the concept and named entities from the text. In the third step of connection connect the named entities to Linked Data sources and in the identification stage identifies

the conflicting steps and resolves them using an algorithm. In the last step of selection selected the semantic relations from that connect the concepts into single taxonomy. This approach is applicable any domain.

In [14] presented a new method to gain knowledge in a domain from unstructured text. This method is comprises of two sub processes called concept extraction and taxonomic relation extraction. In the concept extraction the concept involved in the text are extracted using clustering algorithm and in taxonomic relation extraction defines the relation between the taxonomies. These approaches included contextual information and WordNet synsets to build an extended query set. Then the extended query set is sent to a web search engine and hypernym for a term is obtained by processing the returned pages of search engine.

In [15] implemented an algorithm for automatic building of taxonomy for a vertical domain. In this algorithm the taxonomies are building through the seed entities and continued the process by mining presented source domains for new entities correlated with these seed entities. The taxonomies are constructed using new entities which are created by using machine learning of syntactic parse trees that created commonalities between search results. Thus these commonality expressions created new entities at the subsequent iteration. To equivalent natural language expressions between target and source domains, use syntactic generalization, a process which determines a set of maximal common sub-trees of constituency parse trees of these expressions.

In [16] developed a two-step method for determining estimation of statistical Ontology Learning (OL) algorithms that influences existing biomedical ontologies as suggestion standards. In the first step optimum parameters are created. In the second step, human judges with expertise in ontology development to estimate each candidate proposed by the algorithm organized with the optimum parameters previously established.

In [17] proposed an ontolearn Reloaded to automatic induction of taxonomy from number of documents and websites. In this approach learn the concepts and relations of document 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 novel weight policy and optimal branching.

In [18] proposed a new method for key concept extraction which plays an important role in ontology learning. The method is called CFinder which extracts the noun phrases from the collection of corpus document. The noun phrases are extracted by linguistic patterns in noun phrases the based on Part-Of-Speech (POS) tags. It perform as candidates for key concepts and the weight of the candidates was calculated by collective the statistical knowledge and domain specific knowledge which describes the relative importance within the domain it also consider the inner structural pattern of the candidates to calculate the weight.

In [19] presented a framework called Automatic Taxonomy Construction from Text (ATCT). This framework automatically creates taxonomy of domain from text documents. Initially take out terms from a document and then used filtering approach to select the terms that are more relevant to a domain. Further word sense disambiguation technique is applied on the filtered terms to disambiguate by means of semantics and from the disambiguation technique concepts are generated. Finally make use of submission technique to determine the broader narrower relationship from the concepts in a text corpus.

In [20] introduced an unsupervised computer aided tool to build taxonomies automatically which improves the performance of text classification. This tool uses the semantic knowledge base of Wikipedia where the Wikipedia category graph to determine the relationship between categories to automatic construction of taxonomy and classification schemes to classify collection of unstructured documents. In this tool first form a cluster with the related documents and then extract key phrase from the number of cluster formed which defines the main concept of each cluster which is done automatically with the help of Wikipedia category graph.

In [21] proposed a new taxonomy learning approach to build high associative strength among the concepts called TaxoFinder. Initially it found domain specific concepts from the domain text corpus. Based on the relationship between the concepts TaxoFinder build a graph and it measures the associative strength among the concepts which is main goal of TaxoFinder. The associative strength determines how strongly the concepts are associated in the graph which is based on similarities and spatial distance between sentences. Finally graph analytic algorithm is used to induce taxonomy from the graph.

ISSN (online) : 2277-5390
ISSN (Print) : 2277-5382

1.1. Comparison of various techniques based on parameters used

Ref. No	TITLE	PARAMETERS USED
[2]	Inductive probabilistic taxonomy learning using singular value decomposition	Accuracy for 100 pairs Direct probabilistic model =0.290 mixed probabilistic model =0.510 Accuracy for 1000 pairs Direct probabilistic model=0.269 mixed probabilistic model =0.322
[3]	A Semi-Supervised Method to Learn and Construct Taxonomies using the Web	Precision (induced vehicle taxonomy) = 0.99 Recall (induced vehicle taxonomy) = 0.60
[4]	Learning taxonomic relations from a set of text documents	For likey Taxonomic precision = 0.685 Taxonomic recall = 0.841 Taxonomic f-measure = 0.755
[5]	Probabilistic Topic Models for Learning Terminological Ontologies	For pLSA model Recall (LSHL+JS) = -0.58069 Recall (GSHL+JS) = -0.46031 For LDA model Recall (LSHL+JS) = -0.75429 Recall (GSHL+JS) = -0.75317
[6]	OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information	Hit rate for Financial Glossary = 60 For ASFA thesaurus = 40
[7]	Taxonomy induction based on a collaboratively built knowledge repository	For cyc Coverage = 1.6 Novelty = 99.2 Extracov = 28.2 For WordNet Coverage = 8.7 Novelty = 99.3 Extracov = 211.6
[8]	TaxoLearn: a Semantic Approach to Domain Taxonomy Learning	Precision (PMI) = 0.69 Recall (PMI) = 0.21
[9]	Taxonomy Induction Using Hierarchical Random Graphs	Tree correlation (HRG) = 0.412 Tree correlation (Brown) = 0.181 Tree correlation (Agglo) = 0.274 Tree correlation (Agreement) = 0.511
[10]	Automatic Taxonomy Construction from Keywords	Likelihood (BRT) = -1.441× $10^6 \pm 1.637 \times 10^5$ Likelihood (Knn-BRT) = -1.392× $10^6 \pm 1.053 \times 10^5$ Likelihood (spilltree-BRT) = -1.473× $10^6 \pm 1.837 \times 10^5$ Likelihood (spilltree-BRT) = -1.484× $10^6 \pm 1.348 \times 10^5$
[11]	Constructing Task-Specific Taxonomies for Document Collection Browsing	EMIM (PDistOpt) = 5.2 Reach time (PDistOpt) = 5.2
[12]	Domain taxonomy learning from text: The subsumption method versus hierarchical clustering	Average (cluster linkage) = 0.5432 TF Window (size:17) = 0.6916 TF Subsumption(t=0.25) = 0.6296 TF
[13]	Constructing a Focused Taxonomy from a Document Collection	Error = 229 Rate = 26.4%

[14]	Learning concept hierarchies from textual resources for ontologies construction	F-measure (LonelyPlanet) = 0.74 (concept extraction) 0.67 (taxonomic relation extraction)
[15]	Transfer learning of syntactic structures for building taxonomies for search engines	For Relevancy of re-sorting by using taxonomy and generalization, %, averaging over 20 searches Accuracy = 93.6
[16]	Formative Evaluation of Ontology Learning Methods for Entity Discovery by Using Existing Ontologies as Reference Standards	For NCIT entities precision = 51% F-measure = 0.15
[17]	OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction	Precision For finance domain TREE =93.6% DAG = 93%
[18]	CFinder: An intelligent key concept finder from text for ontology development	Average precision (CFinder) = 0.662 F-measure (CFinder) = 0.53
[19]	A semantic approach for extracting domain taxonomies from text	For word sense disambiguation technique Specific precision = 0.1163 Specific recall = 0.0557 Taxonomic precision = 0.8190 Taxonomic recall = 0.5837
[20]	Unsupervised Data Driven Taxonomy Learning	Precision = 88.6% Recall = 81.2%
[21]	TaxoFinder: A Graph-based Approach for Taxonomy Learning	TaxoFinder for EMD domain Taxonomic precision = 0.58 Taxonomic recall = 0.63 Taxonomic F-measure = 0.61

2. Conclusion

There are various approaches and techniques are used to learn taxonomy to classify the data and for knowledge acquisition, sharing and for application development. Among them TaxoFinder a graph based approach for taxonomy learning to find a good taxonomy which maximize the associative strength among the concepts.

3. References

1. S. Porna Sai, S. Udhaya, R. Suganya, K.S. Sangeetha. Supporting Privacy Protection in Personalized Web Search - A Survey. *Indian Journal of Innovations and Developments. 2014,* 3(3), 45-49.
2. F. Fallucchi, F. M. Zanzotto. Inductive probabilistic taxonomy learning using singular value decomposition. *Natural Language Engineering*. 2011, 17(01), 71-94.
3. Z. Kozareva, E. A. Hovy. semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2010, 1110-1118.
4. M. S. Paukkeri, A. P. García-Plaza, S. Pessala, T. Honkela. Learning taxonomic relations from a set of text documents. In *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference.* 2010, 105-112.
5. W. Wang, P. M.Barnaghi, A. Bargiela. Probabilistic topic models for learning terminological ontologies. *IEEE Transactions on Knowledge and Data Engineering*. 2010, 22(7), 1028-1040.
6. I. Novalija, D. Mladenić, L. Bradeško. OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information. *Knowledge-Based Systems*. 2011, 24(8), 1261-1276.

7. S. P. Ponzetto, M. Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*. 2011, 175(9-10), 1737-1756.
8. E. A. Dietz, D. Vandic, F. Frasincar. Taxolearn: A semantic approach to domain taxonomy learning. In *Web Intelligence and Intelligent Agent Technology (WI-IAT).* 2012; 1, 58-65.
9. T. Fountain, M. Lapata. Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.2012, 466-476.
10. X. Liu, Y. Song, S. Liu, H. Wang. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, 1433-1441.
11. H. Yang. Constructing task-specific taxonomies for document collection browsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2012; 1278-1289.
12. J. De Knijff, F. Frasincar, F. Hogenboom. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data & Knowledge Engineering* 2013; 83, 54-69.
13. O. Medelyan, S. Manion, J. Broekstra, A. Divoli, A. L. Huang, I. H. Witten. Constructing a focused taxonomy from a document collection. In Extended Semantic Web Conference. 2013, 367-381.
14. A. B. Rios-Alvarado, I. Lopez-Arevalo, V. J. Sosa-Sosa. Learning concept hierarchies from textual resources for ontologies construction. *Expert Systems with Applications*. 2013; 40(15), 5907-5915.
15. B. A. Galitsky. Transfer learning of syntactic structures for building taxonomies for search engines. *Engineering Applications of Artificial Intelligence*. 2013, 26(10), 2504-2515.
16. K. Liu, K. J. Mitchell, W. W. Chapman, G. K. Savova, , N. Sioutos, D. L. Rubin, R. S. Crowley. Formative evaluation of ontology learning methods for entity discovery by using existing ontologies as reference standards. 2013; 52(4), 308-16.
17. P. Velardi, S. Faralli, R. Navigli. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. 2013, 39(3), 665-707.
18. Y. B. Kang, P. D. Haghighi, F. Burstein. CFinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications*. 2014; 41(9), 4494-4504.
19. K. Meijer, F. Frasincar, F. Hogenboom. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*. 2014; 62, 78-93.
20. M. M. Hosny, S. R. El-Beltagy, M. E. Allam. Unsupervised Data Driven Taxonomy Learning. In *2015 First International Conference on Arabic Computational Linguistics.* 2015; 9-14.
21. Y. B. Kang, P. D. Haghigh, F. Burstein. TaxoFinder: A Graph-Based Approach for Taxonomy Learning. *IEEE Transactions on Knowledge and Data Engineering*. 2016; 28(2), 524-536.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

S. Sritha, B Mathumathi. A Survey on Various Approaches for Taxonomy Construction. *Indian Journal of Innovations and Developments.* 2016; 5 (6), June.