

## Knowledge Summarization for Scalable Semantic Data Processing

Zaiyue ZHANG<sup>1,†</sup>, Zhisheng HUANG<sup>2,3</sup>, Xiaoru ZHANG<sup>3</sup>

<sup>1</sup> School of Computer Science and Engineering, Jiangsu University of Science and Technology, China

<sup>2</sup> Computer Science Department, Vrije Universiteit Amsterdam, The Netherlands

<sup>3</sup> School of Computer Science and Engineering, Jiangsu University of Science and Technology, China

### Abstract

Scalable semantic data processing has become a crucial issue for practical applications of the Semantic Web. In this paper, we propose an approach of scalable semantic data processing by knowledge summarization. The main idea is to express scalable semantic data on different abstraction and summarization levels to reduce their cardinalities, so that they can be processed efficiently. The notion of knowledge summarization is inspired from various techniques in granular computing and text summarization in computational linguistics. In this paper, we will present a formal framework of knowledge summarization for the Semantic Web and discuss how it can be used to improve the scalability of semantic data processing.

*Keywords:* Knowledge Engineering; Logic; Ontology; the Semantic Web

### 1. Introduction

Since the birth of the World Wide Web, tens of billions of web pages have emerged during less than two decades. Till October 2009, the indexable web contains at least 21.21 billion pages (<http://www.worldwidewebsite.com/>). This trend will continue steadily at least in coming decades. Web information has become the most important resources of human knowledge and information. That appeals for technologies which can deal with extremely large-scale data on the Web efficiently.

The Semantic Web provides a foundation for automatic processing of information resources on the World Wide Web. Ontology technology, the associated knowledge processing technology for the Semantic Web, provides a unified method of data representation and semantic processing foundations for modern information systems, because nowadays information processing systems will be inevitably linked with web information resources. Thus, scalable semantic data processing becomes a crucial issue for nowadays information processing systems.

In this paper, we will propose an approach to dealing with the scalability by knowledge summarization. The main idea is to express scalable semantic data on different abstraction and summarization levels to reduce their cardinalities, so that they can be processed efficiently. The idea of knowledge summarization is inspired by various techniques in granular reasoning and text summarization. We believe that the introduction of the approaches which are inspired by human thinking and knowledge processing capabilities will provide effective methods for scalable semantic data processing. Granular computing, which simulates

---

<sup>†</sup> Corresponding author.

Email addresses: [yzzjzy@sina.com](mailto:yzzjzy@sina.com) (Zaiyue ZHANG)

the human mind to deal with large scale data, is promising because we human beings are always able to deal effectively with scalable information and knowledge efficiently. Human beings can deal with the information in their minds with different abstraction and summarization levels. It would be possible to provide new technical methods to improve scalable semantic data processing.

In this paper, we consider semantic data as description-logics-based ontologies. We will propose a general framework of knowledge summarization for the description-logics-based semantic data processing. We will explore the proposed approach with variant granularities and discuss how it can be used to improve the scalability of semantic data processing.

The rest of the paper is organized as follows. Section 2 overviews the related work briefly, which includes granular computing and existing data summarization techniques. To make the paper self-contained, Section 3 provides the preliminaries of description-logics-based ontologies. Section 4 presents the general framework of knowledge summarization for the Semantic Web. Section 5 investigates how we can measure a knowledge summarization by the notion of summarization rate. Section 6 explores how a knowledge summarization can be achieved from the perspective of granularity, i.e., by different fine-grained modularizations of an ontology. Section 7 discusses the future work before concluding the paper.

## 2. Related Work

### 2.1. Granular Computing

Granular computing-related concepts first appeared in 1979 by Zadeh's article entitled "Fuzzy sets and information granularity" [13]. In that paper, Zadeh argues that the concept of information granular exists in many areas, such as the notion of decomposition and division in automata and system theory, the notion of uncertainty in Optimal Control, etc. Since the notion of information granularity was proposed, it aroused great interest of researchers. In [4], Hobbs discusses the decomposition and consolidation of granularity, as well as access methods of different sizes of granules, and proposes a model of granularity. In [9], Lin proposes the concept of granular computing in the view of a broad point. Through the study of granular computing models under binary relations (o-domain system, rough sets and belief functions), he discusses the problems of granular computing, such as granular structure, granular representations, granular applications based on neighborhood system [10]. In [12], Yao proposes the granular computing with neighborhood systems and investigates the applications of granular computing in machine learning, data analysis, data mining, rules extraction, intelligent data processing and granular logic. There have been a lot of researches in granular computing ([http://en.wikipedia.org/wiki/Granular\\_computing](http://en.wikipedia.org/wiki/Granular_computing)). Because of the page limitation we would not provide a lengthy discussion on granular computing in this paper.

Recently the work of applying granular computing to improve the scalability of reasoning in the Semantic Web has emerged. Some preliminary work on applying the idea of granular reasoning or more generally granular computing to reasoning in the Semantic Web has been investigated by the Large Knowledge Collider Project (<http://www.larkc.eu>) in [14]. Several experiments of granular reasoning with large scale semantic data have been reported in [14]. The results show that granular reasoning is a promising approach for scalable semantic data processing.

### 2.2. Data Summarization

In text retrieval, the research of automated text summarization has a long history. It is mainly divided into the following two categories: 1) Based on the notion of concept dependency representation, find the content relevance of information to produce a summarization [7]; 2) Based on the method of knowledge

classification, develop a formal method of knowledge representation to search for important conjunctive words to generate a text summarization [3].

In recent years, the method of data summarization has also been used in semantic data processing, such as a large-scale ontology reasoning techniques developed by IBM [2], and the method of ontology summarization which is developed in [11,8]. However, the existing information summarization methods still are problem specific. There exists no a general and formal framework of information summarization for the Semantic Web.

### 3. Preliminaries

An ontology typically consists of a hierarchical description of important concepts in a domain, along with descriptions of the properties of each concept, and constraints on these concepts and properties. Given a vocabulary, an ontology can be viewed as a set of formula, alternatively called axioms. In this paper, we consider an ontology as a set of axioms which are specified in Description-Logics (DLs)-based ontologies. Description Logics are a family of class-based (concept-based) knowledge representation formalisms, equipped with well-defined model-theoretic semantics [1].

The Web Ontology language OWL has become a Web standard for ontology specification (<http://www.w3.org/TR/owl-guide/>)[5]. The Description Logic  $\mathcal{SHOIQ}(D)$  is the semantic counterpart of the ontology language OWL-DL.

Let  $\mathcal{K}$  be a Description Logic,  $C, D$   $\mathcal{K}$ -concepts,  $R, S$   $\mathcal{K}$ -roles, and  $a, b$  individuals. An interpretation (written as  $\mathcal{I}$ ) of an ontology consists of a domain  $\Delta^{\mathcal{I}}$  (a nonempty set), and an interpretation function (written as  $\cdot^{\mathcal{I}}$ ), which maps each individual name  $a$  to an element  $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ , each concept name  $CN$  to a subset  $CN^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$  of the domain and each role name  $RN$  to a binary relation  $RN^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ .

The interpretation function can be extended to give semantics to  $\mathcal{K}$ -concepts and  $\mathcal{K}$ -roles, which are concepts and role descriptions, built by  $\mathcal{K}$ -constructors. Example concept constructors of  $\mathcal{SHOIQ}(D^+)$  are  $\neg C, C \sqcap D, C \sqcup D, \exists R.C, \forall R.C, \geq nR, \leq nR$  and  $\{a\}$ , where  $n$  is a natural number. A  $\mathcal{K}$ -ontology (or simply ontology)  $O$  is a finite set of axioms of the following forms (For the direct model-theoretic semantics of  $\mathcal{SHOIQ}(D)$  we refer the reader to [6]): concept inclusion axioms  $C \sqsubseteq D$ , stating that the concept  $C$  is a subconcept of the concept  $D$ , transitivity (abstract) role axioms  $\text{Trans}(R)$ , role inclusion axioms  $R \sqsubseteq S$  and  $T \sqsubseteq U$ , concept assertions  $C(a)$ , role assertions  $R(a, b)$  and individual (in)equalities  $a \approx b$  ( $a \not\approx b$  respectively). In an ontology, we use  $Tbox(RBox, ABox)$  to refer to the set of concept (role, individual, respectively) axioms.

Given an axiom  $\varphi$ , an ontology  $O$  entails  $\varphi$ , written as  $O \models \varphi$ , iff, for all interpretations  $I$  of  $O$ , we have  $I$  satisfies  $\varphi$ . An ontology  $O_1$  entails an ontology  $O_2$ , written as  $O_1 \models O_2$ , iff, for all interpretations  $I$  of  $O_1$ , we have  $I$  satisfies all of the axioms  $\varphi \in O_2$ . For description-logic-based ontologies, in this paper, we will focus on their concept inclusion axioms (i.e., T-box), role inclusion axioms (i.e., R-box), and their concept assertions and role assertions (i.e., A-box). Without loss of generality, in this paper, we define an ontology formally as follows:

**Definition 1.** An ontology  $O$  is defined as a tuple as follows:  $O = \langle \mathcal{C}, \mathcal{P}, \mathcal{I}, \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$ , Where  $\mathcal{C}$  is the set of all concepts of the ontology,  $\mathcal{P}$  is the set of all roles of the ontology, and  $\mathcal{I}$  is the set of all individuals of the ontology,  $\mathcal{T}$  is the T-box which is a set of concept inclusion axioms such as  $C \sqsubseteq D$ ,  $\mathcal{R}$  is the R-box which is a set of role inclusion axioms  $R \sqsubseteq S$ , and  $\mathcal{A}$  is the A-box which is a set of concept assertions  $C(a)$  and role assertions  $R(a, b)$ .

#### 4. A Framework of Knowledge Summarization for the Semantic Web

As discussed above, human beings can always deal with scalable information very effectively. We propose that a knowledge summarization would satisfy the following basic principles:

**Maximal coverage principle.** A summarization should cover the data maximally. Namely, if we consider a coverage as a function which maps from an original data set into another data set (i.e., a summarization of the original data set), a total function is preferred to a partial function.

**Minimal redundancy principle.** A summarization should have redundant expression minimally. Namely, a summarization with less cardinality of an original data set is preferred to a summarization with more cardinality.

**Variable adjustment principle.** That states that the cardinalities of summarization could be adjusted for various requirements.

In real life, a simple and intuitive example of the variable adjustment principle is that we can summarize a lengthy television series as an outline of the story, or a paragraph, or even simply a word, based on different requirements. Similarly, in knowledge management, we believe that an important means for solving scalable semantic data reasoning is using granular computing to explore theories and techniques of knowledge summarization, to summarize scalable knowledge as knowledge expression at different abstraction levels, so that they can be expressed, manipulated and reasoned for different needs. In the following, we will propose a formal framework of knowledge summarization for scalable semantic data processing. As we have argued above, in this paper, we consider semantic data as description-logics-based ontologies which consist of a T-box which states the concept hierarchy of an ontology, an R-box which states the role hierarchy of an ontology, and an A-box which is an assertive part of an ontology. In the following, we will use terminologies, knowledge, ontology, and data set interchangeably.

Considering the basic principles of a knowledge summarization discussed above, in the following we will propose several formal definitions about ontology summarization which are developed based on the basic notions of description logics. Namely a summary of an ontology can be considered as an ontology which preserves the logical features which have been entailed by the original ontology, however, with less cardinalities of the expressions.

Let  $O = \langle \mathcal{C}, \mathcal{P}, \mathcal{I}, \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$  and  $O' = \langle \mathcal{C}', \mathcal{P}', \mathcal{I}', \mathcal{T}', \mathcal{R}', \mathcal{A}' \rangle$  be two ontologies, we are going to introduce an onto function (i.e., surjection) which maps all of the concepts in the original ontology  $O$  into a summarization ontology  $O'$  (i.e., a summary) such that the function is sufficient enough to preserve the concept hierarchy of the original ontology in the summarization ontology.

**Definition 2 (Concept Mapping).** A concept mapping  $f_C$  from an ontology  $O$  into another ontology  $O'$  is an onto mapping  $f_C : \mathcal{C} \rightarrow \mathcal{C}'$ .

Similarly we can define an onto role mapping  $f_R$  and an onto individual mapping of two ontologies  $f_I$ . Furthermore, we define a mapping  $f$  from the ontology  $O_1$  to the ontology  $O_2$  as the union of the concept mapping  $f_C$ , the role mapping  $f_R$ , and the individual mapping  $f_I$  of the two ontologies. Given a mapping  $f$  from the ontology  $O$  to the ontology  $O'$ , we extend the mapping into the composite concepts, roles and nominals as follows.

$$\begin{aligned}
f(C) &= f_C(C) && \text{if } C \in \mathcal{C}. \\
f(R) &= f_R(R) && \text{if } R \in \mathcal{P}. \\
f(a) &= f_I(a) && \text{if } a \in \mathcal{I}. \\
f(\neg C) &= \neg f(C). \\
f(C \sqcap D) &= f(C) \sqcap f(D). \\
f(C \sqcup D) &= f(C) \sqcup f(D). \\
f(\exists R.C) &= \exists f(R).f(C). \\
f(\forall R.C) &= \forall f(R).f(C). \\
f(\leq nR) &= \leq n f(R). \\
f(\{a\}) &= \{f(a)\}.
\end{aligned}$$

However, for the  $\geq$  operator, we cannot consider the function with the following property simply:  $f(\geq nR) = \geq n f(R)$ , because the function  $f$  may map two individual  $b_1$  and  $b_2$  into a single  $b$  (i.e.,  $f(b_1) = b$  and  $f(b_2) = b$ ) so that two pairs  $\langle a, b_1 \rangle$  and  $\langle a, b_2 \rangle \in \mathcal{R}$  are mapped into a single pair  $\langle f(a), b \rangle \in \mathcal{R}'$ . In that scenario, it is easy to see that the semantic condition  $\geq n f(R)$  cannot be guaranteed after the mapping.

In order to solve that problem, we define the extended mapping for the  $\geq$  operator as follows:

$$f(\geq nR) = \geq 1 f(R).$$

It is not too hard to see that if there exist at least  $n$  pairs  $\langle a, b_1 \rangle, \langle a, b_2 \rangle, \dots, \langle a, b_n \rangle \in \mathcal{R}$  where  $b_1, b_2, \dots, b_n$  are different, we can always conclude that there exists at least one pair  $\langle f(a), b \rangle \in \mathcal{R}'$  after any mapping  $f$ .

Furthermore, we extend the mapping over the concept inclusion axioms, role inclusion axioms, and the instance axioms recursively as follows:

$$\begin{aligned}
f(C \sqsubseteq D) &= f(C) \sqsubseteq f(D) && \text{where } C, D \text{ are concepts.} \\
f(R \sqsubseteq T) &= f(R) \sqsubseteq f(T) && \text{where } R, T \text{ are roles.} \\
f(C(a)) &= f(C)(f(a)) && \text{where } C \text{ is a concept,} \\
&&& \text{and } a \text{ is an individual.} \\
f(R(a, b)) &= f(R)(f(a), f(b)) && \text{where } R \text{ is a role,} \\
&&& \text{and } a, b \text{ are individuals.}
\end{aligned}$$

We can prove that the following formal properties hold for a mapping  $f$  from ontology  $O$  into ontology  $O'$ :

**Proposition 1 (Axiom Preserving).**

*i) Concept hierarchy preserving.*

For any concept  $C_1 \in \mathcal{C}$  and any concept  $C_2 \in \mathcal{C}$ , if  $O \models C_1 \sqsubseteq C_2$ , then  $O' \models f(C_1) \sqsubseteq f(C_2)$ ;

*ii) Role hierarchy preserving.*

For any role  $p_1 \in \mathcal{P}$  and any role  $p_2 \in \mathcal{P}$ , if  $O \models p_1 \sqsubseteq p_2$ , then  $O' \models f(p_1) \sqsubseteq f(p_2)$ ;

*iii) Instance preserving.*

For any individual  $a_1 \in \mathcal{I}$  and any individual  $a_2 \in \mathcal{I}$ , and any concept  $C \in \mathcal{C}$ , and any role  $p \in \mathcal{P}$ , if  $O \models C(a_1)$ , then  $O' \models f(C)(f(a_1))$  and if  $O \models p(a_1, a_2)$ , then  $O' \models f(p)(f(a_1), f(a_2))$ ;

Furthermore, we have the following properties about the cardinality of the ontologies:

**Proposition 2 (Cardinality Change).**

*i) Concept Cardinality.*  $|\mathcal{C}'| \leq |\mathcal{C}|$ .

*ii) Role Cardinality.*  $|\mathcal{P}'| \leq |\mathcal{P}|$ .

*iii) Individual Cardinality.*  $|\mathcal{I}'| \leq |\mathcal{I}|$ .

*iv) T-box Cardinality.*  $|\mathcal{T}'| \leq |\mathcal{T}|$ .

v) **R-box Cardinality.**  $|\mathcal{R}'| \leq |\mathcal{R}|$ .

vi) **A-box Cardinality.**  $|\mathcal{A}'| \leq |\mathcal{A}|$ .

vii) **Cardinality in General.**  $|\mathcal{C}' \cup \mathcal{P}' \cup \mathcal{I}'| \leq |\mathcal{C} \cup \mathcal{P} \cup \mathcal{I}|$ .

Thus, we can define a concept summary of the ontology  $O$  as an ontology  $O'$  if there exists a mapping  $f$  from  $O$  onto  $O'$  which leads to a less concept cardinality, i.e.,  $|\mathcal{C}'| < |\mathcal{C}|$ .

Similarly, we can define a role summary of the ontology  $O$  with a less role cardinality, and an individual summary of the ontology  $O$  with a less individual cardinality.

**Definition 3 (Summary).** An ontology  $O'$  is said to be a summary of another ontology  $O$  iff there exists a mapping  $f$  such that  $O'$  is a concept summary of  $O$  or  $O'$  is a role summary of  $O$  or  $O'$  is an individual summary of  $O$ .

It is straightforward to see that we have the following formal property:

**Proposition 3 (Less Cardinality Property).** If ontology  $O'$  is a summary of ontology  $O$ , then  $|\mathcal{C}'| < |\mathcal{C}|$  or  $|\mathcal{P}'| < |\mathcal{P}|$  or  $|\mathcal{I}'| < |\mathcal{I}|$ .

The mapping  $f$  above is said to be a summarization function from the ontology  $O$  into  $O'$ . Alternatively, we can say that ontology  $O'$  is a summary of ontology  $O$  with respect to the function  $f$ .

Note that the notion of the ontology summarization is different from the notion of ontology mapping which has been well studied and proposed in the Semantic Web, because of the following differences:

- Ontology summarization is to introduce a mapping such that a new ontology (i.e., a summarization ontology) can be created, whereas the traditional notion of ontology mapping concerns only a partial mapping between two existing ontologies.

- Ontology summarization would lead to a new ontology with less cardinality, whereas the traditional notion of ontology mapping has no such a requirement on the cardinality difference.

- Ontology summarization concerns the concept hierarchy preservation between two ontologies, whereas the traditional notion of ontology mapping does not necessarily require the concept hierarchy preservation.

Here are some examples of ontology summarization.

*Example 1(Trivial Summarization).* Consider a summarization function  $f$  which assigns the universal concept  $\top$  to all the concepts, namely,  $f(C) = \top$  for any  $C \in \mathcal{C}$ ,  $f(p) = p$  for any  $p \in \mathcal{P}$ , and  $f(a) = a$  for any  $a \in \mathcal{I}$ .

It is easy to see that that the mapping can preserve all of the T-box, R-box, and A-box with less cardinality. However, that summarization is trivial.

Although the example above is trivial, it states the fact that seeking for an ontology summarization does not necessarily mean that it has to check all the data in an infinitely countable data set. A simple mapping of two ontologies may be sufficient to reduce the computational cost significantly, of course with a price that a lot of information which have been implied in the original ontology may be lost in the summarization.

The following is an example of ontology summarization which does not suffer from a significant information loss, however, with a trivial reduction of the cardinality:

*Example 2(Concept Representation Summarization).* Consider a summarization function  $f$  which assigns a “representative” concept  $[C]$  to all of the logical equivalent concepts, namely, for any equivalence concept set  $[C] = \{D : O \models C \equiv D\}$ ,  $f$  assigns a new concept name  $[C]$  to any concept  $D \in [C]$ . Thus,  $f(D) = C$  if  $D \in \{D' : O \models C \equiv D'\}$  for any  $D \in \mathcal{C}$ ,  $f(p) = p$  for any  $p \in \mathcal{P}$ , and  $f(a) = a$  for any  $a \in \mathcal{I}$ .

It is easy to see that that the mapping can preserve all of the T-box, R-box, and A-box with less cardinality, if there exist some non-trivial equivalent concept sets  $[C]$ , i.e.,  $||[C]|| > 1$ .

## 5. Summarization Rate

Knowledge summarization is introduced to reduce the cardinalities of the targeted ontologies. Naturally, we are much interested in the metrics which can measure how much the cardinalities of ontologies have been reduced after the summarization. Therefore, we propose the following notions of summarization rate.

**Definition 4 (Concept Summarization rate).** *The concept summarization rate  $SR_C(O, O', f)$  is defined as the inverse ratio of the concept cardinality of the new ontology to the concept cardinality of the original ontology. Namely,  $SR_C(O, O', f) = 1 - (|C'|/|C|)$ .*

That means that the less cardinality reduction after a summarization is achieved, the higher summarization rate is. If a summarization rate is zero, then that means that it is a trivial summarization, because it does not lead to a reduced cardinality. A summarization rate which is larger than 0 is called a non-trivial summarization rate, otherwise it is called a trivial summarization rate.

We can define the role summarization rate and the individual summarization rate as follows:

**Definition 5 (Role Summarization rate).** *The role summarization rate  $SR_P(O, O', f)$  is defined as the inverse ratio of the role cardinality of the new ontology to the role cardinality of the original ontology. Namely,  $SR_P(O, O', f) = 1 - (|P'|/|P|)$ .*

**Definition 6 (Individual Summarization rate).** *The individual summarization rate  $SR_I(O, O', f)$  is defined as the inverse ratio of the individual cardinality of the new ontology to the individual cardinality of the original ontology. Namely,  $SR_I(O, O', f) = 1 - (|I'|/|I|)$ .*

However, non-trivial summarization rates above do not necessarily mean that the cardinalities of T-box, R-box, and A-box have been reduced. Thus, we need the following summarization rates which can be used to measure the cardinality differences on T-box, R-box, and A-box:

**Definition 7 (T-box Summarization rate).** *The concept summarization rate  $SR_T(O, O', f)$  is defined as the inverse ratio of the T-box cardinality of the new ontology to the T-box cardinality of the original ontology. Namely,  $SR_T(O, O', f) = 1 - (|T'|/|T|)$ .*

Similarly, we have the following definitions about the R-box summarization rate and the A-box summarization rate:

**Definition 8 (R-box Summarization rate).**  $SR_R(O, O', f) = 1 - (|R'|/|R|)$ .

**Definition 9 (A-box Summarization rate).**  $SR_A(O, O', f) = 1 - (|A'|/|A|)$ .

More generally we are interested in the summarization rate which involves the cardinality of the union of the concept set, the role set, and the individual set:

**Definition 10 (Summarization rate).** *The summarization rate  $SR(O, O', f)$  is defined as the reduced cardinality measure with respect to the union of  $\mathcal{C}$ ,  $\mathcal{P}$  and  $\mathcal{I}$  as follows:*

$$SR(O, O', f) = 1 - (|C'| \cup |P'| \cup |I'|) / (|C| \cup |P| \cup |I|).$$

Naturally, we are also interested in the summarization rate with respect to the union of the T-box, R-box, and A-box as follows:

**Definition 11 (Box Summarization rate).**  $SR_B(O, O', f) = 1 - (|T'| \cup |R'| \cup |A'|) / (|T| \cup |R| \cup |A|)$ .

The approach of summarization in the previous sections has not yet considered the variable adjustment principle, the third basic principle of knowledge summarization. In the following we will propose the approach of summarization which supports various requirements on the summarization for variable cardinalities, a natural solution which is inspired from granularity.

## 6. Knowledge Summarization by Granularity

In this section, we will explore how a knowledge summarization can be achieved from the perspective of granularity, i.e. by different fine-grained modularizations of an ontology. Actually a summarization function can be considered as some kind of modularization of an ontology, however, a summarization concerns the preserving of concept hierarchy and role hierarchy, etc. A summarization function  $f$  of an ontology  $O$  into  $O'$  is said to be more fine-grained than another summarization function  $f'$  of  $O$  into  $O'$  if the function  $f$  is more fine-grained partition over  $O$  than  $f'$  (by the definitions in the set theory). Thus, we have the following definition:

**Definition 12 (Granular Summarization Method).** A granular summarization method of ontology  $O$  is to find a set of mapping  $\{f_1, \dots, f_n\}$  from  $O$  into a set of ontologies  $\{O_1, \dots, O_n\}$  such that

- $f_i$  is a summarization function from  $O$  into  $O_i$  for any  $i \in \{1, \dots, n\}$  and,
- $f_i$  is more fine-grained than  $f_{i+1}$  for any  $i \in \{1, \dots, n-1\}$ .

By the definitions, it is easy to see that the following formal properties hold:

**Proposition 4.** For a granular summarization method  $\{f_1, \dots, f_n\}$  of ontology  $O$  into a set of ontologies  $\{O_1, \dots, O_n\}$ ,

*i) Decreasing Cardinality.* The cardinalities of summarization ontologies are decreasing. Namely,  $|O_1| > \dots > |O_n|$ .

*ii) Increasing Summarization Rate.* The summarization rates of ontologies are increasing. Namely,  $SR(O, O_1, f_1) < \dots < SR(O, O_n, f_n)$ .

*iii) Monotonic Partition.* For any  $i \in \{1, \dots, n-1\}$ , and any  $C, D \in \mathcal{C}$ , if  $f_i(C) = f_i(D)$ , then  $f_j(C) = f_j(D)$  for  $j > i$ .

Properties *i)* and *ii)* above state that a granular summarization method is to seek for a decreasing cardinalities of summarization with increasing summarization rate. Property *iii)* states that the partitioning of granularity with a granular summarization method is monotonic. Namely, if two concepts are mapped into the same granule of a partition, then they would be located at the same granule at any sequential partition.

Figure 1 is an example how a granular summarization method can be introduced for ontology summarization with different levels of granularities. On the left side of the figure the ontology Living Thing is presented with its concept hierarchy. A mapping is introduced to merge the lowest concepts (i.e., the concepts subsume no other concepts) such as *Person*, *Dog* and *Cat* so that they are mapped into their parent concept, i.e., *Vertebrate*. Then, we can further merge other concepts so that they can be mapped into a new ontology with less cardinality. The granularity hierarchy is represented on the right side of the figure. Although the idea of merging children concepts so that they can be mapped into their parent concept is trivial, it illustrates very well how different granularities with reducing the cardinalities can be achieved.

## 7. Discussion and Conclusions

In this paper we have argued that knowledge summarization is a promising approach to scalable semantic data processing. We have presented a general framework of knowledge summarization for description-logics-based ontologies, proposed formal definitions of ontology summarization, and investigated how a knowledge summarization can be measured by using their summarization rate. Furthermore, we have explored knowledge summarization from the perspective of granular computing, and proposed some ideas for knowledge summarization with variable granularities.

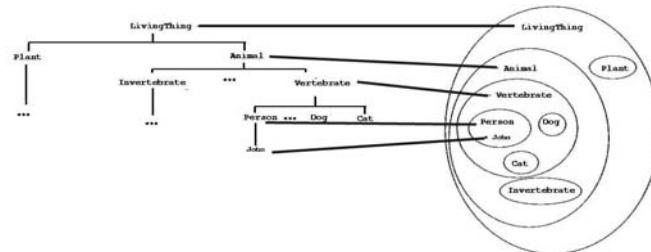


Fig.1 Example of Ontology Summarization with Variable Granularities

The work reported in this paper would provide a foundation for further exploration of efficient approaches for scalable semantic data processing. The future work of this paper includes:

- Develop more meaningful and non-trivial mapping functions for knowledge summarization.
- Explore variant knowledge summarization techniques, including partial mapping and inverse mapping for knowledge summarization.
- Conduct experiments and evaluation with some realistic and large scale ontologies for knowledge summarization.
- Accommodate more techniques from granular computing.
- Investigate the knowledge summarization techniques under the environments with other features of semantic data processing, such as dynamics, inconsistency, fuzzyness, and context-dependence.

### Acknowledgments

This paper is an extended and revised version of our paper entitled “Towards Scalable Semantic Data Processing by Knowledge Summarization”, which appears in the Proceedings of 2010 International Colloquium on Computing, Communication, Control, and Management (CCCM 2010). We thank the CCCM2010 organization committee for their kind permission on this extension and revision. The work reported in this paper was supported by National Natural Science Foundation of China (Grant No.60773059).

### References

- [1] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
- [2] Julian Dolby, Achille Fokoue, Aditya Kalyanpur, Li Ma, Edith Schonberg, Kavitha Srinivas, and Xingzhi Sun. Scalable conjunctive query evaluation over large and expressive knowledge bases. In *Proceedings of ISWC2008*, 2008.
- [3] D. Fum, G. Guida, and C. Tasso. Evaluating importance: A step towards text summarization. In *Proceedings of AAAI85*, pages 840–844, 1985.
- [4] J.R. Hobbs. Granularity. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*. Los Angeles, CA, 1985.
- [5] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *Journal of Web Semantics*, 1(1), 2003.
- [6] I. Horrocks, U. Sattler, and S. Tobies. Practical Reasoning for Very Expressive Description Logics. *Logic Journal of the IGPL*, 8(3):239–263, 2000.
- [7] W. Lehnert. Plot units and narrative summarization. *Cognitive Science*, (5):293–331, 1981.
- [8] Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. Summarization through structure learning with diversity, coverage and balance. In *Proceedings of WWW2009*, 2009.

- [9] T. Y. Lin. Neighborhood systems and relational database. In Proceedings of CSC'88. New York, 1988.
- [10] T. Y. Lin. Granular computing : structures, representations, applications and future directions. In Proceedings of 9th International Conference. RSFDGrC, 2003.
- [11] Thanh Tran, Haofen Wang, and Peter Haase. Searchwebdb: Data web search on a pay-as-you-go integration infrastructure. In Technical Report. Universitt Karlsruhe, 2008.
- [12] Y.Y. Yao. Granular computing using neighborhood systems, advances in soft computing: engineering design and manufacturing. In Proceedings of WSC3). London, 1999.
- [13] L. A. Zadeh. Fuzzy sets and information granulation. In Advance in fuzzy set theory and applications. North-Holland Publishing, 1979.
- [14] Yi Zeng, Yan Wang, Zhisheng Huang, and Ning Zhong. Unifying web-scale search and reasoning from the viewpoint of granularity. In Proceedings of AMT'09, 2009.