Data Semantic Extraction applied to the Higher Education Institutions

Caio Saraiva Coneglian[1], Elvis Fusco[1], José Eduardo Santarém Segundo[2]

[1]Departamento de Ciência da Computação – Centro Universitário Eurípides de Marília (UNIVEM)
Av. Higino Muzi Filho, 529 – 17525-901 – Marília – SP – Brasil

[2]Pós-Graduação em Ciência da Informação – Universidade Estadual Paulista (UNESP)
Av. Higino Muzi Filho, 737 – 17525-000 – Marília – SP – Brasil

caio.coneglian@gmail.com, fusco@univem.edu.br, santarem@usp.br

Abstract. *In the context of architectures of Information Retrieval, this research aims on implementing a semantic extraction agent for higher education institutions Web environment, allowing information finding, storage, processing and retrieval, such as those from the Big Data context produced by several informational sources on the Internet, serving as a basis for the implementation of information environments for decision support. Using this method, it will be possible to verify that the agent and ontology proposal addresses this part and can play the role of a semantic level of the architecture.*

1. Introduction

The massive diffusion of generated data is testing the ability of the most advanced techniques of information storage technological, treatment, processing and analysis. The areas of treatment and information retrieval are being challenged by the volume, variety and velocity of semi-structured and unstructured complex data, offering opportunities for adding value to business-based information providing organizations a deeper and precise knowledge of their business.

Opportunities to add value to the businnes-based information arise due to both the internal and external environment. Hence, there is a need for a new approach to structure Information Technology (IT) companies to transform data into knowledge, which cause broader impact.

To aggregate and use information that are scattered in the internal and external environments of organizations, there is the concept of Competitive Intelligence, which according Fleisher and Blenkhorn [2001] is a process by which organizations gather actionable information about competitors and the competitive environment and, ideally, apply it to their decision-making and planning processes in order to improve their performance. A proactive informational process leads to a better decision, whether strategic or operational, in order to discover the forces that govern the business, reduce risk and drive the decision maker to act in advance, besides protecting the generated knowledge.

In the current scenario of the information generated in organizational environments, especially in those who have the Internet as a platform, there is data that, due to its characteristics, is classified as Big Data.

In the literature, Big Data is defined as the representation of the progress of human cognitive processes, which generally includes data sets with sizes beyond the capacity of current technology, methods and theories to capture, manage and process the data within a specified time [Graham-Rowe, et. al. 2008]. Douglas [2012] defines Big Data as the high volume, high speed and/or high variety of information that require new ways of processing to allow better decision making, new knowledge discovery and process optimization.

In the process of information search for Competitive Intelligence and Big Data robots, data mining techniques on the Internet are used.

According to Deters and Adaime [2003] robots are systems that collect data from the Web and assemble a database that is processed to increase the speed of information retrieval.

According to Silva [2003], the extraction of relevant information can rank a page according to a domain context and also draw information structures them and storing them in databases. To add meaning to the content fetched, the robots are associated with Web search semantic concepts, which let the search through a process of meaning and value, extracting the most relevant information.

The ontology in the philosophical context is defined by Silva [2003] as part of the science of being and their relationships; in this sense, the use of ontologies is essential in the development of semantic search robots, being applied in Computer Science and Information Science to enable a smarter and closer search to the functioning of the cognitive process of the user so that data extraction becomes much more relevant.

Thus, an agent presents itself as a solution to retrieve information on the web by semantic means. Currently, the content is organized in a jointly manner, in which syntactic structures do not have semantic data aggregation. In this sense, the role of the agent is to extract the information from the content and use syntactical ontology to achieve semantic relations and apply them to retrieval information.

This research aims to implement a semantic agent for searching on the Web and allowing the retrieval, storage and processing of information, i.e., Big Data from various informational sources on the Internet. Such semantic agent will be the main mechanism for building a computational architecture that transforms disaggregated information on an analytical environment of strategic, relevant, accurate and usable knowledge to allow managers the access to opportunities and threats in the field of higher education institutions, based on concepts of competitive intelligence. The semantics of the agent will be built using ontological structures.

To achieve this goal, the Semantic Agent will be built using the domain of higher education institution, addressing the problem related to scientific research.

In section II is shown on the related work. In section III it is explained how it works the architecture and the differential of this architecture for the others that already exist. Section IV explains what was the basis for the construction of ontology. Already

in section V explains the entire process of building and operating a semantic agent within the architecture. And finally in section VI is spoken of the conclusions and the next jobs.

2. Related Work

Information retrieval architectures with use of agents have been proposed by some other studies, where conducted ways to retrieve information for the subsequent use of this information in any scenario.

In this sense, Beppler [2005] proposes a type of architecture of information retrieval. This recovery occurs only by analyzing documents, and removing and storing information, without observing the existing context, e.g., using syntactic analyses. This proposal is interesting because it can retrieve information in a successful way but is limited because it only a syntactic search, which narrows the results obtained by the architecture.

Already Wiesner [2008] proposes a semantic solution for this issue, using for making ontologies the recovery of information. This proposal already assembles an architecture that uses a semantic solution using ontologies to assemble a base of knowledge integration, using an agent to make associations and integration of such knowledge. So this research can do a lot of good associations required for each knowledge, but the semantics are limited because only the Association of information and cannot determine what each represents information, and if it in fact will be of real value for that particular domain.

3. Information Retrieval in Big Data

The traditional information systems are unable to cope efficiently with all new data sources and multiple contexts of information that have mainly the Internet as a platform.

Problems are encountered in retrieving, standardising, storing, processing and usage of information generated by various heterogeneous sources that are the basis for enabling systems for decision support organizations.

In this context, it is questioning whether the computing environments of information actually present completely all relevant information to decision makers in organizations. The solution proposed in this paper is to create an architecture for information retrieval in the context of Big Data as seen in Figure 1.

This architecture was proposed so that the recovery of the information can be made using the semantic space. The architecture has the agent as the structure of information retrieval and integrates with all elements and layers of the architecture.

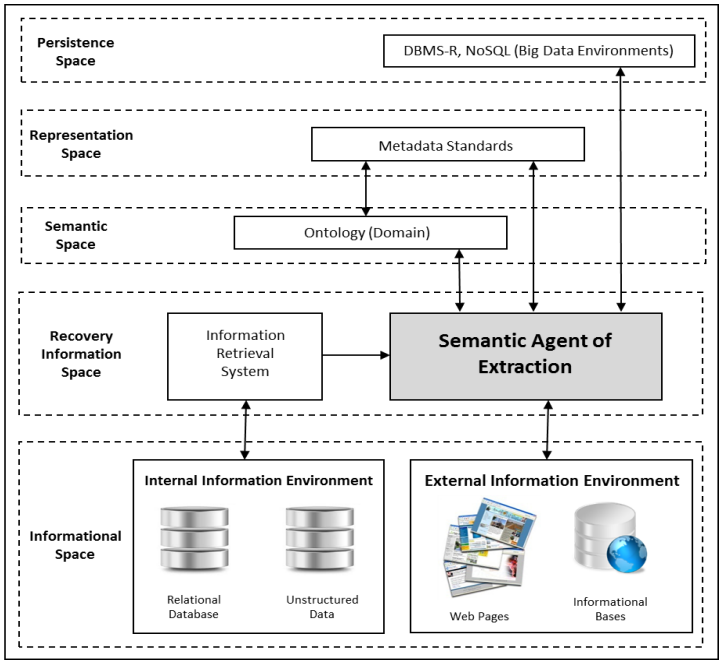

Figure 1. Architecture Context of Semantic Agent of Extraction

This architecture will be used at an institution of higher education, because in this area there are many information not being used several times in competitive intelligence. This architecture distinguishes itself from others by doing all the recovery information using the same domain, therefore, only will be extracted information relating to the problem in accordance with the defined ontology.

4. Ontology

To set a network of Semantic Web, ontologies have been used frequently [Falbo, et al. 2003]. According to Clark [1999], an ontology is organized into concept hierarchies, because it cannot reflect optimally specific formalism, then it is possible to consider an ontology as the embodiment of the knowledge level.

To Mizoguch [2003], there are several types of ontologies, and can be outlined below:

- Upper Ontology: this type of ontology serves to explain what exists in the world. And most are used to represent large knowledge bases;
- Domain Ontology: is a more specific area of knowledge;
- Task Ontology: This ontology serves to solve a specific problem of a domain;
- Heavy-Weight Ontology: these ontologies are much more defined, have well-defined rules, be very careful when conceptualizing the world;
- Light-Weight Ontology: this type need not be precise as large in the conceptualization.

Noy and McGuiness [2001] explains the seven steps that are required to build an ontology: 1. Determine the domain and scope of the ontology; 2. Consider reusing existing ontologies; 3. Enumerate important terms in the ontology; 4. Define the classes and the class hierarchy; 5. Define the properties of classes—slots; 6. Define the facets of the slots and; 7. Create instances.

The focus of the use of ontology in this case is for being the semantics of the agent. The agent will acquire the information from web pages, and then the data will be submitted to the implemented ontology.

5. Semantic Agent of Extraction

The creation of a software agent that aggregates semantically information available on the web in a given domain can bring grants to a computational platform for creating an information environment for decision support giving a through broader view of the internal and external scenarios of information relevance in organizational management.

In this context, we understand the extreme importance of using agents to extract data through scrapper semantic search with the use of technologies like NoSQL persistence in information processing with characteristics of Big Data, essential in the recovery, storage, processing and use of various types of information generated in these environments of large volume data sets on Competitive Intelligence.

In the context of the architecture presented in Figure 1, this research are dealing the problem of automatic and semantic information extraction of web environments that have as informational sources: web pages, web services and database with the development of the agent semantic of data extraction.

This agent should communicate with internal and external information spaces of Big Data basing their search on ontological rules on a metadata standard to perform the semantic extraction of the domain proposed and supported by other systems in a broader context of Information Retrieval.

From this semantic search, the scrapper actuates as a tooling strategy in the search and find the information that really add value to the decision-making process. Inside a huge and massive data structure scattered throughout the web, it is essential that the search engines do not support only syntactic structures of decision in information retrieval, but also in investigations of the use of semantic extraction agents.

Our research uses the domain of higher education institutions as a case study to apply the proposed computing platform in the architecture described in Figure 1. For the development of the prototype of the ontology, we used the issues of scientific research within educational institutions, such as notices, grants, funding agencies, search directories, events, journals, among others.

5.1 Ontological Conceptual Notation

To create the ontology, first it was necessary to check within the domain of scientific research in higher education institutions, which are the classes that are involved in this issue.

It was checked what were these classes, and analyzed the hierarchy between them, based on the experience of the authors ' research and other researchers, and thus was sealed the hierarchy between the classes seen in Figure 2, using mind maps to represent the ontology.

Figure 2. Class Hierarchy of Ontology

It was carried through the mental maps of the ontology schema and was drafted this conceptual notation of ontology, using the Protégé software, as shown in Figure 3. It was built the relationship between classes. In this figure the dotted arrows are properties of objects in each class, i.e., when a dotted arrow goes from one class to another, means that the class from which emerged the arrow contains an object of destination class of the arrow.

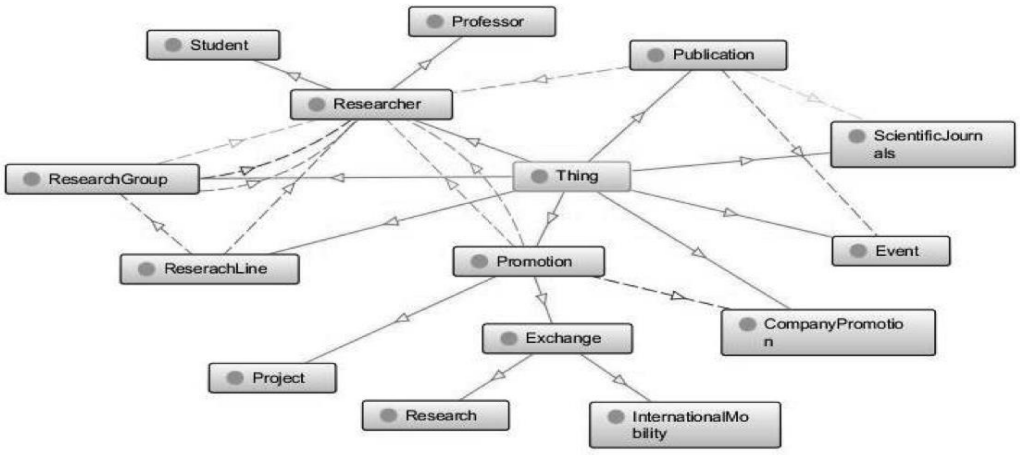

Figure 3. Relationship of classes of ontology of scientific research

After created the class hierarchy of the ontology has been defined all the properties of each class. The properties are shown in Table 1.

Table 1. Slots of ontological classes

Classe	Propriedades	Tipo
Publication	publication_year	int
	publication_class	string
	publication_title	string
	publication_key_words	string
	publication_type	string

	publication_city	string
ScientificJournals	journal_name	string
	journal_edition	int
	journal_editor	string
	journal_page	String
Event	event_name	string
	event_edition	string
	event_organizer	string
	event_date	date
	event_type	string
Book	book_name	string
	book_chapter	string
	book_date	date
ResearchGroup	group_name	string
	group_area	string
ReseachLine	line_name	string
Researcher	researcher_name	string
	researcher_title	string
	researcher_curriculum	string
	researcher_internal	boolean
Professor	professor_title	string
Student	student_registry	string
Promotion	promotion_value	float
Exchange	exchange_duration	int
	exchange_start	date
Research	research_title	string
	research_area	string
InternationalMobility	mobility_university	string
	mobility_country	string
Project	project_area	string
CompanyPromotion	company_name	string
	company_type	string

The agent will act on this proposed ontology that this scenario is called Task Ontology, according Mizoguch [2003].

It is an ontology that solves a specific problem within a domain, that is, solves the problem of scientific research within the domain of an institution of higher education.

We implement the ontology in Protégé software, creating the class diagram and its properties, being implemented in a file OWL (Web Ontology Language) (WC3, 2002H).) There after the Owl2Java tool transformed OWL in classes Java (Java, 2004); thus making the implemented ontology.

5.2 Semantic Agent Working Method

The agent for performing searches captures information through pre-defined web pages and uses the ontology to classify and make a semantic search.

The figure 4 shows the operating scheme of the semantic extraction agent, which shows the entire process carried out by the system since the retrieval of the information to storage in the database.

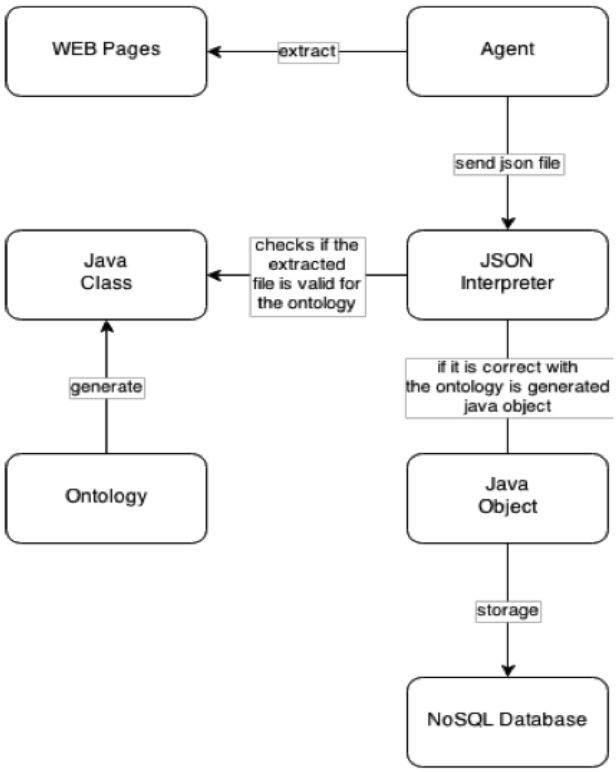

Figure 4. Schematic of operation of agent

The following describes the sequence of tasks performed by the agent

5.2.1 Information Extraction

The agent will extract pages previously defined data pages. The list of pages that are extracted information were based on domain ontology represents. All of this information is extracted from the HTML format and transformed into a JSON file.

All the extracted data will be placed in a JSON file, and for this there must be a standard metadata to identify and classify each item of information to assist in this process, we use the Dublin Core metadata standard.

The Dublin Core assists as well, can catalog the various information contained in a web page, and thus define the characteristics of each page and this way is possible to use the ontology in a semantic search over these metadata.

5.2.2 JSON Interpreter

JSON interpreter has the function of receiving the JSON file generated by the agent, and interpret to verify that file fits within the ontology.

This process, the interpreter, verifies that the information that is contained within that file, is compatible with the ontology, i.e. will check the semantics of the document, if it indeed fits the ontology

If the JSON interpreter to recognize that given how useful, he will create a Java object based on the Java class created by ontology, so that it can be stored in the database.

5.2.3 Use of Ontology for Information Retrieval

Having the information of web pages with their metadata defined, the use of ontology becomes more viable, because from there the information can be classified by ontology implemented by Owl2Java software, and thus obtaining classified information have each according to their class ontology, and may thereafter be used for competitive intelligence.

5.2.4 Storage in the Database

Because the information is not structured and don't know exactly what will be given and what is the structure of each type of information, by the use of a non-relational database MongoDB. Since the MongoDB is a document-oriented database, and do not need to have a well-defined schema for guards information.

And from the Java object that comes from the JSON parser is done the storage in the database.

6. Conclusions

From this research it was observed that the use of ontology to trailer search agent is an effective way to get retrieve value information, and manage to get an efficient competitive intelligence.

The ontology can be effective in this process, because it becomes a way to organize the information semantically, and thus, only meaningful information will be used.

In addition, that was achieved through the implementation of the ontology and use it as a reference search for the agent.

Having access to information from your business domain is a fundamental requirement for management and decision making in organizations.

An Information Retrieval system has the ability to provide relevant information for accessing Web sites and services, it is necessary the existence of software agents that add semantic information from various informational sources for a specific domain.

Although the term semantic web be used for a few years, there is still a limitation of its use, because much of the web is still organized in a syntactic way, this semantic agent is a solution to this problem, because the semantics are treated off the web, through the ontology and achieving achieve their results.

7. Acknowledgements

The work presented in the paper was supported by the FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo), process 2013/16369-3.

References

Bepler, F. D. et al. (2005). "Uma Arquitetura Para Recuperação De Informação Aplicada Ao Processo De Cooperação Universidade-Empresa" ("An Architecture for Retrieval of Applied Information In Case Of University- Industry Cooperation"), São Paulo.

Clark, D. (1999). "Mad cows, meta-thesaurus and meaning, IEEE Intelligent Systems".

Davies, J.; Fensel, D. and Van Harmelen, F. (2003). "Towards The Semantic Web: Ontology-Driven Knowledge Management", John Wiley & Sons Ltd.

Deters, J. I. and Adaime, S. F. (2003). "Um estudo comparativo dos sistemas de busca na web" ("A comparative study of search systems on the web"), Anais do V Encontro de Estudantes de Informática do Tocantins. Palmas, TO, 189- 200.

Diana, M. and Gerosa, M. A (2010). "NOSQL na Web 2.0: Um Estudo Comparativo de Bancos Não-Relacionais para Armazenamento de Dados na Web 2.0" ("NoSQL Web 2.0: A Comparative Study of Non-Relational Data Storage Benches for Web 2.0"). São Paulo.

Douglas, L. (2012) "The Importance of 'Big Data': A Definition." Gartner (June 2012).

Falbo, R. D. A., Natali, A. C. C., Mian, P. G., Bertollo, G., & Ruy, F. B., (2003). "ODE: Ontology-based software Development Environment". IX Congreso Argentino de Ciencias de la Computación, p. 1124-1135, La Plata, Argentina.

Fleisher, C. S. and Blenkhorn, D. L. (2001). "Managing Frontiers in Competitive Intelligence". Greenwood Publishing Group.

Graham-Rowe, D. et. al. (2008). "Big data: science in the petabyte era". Nature, 455, 1-50.

Gruber, T. R. (1995). "Towards Principles for a Design of Ontologies Used for Knowledge Sharing", International Journal of Human and Computer Studies, 907-928.

Mizoguchi, R. (2003) "Tutorial on Ontological Engineering". NEW GENERATION COMPUTING- TOKYO- 21.4, 363-364.

Noy, N. F. and McGuiness, D. L. (2001) "Ontology Development 101: A Guide to Creating Your First Ontology". <http://www.ksl.stanford.edu/people/dlm/papers/ontology- tutorial-noy-mcguinness-abstract.html> [retrieved: 05/27/2014].

Owl2Java.<http://www.incunabulum.de/projects/it/owl2j ava> [retrieved: 03/14/2014].

Prescott, J. E. (1995). "The Evolution of Competitive Intelligence", International Review of Strategic Management v. 6, 71-90.

Protégé Stanford University. <http://protege.stanford.edu/> [retrieved: 03/10/2014].

Silva, T. M. S. (2003) "Extração De Informação Para Busca Semântica Na Web Baseada Em Ontologias" ("Information Extraction for Semantic Search In Web Based On Ontology"). Florianópolis.

Wiesner, K. et al. (2008) "Recovery Mechanisms for Semantic Web Services". DAIS 2008, LNCS 5053, pp. 100– 105.