Engineering Applications of Artificial Intelligence 63 (2017) 108–125

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

# Semi-automatic terminology ontology learning based on topic modeling

Monika Rani *, Amit Kumar Dhar, O.P. Vyas

*Department of Information Technology, Indian Institute of Information Technology, Allahabad, India*

ARTICLE INFO

*Article history:*
Received 18 August 2016
Received in revised form
17 February 2017
Accepted 9 May 2017

*Keywords:*
Ontology Learning (OL)
Latent Semantic Indexing (LSI)
Singular Value Decomposition (SVD)
Probabilistic Latent Semantic Indexing (pLSI)
MapReduce Latent Dirichlet Allocation (Mr.LDA)
Correlation Topic Modeling (CTM)

ABSTRACT

Ontologies provide features like a common vocabulary, reusability, machine-readable content, and also allows for semantic search, facilitate agent interaction and ordering & structuring of knowledge for the Semantic Web (Web 3.0) application. However, the challenge in ontology engineering is automatic learning, i.e., the there is still a lack of fully automatic approach from a text corpus or dataset of various topics to form ontology using machine learning techniques. In this paper, two topic modeling algorithms are explored, namely LSI & SVD and Mr.LDA for learning topic ontology. The objective is to determine the statistical relationship between document and terms to build a topic ontology and ontology graph with minimum human intervention. Experimental analysis on building a topic ontology and semantic retrieving corresponding topic ontology for the user's query demonstrating the effectiveness of the proposed approach.



## 1. Introduction

Web 2.0 allow a user to participate in information sharing by writing reviews, comments, and feedback on sites not like the Traditional Web (Web 1.0: 1990–2000) where user are passive information receivers. But there are certainly significant problems in the Web 2.0 (2000–2010) which lead to the birth of Web 3.0 (2010–2020). For example, in web 2.0 (Web of Document) content is not machine-readable format. Information on the web is in a heterogeneous source format (HTML, XML, etc.) and thus can't be used for purpose of integration, analysis, and intelligent data analysis. Berners-Lee et al. (2001) gave a vision of Semantic Web (Web 3.0) which can connect data semantically in machine-readable format, provide common vocabulary using ontologies. Basically, Web 3.0 is an extension of Web 2.0 which aim to provide a common framework for sharing data across application boundaries. Web 3.0 also called the Web of Data (WoD) whereas Web 2.0 was known as the web of documents. We are moving towards the era of Web of Documents to Web of Data. Web 2.0 perform syntactic search whereas Web 3.0 capable of doing the semantic search using semantic web and ontologies technology. Web of Data (Web 3.0) provides one more level deeper information thus it can answer more complex queries. Also, Web 3.0 allow ordering and structuring of content, provide machine-readability, content reusability and facilitate agent interaction on the web using ontologies (Gruber, 1993). Ontologies are the pillar of semantic web used for knowledge representation. Ontology Learning (OL) greatly helps ontology engineers for building their own ontologies for a particular domain. The steps for OL are Import & reuse, Extract, Prune, Refine and Validation (Maedche, 2012). Different approaches for OL can be classified based on the type of knowledge resources for which to learn ontology (structured, semi-structured, unstructured format of data), Level of automation (semi or full automated), Learning target (concepts, Taxonomy, Conceptual relation, Attributes, Instances, Axioms), purpose (creating/updating), Learning techniques (Linguistics, Statistical and Hybrid approaches) (Al-Arfaj and Al-Salman, 2015).

The present research focuses on Ontology Learning (OL) disabilities such as: automatic conversion of Text to Ontology (Text to Onto) (Wong et al., 2012), Never Ending Language Learning (NELL) (Carlson et al., 2010) as data is continuously increasing on the web and database, Open Information Extraction (OIE) etc. So, in continuation of that our main focus is to learn terminology ontology based on topic modeling and then provide primarily step for semantic-based query retrieval (Topic and Words Detection). In this paper, two topic modeling algorithm are explored namely: Latent Semantic Indexing (LSI) & Singular Value Decomposition (SVD) and MapReduce Latent Dirichlet Allocation (Mr.LDA) for learning terminology ontology. Our target is to achieve terminology ontology from textual data using topic modeling and primarily provide semantic-based

* Corresponding author.
*E-mail address:* monikarani1988@gmail.com (M. Rani).

http://dx.doi.org/10.1016/j.engappai.2017.05.006
0952-1976/

query retrieval for Topic and Words Detection for a domain. Though there are powerful search engines like Google, Bing, Yahoo!, etc. but they are limited in their capability of retrieving a relevant list of the Web of Document (WoD) rather than required data. Also, unfortunately, most of the textual data available on the internet are not machine-readable format. It is also quite cumbersome to organize and access as textual data which is continuously growing per second. Existing search engine as retrieval result brings out as the Web of Document (WoD) whereas if we classify the large corpus into ontology then search required content (topic and words) which is the Web of Data (Web 3.0) is time-saving and cost of searching. Therefore, our approach is to build semi- automatic terminology ontology using topic modeling algorithm (LSI & SVD and Mr.LDA) to classify topics and associated words for knowledge management and semantic retrieval.

Ontology learned can be used in various fields as they have the ability to reduce communication gap by providing the common vocabulary, in association with agents provide a personal assistant for real-time applications also classify and recommend content for the user. The proposed terminology ontology is based on topic modeling (Mr.LDA model) can be used in other fields of research examples: Semantic Retrieval (Tran et al., 2007), Semantic Web in E-learning (Rani et al., 2015), Data Enrichment & Mining (Dou et al., 2015; Abedjan, 2014), Topic Detection and Tracking, Natural Language Processing (NLP), Semantic search over heterogeneous networks (Tang et al., 2011; Hogan, 2011), Knowledge Engineering and Management, Electronic Commerce, Sentimental Analysis, Scientific Exchange, Bio-informatics, Biomedical, Human Computer Interaction. Also, build ontology is beneficial for Ontology-based Association Rule Mining (ARM), Classification, Clustering, Link Prediction, Information Retrieval, Recommendation System, etc.

The rest of this paper is organized as follows. Section 2 describes Ontology Learning (OL), Latent Semantic Indexing (LSI), probabilistic Latent Semantic Indexing (pLSI), MapReduce Latent Dirichlet Allocation (Mr.LDA). Section 3 introduces a general description of the explored topic modeling (LSI & SVD and Mr.LDA). Experimental results and conclusion is reported in Sections 4 and 5 respectively.

## 2. Review of related works

Topic Modeling is a form of text mining, where we can retrieve required text from the large corpus. Topic Modeling uses various algorithms or a modeling approach to organizing, summarize large corpus and retrieve require text. In this section, some related works, namely those regarding Ontology, Topic Models (Latent Semantic Indexing (LSI) & Singular Value Decomposition (SVD) and Latent Dirichlet Allocation (LDA)), are briefly reviewed.

### 2.1. Ontology

Ontology is the study of semantics, existing in the world, which can be formally defined as, a formal and explicit specification of a shared conceptualization (Gruber, 1995). It means ontology explicitly define the rich relation between concepts which are machine readable and sharable among a group of people. Since the introduction of Web 3.0, there has been an exponential rise in the amount of data that is accumulated each day. This data has to be well defined and explicitly represented so that it can be shared and used by humans and machines, enabling them to work together in a better way.

#### 2.1.1. An Ontology generally consists of

- Individuals (aka instances): consider every existing object, example: you and me.
- Concept (aka classes): a group of existing objects example: person, organization, etc.
- Attributes: describe the property of concept example: height, weight, etc.
- Relationships: represent the relation by which two concepts are associated, example: students are affiliated with IIIT-Allahabad College.

#### 2.1.2. Following are the advantages of making an ontology

- An ontology provides us with a common vocabulary for a domain (Cakula and Salem, 2013).
- Merge and expansion of ontologies based on their metadata are easy.
- An ontology defines content unambiguously.
- To separate operational knowledge from domain knowledge.
- The ontology allows re-use of content represented in it (Jovanović et al., 2007).
- Ontology provides ordering and structuring of the content store in it (Dzemydiene and Tankeleviciene, 2008).
- A rule can be added to ontology to infer new knowledge.
- Integrate content from a heterogeneous source.
- Effective information sharing, storage, and retrieval of content (text corpus).
- Agent interaction to share content store in an ontology (Jekjantuk and Hasan, 2007).

#### 2.1.3. Ontology classification: by level of generality

Guarino categorized ontology according to the level of generality (Guarino, 1998):

- Top-level ontology: Its main concerns on general concepts like time, event, action, matter, etc.
- Domain ontology: It provides a common vocabulary to a domain, so various domain knowledge can be interpreted and exchange. An example of domain ontology: Music ontology, Food ontology, Geo ontology, Gene ontology, etc.
- Task ontology: It is based on activity or task example selling or diagnosing.
- Application ontology: Ontologies are built for a specific purpose to share knowledge modeling among various domains.

#### 2.1.4. Ontology classification: by level of formality

- Informal ontology– It is a taxonomy, example: web directories (Yahoo! Directory), glossary directory, etc. Ontology is rich in term of expressing the relationship, than taxonomy. Taxonomy represents only a hierarchical arrangement of the group.
- Formal ontology– To build a formal ontology, OWL formal language is considered like OWL 1 and OWL 2. Example Cyc and DOLCE.
- Semi-formal ontology– schema structure is considered semi-formal ontology as RDFS language is used rather than a formal language like OWL.

#### 2.1.5. Ontology can be categories on the basis of purpose into classification ontologies and descriptive ontologies

- Classification Ontologies: The document stores are huge and increase with the time on the internet. To classify this document on the basis of relations between terms appropriate hierarchy is used. Searching of a document using the title, subject, and an author can be easily done using classification ontologies.