IJCSE International Journal of Computer Sciences and Engineering Open Access
Technical Paper Vol.8, Issue.4, Apr 2020 E-ISSN: 2347-2693

TaxoFinder A Graph-Based Technique for Taxonomy Learning

Abhijeet Ashokrao Kadam[1*], Shivputra Guruling Swami[2]

[1]Dept. of Computer Science & Engg. M.S. Bidve Engineering College, Latur
[2]Dept. of Computer Science & Engg. M.S. Bidve Engineering College, Latur

Corresponding Author: abhijeet.kadam13@gmail.com

DOI: https://doi.org/10.26438/ijcse/v8i4.129132 | Available online at: www.ijcseonline.org

Received: 22/Mar/2020, Accepted: 13/Apr/2020, Published: 30/Apr/2020

Abstract- Taxonomy is an essential process for gaining, sending, and classifying information, and also creating and using applications in several fields. To minimize humans, work to form the taxonomy learning from scratch and then increase the consistency of the taxonomy, now we suggest an approach to taxonomy learning, called TaxoFinder. TaxoFinder does three stages to construct a taxonomy automatically. Next, it distinguishes notions which are specific to the domain from a corpus of text. Later, it develops a graph describing how these definitions are connected at once depending on their co-occurrences. We will provide a technique for calculating strengths of associative between the concepts as the main method in TaxoFinder, which proves the strength and how tightly they have associated in the graphs, Using their similarities and spatial differences in sentences. Then lastly, have the TaxoFinder which uses a graph-analytical algorithm to trigger a taxonomy. TaxoFinder attempts to construct a taxonomy in such a way that to create a taxonomy, it enhances the associative strengths between the concepts in the graph. We test TaxoFinder on three separate domains using the gold standard evaluation: Mass-meetings emergency response, autism research and disorder domains. We evaluate TaxoFinder as the very effective subsumption method in this development, and it reveals that TaxoFinder was an efficient solution that successfully outperforms the subsumption process.

Keywords: Knowledge searching, Taxonomy learning, Taxonomy, TaxoFinder, keyword phrases

I. INTRODUCTION

In todays, Taxonomies are the secret to effective domain applications, including information retrieval (IR), knowledge search, and classification. Specifically, by considering the ever-increasing the level of digital text data each year, text-learning taxonomy is an initial research field for improving these applications. In a particular domain, the aim of taxonomy learning is to construct a taxonomy instantly or pseudo-automatically by finding Domain concepts (and then referred to as concepts) and their taxonomic partnerships with other relevant knowledge, if necessary, from the domain content corpus. An important feature of a taxonomy is that it enables the representation of closely linked concepts together, and the path in between two concepts that demonstrate how they are lexically linked within the domain.

A taxonomy is sometimes called to as the 'backbone of an ontology' built using the significant relationship between 'is-a.' Because of this association, taxonomy learning is often seen as a prerequisite for ontology learning, which aims at extracting concepts, relationships and often hypotheses about both the concepts to construct ontology. Creating a taxonomy manually presents a great task that needs a tremendous amount of human time and energy. Taxonomy research uses techniques built in the areas of natural language processing (NLP), information retrieval, and machine learning (ML), in a trial to minimize human effort and create a high-quality taxonomy.

II. EXISTING SYSTEM

Current techniques to testing learned taxonomies can be categorized as four groups depending on how assessment is performed Application oriented or also task based assessment measures the consistency of trained taxonomies in the context of applications by analysing their effect on the performance enhancement dimension of applied applications. Between such a learned taxonomy and a topic-specific corpus denoting targeted domain intelligence. Normally it tests terminology reportage of the learned taxonomy with regard to key words derived from the corpus Domain-expert assessment depends on human judges with appropriate domain knowledge to determine the accuracy of the learned taxonomies. Trends that satisfy constraints on derived scores were called emerging trends, and comparison sets for subgroup descriptions correlated Patterns, patterns of prejudice, and laws of interest. In this case one might not be interested to find all patterns that satisfy the constraints. Instead, when a set of patterns is created, one might be interested in finding top k scoring patterns, or finding top k patterns, for example, in the training data, this can be achieved in two ways.

III. PROBLEM STATEMENT

- Consider also the effect of negative training examples on patterns to identify unclear (noisy) patterns.
- Pattern evolution can be referred to as the method of updating vague patterns.

- Only the concept of a pattern shifts supports throughout the pattern.
- It suffers from polysemia and synonymy problem s.

IV. PROPOSED SYSTEM

Using a concept extractor, it distinguishes the concepts from the domain corpus. The production of this phase involves the ranked list of concepts as per the specific domain and sentence details in that every concept seems in the corpus in every text. Second, TaxoFinder forms a CGraph that describes the way of concepts are combined depending on their pre-occurrences. Their associative power is quantified using the statement similitude and sentence distance measurements. Eventually, a taxonomy using a graph analytical algorithm is built from the CGraph. The collection of the concept is the principal process of taxonomy learning given a subject corpus. When extracted concepts are sort of irrelevant, a taxonomy could not accurately indicate domain knowledge, as these irrelevant concepts can appear to establish irrelevant taxonomic connections.

ADVANTAGES

- Bag of words method is how to pick a limited range of features from a large set of words or terms to increase the effectiveness of the device and avoid overfitting
- Calculation of two strings to find string measurement.

V. TECHNOLOGY USED

Microsoft's new technology development platform, .NET Framework, is Microsoft's first programming system developed from the bottom to the top of internet programming. While. NET was not only planned for Internet development as well as for window expansion, its developments have been inspired by the limitations of current tools and methods for network creation.

DATA BASE USED

A database resembles a data file, because it is a storage location. A sql database does not directly transmit information as a data file; the user now runs an application and that application will accesses the data which is coming from the sql database and it represents simply and comprehensibly to them. We make use of the 2005 microsoft sql Server within this project.

About Microsoft SQL Server 2005

Microsoft SQL Server is a client or server relational database, based on Structured Query Language (SQL). Each of these words describes a specific piece of SQL Server architecture.

VI. IMPLEMENTATION

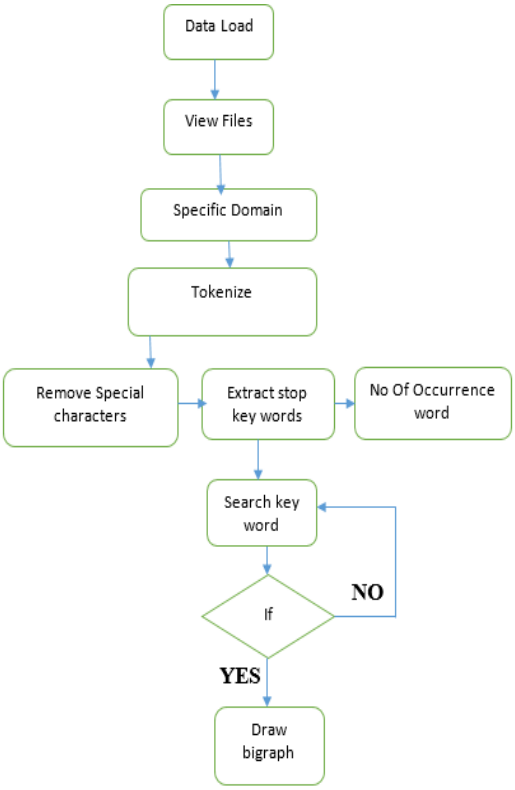

Fig.1

In the implementation we have used following modules such as

- Data loading
- Extract keywords
- Similarity of the data
- Represent relationship

MODULES DESCRIPTION

Data Loading

- Data loading is process of input that load text file that contains information
- It has some attributes like domain name, title, year etc.,
- It support different format of files like text file and pdf files
- Load data are stored to process the task for mining text.

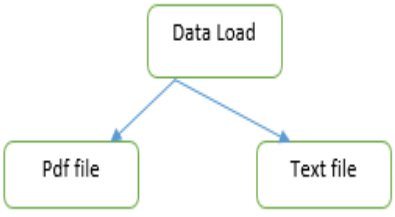

Fig.2

Extract Keywords

- Mining is Text mining process, also known as text data mining, loosely similar to text analytics, refers to the method of deriving high-quality text information.
- Strong quality knowledge is usually obtained from the analysis of patterns and trends through methods such as learning statistical patterns.
- From data we can extract keywords from the database it removes special characters

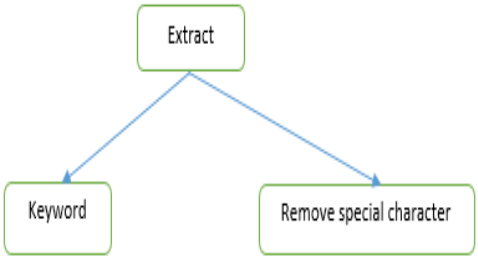

Fig.3

Similarity of Data

- From the extracted keywords all keywords are stored in database and calculate two data to match of words
- First we analysis two data from the data loading it show the how many characters are not matched from the data
- It also count number occurrence from data loading.

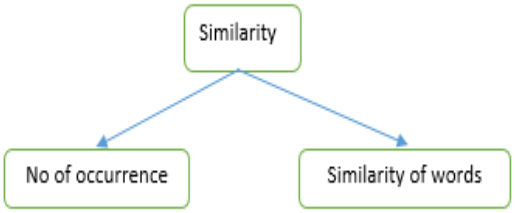

Fig.4

Represent Relationship

- In Text mining relationship are represented by bigraph to identify the words are occur in particular file
- If consider one words are occurs in both files show it represent graph by arrow
- Once search data from the particular files and gives exact result to the relationship

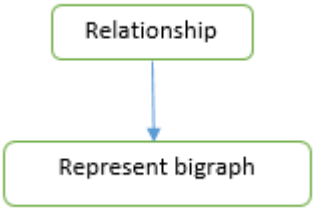

Fig.5

VII. CONCLUSION

Taxonomy learning technique, TaxoFinder, and demonstrated its efficacy on three separate domains against a recent method of sub Sumption. TaxoFinder begins to construct a graph, CGraph, that presents the theories extracted from both the corpus of a subject and its associative abilities.
(1) The probability that concepts would co-occur in a moving window, i.e. set of concurrent sentences and (2) The distance and similarities between the statements in which these definitions coexist. We used the CGraph's analytical algorithm to cause a taxonomy that tries to enhance the entire associative capacity between concepts to achieve a good taxonomy. Our evaluation has shown that TaxoFinder is a very good method for learning taxonomy; Significantly outshining the 99.9 per cent confidence subsumption cycle using a typical three-domain gold check method.

VIII. FUTURE SCOPE

We expect to test TaxoFinder with different criteria in our future research e.g., computational complexity, and compared TaxoFinder with hierarchical clustering methods that produce linked and deep taxonomies. We also intend to learn as many concepts as possible as the TaxoFinder data, rather than using a set number of concepts. In addition to an MST algorithm, we will attempt to apply specific graph analytical methods (e.g., local node connectivity) to build a taxonomy. Also it would be important to research the integration of Word2-Vector 9 into TaxoFinder as an alternate technique for learning the relationships between concepts.

REFERENCES

[1] K. Meijer, F. Frasincar, and F. Hogenboom, “A semantic approach for extracting domain taxonomies from text,” Decision Support Syst., vol. 62, pp. 78–93, 2014.
[2] W. Wong, W. Liu, and M. Bennamoun, “Ontology learning from text: A look back and into the future,” ACM Comput. Surv. vol. 44, no. 4, pp. 20:1–20:36, Sep. 2012.
[3] M. A. Hearst, “Automatic acquisition of hyponyms from large text corpora,” in Proc. 14th Conf. Comput. Linguistics, 1992, vol. 2, pp. 539–545.
[4] P. Pantel and M. Pennacchiotti, “Espresso: Leveraging generic patterns for automatically harvesting semantic relations,” in

Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meet. Assoc. Comput. Linguistics, 2006, pp. 113–120.
[5] X. Liu, Y. Song, S. Liu, and H. Wang, "Automatic taxonomy construction from keywords," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1433–1441.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, Introduction to Algorithms, 2nd Ed. New York, NY, USA: McGraw-Hill, 2001.
[12] K. Dellschaft and S. Staab, "Strategies for the evaluation of ontology learning," in Proc. Conf. Ontol. Learn. Population: Bridging Gap Between Text Knowl, 2008, pp. 253–272.
[13] 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.
[14] S. P. Ponzetto and M. Strube, "Taxonomy induction based on a collaboratively built knowledge repository," Artif. Intell. , vol. 175, no. 9-10, pp. 1737–1756, Jun. 2011.
[15] A. B. Rios-Alvarado, I. Lopez-Arevalo, and V. J. Sosa-Sosa, "Learning concept hierarchies from textual resources for ontologies construction, " "Expert Syst. Appl., vol. 40, no. 15, pp. 5907–5915, Nov. 2013.

