OntoWM: an Ontology for Unification and Description of Web Mining

Khaled Benali
Lab systems, Networks, Databases, SNDB.
University of Science and Technology of Oran
USTO
2 C 18 Debdaba Bechar Algeria 08000
Benalikhaled2013@yahoo.fr

Sidi Ahmed Rahal
Lab systems, Networks, Databases, SNDB.
University of Science and Technology of Oran
USTO
ORAN, BP 1505 El Mnaouer
Rahalsa2001@yahoo.fr

Abstract – This article is concerned with the merging of two active research domains: Knowledge Discovery in Databases (KDD) and Knowledge Engineering (KE) with a main interest in Ontology. In KDD, we need to unify the domain of web mining. To overcome this drawback, several methods have been proposed in the literature. So, we propose an ontology, named OntoWM which includes definitions of basic Web Mining entities, such as tasks, algorithms... to describe the spots and the basic entities of the web mining in order to share common understanding of this method and explain what is considered as implicit.

Keywords – Web Mining, Ontology, KDD, KE

1. INTRODUCTION

In recent years use of term ontology has become prominent in the area of computer science research and the application of computer science methods in management of scientific and other kinds of information. In this sense the term ontology has the meaning of a standardized terminological framework in terms of which the information is organized [1]. "Data Mining (DM) is an emerging field that covers a wide range of application domains, such as marketing, finance, e-commerce, biology and privacy among the others. [2]. Among the knowledge models used in DM, Web mining (WM) consists of a set operations defined on data residing on WWW data servers. The reference [3] defines web mining as "…the discovery and analysis of useful information from the World Wide Web". Such data can be the content presented to users of the web sites such as hypertext markup language (HTML) files, images, text, audio or video. Also the psychical structure of the web sites or the server logs that keep track of user accesses to the resources mentioned above can be targets of web mining techniques. Web mining is mainly categorized into two subsets namely web content mining and web usage mining [3]. While the content mining approaches focus on the content of single web pages, web usage mining uses server logs that detail the past accesses to the web site data made available to public.

While KDD and data mining have enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework that would cover and unify the data mining domain. The present lack of such a framework is perceived as an obstacle to the further development of the field. In [4], Yang and Wu collected the opinions of a number of outstanding data mining researchers about the most challenging problems in data mining research. Among the ten topics considered most important and worthy of further research, the development of a unifying framework for data mining is listed first. One step towards developing a general framework for data mining is constructing ontology of data mining.

In this article we will create our new ontology model (OntoWM) to unify Web mining and to represent web mining tasks, and define the semantics of the relationships between entities of web mining.

2. OntoWM

For the development of our ontology, we tried to follow the steps proposed in [5].

Step 1. Determination of the domain and scope of the ontology

Our application requires an Ontology of Web Mining (OntoWM), which should allow us to describe the spots and the basic entities of the web mining in order to share common understanding of this method and explain what is considered as implicit. Therefore we can limit our study to the description of spots and basic entities of WM. OntoWM will be used to help users and researchers of the web mining to understand the elements and steps of this method.

Step 2. Enumerate important terms in the ontology

In this step we write down a list of all terms (figure 1), with an expert in the method of Web Mining, we extracted, using a domain expert, more than 250 relevant terms. For example, important Web-Mining-related terms will include: Web Usage Mining, Web Content Mining, Web Log Mining, Semantics Web, Web Robots, Software Agent, Spiders, Log, History User, Softbots, Data Cleaning…

Nbr of terms	Terms	Nbr of terms	Terms
Web Mining			
58	Virtual Visitor	87	Web Robots
59	Web Content Mining	88	Spiders
60	Web Crawlers	89	Agents
61	Web Log Mining	90	Crawlers
62	Web Mining	91	Wanderers
63	Web Mining Techniques	92	Meta Crawlers
64	Web Search	93	TALN Technics
65	Web Server	94	Stemming
66	Web Structure Mining	95	Removing StopWords

Concepts Termes

Fig. 1. Part of the extracted terms

Step 3. Define the classes and the class hierarchy (Conceptualization and Ontologization)

We usually start by defining classes. From the list created in Step 2, we select the terms that describe objects having independent existence rather than terms that describe these objects. These terms will be classes in the ontology and will become anchors in the class hierarchy. We organize the classes into a hierarchical taxonomy. We identified the types of concepts in the field of Web Mining, drawing on data from sources located on ([6], [7]… this is our text corpus). The study of the domain revealed more than 200 concepts concerned.

For example (figure 2), from the terms (History-User, Log-User) were selected, by means of a domain expert, the concept candidate "Log-User". From the terms (Web-Usage-Mining, Web-Log-Mining) were selected, by means of a domain expert, the concept candidate "Web-Usage-Minig".

Term 1	Term 2	Term 3	Concept Chosen
Web Usage Mining	Web Log Mining		Web Usage Mining
History-User	Log-User		Log-User
Softaware-Agent	Softbots	Milti-Agent-System	Softbots
Pre-treatement	Pre-Processing		PreProcessing
Identification-Visitor	History-User	Log-User	Log-User

Fig. 2. Example of conceptualization

Classification, we then classified the types of concepts in classes and subclasses, thus forming a class hierarchy with root class: Web-Mining. These classes are the concepts of our ontology. We selected several kinds of concepts: Algorithms, Application-Domains, Tasks, Web-Usage-Mining, Web-Structure-Mining, HITS, Page-Rank...

To establish the hierarchy of classes, we conduct from top to bottom starting with the most general concepts and ending with the specialization of concepts. Therefore, we start with the most general classes, namely: Web-Mining, Algorithms, Application-Domains, Tasks, Basic-Concepts, … We then refined each class. For example, the class Fichie-Log-Format was refined by the concepts: Common-Log-Format, Extended-Log-Format and the class Text-Classification which has been specializing in sub-concepts: Automated-Filtering, Text-Categorization.

Operationalization (use of OWL)

To build our ontology "OntoWM.owl" we used the representation language OWL. OWL[1] is one of the languages most used in the construction of ontology. The ontology presented was performed through the use of the editor "Protégé" open source distributed by the University of Stanford[2] Medical Informatics. Protégé allows, through its

[1] w3c : http://www.w3.org
[2] Available at: http://protege.stanford.edu/

GUI automatic generation of code corresponding to the OWL ontology.

"Owl: Thing" is a predefined class. Every OWL class is a subclass of owl: Thing. Figure 3 is graphical representations (screenshots) of the class hierarchy of our ontology, produced using the tools OWLViz[3] and OntoGraf[4].

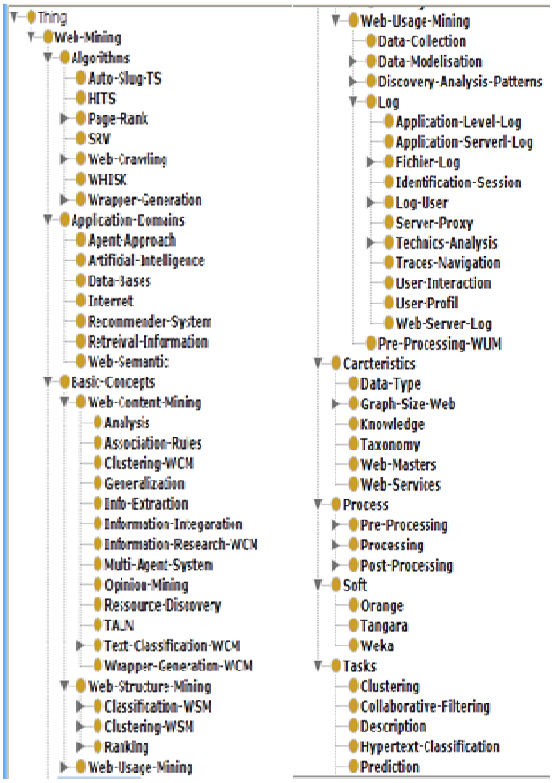

Fig. 3. General structure of our ontology OntoWM.owl

Step 4. Define the properties of classes

The classes alone will not provide enough information to answer the competency questions from Step 1. Once we have defined some of the classes, we must describe the internal structure of concepts. Most of the remaining terms are likely to be properties of these classes. For each property in the list, we must determine which class it describes. These properties become slots attached to classes. Thus, the Doc-Node class will have the following slot: NBR-Node (Figure 4).

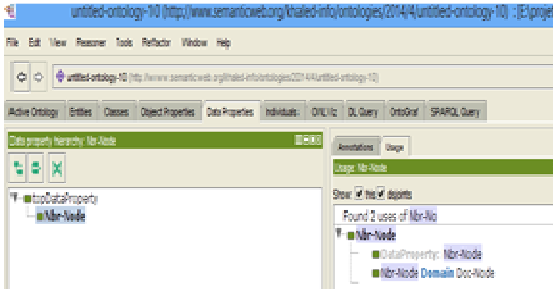

Fig. 4. The data properties

Step 5. Create instances

The last step is creating individual instances of classes in the hierarchy. The tab "Individuals" can create instances and assign properties. For example, Weka1.0 is an instance of the class Weka. On the screen presented (Figure 5, it is possible to edit the information about the individual "Weka1.0".

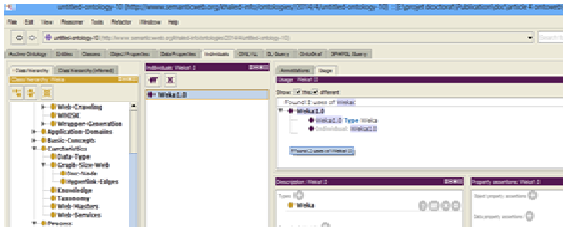

Fig. 5. Individual "Weka1.0"

3. Experimental Results

In terms of this section, we discuss the experimental evaluations and tests applied in our ontology. Assessment tests are performed on a machine with an Intel I5 2,53 GHz, 4 GB of memory under Windows 8. The platform used is Protégé 4.3.0 (with OWL 3.4.2, the reasoner FaCT++1.6.2[5], the reasoner HermiT 1.3.7, OWLViz 4.1.2 and OntoGraf 1.0.1).

3.1. Richer Knowledge representation

In this paper, and with this approach:
1. We have contributed to the advancement of research in the field of Web Mining using an ontology which is characterized by :
- 7 main parts: Application-Domains, Tasks, Caracteristics, Basic-Concepts, Process, Algorithms, Soft)
- 7 levels deep
- More than 200 concepts
- .....

[3] OWLViz is designed for use with Protégé OWL editor plugin. This tool allows you to view the hierarchy of classes in an OWL ontology

[4] OntoGraf gives support for interactively navigating the relationships of your OWL ontologies

[5] Fast Classification of Terminologies

2. We have in our ontology provides a richer representation of knowledge generally accepted in this field (The classification of the basic elements of the web mining by axes (Tasks, Algorithms …) will make it easier for a user or researcher to understand this method).

3.2. Coherence of the ontology model

We tested our ontology according to the criteria Furst [8] by the reasoner Fact + + (installed with Protégé 4, Fig 6) and the RDF W3C validator that allow us therefore also ensure that our OWL document follows the syntax of RDF (Fig 7), which already gives a first indication of the validity of our ontology that allows us to validate the consistency of the model associated with the ontology. So our ontology is characterized by (criteria of Gruber [9]):

• The clarity and objectivity of the definitions, which must be independent of any implementation choices, extensibility, no cycle (that is to say, loop definition) and no redundancy concepts and relationships.

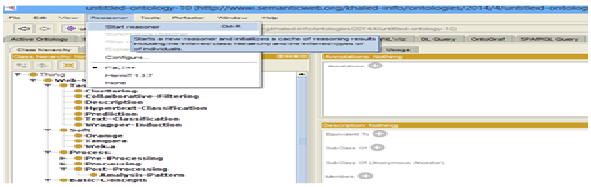

Fig. 6. Formal Validation of our ontology "OntoWM.owl" by the reasoner FaCT + +

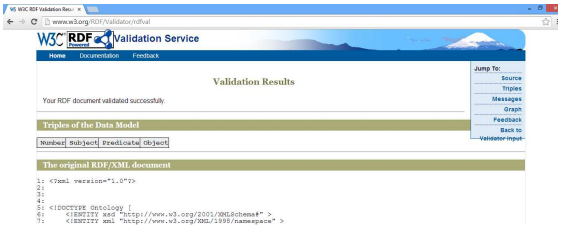

Fig. 7. Syntax validation of our ontology "OntoWM.owl" by the W3C RDF validator

4. Conclusion and Perspectives

In this paper, we have outlined ontology and web mining issues and requirements. A state-of-the-art review considering the main existing (in last 7 years) “web Mining-Ontology” approaches and tools was presented.

In this work, we present a proposal for a new ontology of Web Mining "OntoWM". OntoWM is the first ontology that describes the field of web mining in detail (different types of tasks and basic entities of the method of Web Mining).

OntoWM will be used to help users or researchers of the web mining to understand the elements and steps of this method. We have the ability to use this ontology as such, or to host in a site and make calls from the corresponding URI.

With our ontology OntoWM, we proposed a new unified and standard architecture based on ontologies for the terms most used in Web Mining.

The ontology developed (OntoWM.owl) is considered incomplete (we need to populate the proposed classes of Web Mining entities) and still needs to be improved throughout its life cycle. Our perspective is to improve and continue this work to enrich the ontology with new concepts and add inference refining properties and restrictions.

5. REFERENCES

[1] B. Smith, “Ontology. In: Blackwell Guide to the Philosophy of Computing and Information,” Oxford Blackwell, (Malden, 2003, 155–166).

[2] Bellandi A ., Furletti B., Grossi V., and Romei A., “Pushing constraints in association rule mining: an ontology-based approach,” Proceedings in IADIS International Conference WWW/Internet (Vila Real, Portugal, Year of Publication:2007).

[3] Cooley, R. and Mobasher, B. and Srivastava, J., ”Web mining: Information and pattern discovery on the world wide web,” In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Los Alamitos.

[4] Q. Yang, and X. Wu, “10 chalenging problems in data mining research,” International Journal of Information Technology & Decision Making, Vol. 5, No. 4 PP. 597–604 , 2006.

[5] N. F. Noy, and D. L. McGuinness,”Ontology Development 101: A Guide to Creating Your First Ontology,” (Stanford University, Stanford. 2004).

[6] B. Liu. “Web DataMining Exploring Hyperlinks, Contents, and Usage Data,” Book, ISBN-10 3-540-37881-2 Springer Berlin Heidelberg New York. 2007

[7] H. Yilmaz, “Using ontology based web usage mining and object clustering for recommendation,” Master thesis, the graduate school on natural and applied sciences of middle east technical university, 2010.

[8] F. R Furst, ”Contribution à l'ingénierie des ontologies: une méthode et un outil d'opérationnalisation,”PhD thesis, University of Nantes, France, 2004.

[9] T. R. Gruber, “Towards principles for the design of ontologies used for knowledge sharing,” International Journal of Human-Computer Studies, Vol. 43, n. 5-6, pp 907 – 928, 1995.