

Enabling Semantic Web Forums using Dynamic Representation Schemas

L.Lella, G.Tummarello, C.Morbidoni

Dipartimento di Elettronica, Intelligenza Artificiale e Telecomunicazioni

Gruppo di Intelligenza Artificiale

Universita' Politecnica delle Marche (Ancona, ITALY)

info@semanticweb.deit.univpm.it

Abstract

Traditional web forums allow users to post messages and reply to existing ones. As messages often relate to multiple subjects, the tree like structure created by the simple “reply to” operations are not satisfactory. In this paper we tried experimenting with automatic creation of RDF annotations to link between textual messages that spawn beyond the simple forum message threading. We call this model “semantic web forum”. The creation of linking annotations between messages is performed using a modular knowledge acquisition system based on cognitive criteria. Peculiar characteristic of the system is the use of scale free graph models and the ability to work incrementally as more messages are inserted into the system. The final representation has been compared with analogous schemas obtained from human subjects. Test shows that for proper parameters the added annotations remain highly relevant and concise with respect to the number of messages in the forum.

1. Introduction

Traditional web forums allow users to post messages and reply to existing ones. This creates the well known discussion “threading” structure where messages and replies form a tree. Web forum software as well as Newsgroup readers often provide explicit support for browsing through this structure. This is of course highly meaningful as it allows a reader to follow the replies in their natural order. However, as the discussion topic drift wider than the starting subject, it is often the case that important information are posted about topics which are fundamentally

unrelated to the starting one. Furthermore messages could sometime be referring to other messages in the same mailing list without being in the same branch of the threading tree structure. In this paper we tried experimenting with automatic creation of RDF annotations to link between textual messages thus overcoming the limits of the simple forum message threading. We call this model “semantic web forum”.

The creation of linking annotations between messages is performed using a textual analysis procedure derived from a modular knowledge acquisition system based on cognitive criteria which we describe in starting from the next section.

2. Dynamic representation schemas

Classical representation forms, as semantic networks [4] frames [17] and scripts [18], are not suited to represent knowledge as human mind does. Infact human mind seems to generate contextualized structures that are adapted to the particular context of use, represented by the goals that must be achieved and the semantic or situational context [8].

The networks of propositions, or *knowledge nets*, introduced by Kintsch and Ericsson [9] are an alternative formalism that combines and extends the advantages of the symbolic representations avoiding their limits.

Knowledge nets are made by nodes which represent proposition having a predicate-argument structure with time and location slots. Those atomic propositions are linked by weighted unlabeled arcs.

In this way the absolute – it could be said linguistic – meaning of a proposition is given by the exact position of the corresponding node within the

network, i.e. by the strengths of the links that connect it to the neighbour nodes.

But from a psychologic point of view, only the nodes that are activated, i.e. maintained in the so called *working memory*, contribute to specify the actual meaning of a node. So the contextualized meaning of a concept is not given by a static and fixed set of definitions and relations, but is built every time in the working memory by the activation of a subset of the neighbour propositions, the immediate ones as well as others many steps apart.

The context of use given by the objectives, the semantic and situational state determines which nodes have to be activated and stored in the working memory.

2.1 The LTWM model

In order to specify the activation modalities Kintsch and Ericsson [10] have introduced the concept of Long Term Working Memory (LTWM). This is the activated part of the Long Term Memory (LTM) that is the entire network of propositions and represents all the acquired knowledge. The LTWM is generated by the short term part of the working memory (STWM), that is the representation of the analyzed information, by fixed and stable memory structures called retrieval cues that link the objects present in the STWM to other objects present in the LTM. Kintsch has developed two methods for the definition of the LTWM.

The first, defined with Van Dijk [21], is a manual technique that starts from the propositions present in the text (micropropositions) and using some organizing rules arrives to the definition of macropropositions and macrostructures and even to the definition of LTWM.

The second is based on the latent semantic analysis (LSA) [11]. This technique can infer, from the matrix of co-occurrence rates of the words, a semantic space that reflects the semantic relations between words and phrases. This space has typically 300-400 dimensions and allows to represent words, phrases and entire texts in a vectorial form. In this way the semantic relation between two vectors can be estimated by their cosine, a measure that according to Kintsch can be interpreted as a correlation coefficient.

This latter solution to the problem of the definition of LTWM puts a great and inevitable technical

problem. How many objects must be retrieved from the semantic space for every word present in the text? In some cases, when the textbase, i.e. the representation obtained directly from the text, is sufficiently expressed, the retrieval of knowledge from the LTM is not necessary. In other cases a correct comprehension of the text, or the relative situation model, requires the retrieval of knowledge from the LTM.

After the creation of the LTWM the integration process begins i.e. the activation of the nodes correspondent to the meaning of the phrase. Kintsch uses a diffusion of activation procedure that is a simplified version of the one developed by McClelland and Rumelhart [14]. First an activation vector is defined whose elements are indexed over the nodes of LTWM. Any element's value is "1" or "0" depending on the presence or the absence of the corresponding node in the analyzed phrase (i.e. in the STWM). This vector is multiplied by the matrix of the correlation rates (the weights of the links of the LTWM) and the resulting vector is normalized. This becomes the new activation vector that must be multiplied again by the matrix of the correlation rates. This procedure goes on until the activation vector becomes stable. After the integration process, the irrelevant nodes are deactivated and only those that represent the situation model remain activated.

There is also another problem that must be considered. Theoretically the position occupied by a word in the LTWM is determined by a lifetime experience, i.e. by the continuous use that is made of it. Obviously this kind of knowledge cannot be reached practically and Kintsch build his semantic space using information taken from a dictionary.

Furthermore the construction-integration process does not always assure the semantic disambiguation of the analysed phrase.

The use of an external dictionary as WordNet [16] and of particular disambiguation procedures can overcome the last two limits.

Instead the first problem can be fully solved only by dropping the intermediate representation of the semantic space and by developing new methods for the direct formation of networks of concepts and propositions.

The creation and the updating of a knowledge net after the analysis of new information should be supported by an existing general ontology from which the relations between the disambiguated concepts can be retrieved.

The knowledge acquisition system we have used is capable to obtain a representation that corresponds almost exactly to the content of the analyzed text. The encountered words are connected in order to form a graph of relations that is used to update the system knowledge base which is structured as an associative network, i.e. a graph with weight and not labeled links [15].

3. Architectural overview

The system is based on the theory of the LTWM by Kintsch and Ericsson, but it has been chosen a different implementation that leads to the direct creation of the LTWM nodes and the retrieval cues. Its architecture is shown in figure 1.

A part of the analyzed text is selected by the use of a window. Its content is filtered by the syntactic module of WordNet and a suitable stoplist of words.

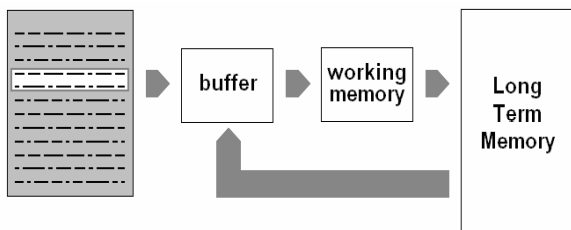


Figure 1. The overall structure of the proposed architecture

The not filtered content of the buffer must be codified on the basis of the context before being elaborated by the working memory block. The context represents the theme, the subject of the processed text and for its correct characterization not only the information present in the document must be considered, but also the one that can be retrieved from the structure representing the knowledge accumulated during the analysis of the previous documents presented to the system (Long Term Memory).

For the implementation of the working memory block, self organizing networks with suitable procedures for the labeling of their nodes could be used, but this solution requires a lot of computational time, especially for the analysis of entire repositories of documents.

So have been used alternative models based on the theory of scale free graphs [2] for the implementation of an associative network.

The graph theory dealt with regular graphs until the 50s. Subsequently random graphs were introduced [7]. They were the first simple forms of complex graphs that had ever been studied.

Their model started with a network made by N isolated nodes. Successively each pair of nodes could be connected with a probability p , leading to a graph having approximately $pN(N-1)/2$ links.

But this model was still far from real networks present in nature and artificial systems. So scientists defined other models characterized by an higher complexity level.

The actual models have three main features.

First their “small world” structure. That means there is a relatively short path between any two nodes [22].

Second their inherent tendency to cluster that is quantified by a coefficient that was introduced by Watts and Strogatz. Given a node i of k_i degree i.e. having k_i edges which connect it to k_i other nodes, if those make a cluster, they can establish $k_i(k_i-1)/2$ edges at best. The ratio between the actual number of edges and the maximum number gives the cluster coefficient of node i . The clustering coefficient of the whole network is the average of the all individual clustering coefficients. In a random graph the clustering coefficient is $C = p$. In real networks the clustering coefficient is much larger than p .

Actual graph models are also characterized by a particular degree distribution. While in a random graph the majority of the nodes have approximately the same degree close to the average degree, the degree distribution $P(k)$ of a real network has a power-law tail $P(k) \sim k^{-\gamma}$. For this reason these networks are called “scale free” [1].

Recently it has been found that human knowledge seems to be structured as a scale free graph [20]. Representing words and concepts with nodes, some of these (hubs) establish much more links compared with the other ones.

This particular conformation seems to optimize the communication between nodes. Thanks to the presence of the hubs, every pair of nodes can be connected by a low number of links in comparison with a random network with the same dimensions. The definition and the eventual updating of a scale free network does not require a lot of time and the execution of particular processes, as the diffusion of the activation signal, is very fast.

necessary to make useful previsions about its evolution.

In the scale free graph models proposed by literature at each temporal step M new nodes are added to the graph, with M defined beforehand. These M nodes generally establish M links with M old units of the network. In the system that we have developed, after the analysis of a new document the links related to an unknown number of nodes of the LTM network are updated on the basis of the content of the WM. This number depends on the analysed document because it is the number of the words that have not been filtered by the stoplist.

Another important difference with other scale free models presented in literature [5] is the particular fitness function that is used. This function does not depend on a single node but on the considered pair of nodes. If this value is chosen as proportional to the weights of the LTM associative network, the fitness value of a word is not constant but depends on the other word that could be linked to it. For example the noun "house" should present for the link with "door" a fitness value greater than the ones presented for the links with "person" and "industry".

In [6][12] it has been demonstrated that both the WM graph and the LTM associative network are structured as scale free graphs and the structural parameters are really similar to the ones of other knowledge representations made by humans (WordNet, Roget's Thesaurus) or obtained by humans (associative network built after experiments of free words associations).

It has been also computed the coherence rate of the obtained representations. The coherence rate is obtained by correlating the LTM ratings given for each item in a pair with all of the other concepts. Schvaneveldt has proved that the coherence rate in humans often correlates with expertise and coherence rates below 0.2 denote that the test has not been performed seriously.

The average coherence rate (0.45) has confirmed that the conceptualization, i.e. the evolution of the associative network, was made by the system on the basis of a precise inner schema.

To evaluate the correctness of this schema LTM associative network have been compared with representations obtained from a group of human subjects [13].

The subjects were asked to read the same medical article examined by the system, assigning a rate of

relatedness to each pair of words that were considered by the system. A Pathfinder analysis [19] was performed on the relatedness matrices provided by human subjects and the LTM matrix, in order to extract the so called "latent semantics", i.e. other implicit relations between words. The obtained matrices were compared using a similarity rate determined by the correspondence of links in each pair of networks.

The representation of the system is more similar to the one of the first two groups which has the highest coherence rate, and the probability to have exactly this number of links in common by chance being very low (0.248%).

4. Automatic Message Annotation

We will now see how this knowledge acquisition model can be used to annotate messages in a web forum. The idea is to organize the messages in an alternative "graph like" threading based on semantic similarities.

We first define a similarity criterion. Given two graphs representations of two different messages A and B we select the corresponding nodes. We then consider the shared links. A simple similarity measure is obtained by considering the ratio between the shared links and the total number of links present in the message.

This kind of relation is not commutative. This comes from the fact that the semantic "payload" of a message can vary in size. As an example we consider a complex text (e.g. this paper) and a shorter one that is simply a subset (e.g. this chapter). It is clear that, based on the above definition, the similarity ration of the chapter to the paper is 1 while is much lower the one relating the paper to the subject.

As a test bench we have considered 200 of the most recent messages of the Yahoo discussion group about the Semantic Web. Given the incremental nature of the algorithm, the messages have been analyzed according to the insertion order. To simplify the final RDF representation of the messages we have considered only the relation with a similarity measure exceeding a previously set normalized threshold. This threshold determines the sheer amount of "similarity links" between the messages..

As shown in figure 3, the number of detected relations between the analyzed messages strongly

increases with a threshold of 1.5, while it is quasilinear given a threshold of 2.0. It is clear how important it is to either properly select or properly regulate said threshold value so that the created similarity links remain concise. We report obtaining positive preliminary results, although a proper, human judgment based, study is necessary to provide a formal quality assessment. A few specific annotation examples derived from the experimental corpus are shown in the following section.

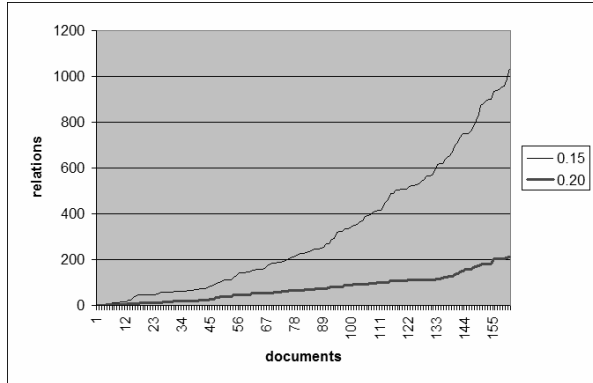


Figure 3. Number of relations detected for different numbers of messages and threshold

5. RDF Syntax and example results

To express the results of this task we have create an ontology, a simple RDF schema, defining the property “#similarTo” corresponding to this similarity relation. This property has been used to point a blank node that is connected by the property “#rate” to a literal representing the similarity rate.

The blank node is also connected by the property “#name” to the resource corresponding to the considered message. The RDFS definitions of these properties are shown in figure 4.

```
<?xml version="1.0" encoding="WINDOWS-1252"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

  <rdf:Property rdf:about="...#similarTo">
    <rdfs:isDefinedBy rdf:resource="...graph#">
    <rdf:type rdf:resource="http://.../22-rdf-syntax-ns#Property" />
    <rdfs:label>similarTo</rdfs:label>
```

```
<rdfs:range rdf:resource="http://.../rdf-schema#Resource" />
<rdfs:domain rdf:resource="http://.../rdf-schema#Resource" />
</rdf:Property>

<rdf:Property rdf:about="...#name">
<rdfs:isDefinedBy rdf:resource="...graph#">
<rdf:type rdf:resource="http://.../22-rdf-syntax-ns#Property" />
<rdfs:label>name</rdfs:label>
<rdfs:range rdf:resource="http://.../rdf-schema#Literal" />
<rdfs:domain rdf:resource="http://.../rdf-schema#Resource" />
</rdf:Property>

<rdf:Property rdf:about="...#rate">
<rdfs:isDefinedBy rdf:resource="...graph#">
<rdf:type rdf:resource="http://.../22-rdf-syntax-ns#Property" />
<rdfs:label>rate</rdfs:label>
<rdfs:range rdf:resource="http://.../rdf-schema#Literal" />
<rdfs:domain rdf:resource="http://.../rdf-schema#Resource" />
</rdf:Property>

</rdf:RDF>
```

Figure 4. RDFS definitions of the properties #similarTo, #name and #rate

Given the incremental nature of the algorithm, the messages have been analyzed according to the insertion order. To simplify the final RDF representation of the messages we have considered only the relation with a similarity measure exceeding a previously set normalized threshold.

In figure 5 there is an example of graph representation of the relation between two messages (no.10 and no.11) regarding the problem of the merging between ontologies.

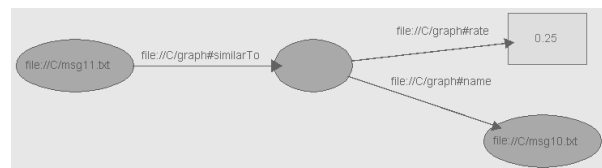


Figure 5. An example of similarity relation between the graph representation of two messages

In fig. 6 are represented the relations between the message no.7 and the message no.3. The first is a job announcement by an organization dealing with knowledge integration and management. The second

is a short course regarding the use of a tool for the managing of ontologies in OWL and RDF(S).

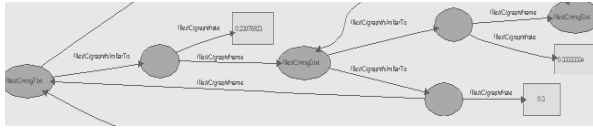


Figure 6. An example asymmetric similarity relations between two analyzed messages

The message no.7 is more general than the message no.3 because the job requirements comprehend other skills in addition to the experience with Semantic Web tools and languages.

In figure 7 there is an example of two messages (no.3 and no.5) having the same level of generality.

The messages present two different tools for the managing of ontologies in RDF(S), and the similarity rates between them have symmetric similarities values.

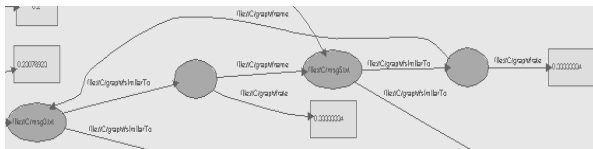


Figure 7. An example of symmetric similarity relations between two analyzed messages

5. Conclusions

In this paper we have suggested how textual analysis can be used to create similarity links between messages in a web discussion. This would enable a new scenario where the messages are organized beyond the classical tree “threading” structure given by the simple “reply” mechanism. An innovative system for the acquisition of knowledge from texts has been presented and used for this purpose.

First experimental results confirms the validity of the final representation, while much broader results are expected once the system is inserted in the DBin project. (see www.dbin.org), a semantic web client allowing textual annotations of URIs.

6. Acknowledgment

The authors are grateful to Prof. Ignazio Licata (Istituto di Cibernetica Non-Lineare per lo Studio dei

Sistemi Complessi, Marsala(TP) - Italy) for helpful discussions, comments and criticisms.

7. References

- [1] R.Albert, A. Barabasi, “Topology of evolving networks: Local events and universality”, *Phys. Rev. Lett.* n.85, 2000, p.5234.
- [2] R. Albert, A. Barabasi, “Statistical Mechanics of Complex Networks”, *Rev. Mod. Phys.* no.74, 2001, pp.47-97.
- [3] G. Bianconi, A. Barabasi, “Bose-Einstein Condensation in Complex Networks”, *Phys. Rev. Lett.* vol. 86, no. 24, 2001.
- [4] A.M. Collins, M.R. Quillian, “Retrieval from semantic memory”, *Journal of Verbal Learning and Verbal Behaviour* 8, 1969, pp.240-247.
- [5] S.N. Dorogovtsev, J.F.F. Mendes. “Evolution of networks”, arXiv: cond-mat/0106144, submitted to *Adv. Phys.*, 2001.
- [6] A. Dragoni, L.Lella, G.Tascini, W.Giordano, “Knowledge Extraction Using Dynamical Updating of Representations”, *Proceedings of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data (Romand 2004)*, Geneve, August 2004.
- [7] P.Erdos, A. Renyi, “On Random Graphs”, *Publ. Math. Debrecen* 6, 1959, p. 290.
- [8] W. Kintsch, *Comprehension. A Paradigm for Cognition*, Cambridge University Press, 1998.
- [9] W. Kintsch, “The Representation of Knowledge in Minds and Machines”, *International Journal of Psychology* 33(6), 1998, pp.411-420.
- [10] W. Kintsch, V.L. Patel, K.A.Ericsson, “The role of long-term working memory in text comprehension”, *Psychologia* 42, 1999, pp.186-198.
- [11] T.K. Landauer, P.W. Foltz, D. Laham, “An Introduction to Latent Semantic Analysis”, *Discourse Processes* 25, 1998, pp.259-284.
- [12] I.Licata, G.Tascini, L.Lella, W.Giordano, “Scale Free Graphs in Dynamic Knowledge Acquisition”, *Proceedings of the 3rd Conference on Systemics AIRS 2004*, Castel Ivano (Trento, Italy), October 2004.
- [13] I.Licata, L.Lella, W.Giordano, “From ontologies to ontogenetic models, extracting semantics by knowledge representation dynamical updating”, *Proceedings of Dynamic Ontology*, Trento (Italy), 2004.
- [14] J.L. McClelland, D.E. Rumelhart, *Parallel distributed processing*, Cambridge, MA: MIT Press, 1986.
- [15] D.E.Meyer, R.W. Schvaneveldt, “Facilitation in recognizing pairs of words: Evidence of a dependence

between retrieval operations". *Journal of Experimental Psychology* 90, 1971, pp.227-234.

- [16] G. A. Miller, "Five papers on WordNet", *Cognitive Science Laboratory Report* 43, 1993.
- [17] M. Minsky. *A framework for representing knowledge*. In P.H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill, 1975.
- [18] R.C. Schank, R.P. Abelson, *Scripts, plans, goals, and understanding*, Hillsdale, NJ: Erlbaum, 1977.
- [19] R.W. Schvaneveldt, F.T. Durso, D.W. Dearholt, "Pathfinder: Scaling with network structures", *Memorandum in Computer and Cognitive Science, MCCS-85-9*, Computing Research Laboratory. Las Cruces: New Mexico State University, 1985.
- [20] M. Steyvers, J. Tenenbaum, "The Large-Scale structure of Semantic Networks", Working draft submitted to *Cognitive Science*, 2001.
- [21] T.A. van Dijk, W. Kintsch, *Strategies of discourse comprehension*, New York: Academic Press, 1983.
- [22] D.J. Watts, S.H. Strogatz, "Collective dynamics of 'small-world' networks", *Nature* vol. 393, 1998, pp. 440-442.