Research Paper Engineering

# Semantic Web Mining: Using Ontology Learning

**Nitesh R Pathak** | Assistant Professor ,Thadomal Shahani Engineering College ,Bandra (West)

**ABSTRACT**

In this paper, a new technique has been described to match the ontology with help of pattern matching. Ontology matching is performed over the research papers. A 3-Satges algorithm is proposed and implemented for the same. In the first stage database creation is carried out which contains IEEE research papers, Conferences papers and Journal papers (any). In the 2nd Stage ontology is developed using title and heading of the paper. In order to extract title and headings a new technique is implemented which uses font-size and font-style of the title and heading. According to the font-size and font-style whatever ontology is developed, that forms our pattern. This pattern is matched with the research papers in order to find out whether it can be categorized into IEEE, Conferences or Journals. The main purpose of this paper is to find out whether the research paper is following a specific IEEE standard writing pattern or not.

**KEYWORDS** Ontology, semantic web.

## INTRODUCTION

What is Ontology? The definitions can be categorized into roughly three groups: 1) Ontology is a term in philosophy and its meaning is ``theory of existence". 2) Ontology is an explicit specification of conceptualization. 3) Ontology is a body of knowledge describing some domain, typically common sense knowledge domain. Current World Wide Web (WWW) is a huge library of interlinked documents that are transferred by computers and presented to people. It has grown from hypertext systems, but the difference is that anyone can contribute to it. This also means that the quality of information or even the persistence of documents cannot be generally guaranteed. Current WWW contains a lot of information and knowledge, but machines usually serve only to deliver and present the content of documents describing the knowledge. People have to connect all the sources of relevant information and interpret them themselves.

Semantic web is an effort to enhance current web so that computers can process the information presented on WWW, interpret and connect it, to help humans to find required knowledge. In the same way as WWW is a huge distributed hypertext system, semantic web is intended to form a huge distributed knowledge based system. The focus of semantic web is to share data instead of documents. In other words, it is a project that should provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by World Wide Web Consortium (W3C). Now this documents can group together we can say one particular domain. This document can be mine and generate the particular pattern we say this pattern as our ontology.

Sample research papers for this project are collected from various sources, mining is performed on this sample set which results in generating an ontology. This ontology is then used to match the pattern of other research papers i.e. a check is performed to find out whether new papers follow the pattern developed as a result of ontology generated in previous step.

3-Stages Ontology Pattern Matching is the algorithm that will perform this whole process and will give desired results.

The paper is organized as follows, Section 2) describes the briefly Ontology and semantic web structure. Section 3) describes the Methodology and Research Model. Section 4) describes the Experimental Result. Sections 5) describes the Conclusion of the project. Sections 6) References.

### Ontology and semantic web structure

In general, ontology is the study or concern about what kinds of things exist - what entities there are in the universe. It derives from the Greek onto (being) and logia (written or spoken discourse). It is a branch of metaphysics, the study of first principles or the essence of things.

In information technology, ontology is the working model of entities and interactions in some particular domain of knowledge or practices, such as electronic commerce or "the activity of planning." In this usage, an ontology is a set of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly.

Ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, the ontology of a program can be defined by a set of representational terms. In such ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, ontology is the statement of a logical theory.

Common ontologies are used to describe ontological commitments for a set of agents so that they can communicate about a domain of discourse without necessarily operating on a globally shared theory.An agent commits to ontology if its observable actions are consistent with the definitions in the ontology.

The idea of ontological commitments is based on the Knowledge-Level perspective. The Knowledge Level is a level of description of the knowledge of an agent that is independent of the symbol-level representation used internally by the agent. Knowledge is attributed to agents by observing their actions; an agent "knows" something if it acts as if it had the information and is acting rationally to achieve its goals. The "actions" of agents---including knowledge base servers and knowledge-based systems--- can be seen through a tall and ask functional interface, where a client interacts with an agent by making logical assertions (tell), and posing queries (ask).

Pragmatically, a common ontology defines the vocabulary with which queries and assertions are exchanged among agents. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. The agents sharing a vocabulary need not share a knowledge base; each knows things the other does not, and an agent that commits to ontology is not required to answer all queries that can be formulated in the shared vocabulary. In short, a commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology.

**SEMANTIC WEB**

Following is the structure of the Semantic Web

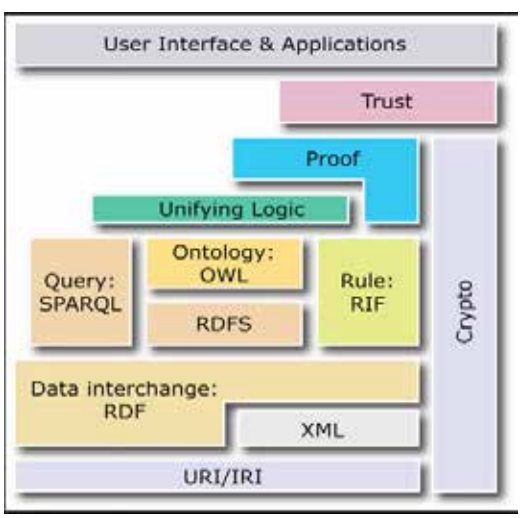


**Figure 1:Semantic Web layered architecture**

**These layers are described as follows:**

**Hyper-Text Web Technologies:-**

These are the bottom layer technologies that are well known in the hypertext web domain. These technologies are used to implement semantic web applications.

**[1] URI[2]**

Universal Resource Identifier (URI) is used to identify the semantic web resources. This unique identification is required so as to provide manipulation with the resources in the top layers.

**[2] UNICODE[2]**

It helps to represent and manipulate text in various languages, thus enabling a bridging of the gap between the human languages and semantic applications.

**[3] XML[2]**

Extended Markup Language (XML) is the markup language that is used to create the semantic web documents in the form of structured data.

Semantic Web Technologies

these are the middle layer technologies, most of which has been standardized by W3C, and can be used to create semantic web applications. All are standardized by the W3C except for RIF/SWRL.

**[1] RDF[2]**

Resource Description Framework (RDF) is the framework that is used to express data in a meaningful way. It expresses data in the form of triples, which is easier to express info in the form of a graph.

**[2] RDFS[2]**

RDF Schema (RDFS) provides the schema, i.e. the vocabulary, for the RDF to maintain a proper structure of the document. It enables to maintain a proper hierarchy of classes and its properties.

**[3] OWL[2]**

Web Ontology Language (OWL) is used to add more meaning, constraints and restrictions to the RDF representation. It expresses the semantics of the RDF statements.

**[4] SPARQL[2]**

SPARQL Protocol and RDF Query Language (SPARQL) is an RDF query language and is used for querying in the database that is represented by the RDF. Querying is done so as to retrieve information by the semantic applications.

**Unrealized Semantic Web Technologies**

These are the top layer technologies that are not yet standardizes or are ideas that needs to be implemented to completely create semantic web applications.

**[1] RIF/SWRL[2]**

Rule Interchange Format/Semantic Web Rule Language (RIF/SWRL) is used to add rules to the RDF data. This enables to represent information that cannot be directly expressed by the OWL.

**[2] Cryptography[2]**

this is to ensure that the statements coming from semantic web applications are from proper sources and this can be implemented using digital signatures of RDF documents.

**[3] Trust[2]**

Trust for statements support: the premises come from trusted sources and relying on formal language to retrieve new information.

**[4] User Interface[2]**

this is the top most layer that will enable the humans to use the semantic web applications.

**The Relationship between Semantic Web and Ontology**

Ontology and the Semantic Web strive to express and enable semantic relations among represented entities. Semantic relations are meaningful associations between two or more concepts, entities, or sets of entities (Khoo and Na, 2006). As a new information representation system, ontology aims to substantiate the rich variety of semantic relations among the concepts it represents – a characteristic that distinguishes it from other representation and organization systems.

Hodge[3] grouped typical information representation systems into three general categories: term lists, classifications and categories, and relationship lists. Term lists emphasize lists of terms usually presented with definitions. Classifications and categories emphasize the creation of subject sets. Relationship lists emphasize the connection between terms and concepts.

**Methodology and Research Model**

**Methodology**

As already been discussed ontology is to be developed which will be used to match the desired pattern over the research documents. 3-stage ontology pattern matching algorithm is explained as under:

**3-Stages Ontology Pattern Matching algorithm.**

This section briefly describe the Algorithm and Research Model

**Algorithm: - 1-stage**

1. Collect the document of the particular domain.
2. Generally document would be research papers. This research paper can be IEEE research paper, Conferences research paper, Journal research paper and any other research paper that is found in internet. But only condition is research paper would be of same domain.
3. Store this papers in data-store (collection of research papers)

Algorithm: - 2-Stage

1. This stage is important phase of algorithm.
2. Whatever document is collected, it is passed to the ontology development system.
3. Ontology Development System- this system extract the document on the basis of specific font. All research papers contain BOLD font-style and High font-size FOR their heading. This style acts as the basis for font selection. Hence documents having desired font style in their heading are extracted.
4. From all research papers same properties are extracted and finally matched with each other to generate the ontology.

Algorithm: - 3-Satge

1. This is final stage of our algorithm; the ontology pattern is matched with other research paper.
2. To do this use again use the Algorithm: - 2-Satge,
3. After extracting the words match this word with ontology.

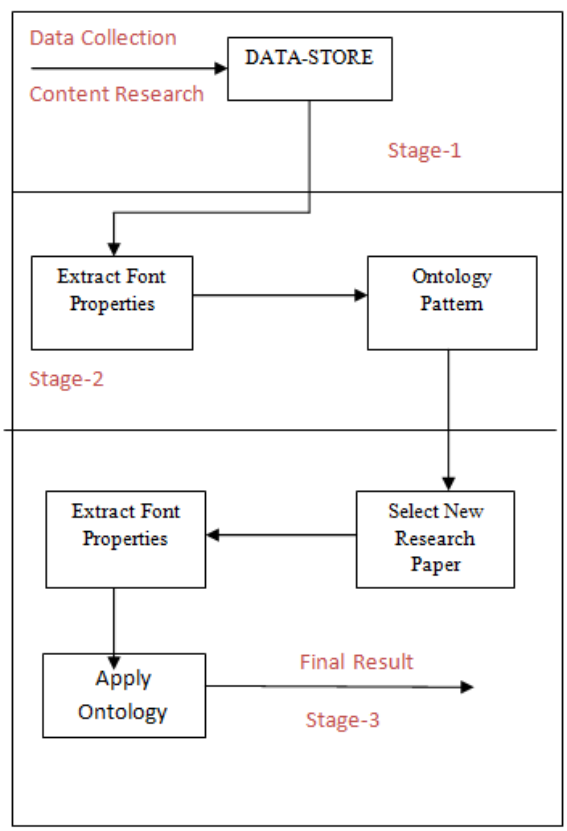


**Figure 2: Research Model**

**Experimental Result**

In this experiment 50 research papers collected from various sources like IEEE research papers, Conferences papers and other Journals papers are considered. Following is the generated ontology pattern.

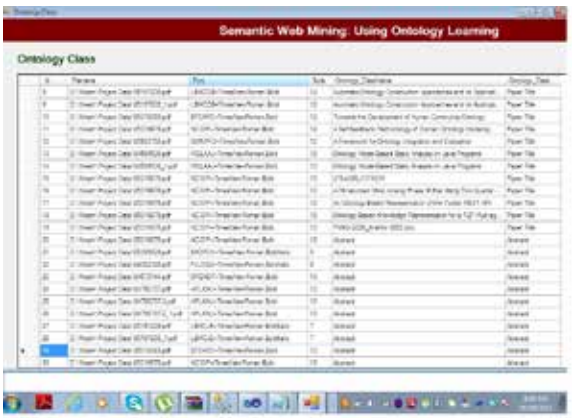


**FIGURE 3:Ontology class**

From the Ontology class the following classes has been discovered.

1. Paper Title
2. Abstract
3. Introduction
4. Conclusion
5. References

Now the new conference paper will be taken as input and the ontology will be applied on his paper to find out whether the given paper is a conference paper or not.

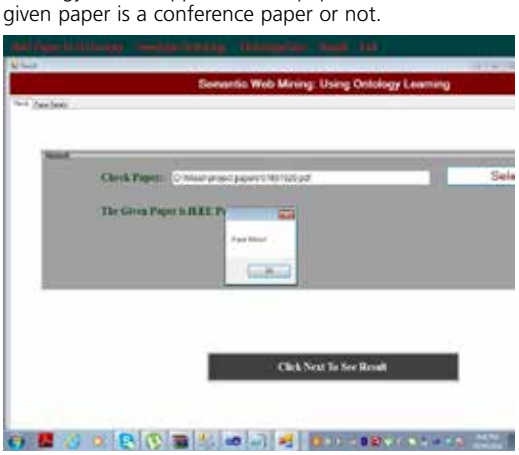


**Figure 4**:**Testing of a new paper**

Now once the new paper has been matched with the ontology generated.the text information of perticular class from that paper can be extracted.

**For eg:**




**Figure 5:Abstract**

The above result shows that the Abstract of a particular paper which was matched with the ontology is extracted. this will help user to perform wide variety of queries.

**CONCLUSIONS**

Different ways to generate the ontology and match it with the new document are discussed. 3-Stages Ontology Pattern Matching is efficient algorithm which stores document, generates ontology pattern and matches the same with the new documents (Research papers). A new system i.e. Ontology Development System is created which works on the different types of research papers like IEEE, Conferences, Journal and other mode of papers. With help of this system we can find out how a particular paper should be treated: IEEE research paper, Conference Paper or Journal paper.

The main aim of the project is to generate the ontology pattern and apply it to the research papers.

**REFERENCES**

1. http://www.ieee.org/index.html | 2. C.S.Bhatia, Dr. Suresh Jain "Semantic Web Mining: Using Ontology Learning and Grammatical Rule Inference Technique" IEEE 2011 IEEE | 3. Anil Kumar Dhiman "Ontology and Ontological Systems for Semantic Webs in Digital Libraries" 7th International CALIBER 2009 | 4. V. Crescenzi, G.Mecca and P. Merialdo, "RoadRunner: Towards automatic data extraction from large web sites" In VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, Roma, Italy, pages 109–118. Morgan Kaufmann, Sept. 1. | 5. A.J Gerber, A.Barnard, A.J Van der Merwe"Towards a Semantic Web Layered Architecture" | 6. A. Maedche, S. Stabb (2001) "Ontology Learning for the Semantic Web" IEEE intelligent Systems, Special Issue on the semantic Web, 16(2). | 7. W3C. The Document Object Model. | http://www.w3.org/DOM |