On Identity and Similarity Growth in Distributed Ontology

Franjo Jović*, Marina Pešut* Alan Jović**

*J.J. Strossmayer University in Osijek - Faculty of Electrical Engineering Osijek
31000 Osijek kneza Trpimira 2b
** Rugjer Bošković Institute Zagreb – Department of Electronics
jovic@etfos.hr

Summary. Ontologies of the dynamic distributed Web enterprises differ. Ontology engineers need a tool for ontology updating that is swift and simple. Therefore a simplistic distributed semantic web model (SDSW) is proposed. Added to the model is an extension of the primary logic operators. Logic operators are viable on domain variables and describe the formation, evolution and updates of the ontologies with conflicting information. Knowledge Base over a specific conflict is formed that uses expert information in order to cover for the logic consistency. The measure of logic consistency is obtained which is based on the extension set of the primary logic operators. Algorithms for checking consistency are thus drastically shortened in the case of large number of object ontology variables.

1. Introduction

The concept presented here is based on the need to uniformly interpret data and knowledge over a vast region of the information such as Web or complex technical objects such as hydrocarbon reserves. Ontologies play a key role in both application fields by providing shared and precisely defined terms that can be used for information description. The operations on the evolution changes and updates of ontologies are important because of the changes in the application requirements and information sources as well /1,2/ This is especially true for Semantic Web's dynamic aspects but is also observable in the hydrocarbon field during its exploration development /3/.

Declarative languages and mechanisms for specifying changes, updates, evolutions and maintenance are required. Each of these languages such as DAML+OIL, OIL /4/ and OWL/5/ are provided with some basic modeling primitives such as classes and roles and mappable onto a corresponding description Language (DL). A DL language called SHOQ(*D*) includes concrete data types and named individuals as well.

In this work we try to extend the basic SHOQ(*D*) concept expression by adding the logic similarity operator (6). By using extended logic operator set a Prefered Knowledge Base is proposed that can be used for evolutions and updates of the searched ontologies. The consistency of the ontology is evaluated when information sources are heterogenous and multiple such as the case of Distributed Web or in complex geological explorations.

2. Extended Concept Expression

2.1. Similarity logic operator

A similarity logic variable is proposed as the fourth logic operator. It is based on the qualitative model concept. A qualitative process and object model is any differential equation given with general parameters. By exchanging general parameters with specific values, i.e. at the very moment of calculation, qualitative analytical model becomes a quantitative analytical model. By this very act it obtains the analytical expressiveness but loses the abstraction.

Still simpler qualitative model is a behavioral similarity of process variables. If for instance by seeking a model for the process variable A one finds another process variable B that behaves in a similar manner then B becomes a candidate for the model of A. Thus

$$A \propto [B = \mathrm{mod}\, el(A)] \quad (1).$$

The sign $\propto$ stands for the similarity operation. But what is similarity? In its discussion of fallability of modern logic, mathematicians state that the only operator lacking in logic is the similarity operator. Thus, if we postulate similarity as the fourth logical symbol $(\vee, \wedge, \neg, \propto)$ and correspondently two-element algebra consisting of the "truth" and "false" values as Σ and ζ, it seems worthy to introduce first order predicate logic with the similarity operation.

It is clear how to define Boolean operations on limited and unlimited qualifiers; the Boolean operations are denoted by extended symbols $(\vee, \wedge, \neg, \propto)$ for each propositional function *F*, i.e.$Q1 \vee Q2$ is a function such that $Q(F) = Q1(F) \vee Q2(F)$. Since Mostowski gave the theory of the propositional functions of qualifiers earlier for $(\vee, \wedge, \neg, \propto)$ /4/, we shall only briefly discuss the same features for the similarity operation.

Here, two qualifiers are available: existential qualifier $\exists$ and general qualifier $\forall$. Introducing a limited qualifier *Qs* and *I* an arbitrary set with $I^* = I \times I \times I\ldots$ its Cartesian power , i.e. the set of infinite sequences $(x1, x2, \ldots)$ with $x_j \in I$ for $j = 1, 2, \ldots$.

A mapping *F* of I^* into $\langle \mathrm{T}, \perp \rangle$ is called a propositional function of *I* provided that it satisfies the following condition: there is a finite number *K* of integers such that if $x = (x1, x2, \ldots) \in I^*$, $y = (y1, y2, \ldots) \in I^*$ and $x_j = y_j$ for $j \in K$, then $F(x) = F(y)$.

Let us define the rank operator $\Re$ as a limited qualifier on *I*. It assigns a rank value to each element x_i of *I* as an individual variable *S* ranging from {1,2, …, *k* } to each functional variable of the degree *k*. The case of individual variable *S* possessing the same rank of its particular elements *i* and *j* can be dropped off by adding a small amount of noise to each particular variable *S*. Each element of two *S* variables will be called an *I*-valuation consisting of rank comparison of correspondent elements. Thus

$$Val_p\,\{Si / Sj\} \;\; \begin{array}{l} = \Sigma \text{ iff } (\Re p\, Si = \Re p\, Sj\} \text{ for all } p \in 1,2,\ldots k \\ = \zeta \text{ otherwise} \end{array} \quad (2),$$

presents the formula that is true in *I* if the equation (2) exhibits logical T and satisfiable in *I* if it exhibits true value for some parts of *p*. The degree of satisfiability can be measured with the rank correlation coefficient.

A latent similarity model *C* of the process variable *A* is a model that satisfies the equation (1), i.e. it is similar to the variable *B* as $C = Model(B) \propto B$. The modeling operation is not a process transitive operation and thus not firmly bounded to *A*. It leads in its extreme to the effects that if increasing B means increasing a certain process value *x* then *C* means decreasing that same value of *x*. Such models are process contradictory.

In order to evaluate expression (2) for $k = 9$ it is worth mentioning that coincidence of having all ranks equal evaluates to 1/9! or approximately to 2 in a million.

Qualitative model is complete when it covers all of the essential process features regarding function A that is being modeled. Qualitative model is completely acceptable when it is convertible into a appropriate quantitative form.

2.2 An extended concept expression

It is assumed that there exists a domain Δ_D of all data types, a set of data types *D*, $d \in D$, $d^D \subseteq \Delta_D$, furthermore the set of concept names **C**, the disjoint union of abstract role names R_A and a concrete role names R_D. There is a finite set *T* of terminological axioms in the form $C \subseteq D$, where *C* and *D* are concept expressions. Let *I* denote the set of individual names. *C*, *R* and *I* are mutually disjoint. The construtors in extended DL are given in Figure 1.

Constructor	Syntax	Semantics
Atomic concept *C*	C	$C^I \subseteq \Delta^I$
Abstract role R_A	R	$R^I \subseteq c \times \Delta^I$
Concrete role R_D	T	$R^I \subseteq \Delta^I \times \Delta_D$
Nominals	{o}	$\{o\}^I \subseteq \Delta^I, \#\{o\}^I = I$
Datatypes *D*	d	$d^D \subseteq \Delta_D$
Conjunction	$C \bigcap D$	$(C \bigcap D)^I = C^I \cap D^I$
Disjunction	$C \bigcup D$	$(C \bigcup D)^I = C^I \cup D^I$
Negation	$\neg C$	$(\neg C)^I = \Delta^I \setminus C^I$
Similarity	$C \propto D$	$(C \propto D)^I = C^I \propto D^I$
Exists restriction	$\exists R.C$	$(\exists R.C)^I = \{x \mid \exists y.(x, y) \in R^I \wedge y \in C^I\}$
Value restriction	$\forall R.C$	$(\forall R.C)^I = \{x \mid \forall y.(x, y) \in R^I \rightarrow y \in \mathrm{C}^\mathrm{I}\}$
At least restristion	≥n R.C	$(\geq n\ R.C)^I = \{x \mid \#\{y \mid (x, y) \in R^I, y \in \mathrm{C}^\mathrm{I}\} \geq n\}$
At most restriction	≤n R.C	$(\leq n\ R.C)^I = \{x \mid \#\{y \mid (x, y) \in R^I, y \in \mathrm{C}^\mathrm{I}\} \leq n\}$
Datatype exists	$\exists T.d$	$(\exists T.d)^I = \{x \mid \exists y.(x, y) \in T^I \wedge y \in \mathrm{d}^\mathrm{D}\}$
Datatype value	$\forall T.d$	$(\forall T.d)^I = \{x \mid \forall y.(x, y) \in T^I \rightarrow y \in \mathrm{d}^\mathrm{D}\}$

Figure 1. Syntax and semantics of the extended DL-concept expression.

The semantics of DL SHOQ(*D*) description logic are defined using an interpretation $I = <\Delta^I, \bullet^I>$, where Δ^I is a nonempty interpretable domain and $\bullet^I$ is an interpretation function. Traditionally a knowledge base cosists of set of assertions about individuals C(a) where an individual a is an instance oif C and roles R(a,b) where a is related to b by the role R.

3. A Discriminatory Knowledge Base

A simplistic distributed semantic web model is based on a knowledge base **K**(*T*,*R*) where *T* is a finite set of terminological axioms in the form $C \subseteq D$ where *C* and *D* are DL concept expressions. *R* is the role box of the knowledge base. If **K** is consistent, satisfactions of elements in **K** are defined according to interpretative notations of the type *I* ~ K where we say that *I* is a model of **K**; the subsumption based on ~ gives a monotonic satisfaction /7/.

By adding a new knowledge to the original knowledge base or by deleting outdated knowledge requires an inconsistency test of the new knowledge base.

3.1 An ordering clue to the SDSW

A simplistic distributed semantic web (SDSW) model introduces a strict partial ordering $\therefore$ as unreflexive anti-simetric and transitive on names operator. Such an operator orders KB rules for a given name. If out of two separate information seeking processes on the same ontology element there appear different rankings such that the interpretation of their difference can lead to KB inconsistency then this feature for that particular ontology is considered as defeated. In such a case with respect to certain measurement (such as the case of the distributed hydrocarbon deposit) or with respect to two corresponding data on the same ontology element the named object has to be discarded from the SDSW. Such a KB is named a Renewed KB (**RKB**).

Definition 1. An **RKB** is evaluated iff for all cases that are described as

- $C \subseteq D$ with respect to x is defeated iff $\exists A \subseteq B \prec C \subseteq D$ such that $A \subseteq B$ with respect to x is applied
- $R \subseteq S$ with respect to (x,y) is defeated iff $\exists R_I \subseteq S_I \prec R \subseteq S$ such that $R_I \subseteq S_I$ with respect to (x,y) is applied.

When an ordering clue is issued on the whole distributed semantic web or on the whole hydrocarbon deposit it comprises ordering control in all relevant semantic variables.

3.2 On inconsistent axioms

An axiom is inconsistent when two orderings with the same similarity ranking point to two different features, i.e.

- there is at least a single binary valued feature that is differently ordered
- there are more binary valued features that are separately differently ordered.

A similarity ranking meaures the information content of the ontology in question by its own proximity matrix /8/. If there are k out of n data in the ontology proximity then there is a flat function of the number of combinations. Thus, the discrimination function in ranking must be the importance of the set of k words in the meaning of the considered ontology. The order in which nine events occure can not be considered stohastic.

Inconsistent axioms are those where the changes in the proximity ranking are mandatory for the occurrence of inconsistency in axioms.

3.3 Construction of the consistent RKB

The requirement for the construction of the consistent Renewed Knowledge Base starts with questionning whether there are some changes observed in any of the ontology elements of the whole regarded domain. The corresponding semantic network should be addressed and changed according to the observed changes. The principal decision parameter is interpretation of the change in view of the corresponding ontology elements of the network by means of extensive questionnaire. In the case of hydrocarbon network it demands the recalibration of the whole deposit field. Particular measurement nodes in the deposit field should be renewed regarding each as an individual name.

4. On Decidability in Extended Concept Expression

An **RKB** is in principle a Prioritized Knowledge Base (**PKB**) /1/ because its development follows from the erradication of rules and axioms that are not strict axioms, i.e that they do not possess defeaters.

Theorem 1. P-satisfaction checking and subsumption checking based on the **RKB** are a decidable problem.

Proof: Acording to the above given description of the **RKB** mechanisms only those axiom rules that are applied with respect to *x* can indeed be used for reasoning. Therefore p-satisfaction checking of a concept based on **RKB** can be reduced to checking ~ satisfiability of the concept based on KB consisting of all the strict axioms that are applied in **RKB**. In this situation the reasoning based on KB is monotonic.

5. Evolution and Similarity Growth in Distributed Ontology

Evolution in an RKB can be caused by growing of the semantic net in the problem domain, simply it can be analogously expressed as a semantic increase in the number of the measured points in the regarded oil-deposit layer resulting in overall increase of the known points in the

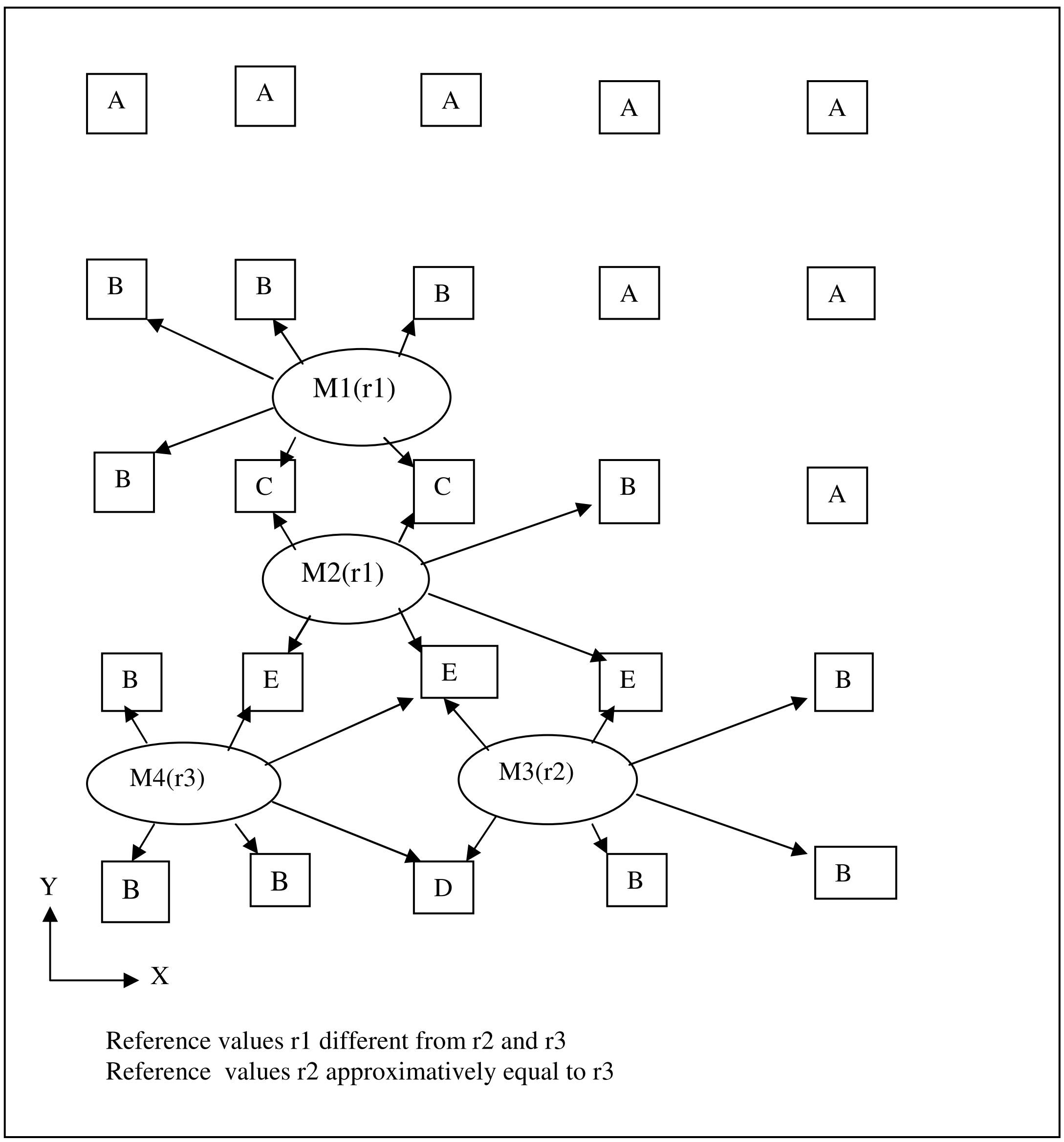

Figure 2. Reference values models and measurement point semantics in a hydrocarbon deposit

field. On the other side it can be increased by means of increasing the ontology domain description. This is an unexpected direction of growth because it could point to a merely semantic petrification of the ontology element. On the other side the increase of the number of ontology elements in the domain semantic net can lead to the possibility of more subtle ontology of the domain. The nature of the language and the entropy in the language development are expected to keep a long-term balance. A distributed ontology is prone to semantic differences because each separate ontology is most probably designed and motivated by different causes and fed with different aspects of the same nominal ontology.

6. An Example from the Hydrocarbon Deposit Semantic: Reservoir Point Estimation

The case of hydrocarbon deposit estimation is presented where the ranking procedure is used at the vicinity of each measurement point for all 30.000 points in the network / 3/. Calibrated measurements were used for identification of points with the same measurement data features. These four measurement models are given with for reference values M1(r1), M2(r1), M3(r2) and M4(r3) in Figure 2, obtained from three reference values. Measurement point outside the measurement semantics are designated with *A*, measurement point with single measurement semantic are designated with *B*, measurement points with compatible reference but with different descriptors are designated with *C*, measurement points with incompatible references are designated with *E* and measurement point with different descriptors and slightly different reference values is designated with *D*, all in Figure 2. Measurement point *A* is not estimable possessing any modeling evidence, as well as *B* because there is no coincidence in the measurement reference. Measurement point *C* is estimable because of non conflicting reference information. Measurement points *D* is feasibly estimable the decision being the matter of the reservoir engineer's standpoint. Measurement point *E* is inestimable because of the conflicting reference information.

A new reference measurement could improve: possibly recovering some of the *B* points into *C* class and giving more information of the *D* point case. It could however less probably diminish the number of estimated points if its values are taken after an enough long interval when the reservoir conditions have been slightly modified.

7. Discussion

Evolution, versioning and transactions are often neglected by the Semantic Web community. The changes and updates of ontologies are important because continuous modifications in applications and requirements are forcing it. Conflicting issues are expectable and the procedures for their operablity must be stated. Introduction of the similarity operator enables some finer operations among names and operators. Still there seems to exist a unresolvability in the ranking of the ontology domain for specific names/points. The example of consistent ontology in the measurement point estimation is presented merely as an illustration of the ontology development and its possible evolution.

References:

1. Y. Ma, B. Jin, Y. Feng, "Dynamic evolutions based on ontologies", Knowledge-Based Systems, 2006.

2. D. Oberle, S. Staab, R. Studer, R. Volz, "Supporting application development in the semantic Web", ACM Transactions on Internet Technology 5 , 2005.
3. F. Jović, "Metoda i postupak za određivanje iznosa poroznosti nalazišta ugljikovodika s pomoću odbacivanja konteksta modelskog sadržaja dominantnih komponenti seizmičkog signala", Hrvatski patentni glasnik, prijava patenta P20040828A, srpanj 2006.
4. I. Horrocks, P. Patel-Schneider, F. van Harmelen, "Reviewing the design of DAML+OIL: a language for the Semantic Web", Proceedings of AAAI'02, pp.792-797, 2002.
5. F. van Harmelen, "OWL Web Ontology Language Reference", available at http://www.w3.org/TR/owl-ref//, 2004.
6. F. Jović, A. Jović, V. Rajković, "A Contribution to the Investigation of Algebraic Model Structures in Quliatative Space", CTS & CIS Computers in Technical Systems Intelligent Systems (MIPRO), pp. 49-53, 2006.
7. F. Baader, A. Borgida, R. Kuesters, D.L. McGuinness, "Matching in Description Logics", Journal of Logic and Computation -- Special Issue on Description Logics, vol.9, no.3, 1999.
8. Z. Radoš, F. Jović, J. Job, "Extracting most frequent Croatian root words using diagram comparison and latent semantic analysis ", Proceedings of the 7th International Conference on Enterprise Information Systems (ICEIS 2005), pp. 370-373, Miami, 2005.