Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PROPOSAL FOR A NEW SYSTEMIC APPROACH OF ANALITICAL PROCESSING OF SPECIFIC ONTOLOGY TO DOCUMENTARY RESOURCES CASE OF EDUCATIONAL DOCUMENTS

[1]MORAD HAJJI, [2]MOHAMMED QBADOU, [3]KHALIFA MANSOURI
Laboratory: Signals, Distributed Systems and Artificial Intelligence (SSDIA)
ENSET, University Hassan II of Casablanca, Morocco
E-mail: [1]morad.hajji@gmail.com, [2]qbmedn7@gmail.com, [3]khmansouri@hotmail.com

ABSTRACT

The purpose of this paper is to present our approach whose target is the combination that results from the Semantic Web and the Business Intelligence with a view to an integration of models stemming from these two areas. This integration, that translates an integrated systemic perception, allows the processing of a domain-specific ontology of interest in order to support decision-making. Our approach aims at exploiting technological advances offered by the architecture of a decision support system to analyze an ontology generated from a corpus of documents related to the field of pedagogy. This ontology, whose generation is based on ontology learning techniques, constitutes a structured source of semantic data that can be analyzed by the suggested system to deduce decisional indicators which can be exploited in the pedagogic decision-making process.

Keywords: *Semantic Web, Business Intelligence, Data Warehouse, Ontology Learning.*

1. INTRODUCTION

With the grandiose expansion that characterizes the contemporary Web, the size and the diversity of its content make it difficult or even impossible to exploit it fully. The scattered and unstructured content lead to inefficient use. Thus, the colossal mass of available information on the web requires adequate resources in terms of the performance of systems provided for the data handling and processing of the web. This requires very huge systems whose main purposes are the manipulation, integration, search, etc. of these data, and therefore vertiginous investments. Therefore, the Semantic Web aims to provide a formal representation to web content to improve the exploitation of the great mass of data [1].

Besides, TIM BERNERS-LEE and al. define Semantic Web as bringing the structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users [1]. The Semantic Web proposes to annotate the web content to allow performing automated treatments exploiting that content and therefore facilitate its use. For example, the Semantic Web provides better performance of the search engine either at the level of precision or at the level of recall. The semantic web is fundamentally based on the notion of Ontology whose main reference definitions are those of Gruber [8] and Studer and al. [9].

On the other hand, decision-making is a fundamental goal of any organization and any management. One of the main problems is to determine relevant information for decision-making. Therefore, it is highly recommended to use Decision Support Systems, which offer tools to evaluate the various alternatives and their impacts on optimal decision-making. To support this decision in the most efficient way, the development of computer systems is an inevitable necessity. One of the characteristics of a Decision Support Systems (DSS) is its architecture bringing tools and models characterized by their high innovation. Anyway, Keen and Scott Morton [3] present the Decision Support Systems (DSS) as systems designed to solve some badly structured decisional problems. The SIAD incorporates statistics, operations research, optimization algorithms and numerical calculations and manage information (databases, files and data flow in the company).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In this context, the research studies on ontologies are limited to the approaches mainly based on fragmented perceptions. Indeed, the works that deal with the construction of an ontology from a corpus are limited to this stage without considering its analytical exploitation [17] [18] [19]. On the other hand, M. Gulić [5], Victoria Nebot and Rafael Berlanga [14] and T. Shiva and al. [20] works are limited to the use of existing ontologies to build a Data Warehouse. Therefore, they tackle partial problems in the treatment process of an ontology.

Our contribution consists of proposing a system model for a documentary corpus analysis to produce decisional indicators supporting the decision-making. Indeed, our approach exploits the overall structure of a decision making support system in order to set up a model of a system reflecting our systemic and integrated perception of the treatment process of an ontology generated from a textual documents corpus. This approach appears to be the first of its kind aiming at the analysis of a documentary corpus taking advantage of the Semantic Web conjugated to the BI. However, our work is limited to the processing of a textual documents corpus but may be extended to non-textual corpus. Furthermore, our approach deals with the analysis of a textual documents corpus stemming from the pedagogic domain. Moreover, during the treatment process, we opt for the intervention of human experts to validate the results of the ontology generation phase and the design of data warehouse phase. However, this human intervention is kept minimal insofar as it consists in correcting and validating the results of these two phases.

2. CONTEXT

Since its polemic birth, the semantic web is experiencing a succession of developments in which the central element is the analysis of ontologies. The current era is experiencing a proliferation of research works addressing an aspect concerning the concept of Ontology. These works can be divided into three phases according to the analysis of ontologies: construction, warehousing and analysis.

Concerning the construction phase of an ontology, there are two main families: the manual construction and semi-automatic construction. Certainly, the manual construction of an ontology is the most reliable one; nevertheless, it is proved to be too tedious and very expensive in terms of effort, resources and time. On the other hand, the second family, referred to as the 'Ontology Learning', refers to the set of approaches whose objective is the semi-automatic construction of ontologies via extraction, generation and acquisition. As part of the Ontology Learning, several approaches have been developed in order to promote the accompaniment and the facilitation of this overwhelming and arduous construction. These approaches can be classified according to the nature of the data sources: construction from textual documents, from dictionaries, from knowledge bases, from semi-structured data and from relational databases. There are a multitude of methods and techniques in each of these approaches.

Several methodologies have been developed with a view to define a roadmap for such a construction. Among the best known, we find Methontology [10], On-To-Knowledge [11], DILIGENT [12] and NeOn [13]. The construction of an Ontology includes a large set of tasks. The automation of this structures achieved through automating the different tasks performed during this phase or at least some of these tasks.

The construction approaches from the text are grouped into several categories. The approaches are outlined based on linguistic, statistical and conceptual relationships techniques. The implementation of each approach gives rise to a multitude of tools including Text2Onto that will be used in this phase of our approach.

The second phase of our approach is warehousing which is based on the concept of the multidimensional model. In this area, there are two main schools: that of the founder father William H. Inmon and that of his rival Ralph Kimball. The first extols the "top-down" approach, while the second one adopts the "bottom-up" approach. The first approach starts by the construction of the global data warehouse toward the specific data marts for each sectorial need. In contrast, the second approach starts by building Data Marts specifically to sectorial needs and then builds the global Data Warehouse. In relation with the end user, the approaches of multidimensional modeling can be classified into three classes: the so-called demand-driven, supply-driven and hybrid approaches. The approaches that are part of "demand-driven" class start with the manual (according to our approach) selection of the 'Facts' of analysis by the designer of the multidimensional model and then the process of transformation continues its sequence to determine the dimensions according to which these facts will be analyzed. On the other side, the approaches of the supply-driven class analyze firstly (according to our approach) the structure of the ontology (T-Box).

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Thus facts and potential dimensions emerge from this structure that will be the designer choice of the multidimensional model. If the first category is characterized by its performance, the second category is characterized by the multiple choice of several facts fled into the complexity of Ontologies.

On the other hand, the design approaches of a Data Warehouse from an ontology fundamentally differ by the level of representation of an ontology. An ontology can be regarded as a database consisting of its structure (T-Box) and the data it contains (A-Box). Therefore, the design of the Data Warehouse is based on the T-Box and loading it on the A-Box.

Finally, the analysis of the constructed Data Warehouse is the final phase of the process of ontology analysis as part of our approach. In the context of decision-making process, there are several ways to restitute the data: data mining, multidimensional analysis called OLAP (OnLine Analytical Processing), dashboards and generating reports. OLAP: Online Analytical Processing is the technology that allows producing descriptive summaries online of data contained in the data warehouses. OLAP is implemented by twelve rules dictated by Codd and his associates [16]. OLAP is based on the hyper-cubes concept whose data structure is specially adapted for extractions and crossings.

3. ARCHITECTURAL MODEL

The major contribution of our approach is the integration of models stemming from the Semantic Web and the Business Intelligence. This integration translates our systematic and integrated perception of the analysis of an Ontology. Figure 1 presents the global schema of the system that we are proposing to implement.

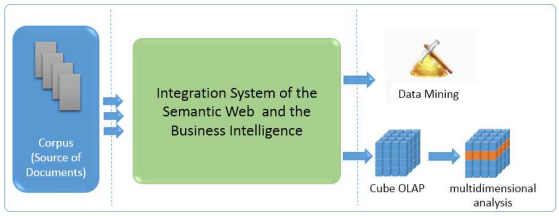

Figure 1: The System Architecture

The architectural model of our approach shown in Figure 1 includes a variety of phases; each phase includes specific models considered among the most relevant in their categories. In fact, our proposal sets up a system that is an intermediate infrastructure between data sources and the analysis of the Ontology built from these data sources. As shown in Figure 2, this infrastructure can be divided mainly into the following phases: A) Data acquisition, B) Construction of Ontology, C) ETL (Extraction-Transformation-Loading) and Data Warehousing and D) restitution of data.

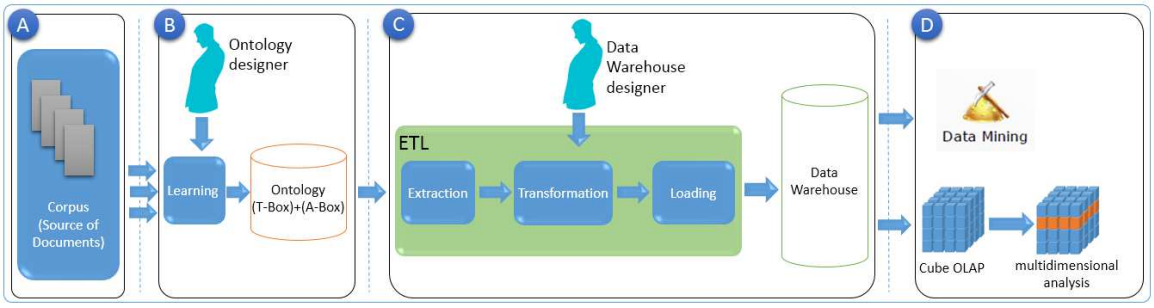

Figure 2: Architecture of the Suggested System for the Integrated Analysis of an Ontology

3.1 Data Acquisition

The constitution of a corpus of documents to establish a data source set of the analysis process is a fundamental part of the proposed architectural model. In fact, the data acquisition part has the vocation to make available documentary sources needed to build the Ontology. This phase consists of collecting and consolidating all documents to constitute the sources of information. The goal targeted during this phase is to build a corpus of documents of various types (.txt, .pdf, .htm, etc.). This corpus, constituting the base of a systemic analysis, is full of information that lead into the great mass of heterogeneous and unstructured data. In other words, this corpus is a raw gold mine which must be processed in order to obtain the most precious material hence the valorization of these documentary sources.

3.2 Construction of the Ontology

By bearing against the potential of Text2Onto tool [6], this phase aims at generating an Ontology of the domain of study from the corpus made up in the previous phase. The choice of Text2Onto tool is

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

justified by its performances. This performances were confirmed by several research works, like the one of Toader Gherasim who performed a comparative study of this tool along with three other tools (OntoLearn, Asium and Sprat) [2] and the work of Jinsoo Park and al. who has performed a comparison of the same tool along with three other tools (DODDLE-OWL, and OntoBuilder and OntoLT) [4]. The tool places several algorithms at the disposal of the designer. Each algorithm is devoted to a specific task of generating Ontology [6]. Ontology designer choses among these algorithms those to use for each task of this generation. The intervention of the designer allows the settings of the tool in addition to the refinement and validation of the intermediate result for the purpose of generation of the final ontology of domain as shown in Figure 3. This ontology composed of the two levels the T-Box (Terminological-Box) and the A-Box (Assertions-Box), constitutes the source of data submitted to the next phase of the analysis process.

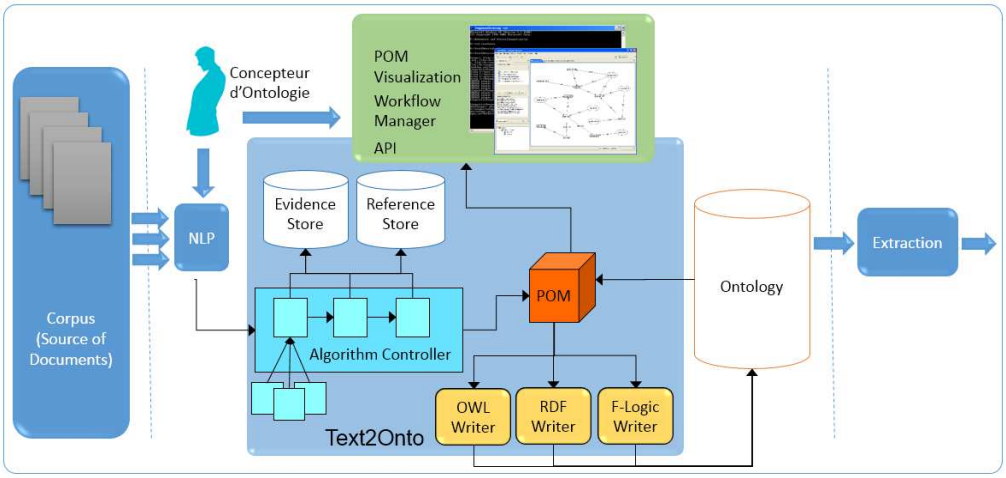

Figure 3: Construction of the Ontology

3.3 ETL and Data Warehousing

The phase of ETL is extremely complex and its implementation is crucial for proper conduct of the following phases of the process. This phase is divided into two parts: the construction of the Data Warehouse and the ETL as illustrated in Figure 4.

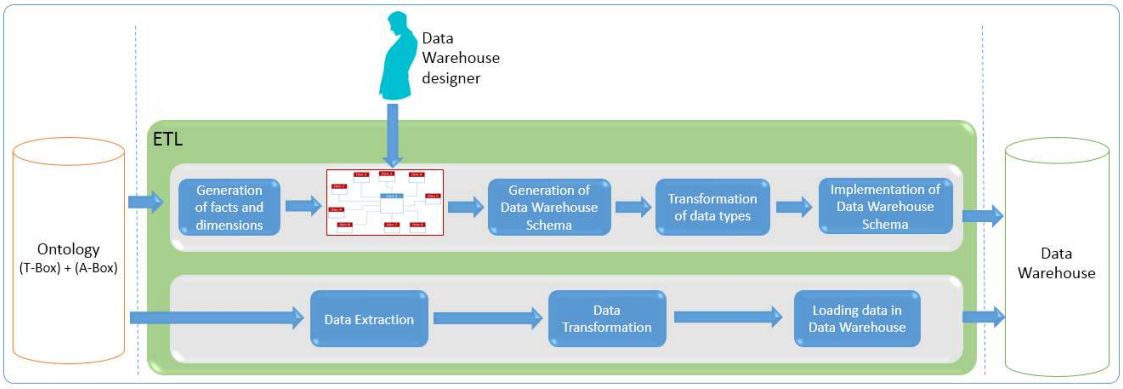

Figure 4: ETL and Warehousing

3.3.1 Construction of data warehouse

The resulting Ontology of the step 'Ontology Learning' is operated to generate firstly an intermediate schema of the Data Warehouse. Among the tables of this schema, the designer choses the 'facts' and 'Dimensions' that best suit with the specification of analysis needs. Following the intervention of the designer, the final schema of the Data Warehouse is automatically generated in the background. Once this schema is generated, it undergoes a transformation of data types stemming from the ontology towards equivalent data types of the chosen DBMS for hosting the Data Warehouse. This result obtained in the form of SQL DDL script is executed by the task 'Implementing the Schema

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of Data Warehouse' for constructing the final structure of the database representing the Data Warehouse according to the ROLAP approach. It is noted that to accomplish this task, we must take into consideration the fact that Ontologies are constructed hierarchically. Among the works, that approach the generation of a Data Warehouse from an ontology can be mentioned the ones of Dario Colazzo and al. [7], M. Gulić [5], Victoria Nebot and Rafael Berlanga [14], Shiva Talebzadeh and al [20] and Mohamed Yassine and Laadidi Bahaj [21].

3.3.2 ETL

This part is assimilated to standard ETL process as part of the architecture of a decision support system. The Ontology produced during the Ontology Learning phase is used yet another time to load data in the Data Warehouse in three stages. The first stage, assimilated to the task of extraction in the context of ETL process, extract the data from the A-Box level of Ontology. The second stage performs the necessary transformations on this extracted data. The last stage loads the transformed data into the Data Warehouse.

3.4 Restitution of Data

Data restitution constitutes the point of interaction between the system and users. The vocation of this phase is the accompaniment of the decision maker for the development and justification of decisions by relying on the decisional indicators developed during this phase. Several techniques are possible to achieve this aim but OLAP and Data Mining remain the best performers. OLAP allows the representation of the analyzed data in the form of multidimensional arrays, while the application of data mining allows the discovery of rules, associations and unknown or hidden trends in data. Indeed, the multidimensional representation allows navigation through the dimensions of Data Warehouse according to the operations Drill-down, Roll-up, etc. While in the context of data mining, the algorithm 'Apriori' [15] allows the discovery of association rules that reflect relationships between the elements of the Data Warehouse.

4. EXAMPLE OF APPLICATION

As mentioned earlier in the title of this article, we are interested in applying our approach to the domain of pedagogy. This domain is characterized by a specific set of pedagogical documents like the description sheets of university courses, the produced results by students in a module, etc. The collection of several pedagogical documents stemming from the academic domain constitutes the documentary corpus exploited to generate a related Ontology to this domain.

4.1 Ontology Construction Phase Result

The application of Text2Onto tool on this corpus generates a domain Ontology. Figure 5 illustrates the representative graph of this Ontology.

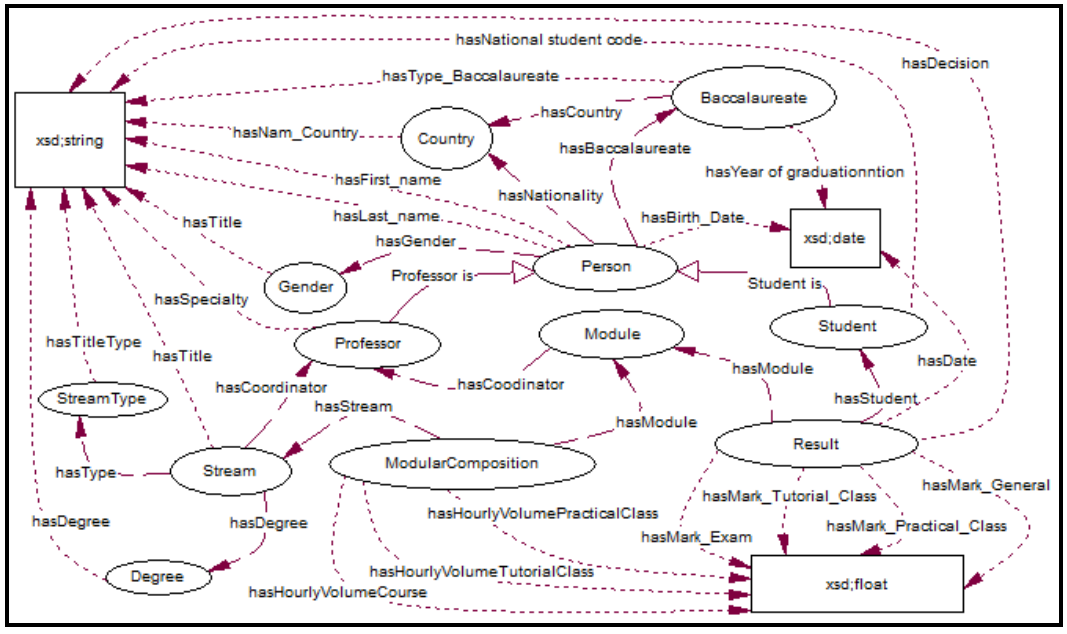

Figure 5: Graph Representing the Ontology

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.2 ETL and Data Warehousing Phase Result

The exploitation of this ontology allows generating a schema of a Data Warehouse that includes one fact table ('Result') and several tables of dimensions: Student, module, etc. Figure 6 presents the schema of this Data Warehouse. We have implemented this schema in the MySQL DBMS in order to obtain the physical Data Warehouse. After the implementation of this schema, we have loaded data, stemming from the Ontology of the domain generated in the step of Ontology Learning, into this Data Warehouse.

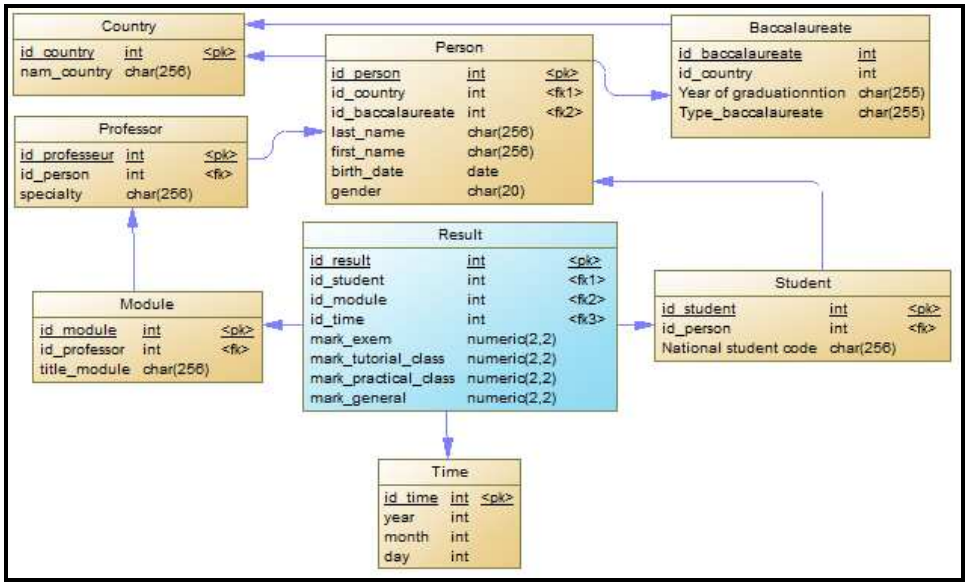

Figure 6: Schema of Data Warehouse

4.3 Data Restitution Phase Result

The data restitution is the latest phase of our model of treatment of an Ontology. In fact, OLAP allows us to navigate in the multidimensional model constructed using Ontology generated from the documentary corpus. While, Data Mining allows us to explore and analyze data in order to discover hidden information useful for decision-making. Figure 7 presents an example of restitution of data using OLAP.

		Mesures			
Student	Module Title	Exam Mark	Practical Class Mark	Tutorial Class Mark	General Mark
All students	All Titles	13,3	13,3	13,3	13,3
	Design and Object Oriented Programming	13,47	12,79	13,13	13,13
	Web technology	13,13	13,8	13,47	13,47
Name_student_1	All Titles	13,13	13,13	13,13	13,13
	Design and Object Oriented Programming	12,12	12,12	12,12	12,12
	Web technology	14,14	14,14	14,14	14,14
Name_student_2	All Titles	13,64	13,64	13,64	13,64
	Design and Object Oriented Programming	14,14	14,14	14,14	14,14
	Web technology	13,13	13,13	13,13	13,13
Name_student_3	All Titles	13,13	13,13	13,13	13,13
	Design and Object Oriented Programming	14,14	12,12	13,13	13,13
	Web technology	12,12	14,14	13,13	13,13

Figure 7: Restitution of Data Using OLAP

In Figure 8, we give the result of the application of the algorithm 'Apriori' for the discovery of association rules concerning the validation of programmed modules, especially those reflecting the precedence relation between these modules. These association rules are very revealing of dependence of certain training modules. Indeed, the validation of the module 'Design and Object

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Oriented Programming' implies the Validation of the module 'Web Technology' with a confidence equal to 100%, Which means that all students who have validated the module 'Design and Object Oriented Programming' have also validated the module 'Web Technology'. This information is extremely important especially for the planning of these two modules insofar as it is preferable that learning of the module 'Design and Object Oriented Programming' occurs prior the learning of the module 'Web Technology' in order to promote the success rate.

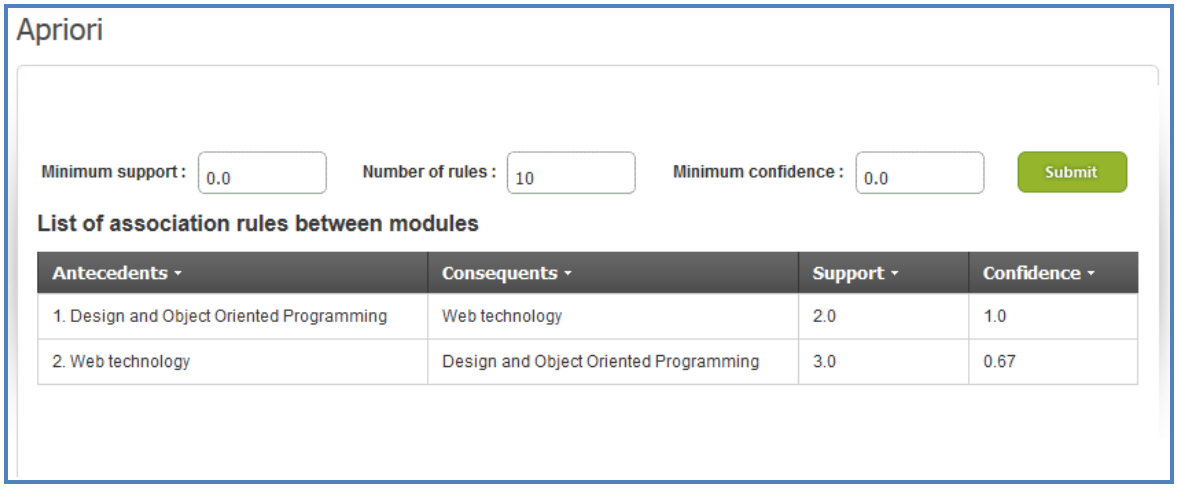

Figure 8: Restitution of Data Using 'Apriori' Algorithm

5. CONCLUSION AND PERSPECTIVES

Our contribution consists in proposing a system model for a documentary corpus treatment to produce decisional indicators supporting the decision-making. Indeed, our approach exploits the overall structure of a decision making support system in order to set up a model of a system reflecting our systemic and integrated perception of the treatment process of a generated ontology from a textual documents corpus. This approach appears to be the first of its kind aiming to analyze a documentary corpus taking advantage of the Semantic Web conjugated to the BI. However, our work is limited to the processing of a textual documents corpus and may be extended to non-textual corpus. Furthermore, our approach is applied for the treatment of a textual documents corpus stemming from the pedagogic domain. Moreover, during the analysis process, we opt for the intervention of human experts to validate the results of the ontology generation phase and the design of data warehouse phase. However, this human intervention is minimal insofar as it consists in correcting and validating the results of these two phases.

Our approach of integrating the Semantic Web and Business Intelligence has been concretized by the proposal of a system model to analyze a documentary textual corpus. Through this model, we have combined the Semantic Web and Business Intelligence through an architecture based on the evolutions carried out in both areas. The proposed model renders clearer and justifiable the decision-making by producing decisional indicators exploiting the documentary corpus of the studied domain. Our approach has been applied to the analysis of a documentary corpus relating to the pedagogical field. With this approach, this documentary corpus has been exploited and appreciated as a raw data mine to extract from it valuable information supporting and justifying the decision-making. Especially, through this approach we have been able to stand out an important educational characteristic defining a precedence rule among several modules of a university course.

The difference between the results of our work and those referenced in [5], [6], [7], [14], [17], [18], [19], [20] and [201], is mainly due to the difference in startup visions. Ours is a holistic approach that covers the overall treatment process of a documentary corpus to directly produce decisional indicators. However, works cited above deal with fragmented aspects of this process limiting itself to the construction of an ontology from a documentary corpus, as in the case of work [6], or limiting itself to the design of a Data Warehouse from an ontology, as in the case of works [5], [7], [14], [17], [18], [19], [20] and [201]. Nonetheless, these works

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

resulted in undeniable and interesting results, but their results remain partial.

This work has opened the horizons for the foundational exploration on which are based the Semantic Web and the Business Intelligence. Indeed, In the future, this work is going to serve us in developing and implementing the suggested model and in trying out the coupling of this model with other disciplines like Big Data, E-Learning and machine learning. Furthermore, the human intervention in the treatment process renders this process semi-automatic. Complete automation of this process is possible based on the concept of expert systems. An extension of our work consists of conducting further works proposing a multi-approaches model for expanding the analytical options to several other domains and for improving the performances process.

REFRENCES:

[1] T. Berners-Lee, J. Hendler and O. Lassila, "The semantic web", Scientific American, vol. 1, 2001, PP. 34–43.

[2] G. Toader, "Détection de problèmes de qualité dans les ontologies construites automatiquement à partir de textes", PhD thesis, University of Nantes, 2013.

[3] Pater G. W. Keen and Michael S. Scott-Morton, "Decision Support Systems: an organizational perspective", Addison-Wesley Publishing, 1978.

[4] J. Park, W. Cho, and S. Rho, "Evaluating Ontology Extraction Tools Using a Comprehensive Evaluation Framework", Data & Knowledge Eng., vol. 69, 2010, PP. 1043-1061.

[5] M. Gulić, "Transformation of OWL Ontology Sources into Data Warehouse", Faculty of Maritime. Studies, Rijeka, Croatia, 2013.

[6] P. Cimiano and J. Völker, "Text2Onto –A Framework for Ontology Learning and Data-driven Change Discovery", Natural language processing and information systems : 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, 2005, PP.227-238.

[7] D. Colazzo, F. Goasdoué, I. Manolescu, and A. Roatis, "Warehousing RDF graphs", Bases de Données Avancées, 2013.

[8] T. Gruber, "A translation approach to portable ontology specifications", Knowledge Acquisition Journal, 1993, PP. 199-220.

[9] R. Studer, V. R. Benjamins and D. Fensel, "Knowledge engineering: principles and methods", IEEE Transactions on Data and Knowledge Engineering, 1998, PP. 161-197.

[10] M. Fernández-López, A. Gómez-Pérez and N. Juriste, "METHONTOLOGY: From Ontological Art Towards Ontological Engineering", Symposium on Ontological Engineering of AAAI, 1997.

[11] S. Staab, H. P. Schnurr, R. Studer and Y. Sure, "Knowledge Processes and Ontologies", IEEE Information Systems, 2001

[12] H. S. Pinto, S. Staab, and C. Tempich, "Diligent : Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies", Proceedings of the 16th European Conference on Artificial Intelligence, 2004, vol. 110, PP. 393–397.

[13] M. C. Suárez-Figueroa, "NeOn methodology for building ontology networks: specification, scheduling and reuse", PhD thesis, Universidad Politécnica de Madrid, 2010.

[14] V. Nebot and R. Berlanga, "Building data warehouses with semantic web data", Decision Support Systems, 2012, PP. 853–868.

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", Very Large Data Bases, 1994, pages 487–499.

[16] E. F. Codd, S.B. Codd, and C. T. Salley, "Providing OLAP to User Analysts: An IT Mandate", Codd E. F. & Associates, Technical report, 1993.

[17] O. Romero and Alberto Abelló, "A framework for multidimensional design of data warehouses from ontologies", Data & Knowledge Engineering (69/11), 2010, PP. 1138-1157.

[18] M. Gulić, "Transformation of OWL Ontology Sources into Data Warehouse", Faculty of Maritime Studies, 2013.

[19] V. Nebot, and R. Berlanga, "Building data warehouses with semantic data", Proceedings of the 2010 EDBT/ICDT Workshops, 2010.

[20] T. Shiva, S. Mir Ali, and S. Afshin, "Automated Creating a Data Warehouse from Unstructured Semantic Data", International Journal of Computer Applications, 2014, PP. 19.

[21] Y. Laadidi and M. Bahaj, " Designing a Data Warehouse from OWL sources", International Journal of Soft Computing and Software Engineerin