Personal Agent for Mapping Elements that Look Alike – Pamela

Ronald Poell

CTO Semantic Network Technologies
Netherlands Organization for Applied Scientific Research (TNO)
ronald.poell@tno.nl

Version 0.5, 24/07/2009

Task Oriented Ontology

The semantic network contains various models each with their own class hierarchies, predicates and properties.
When using the information in the network for a particular purpose, personal and task oriented, you might want to consider things that are "persons" in one model, "man" and "woman" in a second and "employee" in a third as the same kind of things. The same holds for properties and predicates.
Even the "logic" (i.e. the rules you want to apply to infer information) can be specific to your quest for information.

For the things you want to use you might also want to restrict the properties and kinds of relationships (predicates) that are of interest in the context.
Pamela allows you to do so and will build a kind of task oriented ontology, a specific view on all the information available. The agent consist basically for a large part of smart services that, in an autonomous way, try to match model elements and individuals. A second part consists of a user interface that enables the selection the user wants to apply and confirmation of the findings.
Typical technologies that are used by Pamela are: name similarity, used property value range analysis, property set and relationship set similarity, etcetera.

Property Value Range Analysis (PVRA)

Property types in the semantic network have an associated property value type. This value type is only a hind to services and applications exploiting the property how the value should be interpreted (in the implementation they are all strings). Typical property value types are: integer, long, date, string etc.

When comparing properties, one of the comparisons is done on the value type of the property. But as there are only a few of them, although it gives some indication, it is far away of the needed distinguishing capacity we need. We decided to extend the model comparison (property type and property value type) with the value instances. The analysis that looks at the values really used in the network is covered in this chapter.

As always in the semantic network technology, the configuration of the PVRA is context sensitive. There is a standard configuration but every user can have its Pamela(s) execute the PVRA in a specific way. The results define a main value type and possibly a range subtype. See table 1 for the actual 66 default combinations.
The analysis is statistical and uses on a random subset of all available values. The maximum size of the subset is also configurable.
The analysis results are stored in the network associated with the specific Pamela context node for later use. Besides the analysed value type and the range type, whenever applicable, the lowest and highest value encountered are conserved.

An important aspect to bear in mind is that the "value" of the analysis result has its signification in the light of its purpose: to use it in comparisons with other results. If we would have analysed only the value type and not the range type it would have only less distinguishing capacity but would not be wrong. Similar if a range type "Human historical" (Date) would not correspond to your perception of historical, Pamela will analyse other properties covering approximately the same range of dates also as "Human historical" and will establish a value range similarity between these properties. And that is what the PVRA is intended to serve.

Future work
Of course we expect the possible value types and range types to be extended in the future. Properties may have also values that point to other nodes (NID-node identifier). Actually we do not analyse these. In future we will combine several analysis on these kinds of properties. First these properties can be compared to a statement and the property type as a predicate. So we can do a subject and object type analysis on the nodes. Second, the node pointed to can be considered as a node representing a value type. This would be the case e.g. for dates. With a property "is born during" pointing to "Second World War", the latter could be interpreted as a date. A combination of both analysis might give: "This property point to nodes that are for 96% typed as 'Events' that have themselves to properties typed as 'Date' where the property 'Begin date' is always smaller then 'End date'".
The subset selection for the PVRA will be similar to a method used in ecology to define the limits and species richness of an ecosystem (Nested Quadrat Method): start with a specified size and increasing this by a factor 2 until no modifications occur anymore (or all available values are used). The spatial relation from the Nested Quadrat Method will be represented by the network distance between the nodes. This technique should be able to distinguish different node "ecosystems" belonging to different "area's" in the network who we expect to belong to different models.

Type Identification Service (TIS)

As opposed to ontologies there is no class hierarchy in the semantic network. Nodes do not belong to a specific class at a model level. The user will define, within the usage context, to what kind of thing a specific node should belong to (the "type" of the node).

Amongst the services available in the semantic network technologies there is already one configurable "Type Identification Service". The configurations of this standard TIS consist each of a set of predicates designing "is a (kind of)" (hyponym) like relations ships; a set of "class" nodes that are grouped in the type we are looking for and optionally sub-services that analyse subtypes. The "Person" service uses e.g. the "Man" and the "Woman" service, and is itself used by the "Living Creature" service. The API for the TIS allows a request for a list of types a specific node belongs to (above a given threshold) or whether a specific node belongs to a given type or not (also above a given threshold).
Pamela will use these standard services when appropriate.

The Pamela agent will also generate its own kind of type identification service using statistic occurrences of kinds of relationships based on the predicates (with or without associated target types) and occurrences of property types (with or without associated value type ranges). It might e.g. discover that the only nodes that have an "email address" property are designed by the standard TIS as "Person", "Organization" or "Avatar". That "first name" properties are only associated with "Person" and "Animal". The a node that "has written" something is in general (97%) a "Person" and rarely (3%) an "Organisation". Based on these observations the agent can make a Pamela Type

Value type	Range type
Date	Future Near future Recent Computer age Living persons Human historical Historical
Email	-
Float	Zero One Small angle Medium angle Angle Positive Negative
Hexadecimal	-
ISBN	-
Long	Zero One Binary 1 byte Binary 2 bytes Binary 3 bytes Binary 4 bytes Binary 5 bytes Small angle Binary 6 bytes Binary 7 bytes Medium angle Binary 8 bytes Binary angle Positive Negative
Numeric code	-
Phone number	-
String	Identifying 1 length capital code 1 length alphanumeric code 1 length character code 2 length capital code 2 length alphanumeric code 2 length character code 3 length capital code 3 length alphanumeric code 3 length character code 4 length capital code 4 length alphanumeric code 4 length character code 5 length capital code 5 length alphanumeric code 5 length character code 1 or 2 length capital code 1 or 2 length alphanumeric code 1 or 2 length character code 1 to 3 length capital code 1 to 3 length alphanumeric code 1 to 3 length character code Restricted name set (enumeration) Restricted set (enumeration) Long text
Url	-
Unit based	Length Volume Time Weight Currency Speed

Table 1: The standard main value types and range subtypes the property value range analysis produces.

Identification Service (PTIS) stipulating that a node that has an “email address” and a “first name” property is most probably a person and that this conclusion is reinforced if that node has a “has written” relationship.

As Pamela sets up the ontology through examples given by the user, the agent will create for each “class” a new PTIS that will take into consideration not only the predicates and property types the user would like to use in the ontology but also the predicates and property types that occur in the sample nodes but are not retained for the current ontology and which might be useful in the identification process.

Creating the ontology

You start the creation of an task oriented ontology by providing a name for it. Next you'll create the classes you want to use by providing its name and selecting one or more existing nodes that are typical instances of that class.
Pamela will analyse the property types and predicates used in the sample nodes and set up temporary profiles based on the standard type identification services. If none of these services provides a type it will make one based on the example node (s). You can ask Pamela to look for other similar instances that might belong to your class but which have also other properties or predicates and you can add some these to your sample set.
You “tune” the profiles by removing property types and predicates you are not interested in.
Finally you can add other property types and predicates that were not associated with the selected examples (e.g. a new property that has not yet been used).
For the predicates you indicate which classes are potential objects of the relationship.
When you are done with the creation of all your classes you ask the Pamela agent to do its actual work: analyse the semantic network and provide you feedback on property types and predicates you have not selected but that semantically or model based similar. This is a task you can ask to be executed permanently or with a certain interval. Pamela will provide you feedback on its findings and, if you confirm, will extend the ontology.