A STUDY OF DATA MINING TECHNIQUES FOR AUTOMATION ONTOLOGY

Sreedhar Pulipati[1], Dr. Neeraj Sharma[2]

[1]Research Scholar, Department of Computer Science Engineering, of Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

[2]Research Supervisor, Department of Computer Science Engineering, of Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

Abstract

Ontologies have grown in popularity and renown on the semantic web as a result of its extensive usage in Internet-based applications. Ontologies are widely recognised as a useful source of semantics and interoperability in all artificially intelligent systems. In academic circles, the explosion of unstructured internet data has created an urgent need for automated ontology extraction from unstructured text. Several strategies using a range of techniques from diverse domains (machine learning, text mining, knowledge representation and reasoning, information retrieval, and natural language processing) are being presented to help automate the process of acquiring ontologies from unstructured text. — Semantic Data mining is a term used to describe activities that use domain knowledge, such as formal semantics, as part of the process. In the past, several studies have shown the importance of incorporating domain knowledge into data mining. A simultaneous development in Knowledge Engineering, notably formal semantics and Semantic Web ontologies, has widened the domain knowledge family yet more still.

Keywords: *Ontologies, Data mining, Internet-based applications.*

1. INTRODUCTION

It is sometimes called knowledge discovery from databases (KDD), data mining involves finding previously undiscovered and potentially useful information in large datasets. The development of data mining methods has led to significant changes in data analytics and big data during the previous several decades. Data mining uses methods from a wide range of domains, including statistics, artificial intelligence, machine learning, database systems, and others, to evaluate large data sets. Semantic data mining refers to data mining activities that systematically include domain knowledge, notably formal semantics, into the process. Prior research has shown the importance of domain knowledge when it comes to data mining. There are several data mining steps that may benefit from domain knowledge, such as data transformation, feature reduction, selection of methods, post processing, model interpretation, and others.

An intelligent agent (such as a data mining system) should be able to get background information and use that information to learn more quickly. Domain knowledge has been demonstrated to improve data mining in previous semantic data mining studies. Domain knowledge, for example, might aid with pre-processing by removing duplicate or inconsistent data. Knowledge about constraints may be collected using domain knowledge to assist reduce search space and steer the search route when generating patterns and conducting searches.

2. ONTOLOGIES' FUNCTION IN SEMANTIC DATA MINING

Various systems and applications use ontologies in semantic data mining from different viewpoints and techniques. When it comes to data mining, it's not quite apparent how ontology helps. We evaluate previous ontology-based methodologies to identify the three main reasons why ontologies were introduced to semantic data mining:

- ✓ To eliminate the semantic gap between data, applications, data mining algorithms, and data mining outcomes.
- ✓ To supply a priori information to data mining algorithms that either guides the mining process or reduces/constricts the search space?
- ✓ To provide a standardised manner to depict the data mining process, from data pretreatment to mining outcomes.

A. Bridging the semantic gap:

There has been much discussion on the use of domain knowledge in the data mining process in previous semantic data mining studies. To far, experts think there is a knowledge chasm between the underlying datasets, the data mining methods used to extract them, and the resulting mining results at every level of data mining. In data preparation we use a variety of techniques to make sure our data is as clean and as usable as possible. Most of the time, data preparation processes have semantic flaws. Data quality is determined using ad hoc or empirical approaches rather than formal semantics. Outliers and missing data may be found by using principles like scarcity and closest neighbour. Normalization and transformation need knowledge of data semantics to understand the data connections. It's vital to look at the relationship between data characteristics and attributes before finishing data normalisation, for example.

Qualities that are closely related may be combined into a single value. Up practise, domain professionals fill in semantic gaps manually all the time. Data preparation activities have been shown to benefit from using ontologies, on the other hand. Furthermore, the data mining process and the data have semantic differences. To handle data from many domains and situations, data mining algorithms are often developed. Domain-specific information on the other hand contains meanings that are unique to the given context. Generic data mining technologies lack the ability to recognise and leverage semantics across domains and applications. Ontologies may be used to define domain semantics and to annotate data with rich semantics to fill in semantic gaps. Semantic annotation gives the core component of information links to formal semantic descriptions. It's important that the semantics of the source be composed of little chunks like this. Semantic annotation helps semantic data mining by giving formal meanings to the data. The annotated data are especially beneficial for further phases of semantic data mining because they have been elevated to a formal and organised style that links ontological terms and relations. Bridging the semantic gap between data mining discoveries and users has taken a lot of research. Post-pruning and filtering ontologies for association rule mining findings to help with subjective analysis for the post-pruning job

Data mining findings may be represented using ontologies in a semantic rich style, making it simpler to share and reuse them. Information extraction, for example, is the process of automatically gathering structured data from text (IE). Data/text mining yields collections of organised information as well as domain-specific expertise. For information to be organised and machine-readable, ontology representation is a logical choice. For Ontology-Based Information Extraction, this paradigm has been extensively used (OBIE). On top of being appropriately organised, the information obtained using OBIE is also represented in the ontology via predicates that are simple to exchange and reuse..

B. Providing background information and constraints:

Semantic data mining has several challenges, one of which is defining and utilising previous information. As a formal definition of concepts and connections, ontology is a natural way to express the formal semantics of prior knowledge. The embedded prior knowledge has the ability to control and affect all aspects of the data mining process, from preprocessing to result filtering and representation. For example, an RDF hypergraph structure may be used to collect data and ontology information. When constructing a graph, ontologies are employed as a kind of previous knowledge to help shape the structure. The approach turns the hyper graph and weighted hyper edges into a bipartite graph to convey data and ontology in a unified format. Random walks with restarts through the bipartite network are used to establish semantic associations. A trip across ontology-based edges connects the latent semantic connections under data with rich meaning provided by domain knowledge found in ontologies. Because ontology is a collection of concepts and predicates, it may engage in logical reasoning and draw conclusions about the coherence of those concepts. The capacity to derive consistency inferences is typically expressed using constraints

in semantic data mining. The collection of constraints driven by the ontology can identify inconsistent input and outcomes in the preparation stage, the algorithm execution stage, and the result filtering and generating stage, for example, as consistency constraints in many linked classification jobs.

A classification ontology identifies the constraints that must be satisfied when utilising a semi-supervised information extraction strategy to train many information extractors at the same time. When ontology is utilised as a constraint on the collection of extractors, the results are more accurate. Results from association rules mining are post-processed using ontology to ensure consistency. Invalid or inconsistent association rules are pruned and sorted using ontology and an inference engine.

3. MINING WITH ONTOLOGIES

Formally encoded semantics in ontology make it useful for a wide range of data mining tasks. In this section, we examine semantic data mining approaches for a range of use cases, such as association rule mining, classification, clustering, recommendation, information extraction, and link prediction (among others).

A. Association Rule Mining based on Ontologies:

In many different applications, association rule mining is a common data mining job. created an association mining tool that uses ontologies in all four phases of the mining process: data understanding, task design, result interpretation, and internet distribution of results in early work An ontology-based association rule mining approach called Semantic Web searches the ontology for cases that are used in the process of filtering instances using ontology limitations for queries in the association mining process. Due to the ontology query, the search area for association mining is constrained because of the need to reject or characterise intriguing items at a higher abstraction level from the output association rules. The user restrictions include pruning constraints for removing uninteresting things and abstraction requirements for generalising a notion to an ontology.

Checking for uniformity Using ontologies and an inference engine to post-process association rule mining discoveries, faulty or inconsistent association rules may be pruned and removed. It was recently discovered that a bipartite hypergraph model was utilised to link ontology and data in order to detect latent association rules in the data. Random walk-based metrics were presented to evaluate the latent semantic distances between ideas and sentences. Priority is given to the term sets with the shortest semantic distances. The most important word sets are built using linkages that are as strong as possible.

B. Ontology-based Classification:

One of the most common data mining tasks is constructing a model (or function) to describe and recognise different types of data. Semantic data mining uses ontologies to annotate classification labels with the ontology's collection of relations. Ontology annotated classification labels, according to study, may affect not only the labelled data in the classification process, but also handle vast volumes of unlabeled data in the classification task. They used ontology as a restriction on consistency in a variety of classification exercises. At the same time, these problems categorise a huge number of different groups. Ontology establishes the limitations on which tasks may be classified as belonging to certain categories. The unlabeled mistake rate indicates the likelihood that the classifier will categorise the unlabeled data in a way that is contrary to the ontology. Using the classifiers with the lowest unlabeled error rate and hence the best classification consistency, this classification task gives the classification hypothesis. Text documents are automatically classified using an ontology-based method into a dynamically generated collection of relevant topics. A DBpedia-based ontology is used to recognise entities and connections between entities in a text source. A semantic network is built using the collection of relations. Semantic graphs may be used to discover dynamic themes using the HITS technique, which locates the graph's core constituents. The semantic network of a document's ontology is compared to other documents to see how similar they are (topics).

4. WEB MINING TECHNIQUES-BASED ONTOLOGY LEARNING PROCEDURES

Massive amounts of data are being created on the World Wide Web, resulting in an avalanche of fresh data. The semantic web has been presented as a way to deal with the issue of too much information being available at the same

time online. Ontology is critical in defining the semantic web. Today's information systems regard ontology as a representational model that portrays domain-specific knowledge with well stated standards that enable human-computer interaction. Due to the widespread usage of ontology in knowledge-based systems and on the semantic web, efficient techniques for ontology construction are required. For identifying ontological information from various online resources, such as unstructured, semi-structured, and fully structured texts, ontology learning is a fully or semi-automated technique. Nowadays, the vast majority of web content are written in a semi-structured manner. Semi-structured data for identifying meaning for ontology learning has been the subject of very few studies in the literature. Embedded information in semi-structured articles is often overlooked by researchers, as this study reveals (see below). Few recent research projects have concentrated on semi-structured data in ontology construction, with the bulk of them using web mining methods..

Web mining methods have recently been used to develop a number of ontology learning frameworks and approaches. By deriving ontological structures from semi-structured online information, these methods hope to cut down on the amount of time and money it takes to generate and evolve ontologies. In this section, the various ontology learning frameworks and methodologies have been explained and compared in table form based on the most important ontology learning process tasks. Web use data and text mining technologies have been combined to provide a framework for semantic online adaptation based on web usage data. Because of the adaption process, the website's ontology has developed semi-automatically. The log file of the web server, the topology of the site, and the ontology are all inputs into the adaptation process. All of this is being done to make the website's topology and ontology better. Finding new non-taxonomic connections between ontology ideas might be a step in the process of modifying it, according to the author.. These non-taxonomic associations are extracted using a classification strategy based on support vector machines.

The developed ontology has not been assessed using any evaluation methodologies. Using a strategy that blends online content mining with web usage mining, researchers were able to find conceptual links between different semi-structured web sites. The obtained information was used to construct the ontology. Associative mining methods and approaches to natural language processing are used in this study to discover conceptual connections. In the proposed technique, the ontology was examined by comparing it to user and gold standard assessment methodologies. The study's goal was to show how online use mining may have an influence on ontology evolution and website rearrangement. Using the website's web log file to gather information, this technique aims to put newly found knowledge up against the present ontology in order to generate new concepts. There are two techniques used to figure out what the relationships between the ideas are: clustering and sequential pattern mining. The ontology has not been assessed using any evaluation methodology. A framework is created by using ontology-based online use mining. A pre-existing or semi-automatically-built ontology is used to enhance information retrieval.

- **Ontology Learning:**

Ontology is a knowledge representation, distribution, and repurposing approach. Knowledge acquisition has been cited as a key challenge in the ontology construction process by academics in the literature. The term "knowledge acquisition bottleneck" refers to the difficulty that knowledge-based systems have in acquiring the information they need. To assist ontology engineers and overcome the knowledge acquisition barrier, a variety of methodologies have been proposed in the literature. On the other hand, knowledge-based and semantic web systems need an expedient and suitable ontology construction procedure. Ontology learning offers a viable solution to this problem since it allows ontology construction to be fully automated or semi-automated.

One of the most significant advancements in ontology development has been ontology learning. A collection of frameworks used for ontology creation and accommodating semi-automatic ontology from various sources is referred to as a "bundle" in this context. Knowledge acquisition is the focus, and in this research, knowledge acquisition is derived especially from online content and web use statistics. For example, it is an interdisciplinary area that utilises methods from many different academic disciplines such as semantic web, machine learning (ML), logic and knowledge representation as well as philosophy and databases. Semi-automatic ontology building has seen some work in the previous several years. There are a plethora of methods for defining ontology, each with their own advantages and disadvantages. In the ontology learning process, the following six kinds of specialised tasks are shown in Fig. 1:

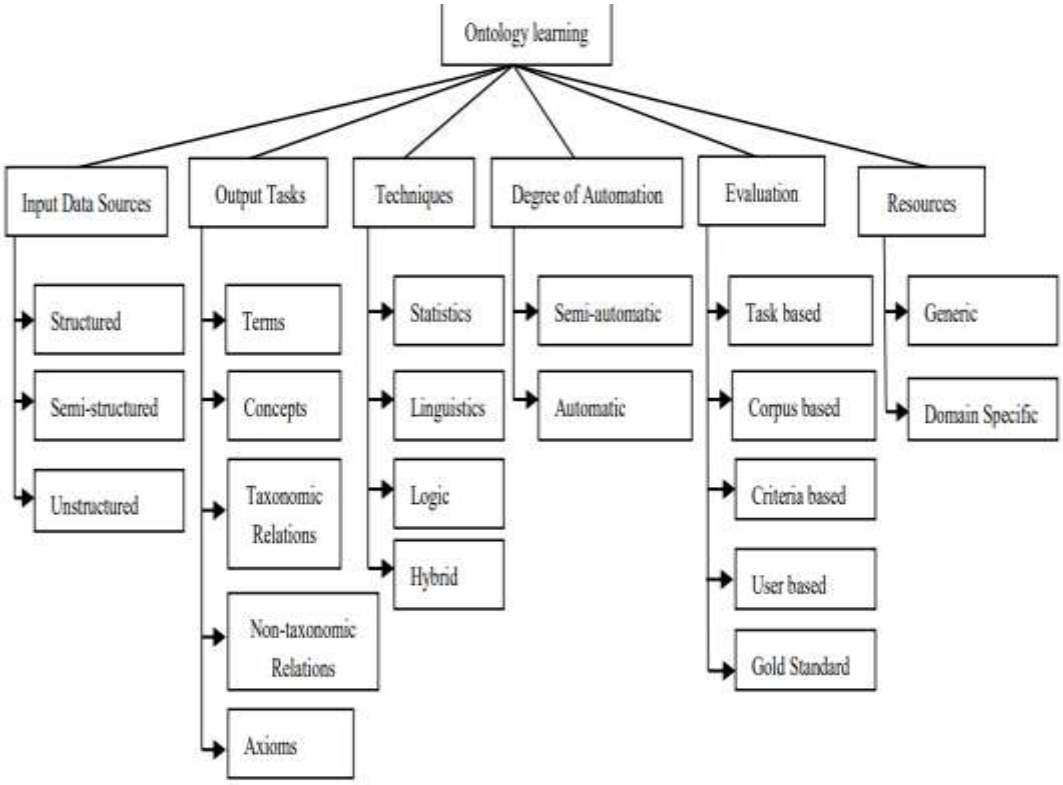

Figure 1 Different Tasks in Ontology Learning Process

1. Input Data Sources:

The ontologies are learned using input data sources. They are categorised as follows:

i. Fully Structured

ii. Semi-Structured

iii. Unstructured Data format

The ontology learning method may employ schema information and current ontologies to extract accessible structured information for ontology creation as candidates for structured data. Other semi-structured data sources utilised in the ontology learning process include WordNet, dictionaries, HTML Tags, XML Tags, and DTD (Document-type definition). Natural language texts are the unstructured data type used in the ontology learning process.

2. Tasks for Output

Figure 1.1 depicts the five kinds of output tasks. Layer cake ontology learning refers to them. When studying an ontology, terms are the foundational building blocks. Any word or combination of words will do. Associated tasks include text preprocessing and text extraction. When providing input text to an ontology learning process, preprocessing is necessary. Relevant keywords are then retrieved from the preprocessed input texts. Concepts may be abstract or tangible, based on reality or purely fictitious. Concepts are created by combining related concepts and giving them meaningful names. To define a notion, you must first identify and integrate all of the many forms that the term might take. With the use of clustering, it is possible to identify multiple variations of a word using various metrics, such as similarity measures and predefined knowledge. Concepts in the ontology are linked together via relations. There are two sorts of relationships: taxonomic and non-taxonomic. The ideas are arranged into a

hierarchical IS-A connection using taxonomic relationships. When it comes to ontology, nontaxonomic relationships reveal nonhierarchical connections between ideas. A self-evident truth is referred to as an axiom. It is via axioms that the correctness of different aspects in a previously specified ontology can be validated and restrictions may be established.

5. CONCLUSION

To address particular problems, the new Apriori algorithm approach makes use of a mix of semantic Web and web mining methods. This research showed how Web Mining methods can be used to construct the Semantic Web and how new semantic structures in the Web can be leveraged to enhance results using Semantic Web Mining. This semantic search utilising an Apriori-based retrieval model has allowed enhanced search capabilities that create a qualitative improvement over keyword-based full-text search by introducing and using finer-grained domain ontologies.

It makes use of DARPA Agent Markup Language or Resource Description Framework, both of which are Semantic Web technologies. Users submit queries, which are processed by a machine learning agent once they've been accepted by the app's server. A comparison of similarity across different ontologies is performed by the agent, and the matched item is subsequently communicated to the user. As a result, it's important to use caution while relying on the internet as a source of information. XML documents created with the apriori method are the only ones of their kind in data mining. A new layer has been added to the mining process domain ontology, giving users easier access to the excavation and the mining results created levels.

6. REFERENCES

1. A., V.; A., Amruta (2016). Semantic Web Mining using RDF Data. International Journal of Computer Applications, 133(10), 14–19. doi:10.5120/ijca2016908022
2. Agha, Salman & Haider, Agha. (2014). An Introduction to Data Mining Technique. IJAETMAS. 3. 5.
3. Ahmad, Hussain & Anwar, Zahid & Shah, Munam. (2017). Data mining techniques and applications — A decade review. 1-7. 10.23919/IConAC.2017.8082090.
4. Al-Hashemi, Idrees & Kalathur, Suresh. (2013). Applying Data Mining Techniques over Big Data.
5. Aloui, Amira & Grissa, Amel. (2015). A New Approach for Flexible Queries Using Fuzzy Ontologies. Studies in Computational Intelligence. 575. 315-342. 10.1007/978-3-319-11017-2_13.
6. Armstrong, Leisa & Diepeveen, Dean & Maddern, Rowan. (2007). The Application of Data Mining Techniques to Characterize Agricultural Soil Profiles. ECU Publications. 70.
7. Atapattu T., Falkner K. and Falkner N. (2017) A comprehensive text analysis of lecture slides to generate concept maps. Comput. Educ., 115, 96–113
8. Bhojani, Shital & Bhatt, Nirav. (2016). Data Mining Techniques and Trends – A Review.