A Method for Building Personalized Ontology Summaries

Paulo Orlando Queiroz-Sousa[1], Ana Carolina Salgado[1], Carlos Eduardo Pires[2]

[1] Universidade Federal de Pernambuco, Brazil
{povqs, acs}@cin.ufpe.br
[2] Universidade Federal de Campina Grande, Brazil
cesp@dsc.ufcg.edu.br

Abstract. In the context of ontology engineering, the ontology understanding is the basis for its further development and reuse. One intuitive effective approach to support ontology understanding is the process of ontology summarization which highlights the most important concepts of an ontology. Ontology summarization identifies an excerpt from an ontology that contains the most relevant concepts and produces an abridged ontology. In this article, we present a method for summarizing ontologies that represent a data source schema or describe a knowledge domain. We propose an algorithm to produce a personalized ontology summary based on user-defined parameters, e.g. summary size. The relevance of a concept is determined through user indication or centrality measures, commonly used to determine the relative importance of a vertex within a graph. The algorithm searches for the best ontology summary, i.e., the one containing the most relevant ontology concepts respecting all the original relationships between concepts and the parameters set by the user. Experiments were done comparing our generated ontology summaries against golden standard summaries as well as summaries produced by methods available in related work. We achieved in average more than 62.5% of similarity with golden standard summaries.

Categories and Subject Descriptors: E.1 [**Data Structures**]: Graphs and networks; E.2 [**Data Storage Representations**]: [Linked representations; Object representation]; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks

Keywords: Ontology, Ontology Summarization, RDF, Graph, Centrality Measure

1. INTRODUCTION

Ontology has played an important role in the development and deployment of the Semantic Web. The ontologies have been used to semantically organize and define the information shared on the web with a remarkable capability to specify a shared conceptualization explicitly and formally [Gruber 1993]. Based on these characteristics, domain ontologies were developed to model a basic structure of knowledge for a specific domain. The modeling of the terms and processes in a specific domain ontology, combined with the ability to reuse ontologies, has made easier the development of new ontologies.

With the reuse of ontological knowledge, more and more ontologies have been developed and expanded in different application domains, such as multi-agent systems [Obitko and Marik 2002] and data integration [Gagnon 2007]. Ontologies are used in different applications and are considered a powerful tool to enable knowledge sharing. In addition, they can be seen as a way to achieve semantic interoperability in heterogeneous distributed systems. According to [Gagnon 2007], ontologies can be used to solve semantic heterogeneity problems in the integration of data sources. The incorporation of semantic knowledge can provide a better semantic understanding of the data. Despite of this, ontologies have the same comprehension problems of any other data source with a complex data schema or large amount of data.

The development of this work was supported by CNPq and FACEPE research grants.

Given the current trend of increasingly large data repositories, with an explosive growth in both data size and complexity of schemas, data source schemas are becoming increasingly difficult to be understood. One possible way to solve this problem is using a summary for providing an overview of the data source schema that highlights its most important portion. This solution can be of particular interest for a user which is querying a complex data source schema since he/she can use a summary of the schema that contains the main information he/she needs. Furthermore, in data integration systems, a summary of a data source schema might be useful to understand its content and to improve some automated integration tasks such as schema matching. For instance, in a Peer Data Management System (PDMS) [Pires et al. 2010], each peer is an autonomous data source that provides a semantic representation of its local schema (an ontology). In order to semantically organize these peers in the overlay network, several comparisons between peer schemas are required which can be costly for the overall process. Therefore, the use of schema summaries in the comparisons instead of the whole schemas can contribute to reduce the cost of the schema matching process. Summarizing schemas has recently been the focus of some research works [Yang et al. 2011].

Ontology understanding is an important process in ontology engineering to support tasks such as ontology selection and reuse [Li and Motta 2010]. This process is a prerequisite in areas such as Linked Data [Poggi et al. 2008], Ontology Learning [Maedche and Staab 2004], and Ontology Selection [Park et al. 2011]. Ontology understanding has been supported by some user-centric technologies such as the ontology visualization [Katifori et al. 2007] and navigation tools [Motta et al. 2011]. The amount, scale, and complexity of ontologies are increasing making difficult its comprehension by human beings. This shows that the tools for ontology visualization and navigation must apply ways to filter large ontologies. Ontology summarization can be a solution to decrease the amount of displayed information, filtering the most important parts of an ontology.

The goal of this work is to propose a method to summarize ontologies through user-defined parameters, in which a user can establish restrictions to produce an ontology summary. We assume that ontologies are written in RDF or OWL and they represent a data source schema or describe a knowledge domain. These kinds of ontologies can be structured in a graph. In our work, an ontology summary corresponds to a subontology of the input ontology under a specified size. The ontology summary contains the most important concepts of the ontology or the most relevant concepts for a user, respecting the original relationships and properties of the input ontology.

The main contributions of this work are: (i) a method for building personalized ontology summaries, (ii) a formula to assess the relevance of concepts combining the centrality and closeness measures, (iii) an algorithm for building a subgraph that contains the most relevant vertexes, and (iv) an evaluation of the proposed summarization algorithm.

The article is organized as follows: Section 2 presents an overview of the proposed ontology summarization method. Section 3 describes the measures used to determine the relevance of concepts in an ontology. Section 4 presents the method for building ontology summaries. Section 5 shows the implementation. Section 6 shows the experimental evaluation. Section 7 presents the related work. Finally, Section 8 presents our conclusions and suggestions for future research.

2. ONTOLOGY SUMMARIZATION

According to [Das and Martins 2007], the definition of "summary" in natural language processing presents the following features: (i) produce a summary from a single document or multiple documents, (ii) preserve important information, keeping the sense, and (iii) generate a small text, no more than half of the original text. In the context of ontology development, the second feature is the main one that differentiates the ontology summarization technique from the others applied to ontologies. The other techniques that also aim to reduce the size or complexity of an ontology do not preserve the most "important" information. Among such techniques we can include ontology partitioning and ontology

modularization. Ontology partitioning divides a large ontology into several subontologies that covers each subtopic of the original ontology [Stuckenschmidt and Klein 2004]. On the other hand, ontology modularization is used to simplify the reuse of small portions that correspond to certain aspects of the original ontology [D'Aquin et al. 2009]. Differently, ontology summarization is defined as "the process of distilling knowledge from an ontology to produce an abridged version for a particular user (or users) and task (or tasks)" [Zhang et al. 2007].

2.1 Scenarios for Ontology Summarization

A typical scenario in which ontology summarization can be applied arises when a user tries to use a semantic search engine. For instance, when searching for an ontology in SWoogle, the system could provide an ontology summary containing only the relevant concepts which are useful for user needs [Sabou et al. 2006]. Another scenario refers to a data sharing system based on semantic knowledge that utilizes ontologies to represent an autonomous data source for organizing and searching data [Pires et al. 2010]. As an example, consider a Semantic PDMS whose architecture and entry points on the overlay network are based on ontologies which promotes semantic search in autonomous data source.

2.2 Generic Method for Ontology Summarization

Figure 1 illustrates an overview of the proposed ontology summarization method which consists of: given an input ontology O to generate a summarized version, denoted as ontology summary OS. Firstly, the relevance of concepts is calculated, based on the parameters and measures that have been defined by the user. A hierarchy of concepts is formed according to their importance in the ontology O (represented by gray shades). The next step is the generation of OS, which corresponds to a subontology of O, concentrating the concepts with the highest relevance and respecting a specified summary size. As the most relevant concepts can be non-adjacent in O, it is possible that less important concepts (lighter gray shades) can be introduced in OS. Such concepts are necessary to maintain the integrity and preserve the relationships between the most relevant concepts of the original ontology. Therefore, OS corresponds to the subontology containing the most relevant concepts properly interconnected, avoiding any user intervention.

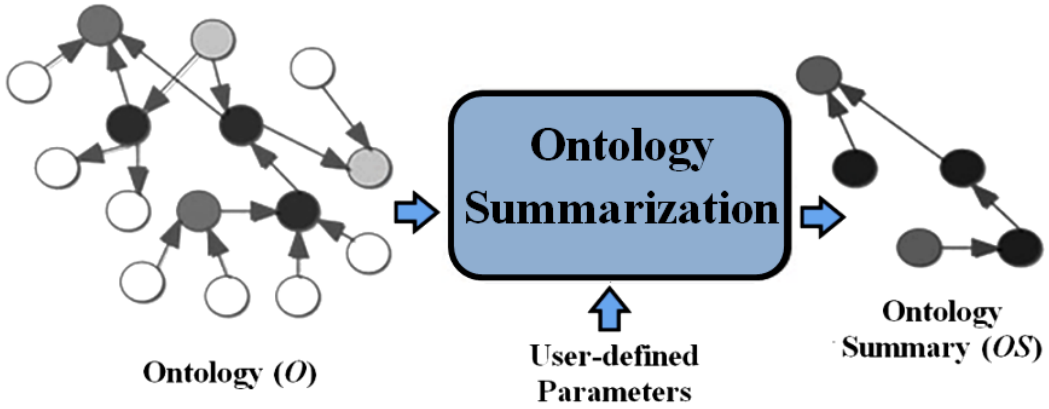

Fig. 1. A generic ontology summarization method

In the method, the ontology O is modeled as a connected directed labeled graph $O = (C, R)$, where $C = (c_1, ..., c_n)$ is a finite set of vertexes (concepts) and $R = (r_1, ..., r_n)$ is a finite set of directed edges (relationships between concepts). A relationship $r_k \in R$ represents a directed relationship between two adjacent concepts c_i and $c_j \in$ C; i.e., $r_k = (c_i \times c_j)$. Two concepts c_i, $c_j \in C$ are adjacent in O if

$\exists\ r_k \in R\ /\ r_k = (c_i \times c_j)$ or $r_k = (c_j \times c_i)$. A directed labeled edge is defined from c_i to c_j if c_i is a direct subconcept of c_j. Similarly, if c_i is a domain concept and c_j its range concept then a directed labeled edge is added from c_i to c_j. Likewise, an ontology summary OS is a subgraph of O such that $OS \subset O$. Formally, $OS = (CS, RS)$, where $RS \subset R$ and $CS \subset C$.

3. RELEVANCE MEASURES

The evaluation of relevance measures is the key to ontology summarization since these measures are responsible to define the relevance of concepts in ontologies. These measures encouraged researches in different areas, such as graph theory and network analysis, to find ways to evaluate concepts in ontologies. In graph theory and network analysis, the most used centrality measures include: Degree, Betweenness, Eigenvector, and Closeness. These measures determine the relative importance of a vertex within a graph, calculating the centrality value of the vertex [Newman 2010]. In the work of [Zhang et al. 2007], a comparative study was conducted with variations of these measures applied in ontology summarization. A description of the measures is provided in Table I.

Table I. A description of centrality measures by [Zhang et al. 2007]

Measure	Description
weighted in-degree (CI)	Based on the Degree Centrality, considers the number of relations of a vertex.
betweenness centrality (CB)	Considers the number of paths passing through a vertex.
weighed PageRank (CP) weighted HITS (CH) focused weighted PageRank (CF)	Based on the Eigenvector Centrality which verifies the importance of a vertex on the basis of its relationships with other vertexes.

The analysis performed by [Zhang et al. 2007] comparing the five different centrality measures showed that: the weighted in-degree centrality had the best result and the measures based on eigenvector centralities had a reasonable result. Therewith, we propose two measures to calculate the relevance of a concept C_n in an ontology O: one inspired in Degree Centrality with weights assigned according to the type of relationships and another one based on Closeness Centrality, which was not evaluated by [Zhang et al. 2007]. These proposed measures will be detailed in the following.

3.1 Centrality Measure

The centrality measure defined by [Pires et al. 2010], based on the Degree Centrality measure, considers the number of relationships between ontology concepts and the types of relationships between them. The types of relationships considered by the work are: *standard* relationships such as *is-a*, *part-of*, and *same-as*; and *user-defined* relationships such as *HasItems* and *authorOf*. Another study using a variation of the Degree Centrality measure was the weighted in-degree proposed by [Zhang et al. 2007], which considers the direction of the relationship to count the input relations of a concept. In this work, we considered the best features of the two measures, using weights to configure the relationship types. The normalized formula defined to measure the centrality of a concept C_n is defined as:

$$C_I(i) = \sum_{(j,i)\in C} r(j,i) \qquad C_O(i) = \sum_{(i,j)\in C} r(i,j)$$

$$Centrality(C_n) = \frac{(W_I \times C_I + W_O \times C_O) \times \left(\frac{n_s \times w_s}{max_s} + \frac{n_{ud} \times w_{ud}}{max_{ud}}\right)}{|C| - 1}$$

where C_I and C_O represent the number of different concepts that maintain input and output relationships with a concept C_n, respectively, and r represents the relationships between concepts. W_I and W_O are, respectively, the user-defined weights of C_O and C_I. n_s and n_{ud} are, respectively, the number of standard/user-defined relationships maintained by C_n. Note that, if C_n keeps more than one relationship with another concept, it is counted once. w_s and w_{ud} are, respectively, the weights of standard/user-defined relationships. max_s and max_{ud} represent the maximum number of standard/user-defined relationships held for a certain concept in the ontology O, respectively. In addition, (i) *Centrality*(C_n) $\in$ [0,1]; (ii) $w_s + w_{ud} = 1$; (iii) $W_I + W_O = 1$.

3.2 Closeness Measure

The Closeness Measure gives emphasis on the concepts that are close to the most relevant concepts, i.e., the measure of the concept C_n is directly proportional to the number of relevant concepts that are close to C_n. The closeness measure requires that the concepts already have a relevance value previously determined. The purpose of this measure is to distinguish the concepts that have the same relevance, by considering the distance for the relevant concepts. Moreover, this measure highlights the concepts which have more chance to be related with important concepts, facilitating the interrelation of them. The formula to provide a value-weighted by distance relationship and relevance of a concept C_n is defined as:

$$Closeness(C_n) = \frac{\sum_{n \in (c - c_n)} \frac{relevance(\mathrm{n})}{distance(C_n,\ n)}}{\sum_{n \in (c - c_n)} 1/distance(C_n,\ n)}$$

In other words, closeness is a weighted average formed by the sum of multiplications of the concepts' relevance by its corresponding weight, which is represented by the inverse of the distance from n to C_n, divided by the sum of these weights. n is a variable that can assume all the concepts of C minus the concept C_n in the ontology O. *Relevance*(n) is the relevance of the current concept on n; *distance*(C_n, n) is the shortest number of relationships between C_n and n. Additionally, *Closeness*(C_n) $\in$ [0,1].

4. BUILDING AN ONTOLOGY SUMMARY

Two tasks are needed to build an ontology summary: identify the key concepts and select them in order to produce a subontology of the original ontology. The first task is accomplished using relevance measures and user-defined parameters whilst the second one is performed by the *Broaden Relevant Paths*(BRP) algorithm, proposed to identify the best path (ontology summary) in a graph (ontology) that represents a set of interrelated vertexes (concepts). In this section, we explain the BRP algorithm and the ontology summarization method.

4.1 Broaden Relevant Paths Algorithm

Broaden Relevant Paths(BRP) is an algorithm to find a path that includes the vertexes of greatest relevance within a graph. To start the algorithm it is necessary to assemble a structure with three lists to manipulate the vertexes: *PathSet*, *NodeSet* and *AdjacentNodes*. The *PathSet* list is used to record the best paths generated by the algorithm ordered according to the quality of the path. In this article, we define two metrics to measure the quality of a path. The first metric is *Relevance Coverage*(RC) that evaluates the proportion of the sum of vertexes' relevance within the path T by the sum of relevance of the vertexes within the original graph O. This metric is defined as:

$$RC(T) = \frac{\sum_{i \subset T}^{n} relevance(\mathrm{i})}{\sum_{i \subset O}^{n} relevance(\mathrm{i})}$$

The other metric is *Relevance Degree*(RD) that assesses the relevance average within the path T by the higher value of relevance in the graph O, explained in the following expression.

$$RD(T) = \frac{averageRelevance(T)}{maxRelevance(\mathrm{O})}$$

A combination of RC and RD using the f-measure formula [Baeza-Yates and Ribeiro-Neto 1999] is described next. We assume that the parameter α is adjusted by the user. The formula is used to order the paths in the *PathSet* list.

$$f\text{ - }measure(T) = \frac{RD(T) \cdot RC(T)}{(1-\alpha) \cdot RD(T) + \alpha \cdot RC(T)}$$

Now, we present how the vertexes are ordered in the *NodeSet* and *AdjacentNodes* lists according to their relevance value. In the *NodeSet* list, the vertexes are ordered by their relevance values. The *AdjacentNodes* list includes the vertexes that maintain relationships to the vertexes of paths contained in the *PathSet* list. The formula used for ordering the vertexes in the *AdjacentNodes* list is based on: a) the number of relationships among a vertex V_n and the paths contained in the *PathSet* list, b) the sum of the relevance values of vertexes in the path T over the number of vertexes contained in the *PathSet* list, and c) the relevance of vertex V_n. This function called *Relation Relevance*(RR) is represented as follows:

$$RR(V_n) = \mathrm{No.}\ Path(V_n) + relevance(V_n) + \sum_{T=1}^{No.Path(V_n)} \frac{sumRelevance(T)}{sumSizePaths}$$

where V_n is a vertex in the *AdjacentNodes* list and T represents a path in the *PathSet* list. No. $Path(V_n)$ is the number of relationships between the vertex V_n and the vertexes in the *PathSet* list. $sumRelevance(T)$ is the sum of the relevance of the vertexes contained in a path T, and relevance(V_n) is the relevance of the vertex V_n. After defining the functions for the ordered list we present the main steps of the BRP algorithm.

4.2 Ontology Summarization Method

The proposed ontology summarization method is composed of parameters, relevance metrics, and the algorithm to build ontology summaries. The main steps of this summarization method are: (1) select the parameters and the ontology, (2) calculate the relevance of the concepts in the ontology, (3) start the BRP algorithm, (4) search for relevant concepts to form paths, (5) add selected concepts to a path in the *PathSet* list, and (6) verify the paths in the *PathSet* list that satisfy the stop conditions. For a better understanding of the ontology summarization method a corresponding flowchart is illustrated in Figure 2.

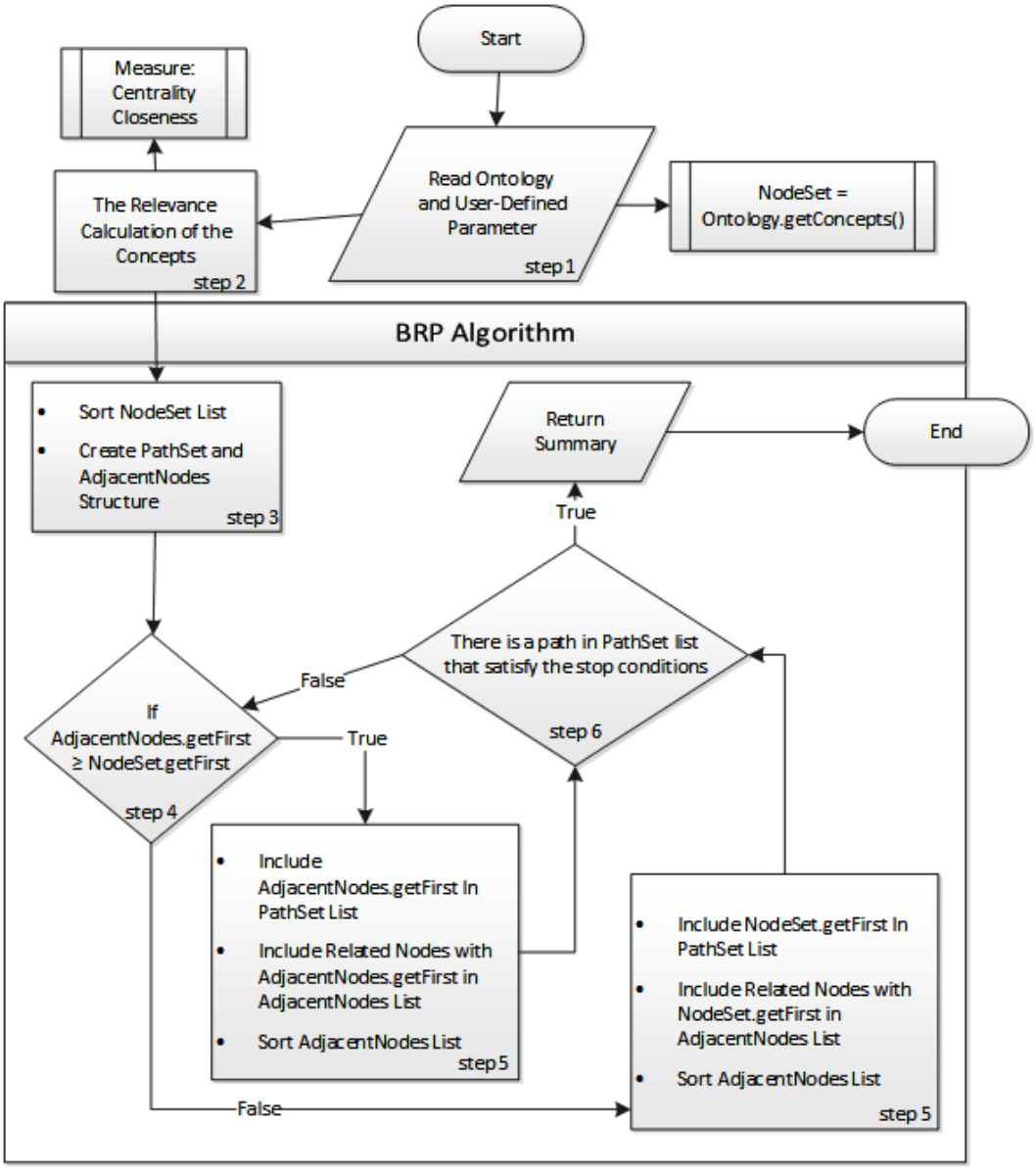

Fig. 2. Flowchart of the ontology summarization method

1) **Select the parameters and the ontology**: in the selection of parameters, the ontology summarization method can be configured to work in two manners:

i. Automatically: considers only the relevance measures to determine the relevant concepts. It can produce an ontology summary with the size determined by the user or automatically. A parameter is used to define the percentage of the ontology's size or a threshold of relevance, e.g. consider the concepts with a relevance value greater than the average of relevant concepts in the ontology;

ii. Semi-automatically: considers the user's opinion and relevance measures to calculate the relevance of the concepts. The user can suggest concepts which are important for him/her and that should be included in the ontology summary. The relevance measures are used to identify the relevant concepts that are close to the concepts chosen by the user. It enables to identify the important concepts, according to the user's opinion, in order to produce a personalized ontology summary.

These two ways of working are defined based on the user-defined parameters. To start the ontology summarization method it is necessary to receive as input an ontology O and a set of parameters. The main configuration parameters are: address of the ontology to be summarized, address to save the ontology summary, ontology summary size, the weights to determine the importance of each centrality metric (centrality and closeness) for the calculation of concept relevance, the weight for *f-measure* to define the quality of the path, the threshold of relevance (it defines the concepts that can take part in the summary because their relevancy value is greater than the threshold), and the list of mandatory concepts that enables the user to inform specific concepts that are important to his/her application.

2) **Calculate the relevance of the concepts in the ontology**: after receiving the configuration parameters, the load of the ontology is performed. We use a directed graph notation with connections labeled to represent an ontology and apply metrics to define the closeness and centrality of the concepts. Each concept C_n that is included in the list of mandatory concepts will have a maximum value of relevance. Thus, the concepts near to the concept C_n will also have an increase of relevance in the closeness metric. After calculating each metric, the following formula is used to determine the relevance of a particular concept C_n in an ontology O:

$$relevance(C_n) = \beta \cdot centrality(C_n) \; + \; \alpha \cdot closeness(C_n)$$

where $relevance(C_n) \in [0,1]$ and the weights $\beta \; + \; \alpha = 1$.

3) **Start the BRP algorithm**: in this step, the Broaden Relevant Paths (BRP) algorithm is started taking as input the ontology O. In the BRP initialization, the concepts are represented by vertexes and the relationships by edges, the *AdjacentNodes* and *PathSet* lists are created, and the *NodeSet* list contains the concepts ordered by relevance.

4) **Search for relevant concepts to form paths**: the BRP algorithm continues with the execution of a loop to find a path that meets the specifications of the parameters and contains the most relevant concepts. In this loop, each cycle consists of choosing a concept C_r that will be added to the *PathSet* list. To this end, it is necessary to compare the first concept C_a in the *AdjacentNodes* list with the first concept C_f in the *NodeSet* list. If the relevance of C_a is equal or higher than the relevance of C_f, then C_a will be the concept C_r chosen to be included in the *PathSet* list. Otherwise, the concept C_f, which is the most relevant concept in the *NodeSet* list in the current moment, will be the concept C_r chosen to be included in the *PathSet* list.

5) **Add selected concepts to a path in the *PathSet* list**: in this step, after choosing the concept C_r, it is checked if there are relations of C_r with the vertexes in the paths contained in the *PathSet* list. If there are connections among them, C_r will be linked to all the concepts contained in the paths which have relations with C_r, forming a greater path with the connected concepts. Otherwise, a new path is created including only the concept C_r in the *PathSet* list. After the inclusion of C_r in the *PathSet* list, the insertion in the *AdjacentNodes* list is performed including all the concepts that are connected to C_r.

6) **Check the paths in the *PathSet* list that satisfy the stop conditions**: this last step checks if there is a summary in the *PathSet* list that has the greatest value of *f-measure* and satisfies the established user specifications. If it is true, the method is finalized and the path is returned as the ontology summary. Otherwise, the loop continues and the execution goes back to step 4.

4.3 An Example of the Ontology Summarization Method with the BRP Algorithm

For a better understanding of the BRP algorithm, Figure 3 illustrates a step-by-step example. The goal is to form an ontology summary that contains 4 concepts. Some issues must be considered: the directed graph represents an ontology with 28 concepts, the vertexes are the concepts, the edges are relationships and properties, and the produced path corresponds to an ontology summary. The steps followed by the BRP algorithm are equivalent to the ontology summarization method. In the illustration, the vertexes are identified by the concatenation of a number and a character. The number represents the relevance of the concept while the character is used to differentiate vertexes with the same relevance. The white color is used to indicate the vertexes that are contained in the *NodeSet* list, red refers to vertexes that are contained in the *PathSet* list, and yellow shades refer to vertexes that are contained in the *AdjacentNodes* list. A shade of dark yellow is used to distinguish vertexes that have more than one relationship with paths in the *PathSet* list. The illustration steps are detailed in the following.

In step 1, the ordering of the *NodeSet* list is executed. In steps 2, 3 and 4, the respective vertexes {7A, 5B, 5A} are selected and added to the *PathSet* list, comparing the relevance value of the first vertex V_S in the *NodeSet* list with the first vertex V_A in the *AdjacentNodes* list. In these steps, the vertex V_S {7A, 5B, 5A} is selected because the *AdjacentNodes* list is empty or its relevance value is higher than V_A {_, 3B, 4A}. The selected vertex V_S is added to the *PathSet* list. Since the vertex V_S does not have relationships with any path in the *PathSet* list, a path is created, containing only V_S. Afterwards, all the vertexes that have a relationship with the vertex V_S are added to the *AdjacentNodes* list. The list is then ordered by the *Relation Relevance*(RR) function.

In step 5, a vertex is selected comparing the relevance value of the vertexes V_S and V_A. The vertex V_S(4A) is selected because its relevance value is higher than V_A(3A). However, since V_S has a relationship with a path in the *PathSet* list, V_S is added to the corresponding path in the *PathSet* list. As a result, the path size is broadened with a relevant vertex. Afterwards, all the vertexes that have a relationship with the vertex V_S are added to the *AdjacentNodes* list.

In step 6, a vertex is selected comparing the relevance value of the vertexes V_S and V_A. The vertex V_S(3D) is not chosen because its relevance values is not higher than V_A(3F). Thus, V_A(3F) is added to the *PathSet* list. Since V_A(3F) has relationships with more than one path in the *PathSet* list, such as {7A} and {5B, 4A}, V_A(3F) joins all the paths, which have relations with it, contained in the *PathSet* list. This creates only one broader path that contains the interrelated vertexes {7A, 3F, 4A, 5B}, depicted in red color. Finally, a path including the 4 most relevant vertexes is generated.

In the BRP algorithm, it is possible to include different stop conditions. In our illustrated scenario, the algorithm stops when a path of size 4 is produced. Another stop condition is to stop when the path contains the concepts previously indicated by the user. This makes the algorithm general and flexible enough to be used in other applications that need to find a relevant path in a graph.

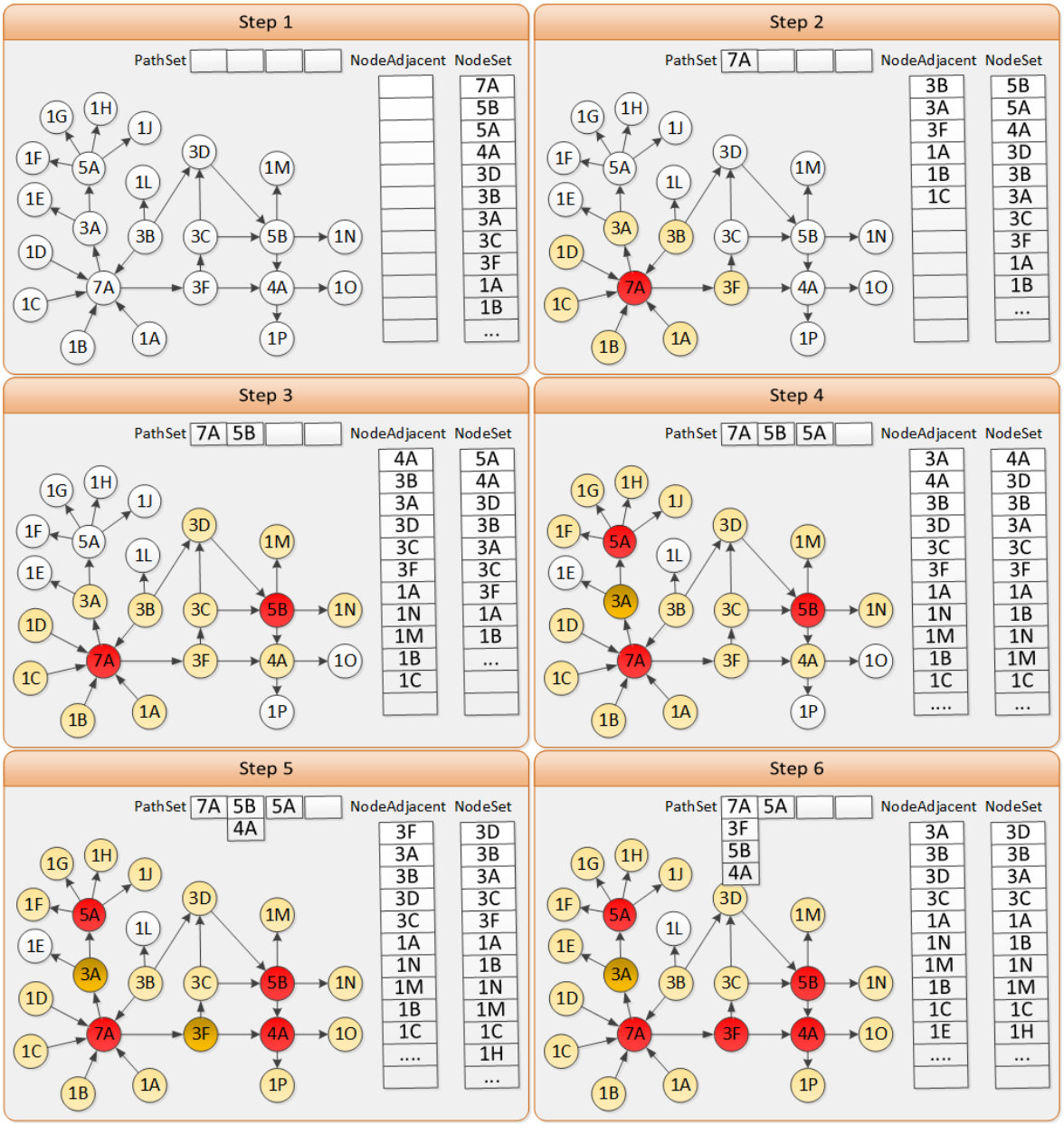

Fig. 3. A step-by-step example of the BRP algorithm

5. IMPLEMENTATION

We have developed an ontology summarization tool to produce automatic summaries of OWL and RDF ontologies. Our tool called OWLSumBRP was implemented in Java and uses the OWL API[1] to manipulate ontologies. A current version of the tool is available for download at a website[2]. The application offers a graphical interface developed using OWL2Prefuse[3]. The whole graph illustrated

[1] http://owlapi.sourceforge.net/
[2] http://www.cin.ufpe.br/ povqs/OWLSumBRP
[3] http://owl2prefuse.sourceforge.net/index.php

in Figure 4 represents the biosphere ontology[4] (87 concepts) whilst the yellow nodes are the relevant concepts and, therefore, form the biosphere summary (20 concepts). Note that all yellow nodes are interrelated in the biosphere ontology, enabling only the related concepts to participate in the biosphere summary. Figure 5 illustrates the generated biosphere ontology summary. It contains the same yellow concepts presented in Figure 4.

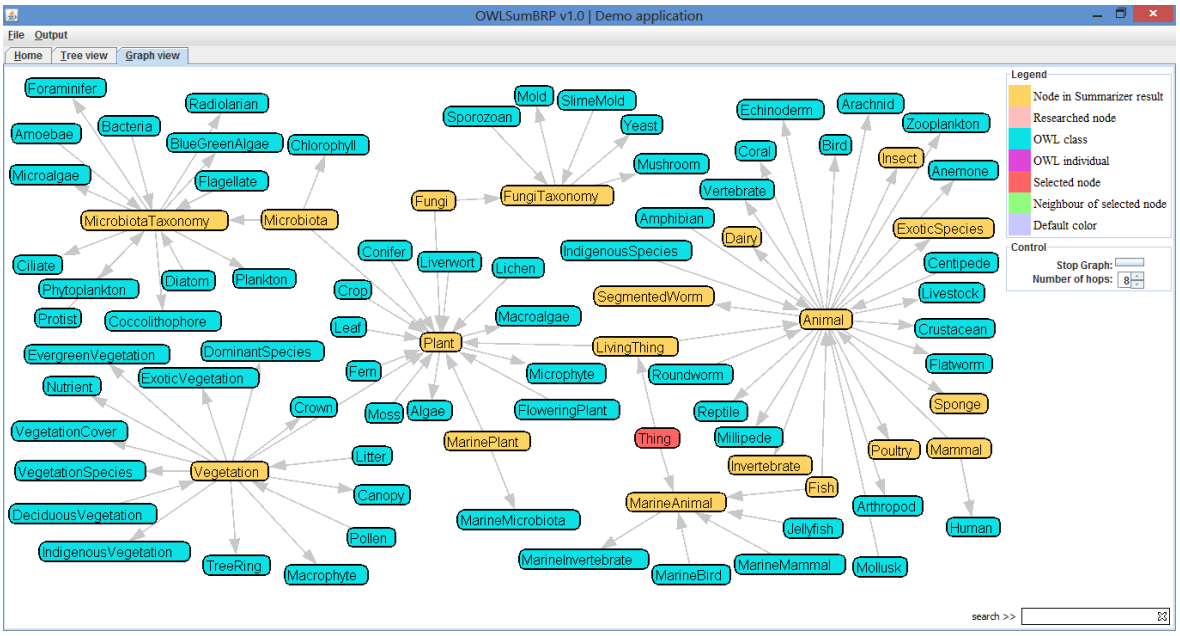

Fig. 4. A graphical visualization of the biosphere ontology in OWLSumBRP

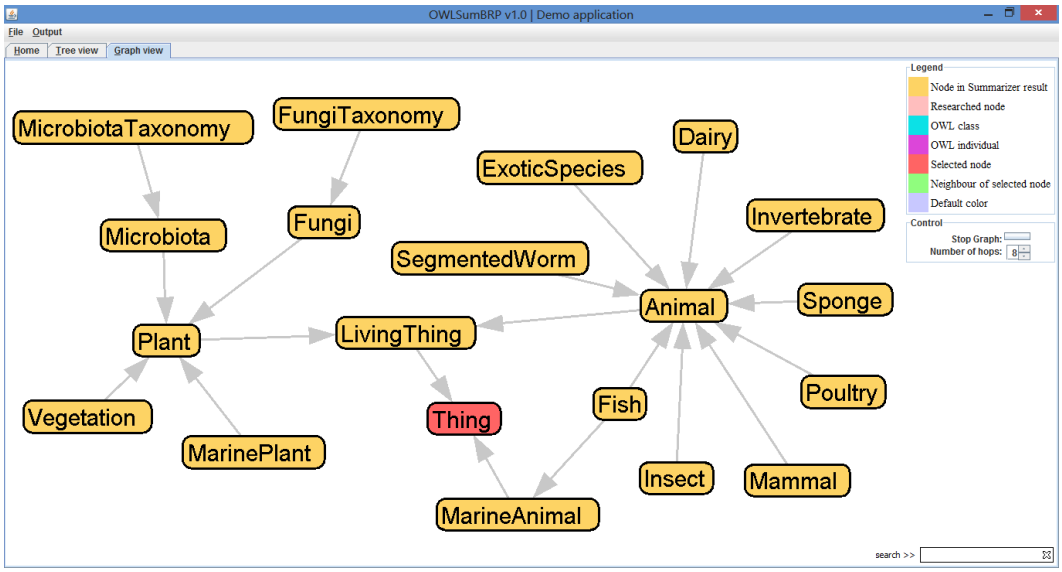

Fig. 5. A graphical visualization of the biosphere ontology summary in OWLSumBRP

[4]http://sweet.jpl.nasa.gov/ontology/biosphere.owl

6. EXPERIMENTAL EVALUATION

In this section, we present an evaluation of the ontology summarization method using the OWL-SumBRP algorithm. During the evaluation we used the same ontologies employed in the experiment performed by [Peroni et al. 2008][Li and Motta 2010]. The experiments performed by [Peroni et al. 2008] used the biosphere, financial[5], and aktors portal[6] ontologies. The experiments performed by [Li and Motta 2010] used the biosphere and financial ontologies. In the work of [Peroni et al. 2008], a method with topological and lexical measures is defined for automatically identifying the key concepts in an ontology. [Li and Motta 2010] propose a similar method with additional measures based on naming of concepts and semantic web search. Our evaluation compares the similarity degree between the ontology summaries produced by the proposed method against the summaries generated by eight human experts with good experience in ontology engineering [Peroni et al. 2008]. The experts were requested to extract 20 key concepts considered as the most representatives of each ontology. The key concepts formed a "gold standard" that included the concepts that agreed in more than 50% by the experts. The concepts are listed in Table II.

Table II. The concepts shared by more than 50% of the experts [Peroni et al. 2008]

Ontology	Number of concepts	Key concepts shared by the experts
biosphere	87	{Animal, Bird, Fungi, Insect, Mammal, MarineAnimal, Microbiota, Plant, Reptile, Vegetation}
financial	188	{Bank, Bond, Broker, Capital, Contract, Dealer, Financial_Market, Order, Stock}
aktors portal	247	{Computing-Technology, Geopolitical-Entity, Event, Organization, Person, Publication, Publication-Reference, Software-Technology}

We compared the "gold standard" against our automatic ontology summaries of size 20 (concepts). The produced summaries and the comparisons between them and the experts' choices are shown in Table III.

Table III. The ontology summaries produced automatically by our method

Ontology	Automatic ontology summaries	% matches with experts' choices
biosphere	**Animal**, Dairy, **Insect**, Poultry, Sponge, **Mammal**, Invertebrate, ExoticSpecies, SegmentedWorm, Fish, **MarineAnimal**, LivingThing, **Plant**, MarinePlant, **Microbiota**, MicrobiotaTaxonomy, **Fungi**, FungiTaxonomy, **Vegetation**, **Reptile**	90
financial	**financial_market**, market, **dealer**, organization, financial_agent, agent, market_agent, supplier, thing, **order**, option_contract, **contract**, financial_instrument, card, security, **stock**, **bond**, tax_exempt_bond, corporate_bond, municipal_bond, government_bond	66.67
aktors portal	Information-Transfer-Event, Intangible Thing, Generic-Area-Of-Interest, Method, Abstract-Information, **Publication-Reference**, Tangible Thing, Information-Bearing-Object, **Person**, Affiliated-Person, Employee, Temporal Thing, **Event**, Generic Agent, **Geopolitical-Entity**, Legal Agent, **Organization**, Address, Oraganization-Unit, Technology	62.5

[5]http://ai.uom.gr/gmarkou/Files/Owl-s/finance_th_web.owl
[6]http://www.aktors.org/ontology/portal

A comparative analysis of our method against the methods of [Peroni et al. 2008] and [Li and Motta 2010] is shown in Table IV. The values refer to the percentage of concept matches between each method and the experts' choices. Regarding the biosphere ontology, our method presented the best result in comparison with the other methods, achieving 90% of hit. Concerning the financial ontology, our method showed the same performance of the other methods with 66.67% of hit. Finally, regarding the aktors portal ontology, our method presented a lower result compared to the result generated by [Peroni et al. 2008] (62.5% against 75% of hit). Unlike the other two works, our method respects the relationships defined in the original ontology and ensures that all concepts are interrelated in the ontology summary. As a result, less relevant concepts can be included in the ontology summary in order to preserve the original relationships between the relevant concepts.

Table IV. A comparative result of ontology summarization methods

Ontology	% matches between [Peroni et al. 2008] with experts' choices	% matches between [Li and Motta 2010] with experts' choices	% matches between our method with experts' choices
biosphere	80	60	90
financial	66.67	66.67	66.67
aktors portal	75	n/a	62.5

In an attempt to improve this last result, a second experiment was carried out in a semi-automatic way. The concept "Computing-Technology" was defined as mandatory concept. This concept was selected to be an important concept about technology in the aktors portal ontology. In other words, we forced the concept to be included in the ontology summary. The resulting ontology summary is presented in Table V.

Table V. The ontology summary produced semi-automatically

Ontology	Semi-automatic ontology summary	% matches with experts' choices
aktors portal	Intangible-Thing, Generic-Area-Of-Interest, Information-Transfer-Event, **Publication-Reference**, Information-Bearing-Object, **Person**, Employee, Temporal-Thing, **Event**, Generic Agent, **Geopolitical-Entity**, Legal-Agent, **Organization**, Time-Interval, Country, Oraganization-Unit, Technology, **Computing-Technology**, **Software-Technology**, Implemented-System, **Publication**	100

Based on this last experiment, we can observe that setting specific concepts as mandatory can improve the results. A significant gain was obtained with the indication of only one concept. We reached a perfect match between the produced ontology summary and the one suggested by human experts. The use of this parameter gives a new perspective allowing the generation of semi-automatic ontology summaries for particular applications.

7. RELATED WORK

Various techniques have been developed for the identification of relevant concepts in ontologies. The authors of [Zhang et al. 2007] propose a method for automatic ontology summarization based on RDF Sentence Graph. Summaries are customizable, i.e., users can specify the length of summaries and navigational preferences. The notion of RDF sentence is proposed as the basic unit of summarization.

An RDF Sentence Graph is proposed to characterize the links between RDF sentences derived from a given ontology. The importance of each RDF sentence is assessed in terms of its centrality in the graph. These authors showed that weighted in-degree centrality measures and several

eigenvector-based centralities have good performance by comparing five different centrality measurements: weighted in-degree, betweenness, weighted PageRank, HITS weighted and weighted PageRank focused. In this method, an ontology is summarized by extracting a set of salient RDF sentences according to a re-ranking strategy.

In the work of [Peroni et al. 2008], the authors propose a user-independent method for automatically identifying the key concepts in an ontology. The method integrates both topological and lexical measures. The topological measures used are: density, that is based on the number of direct subconcepts, properties and instances; and coverage, that is computed on the basis of dissemination of important concepts in the ontology. The lexical measures employed are: statistical popularity, which is computed according to the number of results returned in the search by concept name on Yahoo; and natural categories, that is based on structure and names of concepts. With the use of these measures, the method has been validated empirically by human experts but does not generate a file of ontology summary.

In the article of [Li and Motta 2010], an analysis of the state-of-the-art on methods for ontology summarization is performed. The authors investigate the ontology features which are important in ontology summarization. They evaluate the key concepts through the measures: density and popularity, that had already been defined in [Peroni et al. 2008]; reference, that is computed based on the number of entities collected in the Watson semantic search engine; and name simplicity, that is based on naming of concepts. This method does not generate an ontology summary.

In the work of [Pires et al. 2010], an automatic method to summarize ontologies that represent data schemas in a PDMS system is proposed. To determine the relevance of concepts a combination of two measures is used. Centrality is calculated using the number of relationships between the concepts. Frequency is used as a distinguishing criterion when the ontologies to be summarized are merged ontologies. A detailed description of the summarization method is presented as well as an algorithm that generates summaries with all the interrelated concepts. The algorithm is validated using classical Information Retrieval metrics.

Some notion of centrality is used to calculate the relevance of concepts in all the discussed works. However, none of them explore the proximity of relevant concepts to others, as an evaluation measure, such as closeness measure. Another aspect that differentiates our method from the others is that the generated summaries can optionally include concepts indicated by the user. By combining the closeness measure with user opinion it is possible to produce personalized ontology summaries according to the user needs. The relevance of the concepts can change based on the concepts chosen by the user and the closeness measure. Moreover, our ontology summarization method, differently from the ones proposed by [Pires et al. 2010] and [Zhang et al. 2007], can produce ontology summaries using different stop conditions.

8. CONCLUSIONS AND FUTURE WORK

This work proposes a method to summarize ontologies in a personalized way. The ontology summaries can be produced in two manners: automatically, using relevance measures to define the relevance of concepts, or semi-automatically, using the user's opinion (through configurable parameters) to determine the relevance of concepts. In both ways, to determine the relevance of concepts, a combination of two measures (centrality and closeness) is used. Centrality is calculated using an extended definition of the degree centrality measure, considering weights to configure the value assigned to the types of relationships. Closeness is calculated based on the distance of a concept to the others and on its respective relevance. A detailed description of the summarization method is presented as well as a demonstration of the corresponding algorithm to join the relevant vertexes in a graph. The experiments have shown that the proposed method is able to find good summaries compared to the ones manually generated by human experts and with other proposed ontology summarization methods.

The analysis of the experiments showed satisfactory results reaching an average greater than 62.5% of hit in the automatic production of ontology summary while the semi-automatic way presented 100% of coverage with respect to the concepts chosen by human experts.

In future work, we intend to introduce new measures based on lexical statistics considering the name of the concepts and the results of web searches. We will exploit the different stop conditions of the BRP algorithm. We will perform new experiments with real-world ontologies considering the use of domain specialists to identify the gold standard. Another activity is to apply our method in real-world applications, such as ontology reuse and ontology engineering. We will exploit our method in ontology reuse by importing only the concepts needed for developing a new ontology. Moreover, we intend to apply the BRP algorithm in other scenarios. For instance, in a sensor network, to find a path that contains the most important nodes.

REFERENCES

Baeza-Yates, R. A. and Ribeiro-Neto, B. *Modern Information Retrieval*. ACM Press, Boston, MA, USA, 1999.

D'Aquin, M., Schlicht, A., Stuckenschmidt, H., and Sabou, M. *Modular Ontologies*. Lecture Notes in Computer Science, vol. 5445. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Das, D. and Martins, A. F. T. A Survey on Automatic Text Summarization. *Literature Survey for the Language and Statistics II Course at CMU* vol. 4, pp. 1–31, 2007.

Gagnon, M. Ontology-based integration of data sources. In *2007 10th International Conference on Information Fusion*. IEEE, Quebec, Que, pp. 1–8, 2007.

Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5 (2): 199–220, June, 1993.

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., and Giannopoulou, E. Ontology visualization methods—a survey. *ACM Computing Surveys* 39 (4): 10–48, Nov., 2007.

Li, N. and Motta, E. Evaluations of User-Driven Ontology Summarization. In *Knowledge Engineering and Management by the Masses*, P. Cimiano and H. S. Pinto (Eds.). Lecture Notes in Computer Science, vol. 6317. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 544–553, 2010.

Maedche, A. and Staab, S. Ontology Learning. In *Handbook on Ontologies*, S. Staab and R. Studer (Eds.). Springer, Karlsruhe, Germany, pp. 173–190, 2004.

Motta, E., Mulholland, P., Peroni, S., D'Aquin, M., Gomez-Perez, J. M., Mendez, V., and Zablith, F. A novel approach to visualizing and navigating ontologies. In *ISWC'11 Proceedings of the 10th international conference on The semantic web - Volume Part I*. Springer-Verlag, Berlin, Heidelberg, pp. 470–486, 2011.

Newman, M. *Networks: An Introduction*. Oxford University Press, New York, NY, USA, 2010.

Obitko, M. and Marik, V. Ontologies for multi-agent systems in manufacturing domain. In *Proceedings. 13th International Workshop on Database and Expert Systems Applications*. IEEE Comput. Soc, Prague, Czech Republic, pp. 597–602, 2002.

Park, J., Oh, S., and Ahn, J. Ontology selection ranking model for knowledge reuse. *Expert Systems with Applications* 38 (5): 5133–5144, May, 2011.

Peroni, S., Motta, E., and Aquin, M. Identifying Key Concepts in an Ontology, through the Integration of Cognitive Principles with Statistical and Topological Measures. In *The Semantic Web*, J. Domingue and C. Anutariya (Eds.). Lecture Notes in Computer Science, vol. 5367. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 242–256, 2008.

Pires, C. E., Sousa, P., Kedad, Z., and Salgado, A. C. Summarizing ontology-based schemas in PDMS. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*. IEEE, Long Beach, CA, pp. 239–244, 2010.

Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., and Rosati, R. Linking data to ontologies. In *Journal on data semantics X*, S. Spaccapietra (Ed.). Springer-Verlag, Berlin, Heidelberg, pp. 133–173, 2008.

Sabou, M., Lopez, V., Motta, E., and Uren, V. Ontology selection: ontology evaluation on the real Semantic Web, 2006.

Stuckenschmidt, H. and Klein, M. Structure-Based Partitioning of Large Concept Hierarchies. In *The Semantic Web ISWC 2004*. Springer Berlin Heidelberg, Hiroshima, Japan, 21, pp. 289–303, 2004.

Yang, X., Procopiuc, C. M., and Srivastava, D. Summary Graphs for Relational Database Schemas. *Proceedings of the VLDB Endowment* 4 (11): 899–910, 2011.

Zhang, X., Cheng, G., and Qu, Y. Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th international conference on World Wide Web (WWW'07)*. ACM Press, New York, New York, USA, pp. 707, 2007.