Semantic Agent in the Context of Big Data
Usage in Ontological Information Retrieval in Scientific Research

Caio Saraiva Coneglian[1], Elvis Fusco[2] and José Eduardo Santarem Segundo[3]
[1]*Information Science Department, Universidade Estadual Paulista (UNESP), Marília, São Paulo, Brasil*
[2]*Computer Science Department, Centro Universitário Eurípides de Marília (UNIVEM), Marília, São Paulo, Brasil*
[3]*Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brasil*

Keywords: Semantic Web, Semantic Agent, Ontology, Big Data.

Abstract: The evolution of information technology caused an expansion in the amount of data available on the internet. Moreover, such developments demanded that new tools were created to allow processing at high velocity, trying various informational sources. In this context, in flocking to the three V (Velocity, Variety and Volume), emerged the phenomenon called Big Data. From the emergence of this phenomenon, the need to generate new architectures that allow that users, enjoy the high volume of data spread throughout the Web. One way to improve the processes carried out, insert the question of semantic information processing, in which the use of domain ontologies can expand as computational agents interpret the meaning of the data. Thus, this paper aims to present a proposal for architecture that places the elements of Big Data and semantic, seeking to insert a model that is adapted to the current computing needs. As proof of concept performed the implementation of the architecture, exploring the question of scientific research, where a user is assisted to find relevant information in academic databases. Through the implementation, it was found that the use ontologies in a Big Data architecture, significantly improves the recovery of information performed by computational agents.

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 business-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 (Coneglian and Fusco, 2015). 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

Coneglian, C., Fusco, E. and Segundo, J.
Semantic Agent in the Context of Big Data - Usage in Ontological Information Retrieval in Scientific Research.
DOI: 10.5220/0005875703240330
In *Proceedings of the International Conference on Internet of Things and Big Data (IoTBD 2016)*, pages 324-330
ISBN: 978-989-758-183-0

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 search to scientific papers.

This paper is a continuation of the work presented in the papers Coneglian, Fusco and Botega (2014) and Coneglian and Fusco (2014).

2 INFORMATION RETRIEVAL IN BIG DATA

The creation of a software agent that semantically aggregate the information generated by the IoT devices can bring to a computational platform, subsidize their implementation of an information environment to support the decision to give a broader view of internal and external scenarios of information relevance in organizational management.

In this context, means the extreme relevance of use data extraction agents through semantic search robots with the use of metadata standards, technologies and means for the achievement of interoperability, being essential in recovery, storage, processing and use of various types of information generated in these environments of large volume of data through the various devices that remain connected through the internet.

To this end it was proposed an Information Recovery architecture in the context of Big Data as seen in Figure 1.

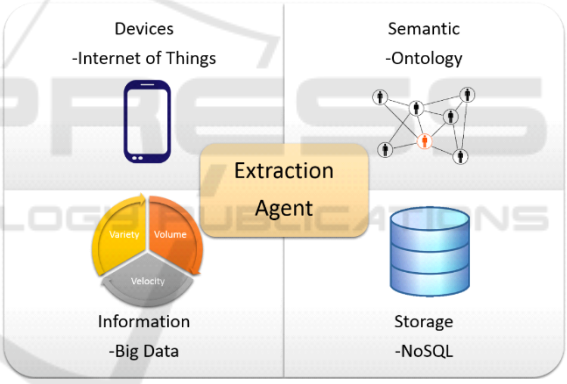

Figure 1: Architecture Context of Semantic Agent of Extraction.

In the context of the presented architecture, this research relates the various network-connected devices called IoT, with information environments on the Web that show the Big Data features, due to the great variety, the volume and the velocity with which it is generated and processed data . Data generated from such devices also create a high volume, having thus Big Data characteristics. All this data is unstructured and disorganized, needing a tool that links and enter meaning and insert the various information generated and stored. To this end, the center of the architecture has an agent that extracts data both information environments as data generated by the devices, and through the use of ontologies, insert meaning to information, so that treatment and

making inferences from the collected data to present more efficiently. Finally, the proposed architecture of the storage information generated and collected and semantically related, in a NoSQL database structure.

The use of ontologies is given through the agent, which after extraction of information, query an ontology that addresses the specific domain in which those data are inserted. From this semantic representation structure, data is rearranged, making the structured, and thus, the possibility of inferences made based on data that have been extracted.

3 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 database discipline, to perform the search for scientific papers

3.1 Ontology

To create the ontology, first it was necessary to check within the domain of database discipline, 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. To build the ontology was used Noy and McGuiness (2001) methodology, that 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.

It was drafted this conceptual notation of ontology, using the Protégé software. It was built the relationship between classes.

The agent will act on this proposed ontology that this scenario is called Task Ontology, according Mizoguchi (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 database discipline.

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. The figure 2 represents the development ontology.

3.2 Working Method of the Semantic Agent

The search agent captures information by means of pre-defined web pages and uses the ontology to classify and make a semantic search.

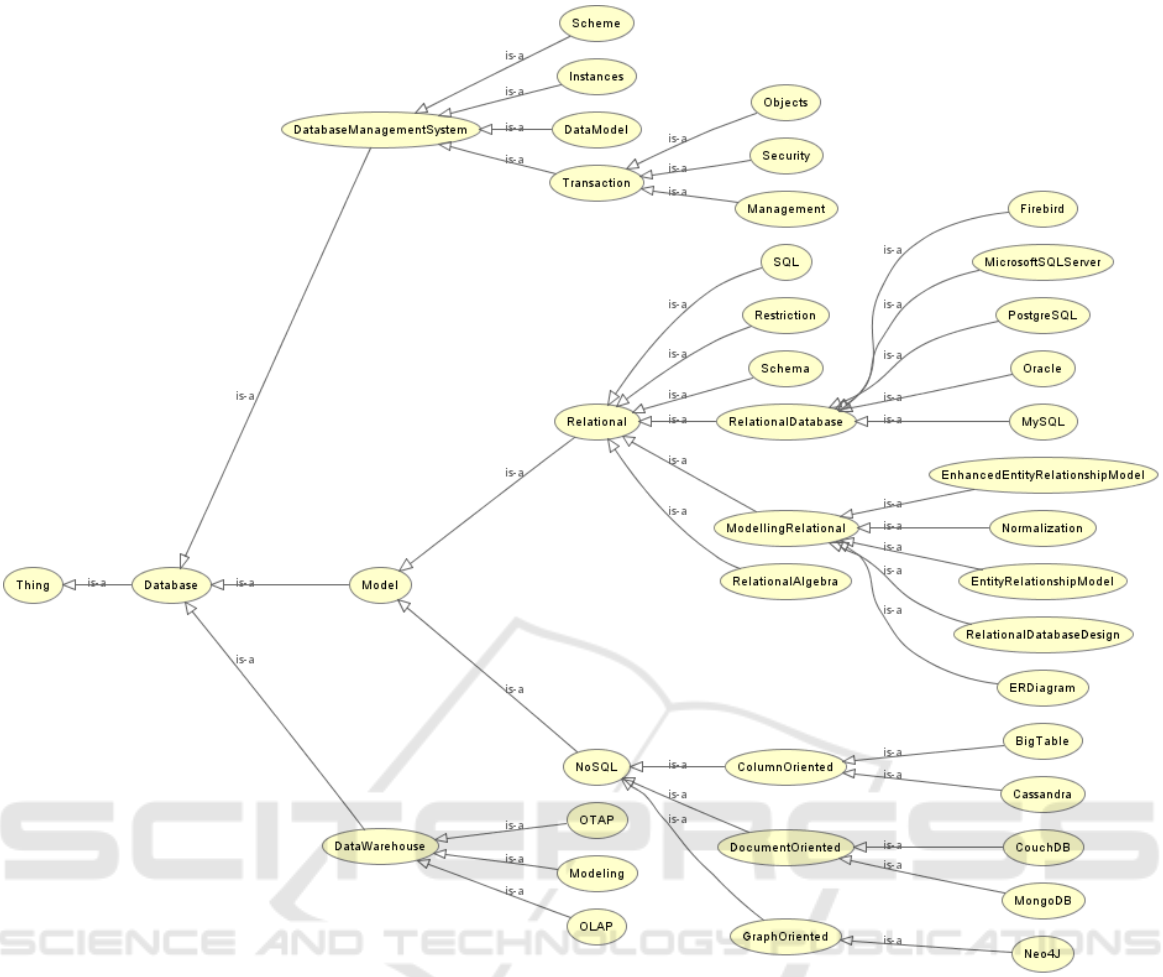

Figure 2: Ontology.

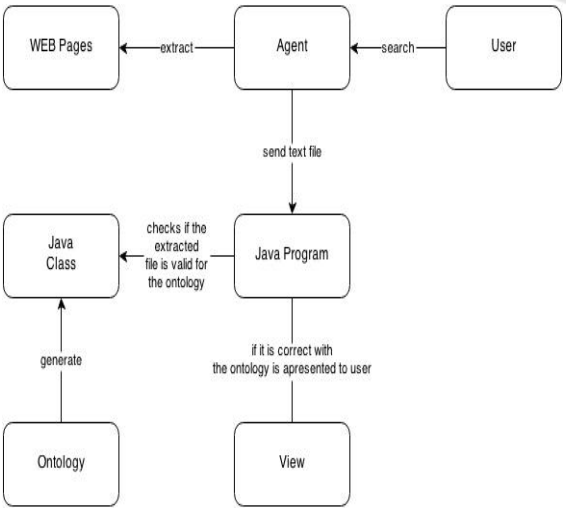

Figure 3: Scheme of operation of agent.

Figure 3 shows the process made by the system.

The user performs a search on any subject, the agent extracts of databases summaries relating to this topic.

These summaries will go through a process where they will be analyzed, taking into consideration the words contained in this summary are present in the field that sought theme. This will be possible using a built ontology, which deals with a specific topic in the area of scientific research.

The sequence of tasks performed by the agent is described as follows:

3.2.1 Information Extraction

The agent extracts the abstracts from IEEE Xplore page (http://ieeexplore.ieee.org) based on the research the user carries out. Based on the location of the abstracts in the HTML page, the agent extracts the information and transforms it into a String chain.

The agent process is divided into three phases: search in the page, extraction of the titles and

abstracts, and return of a list of the papers to the main program.

- Search in the HTML page: this first phase is characterized by performing a search in the IEEE Xplore search system, so that the search is characterized by a request to this system inserted in the url, which is the theme that the user want to search. For example, if the user wants to perform a search on Data warehouse, the agent will open a connection, and look at the following address (http://ieeexplore.ieee.org/search/searchresult.jsp ?Newsearch=true&QUERYTEXT = datawarehouse). From this page, the IEEE returns an HTML containing articles related to this topic.
- Extraction of titles and abstracts from the page: after returning from the HTML, the agent extracts the title and the abstract of each paper It is made possible by analyzing the HTML page by checking the tags whose data abstracts and titles are inserted. In this manner, for each item it is create a Java object that contains information about title, abstract and link to the full article. To perform this removal of data inside an HTML page, we used the tool JSOUP, which works as an HTML parser, in other words, working with the HTML page, so that it can extract the class data, tags and structures of HTML.
- Development of a list of articles extracted: finally, the agent creates a list of all the items that were extracted from the HTML page. This list will be used by the main program, which will join the ontology with this information retrieval agent.

Thus, this search robot is able to perform a syntactic extraction of the articles contained in the IEEE Xplore database. Therefore, the search robot retrieves the items that have been indexed by the database, creating a list of all the articles presented to be used in the ontology.

3.2.2 Use of Ontology for Information Retrieval

In order to effectly have the semantics presented, the program makes use of ontology to assess which of the results obtained from the database are actually useful, and related to the context of that search. This integration takes place in five stages:

First, it is checked where the term searched by the user is found within the ontology. For example, if the user performs a Datawarehouse search, the system checks where this term is found within the ontology. It is obtained the hierarchically higher and lower classes to the search term. In the case of Datawarehouse, it is obtained the lower classes: OLAP, OTAP and modeling, as well as the upper class Database. It is then checked in the abstract and title of the searched papers whether it contains or not the words that make part of that hierarchy of the search term. In the case of Datawarehouse, it would be verified if the terms OLAP, OTAP, modeling, Datawarehouse and database are contained within the abstracts and titles of the extracted papers. Next, it is done a comparison between the amount of terms in the hierarchy and those contained in the abstract and the title of the article, thereby resulting in a percentage quantity of terms contained in the hierarchy, which are within the abstract and the title of the article. In the same example, if the terms Database, OLAP, Datawarehouse and modeling are contained within an article, it will contain four of the five terms of the hierarchy, which results in a percentage of 80% of the terms. Finally, it is presented to the user all the items that reached a percentage above 35%.

3.2.3 Presentation to User

The user will be held in an initial screen of search, and then the system will make it from the ontology integration process with user requirements. After the extraction and process of the articles, the system returns to the user a screen containing the articles and links of these articles that the system extracted and considered related to the search performed by the user. This result can be seen in Figure 4, where the names and links are presented so that the user can access the full article.

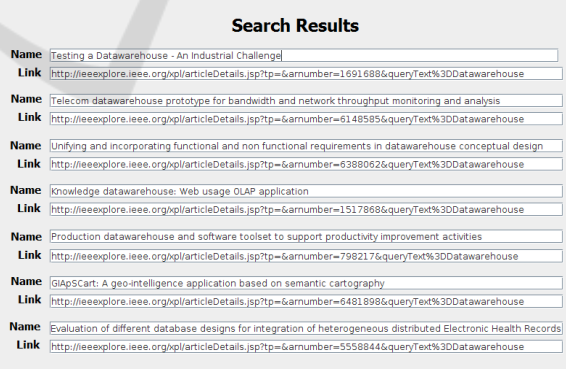

Figure 4: Screen presentation of results.

4 RESULTS

In order to check if the system is extracting and verifying the semantics of extracted articles, a search was made with the user searching for the term

"Datawarehouse".

The Datawarehouse term hierarchy are the words: Database, Datawarehouse, OLAP, OTAP and modeling.

In Table 1, you can view all the titles of the articles that have been extracted from the IEEE Xplore website, the amount of jail terms of ontology that were found in the abstract and title, the relationship between the terms found in the article and the terms of the chain the ontology of the term "Datawarehouse" (in this case will be the percentage resulting of dividing the amount of words found in the ontology by 5, which are the terms contained in the ontology chain hierarchy) and this paper meets or not the minimum requirement of at least 35% of the terms contained in the abstract and title.

Table 1: This caption has one line so it is centered.

Title	*Number of words*	*%*	*Will be presented?*
Testing a Datawarehouse - An Industrial Challenge	2	40	**YES**
Telecom datawarehouse prototype for bandwidth and network throughput monitoring and analysis	3	60	**YES**
Unifying and incorporating functional and non functional requirements in datawarehouse conceptual design	3	60	**YES**
Knowledge datawarehouse: Web usage OLAP application	2	40	**YES**
Datawarehouse and dataspace — information base of decision support syste	1	20	NO
The implementation of datawarehouse in Batelco: a case study evaluation and recommendation	1	20	NO
E-Business Model Approach to Determine Models to Datawarehouse	1	20	NO
Production datawarehouse and software toolset to support productivity improvement activities	2	40	**YES**
A genomic datawarehouse model for fast manipulation using repeat region	1	20	NO
A datawarehouse for managing commercial software release	1	20	NO
Modeling Analytical Indicators Using DataWarehouse Metamodel	1	20	NO
An SLA-Enabled Grid DataWarehouse	1	20	NO
Business Metadata for the DataWarehouse	1	20	NO
A partition-based approach to support streaming updates over persistent data in an active datawarehouse	1	20	NO
Study of localized data cleansing process for ETL performance improvement in independent datamart	1	20	NO
Visualizing Clouds on Different Stages of DWH - An Introduction to Data Warehouse as a Service	0	0	NO
GIApSCart: A geo-intelligence application based on semantic cartography	2	40	**YES**
JISBD 2008 + TELECOM I+D 2008 = INTRODUCTIONS	0	0	NO
Normed principal components analysis: A new approach to data warehouse fragmentation	0	0	NO
Enriching hierarchies in multidimensional model of data warehouse using WORDNET	0	0	NO
The fragmentation of data warehouses: An approach based on principal components analysis	0	0	NO
Evaluation of different database designs for integration of heterogeneous distributed Electronic Health Records	2	40	**YES**
Keynote talk data warehouses: Construction, exploitation and personnalisation	1	20	NO
Security Analysis of Future Enterprise Business Intelligence	0	0	NO
QVT transformation by modeling: From UML model to MD model	1	20	NO

In the case of 25 articles, seven out of them fulfilled the requirements, which are presented to users. This presentation can be viewed in Figure 4.

5 CONCLUSIONS

This paper presents the use of ontologies to improve the Information Retrieval process.

The objective of this research is to add semantics to the information retrieval process by using the information in the context of Big Data to perform a process that adds more value to the searches performed.

In order to prove this goal, it was used the domain of scientific research in which the user, when

performing a search of scientific articles in databases, is faced with the problem of having a very large number of documents, but much of these are not actually useful and do not attend the user's need.

It was then created an ontology and a search robot, and the connection between them was established so that the the initial goal was achieved.

For testing, in a way to assess the actual operation of this process, the search robot was implemented with the ability to extract articles from the IEEE Xplore database, and the ontology has been built with the field of database discipline.

After testing, it was observed that the use of ontology for the search agent is an effective way to obtain valuable information and be able to meet the informational needs of the user.

The ontology can be effective in this case, because it becomes a way of organizing semantic information, and in this manner, only significant information will be presented to the user.

Although the Semantic Web term has already been used for some years, there is still a limitation in their use, because much of the Web is organized in a syntactic form in which most pages are created so that only humans can read what is written without being structured in a way that computational agents can extract the data inside a context with the implied meaning in the HTML.

The extraction agent can extract the documents from the web, and a program can process information by using onthology, thereby presenting the most relevant results.

In this manner, the results obtained by using the prototype developed can substantially narrow down the amount of items presented to users. This research aims, therefore, at making the user get, in a process of Information Retrieval, more significant, quality results. Thus, the user can evaluate more information that is meaningful and does not waste time with data that does not meet their needs.

Therefore, in order to address the issue of how to insert intelligence in the recovery of web pages which do not contextualize their information, the present research proposes that the process of adding semantics to these pages takes place outside the Web, that is, the extraction of pages occurs in a syntactic way, and from what was extracted, information is checked by semantically entering into this process. This method was very efficient since it is in fact able to make a smarter search, which goes beyond simple formulas of searches, which observe only the syntax of the texts, and is able to analyze the context in which the extracted documents are inserted, and then visualize if that document fulfill the user's needs.

ACKNOWLEDGEMENTS

The work presented in the paper was supported by the CAPES and FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo), process 2015/01517-2.

REFERENCES

Coneglian, Caio Saraiva, and Elvis Fusco. "Recuperação da Informação em Ambientes Semânticos: uma ferramenta aplicada à publicações científicas." Journal on Advances in Theoretical and Applied Informatics (ISSN 2447-5033) 1.1 (2015): 30-37.

Coneglian, Caio Saraiva, and Elvis Fusco. Semantic Agent of Informational Extraction on Big Data Ontological Context, eKNOW 2014: The Sixth Internationl Conference on information, Process, and Knowledge Management (2014): 34-37.

Coneglian, Caio Saraiva, Elvis Fusco and Leonardo de Castro Botega. Using Ontological Semantic Agent to Data Extraction in the Context of Big Data: a tool applied to Higher Education Institutions. Proceedings of the International Conference on Semantic Web and Web Services (SWWS) (2014): 1-6.

Fleisher, Craig S., Blenkhorn, David L., "Managing Frontiers in Competitive Intelligence". Greenwood Publishing Group, 2001.

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

Deters, Janice I., Adaime, Silsomar F., "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, 2003.

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

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

Noy, Natalya F., McGuiness, Deborah L., "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: 09/27/2015], 2001.

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

Protégé. Stanford University. <http://protege.stanford.edu/> [retrieved: 10/10/2015].

Owl2Java.<http://www.incunabulum.de/projects/it/owl2java> [retrieved: 11/14/2015].