International Journal of Sensors and Sensor Networks
2017; 5(5-1): 13-17
http://www.sciencepublishinggroup.com/j/ijssn
doi: 10.11648/j.ijssn.s.2017050501.13

Mining the Web for Learning Ontologies: State of Art and Critical Review

Mohamed El Asikri, Salahddine Krit, Hassan Chaib, Mustapha Kabrane, Hassan Ouadani, Khaoula Karimi, Kaouthar Bendaouad, Hicham Elbousty

Department Mathematics and Informatics and Management, Laboratory of Engineering Sciences and Energy, Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Agadir, Morocco

Email address:

meda.asikri@gmail.com (M. El Asikri), salahddine.krit@gmail.com (S. Krit), hchaib@gmail.com (H. Chaib), mustaphakabrane@gmail.com (M. Kabrane), hassan.oudani@gmail.com (H. Ouadani), karimi.khaoula92@gmail.com (K. Karimi), kaouthar.bendaoud@gmail.com (K. Bendaouad), elboustyhicham@gmail.com (H. Elbousty)

To cite this article:

Mohamed El Asikri, Salahddine Krit, Hassan Chaib, Mustapha Kabrane, Hassan Ouadani, Khaoula Karimi, Kaouthar Bendaouad, Hicham Elbousty. Mining the Web for Learning Ontologies: State of Art and Critical Review. *International Journal of Sensors and Sensor Networks.* Special Issue: Smart Cities Using a Wireless Sensor Networks. Vol. 5, No. 5-1, 2017, pp. 13-17. doi: 10.11648/j.ijssn.s.2017050501.13

Received: March 30, 2017; **Accepted**: April 7, 2017; **Published**: May 13, 2017

Abstract: The aim of the paper is to investigate and present the subject of building ontologies using the Semantic Web Mining that is defined as the combination of the two fast-developing research areas Semantic Web and Web Mining.Web mining is the application of data mining techniques to the content, structure, and usage of Web resources and The Semantic Web is the second-generation WWW, enriched by machine-processable information which supports the user in his tasks.. This can help to discover global as well as local structure “models” or “patterns”within and between Web pages and ontology extraction witch is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between those concepts, and encoding them with an ontology language for easy retrieval. As building ontologies manually is extremely labor-intensive and time-consuming, there is great motivation to automate the process. This paper gives an overview of where the two areas meet today, and discuss ways of how a closer integration could be profitable.

Keywords: Semantic Web, Web Mining, Ontology, Konwledge Discovery, Ontology Learning

1. Introduction

Internet is becoming more and more important for nearly everybody as it is one of the newest and most forward-looking media and surely the medium of the future.The problem of the huge volume of information is growing and the extraction of useful knowledge from the Web is becoming very difficult because of unstructured web contents and absence of standardization, this complexity in the treatment and extraction Knowledge of large volumes Web pages make hard to use information in the last decade, a vast amount of approaches have been proposed which combine methods from data mining and knowledge discovery with Semantic Web data. The goal of those approaches is to support different data mining tasks, or to improve the Semantic Web itself. All those approaches can be divided into three broader categories:

a) Using Semantic Web based approaches, Semantic Web Technologies, and Linked Open Data to support the process of knowledge discovery.
b) Using data mining techniques to mine the Semantic Web, also called Semantic Web Mining.
c) Using machine learning techniques to create and improve Semantic Web data

Semantic Web is a recent initiative inspired by Tim Berners-Lee [1] and it is an extension of the current Web that provides an easier way to find, share, reuse and combine information. It is based on machine-readable information and builds on XML technology's capability to define customized tagging schemes and RDF's [2] (Resource Description Framework) flexible approach to representing data. The Semantic Web provides common formats for the interchange of data. It also provides a common language for recording how data relates to real world objects, allowing a person or a machine to start off in one database, and then move through

an unending set of databases of knowledge in order to find various concepts and words to build ontologies and representation in a way that can be understood and consumed by every users [3].

This paper will describe the commonalities of the two areas in order to extract useful and shared Knowledge.

2. Basic Definitions

2.1. Ontology

An ontology defines the terms used to describe and represent an area of knowledge, explicitly Ontologies are used by people, databases, and applications that need to share domain information. Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them. They encode knowledge in a domain and also knowledge that spans domains [5].

More formally, an ontology consists of classes, relationships and attributes. The classes in an ontology are general things (in the many domains of interest). Usually, the names of classes are nouns. The relationships exist among the things, we use two relationships: 'part-of' and 'is-a' in this research. The properties (or attributes) are those the things may have.

We start by introducing the following example fig 1 for an ontology and its representation.

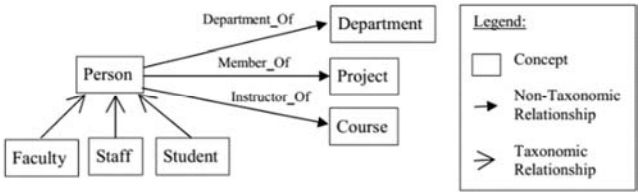

Fig. 1. *Example of an ontology: The University Structure.*

2.2. Semantic Web Mining

The human ability for information processing is limited on the one hand, whilst otherwise the amount of available information of the Web increases exponentially, which leads to increasing information saturation[3]. In this context, it becomes more and more important to detect useful patterns in the Web, thus use it as a rich source for data mining [4].

The research area of Semantic Web Mining is aimed at combining two fast developing fields of research: the Semantic Web and Web Mining. The idea is to improve, on the one hand, the results of Web Mining by exploiting the new semantic structures in the Web; and to make use of Web Mining, on the other hand, for building up the Semantic Web.These two fields address the current challenges of the World Wide Web (WWW): turning unstructured data into machine-understandable data using Semantic Web tools.

As the Semantic Web enhances the first generation of the WWW with formal semantics, it offers a good basis to enrich Web Mining: The types of (hyper) links are now described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, to apply mining techniques which require more structured input.

2.2.1. Semantic Web Content and Structure Mining

In the Semantic Web, content and structure are strongly intertwined. Therefore, the distinction between content and structure mining vanishes. However, the distribution of the semantic annotations may provide additional implicit knowledge. An important group of techniques which can easily be adapted to semantic Web content / structure mining are the approaches discussed as Relational Data Mining (formerly called Inductive Logic Programming (ILP)). Relational Data Mining looks for patterns that involve multiple relations in a relational database. It comprises techniques for Semantic Web Mining like classification, regression, clustering, and association analysis. It is quite straightforward to transform the algorithms so that they are able to deal with data described in RDF or by ontologies.

There are two big scientific challenges in this attempt. The first is the size of the data to be processed (i.e. the scalability of the algorithms), and the second is the fact that the data are distributed over the Semantic Web, as there is no central database server. Scalability has always been a major concern for ILP algorithms. With the expected growth of the Semantic Web, this problem increases as well. Therefore, the performance of the mining algorithms has to be improved, e.g. by sampling.

2.2.2. Semantic Web Usage Mining

Usage mining can also be enhanced further if the semantics are contained explicitly in the pages by referring to concepts of an ontology. Semantic Web usage mining can for instance be performed on log files which register the user behavior in terms of an ontology. A system for creating such semantic log files from a knowledge portal has been developed at the AIFB. These log files can then be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology.

3. The Knowledge Discovery Process

The process model for knowledge discovery processes model comprises five steps, which lead from raw data to actionable knowledge and insights which are of immediate value to the user. The whole process is shown in Fig. 2. It comprises five steps:

1. *Selection* The first step is developing an understanding of the application domain, capturing relevant prior knowledge, and identifying the data mining goal from the end user's perspective. Based on that understanding, the target data used in the knowledge discovery process can be chosen, i.e., selecting proper data samples and a relevant subset of variables.

2. *Preprocessing* In this step, the selected data is processed in a way that allows for a subsequent analysis. Typical actions taken in this step include the handling of missing values, the identification (and potentially correction) of noise and errors in the data, the elimination of duplicates, as well

as the matching, fusion, and conflict resolution for data taken from different sources.

3. *Transformation* The third step produces a projection of the data to a form that data mining algorithms can work on—in most cases, this means turning the data into a propositional form, where each instance is represented by a feature vector. To improve the performance of subsequent data mining algorithms, dimensionality reduction methods can also be applied in this step to reduce the effective number of variables under consideration.

4. *Data mining* Once the data is present in a useful format, the initial goal of the process is matched to a particular method, such as classification, regression, or clustering. This step includes deciding which models and parameters might be appropriate (for example, models for categorical data are different than models for numerical data), and matching a particular data mining method with the overall criteria of the KDD process (for example, the end user might be more interested in an interpretable, but less accurate model than a very accurate, but hard to interpret model). Once the data mining method and algorithm are selected, the data mining takes place: searching for patterns of interest in a particular representational form or a set of such representations, such as rule sets or trees.

5. *Evaluation and interpretation* In the last step, the patterns and models derived by the data mining algorithm(s) are examined with respect to their validity. Furthermore, the user assesses the usefulness of the found knowledge for the given application. This step can also involve visualization of the extracted patterns and models, or visualization of the data using the extracted models.

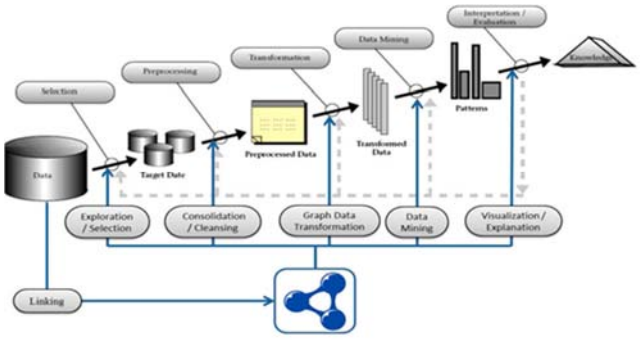

Fig. 2. *Process of knowledge discovery.*

As a means to express knowledge about a domain in the Semantic Web, *ontologies* have been introduced in the early 1990s as "explicit formal specifications of the concepts and relations among them that can exist in a given domain" [6]. For the area of knowledge discovery and data mining, Nigro et al. [7] divide ontologies used in this area into three categories:

a) *Domain ontologies:* Express background knowledge about the application domain, i.e., the domain of the data at hand on which KDD and data mining are performed.
b) *Ontologies for data mining process:* Define knowledge about the data mining process, its steps and algorithms and their possible parameters.
c) *Metadata ontologies:* Describe meta knowledge about the data at hand, such as provenance information, e.g., the processes used to construct certain datasets.
d) It has been already shown that ontologies for the data mining process and metadata ontologies can be used in each step of the KDD process.

4. Example of Use of Ontology-Based Web Mining

Ontology-based Web mining, like traditional Web mining, is useful to many different digital library applications. We can group these applications under the following classes:

a) Improved search to Web data: With additional ontological semantics, Web data can be indexed by their concepts and relationships to support expressive search queries. For example, using the University ontology in Figure 1, we can query faculty member information working on digital library projects by assigning query term "digital library" to the project concept and specifying that faculty related to the
b) qualified projects to be returned in the query results. Such queries may resemble structured database queries except that the data to be dealt with are Web pages. A more expressive query model can support very precise information search and reduce the amount of irrelevant Web information in the results [13].
c) Better browsing capabilities: Similar to searching, Web pages can be browsed based on their ontology concepts and relationships instead of following Weblinks only. If Web pages are the concept instances, relationship instances can be created as some virtual links between Web pages. Other than selecting Web pages belonging to concepts of interest, one can thus navigate the virtual links between Web pages enriching the browsing experience in digital library applications [13]. On the other hand, if some text elements are identified as concept instances and their relationships are extracted, they can also be marked up in Web pages to direct user attentions to the more important text passages.
d) Personalization of Web data access: Personalization aims to find a subset of Web data that matches the interest profile of a user or a group of users. This can be achieved by recommending Web pages or Websites to the user(s), or by filtering Web pages that are of interest to the user(s). For example, this can done by analysing the historical data recording user accesses to Web data, and mining the topics relevant to a user by clustering previously accessed Web pages based on content similarities. When a new Web page is found to be similar to one of the clusters,it can be routed to the user. As Web pages are annotated with ontology entity labels, the grouping of Web pages accessed by a usercan be more effectively done leading to more effective content recommendation.

5. Ontology Learning Process

Ontology learning is used to (semi-)automatically extract whole ontologies from natural language text or Web Pages. The process is usually split into the following eight tasks, which are not all necessarily applied in every ontology learning system [8] [9].

a) Domain terminology extraction
b) Concept discovery
c) Concept hierarchy derivation
d) Learning of non-taxonomic relations
e) Rule discovery
f) Ontology population
g) Concept hierarchy extension

5.1. Domain Terminology Extraction

During the domain terminology extraction step, domain-specific terms are extracted, which are used in the following step (concept discovery) to derive concepts. Relevant terms can be determined e. g. by calculation of the TF/IDF values or by application of the C-value / NC-value method. The resulting list of terms has to be filtered by a domain expert. In the subsequent step, similarly to coreference resolution in IE, the OL system determines synonyms, because they share the same meaning and therefore correspond to the same concept. The most common methods therefore are clustering and the application of statistical similarity measures.

5.2. Concept Discovery

In the concept discovery step, terms are grouped to meaning bearing units, which correspond to an abstraction of the world and therefore to concepts. The grouped terms are these domain-specific terms and their synonyms, which were identified in the domain terminology extraction step.

5.3. Concept Hierarchy Derivation

In the concept hierarchy derivation step, the OL system tries to arrange the extracted concepts in a taxonomic structure. This is mostly achieved by unsupervised hierarchical clustering methods. Because the result of such methods is often noisy, a supervision, e. g. by evaluation by the user, is integrated. A further method for the derivation of a concept hierarchy exists in the usage of several patterns, which should indicate a sub- or supersumption relationship. Patterns like “X, that is a Y” or “X is a Y” indicate, that X is a subclass of Y. Such pattern can be analyzed efficiently, but they occur too infrequent, to extract enough sub- or supersumption relationships. Instead bootstrapping methods are developed, which learn these patterns automatically and therefore ensure a higher coverage.

5.4. Learning of Non-taxonomic Relations

At the learning of non-taxonomic relations step, relationships are extracted, which do not express any sub- or supersumption. Such relationships are e.g. works-for or located-in. There are two common approaches to solve this subtask. The first one is based upon the extraction of anonymous associations, which are named appropriately in a second step. The second approach extracts verbs, which indicate a relationship between the entities, represented by the surrounding words. But the result of both approaches has to be evaluated by an ontologist.

5.5. Rule Discovery

During rule discovery axioms [10] (formal description of concepts) are generated for the extracted concepts. This can be achieved for example by analyzing the syntactic structure of a natural language definition and the application of transformation rules on the resulting dependency tree. The result of this process is a list of axioms, which is afterwards comprehended to a concept description. This one has to be evaluated by an ontologist.

5.6. Ontology Population

At the ontology population step, the ontology is augmented with instances of concepts and properties. For the augmentation with instances of concepts methods, which are based on the matching of lexico-syntactic patterns, are used. Instances of properties are added by application of bootstrapping methods, which collect relation tuples.

5.7. Concept Hierarchy Extension

In the concept hierarchy extension step, the OL system tries to extend the taxonomic structure of an existing ontology with further concepts. This can be realized supervised by a trained classifier or unsupervised by the application of similarity measures.

5.8. Tools

Dog4Dag - an ontology generation plugin for Protégé 4.1 and OBOEdit, ontoStudio fig 3. DOG4DAG is an ontology generation plugin for both Protégé 4.1 and OBO-Edit 2.1. It allows for term generation, sibling generation, definition generation, and relationship induction. Integrated into Protégé 4.1 and OBO-Edit 2.1, DOG4DAG allows ontology extension for all common ontology formats (e.g., OWL and OBO). Limited largely to EBI and Bio Portal lookup service extensions. [11]

OWL ontology languages [12] allow users to write explicit, formal conceptualization of domain models. They full fill the main requirements of ontology languages such as

a) A well-defined syntax
b) A formal semantics
c) Convenience of expression
d) Efficient reasoning support

The subclass relationship between OWL and RDF property OWL builds on RDF and RDF schema and uses RDF„s XML based syntax. OWL can be defined with XML based syntax, Abstract syntax which makes use of RDF and

Graphical Syntax. Proposed architecture applied to generate ontology based on web log which described in web ontology language (OWL).Web authors will be helped by analyzing user "browser" history of web sites which are frequently visited by people and the items which are popular in market and the same is easily identified by this application. Web structure mining has used in developing ontologies and help to retrieving process.

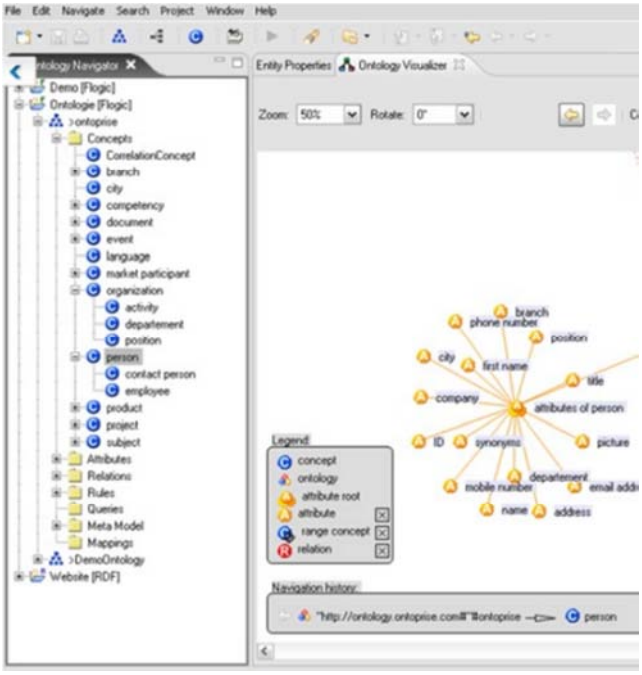

Fig. 3. *Building ontolgy example with ontoStudio.*

6. Conclusion and Outlook

This paper summarizes the use of ontology in Web mining. In particular, we focus on how ontology has been incorporated in Web mining. We discussed how Semantic Web Mining can improve the results of Web Mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques.

We expect that, in the future, Web Mining methods will increasingly treat content, structure, and usage in an integrated fashion in iterated cycles of *extracting* and *utilizing* semantics, to be able to understand and (re)shape the Web. Among those iterated cycles, we expect to see a productive complementarily between those relying on semantics in the sense of the Semantic Web, and those that rely on a looser notion of semantics.

References

[1] T. Berners-Lee, N. Shadbolt, and W. Hall, "The Semantic Web Revisited " IEEE intelligent Systemes, pp. 96-101, 2006.

[2] https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[3] M.-S. Chen, J. Han, and P. S. Yu, Data mining: an overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6):866-883.

[4] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthrusamy, eds., Advances in knowledge discovery and data mining, Menlo Park, California: AAAI Press/ The MIT Press, 1996.

[5] Web-Ontology (WebOnt) Working Group, 2001, http://www.w3.org/2001/sw/WebOnt/.

[6] T. R. Gruber Toward principles for the design of ontologies used for knowledge sharing Int. J. Hum.-Comput. Stud., 43 (5) (1995), pp. 907–928.

[7] H. O. Nigro, S. G. Cisaro, D. H. Xodo Data Mining With Ontologies: Implementations, Findings and Frameworks, Information Science Reference, Imprint of: IGI Publishing, Hershey, PA (2007).

[8] Cimiano, Philipp; Völker, Johanna; Studer, Rudi (2006). "Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text", Information, Wissenschaft und Praxis, 57, p. 315 – 320.

[9] Wong, W., Liu, W. & Bennamoun, M. (2012), "Ontology Learning from Text: A. Look back and into the Future". ACM Computing Surveys, Volume 44, Issue 4, Pages 20:1-20:36.

[10] Völker, Johanna; Hitzler, Pascal; Cimiano, Philipp (2007). "Acquisition of OWL DL Axioms from Lexical Resources", Proceedings of the 4th European conference on The Semantic Web, p. 670 – 685.

[11] Thomas Wächter, Götz Fabian, Michael Schroeder: DOG4DAG: semi-automated ontology generation in OBO-Edit and Protégé. SWAT4LS London, 2011. doi:10.1145/2166896.2166926.

[12] https://www.w3.org/OWL/.

[13] Naing, M.-M, Lim, E.-P., and Chiang, R. H.-L.,"Core: A Search and Browsing Tool for Semantic Instances of Web Sites," Asia Pacific Web Conference (APWeb'05), 2005.