Semantic Agent of Informational Extraction on Big Data Ontological Context

Caio Saraiva Coneglian, Elvis Fusco
Department of Computer Science
UNIVEM –University Center Euripides of Marilia
Marilia, SP - Brazil
caio.coneglian@gmail.com, fusco@univem.edu.br

***Abstract*—The large increase in the production and dissemination of data on the Internet can offer information of high value-added to organizations. This information may be from heterogeneous databases that may not be considered relevant by most systems, e.g., social media data, blogs, and more. If organizations would use such sources, they could build a new management vision known as Competitive Intelligence. In the context of an architecture of Information Retrieval, this research that aims on implementing a semantic extraction agent for the 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.**

Keywords-Big Data; semantic web; semantic scrapper; ontology.

I. 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 far-reaching 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 [1], 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 knowledge generated.

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 [2]. Gartner [3] 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 on the Internet are used; according to Deters and Adaime [5] 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 [6], 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 [6] 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 search smarter and closer 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

 ISBN: 978-1-61208-329-2

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 this paper, the proposed architecture, the test of the ontology and the Semantic Agent are described.

II. 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, standardizing, 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 in full all relevant information to decision makers in organizations.

In this sense, Beppler [16] proposed 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 analysis.

The solution was proposed to create an architecture for information retrieval in the context of Big Data, as seen in Figure 1.

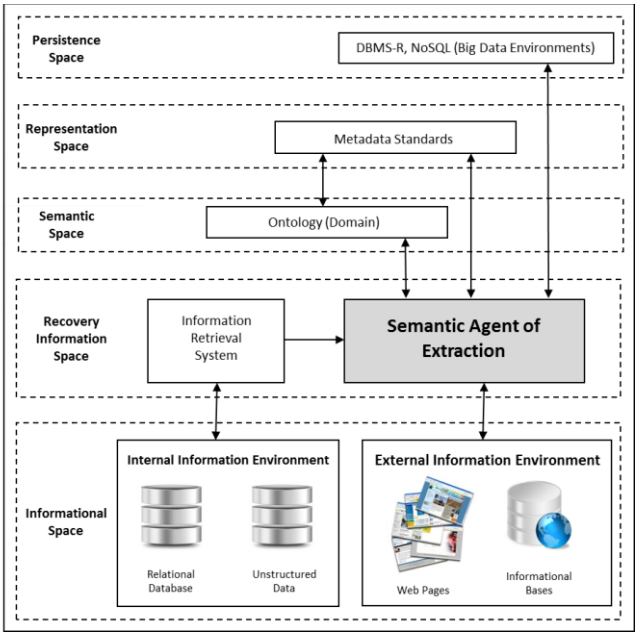

Figure 1. Architecture Context of Semantic Agent of Extraction.

Already Wiesner [17] proposed a semantic solution for this issue, using for making ontologies the recovery of information with a more generic architecture, by this the ontology you can insert semantics during the recovery process.

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.

This architecture will be used at an institution of higher education, because in this area there is much information not being used several times in competitive intelligence.

This architecture distinguishes itself from others, for doing all the recovery information using the same domain, therefore, only information relating to the problem in accordance with the defined ontology is extracted.

III. Ontology

To set a network of Semantic Web, ontologies have been used frequently [13].

According to Clark [10], an ontology is organized into concept hierarchies, because it cannot reflect optimally specific formalisms; then, it is possible to consider an ontology as the embodiment of the knowledge level.

There are several types of ontologies, as outlined below [11]:

- 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 an 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, and
- Light-Weight Ontology: this type need not be precise as large in the conceptualization.

Noy [8] explains the seven steps that are required to build an ontology, these steps are described below: 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 ontology proposal was hatched seeking cover a problem within the domain of higher education institution, which is the issue of scientific research, was used this issue to be able to have a more synthetic ontology, with the focus on extracting value information, why this ontology is not so great; so, one can get a better view about the Semantic Agent, which is the focus of this research.

 ISBN: 978-1-61208-329-2

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 from there pass the data by ontology implemented.

IV. SEMANTIC AGENT OF EXTRACTION

The creation of a software agent that aggregates semantically information available on the web in a given domain can bring to a computational platform grants for the creation of an information environment for decision support giving a 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 [4] 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 comes 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.

The 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, and journals, among others.

To elaborate the conceptual notation of ontology, we used Protégé software [14], as shown in Figure 2, which shows the class hierarchy of the ontology. 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.

The agent will act on this proposed ontology that this scenario is called Task Ontology, according Mizoguch [11]; 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.

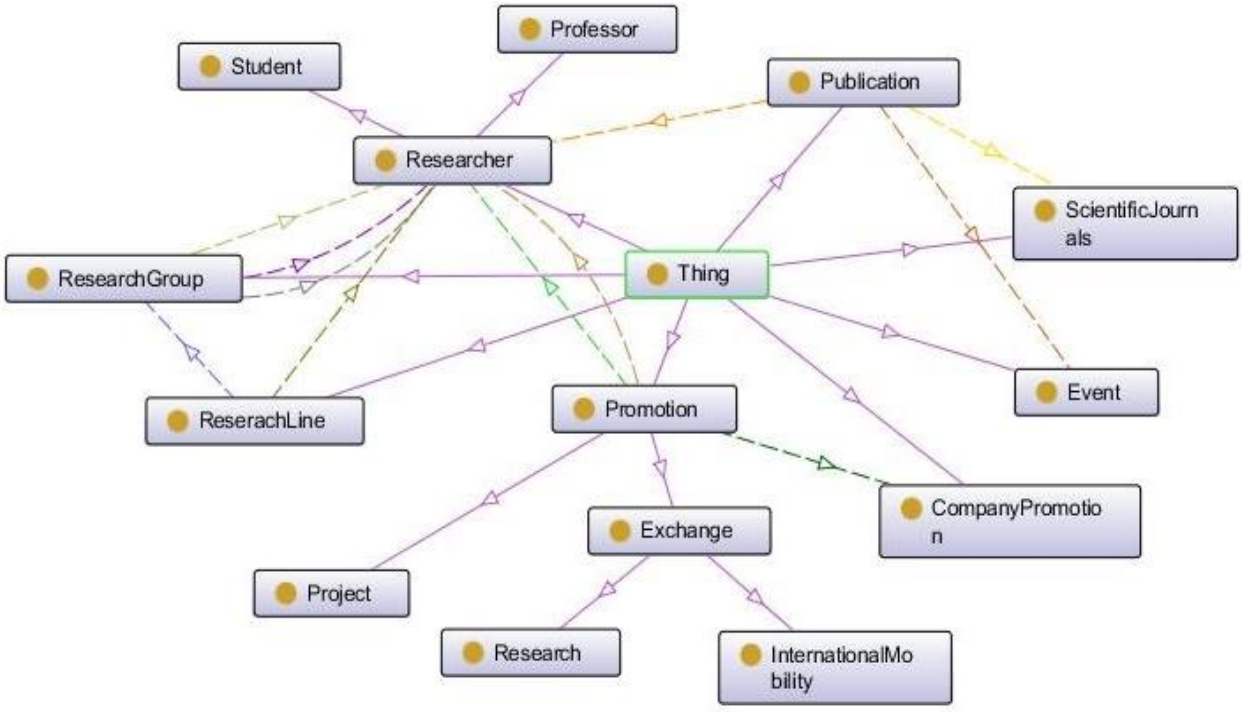

Figure 2. Class Hierarchy of Research Ontology.

 ISBN: 978-1-61208-329-2

V. CONCLUSION AND FUTURE WORK

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.

In this context, the scrapper semantic search enters as a tooling strategy for searching and finding the information that really add value to the decision-making process, because inside a huge and massive structure of data spread across the Web, is essential that the search engines do not support only in syntactic structures of decision, but also in investigations of the use of semantic extraction agents.

Currently, we are performing the implementation of the ontology, and the integration with semantic extraction agent, that is also being created. This process is nearly being finalized. The agents will be tested for verifying if will be able to extract from there will be tests to check if the agent will be able to extract the information that really will add value to competitive intelligence.

By the development of an agent for semantic extraction, authors envision an effective use of information in the Environments Information Retrieval, an effective use of information in the scenarios of Big Data in the field of Higher Education Institutions will be obtained.

REFERENCES

[1] C. S. Fleisher and D. L. Blenkhorn, “Managing Frontiers in Competitive Intelligence”. 2001. Westport.

[2] “Big data: science in the petabyte era”. Nature 455 (7209): 1. 2008.

[3] Gartner, Douglas and Laney, “The importance of big data: A definition”. 2008.

[4] M. Diana and M. A. Gerosa, “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, 2010.

[5] J. I. Deters and S. F. Adaime, “Um estudo comparativo dos sistemas de busca na web” (“A comparative study of search systems on the web”).

[6] T. M. S. Silva, “Extração De Informação Para Busca Semântica Na Web Baseada Em Ontologias” (“Information Extraction for Semantic Search In Web Based On Ontology”). Florianópolis, 2003. <https://repositorio.ufsc.br/handle/123456789/85791> [retrieved: 03/10/2014].

[7] M. A. A. Mesquita, “Web Semântica E Recuperação Da Informação Na Internet : O Que Esperar Do Futuro?” (“Semantic web and information retrieval on the Internet: What to Expect From the Future?” 2010.

[8] N. F. Noy et al, “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: 03/10/2014].

[9] T. R. Gruber, “Towards Principles for a Design of Ontologies Used for Knowledge Sharing”, International Journal of Human and Computer Studies. 1995

[10] D. Clark, “Mad cows, meta-thesaurus and meaning, IEEE Intelligent Systems”. 1999.

[11] R. Mizoguchi, “Tutorial on Ontological Engineering”. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76.6226&rep=rep1&type=pdf> [retrieved: 03/10/2014]

[12] J. Davies, D. Fensel and F. Van Harmelen, “Towards The Semantic Web: Ontology-Driven Knowledge Management”, John Wiley & Sons Ltd, 2003.

[13] R.A. Falbo, et al., “ODE: Ontology-based software Development Environment”. IX Congreso Argentino de Ciencias de la Computación, p. 1124-1135, La Plata, Argentina, Outubro 2003.

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

[15] J.E. Prescott, “The Evolution of Competitive Intelligence”. 1999.

[16] F.D. Bepler, et al. “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”) 2005, São Paulo.

[17] K. Wiesner et al. “Recovery Mechanisms for Semantic Web Services” DAIS 2008, LNCS 5053, pp. 100–105, 2008.

 ISBN: 978-1-61208-329-2