Hybrid Model for Knowledge Representation

Reena T. N. Shetty	Pierre-Michel Riccio	Joël Quinqueton
Doctorate, LGI2P	*Assistant Professor, LGI2P,*	*Professor-LIRMM,*
EMA, Fontenblu, Paris	*Site EERIE, 30035 Nimes*	*Montpellier*
Reena.Shetty@ema.fr	*Pierre-Michel.Riccio@ema.fr*	*jq@lirmm.fr*
Tel: +33 4 66 38 70 39	*Tel: +33 4 66 38 70 48*	*Tel: +33 4 67 41 85 32*

Abstract

In this paper the fundamental idea is top develop and explore an innovative approach of completing human designed networks with that of machine built word networks. This network forms a hybrid method which combines human precision with that of machine computation to form a knowledge representation model. This model in turn encourages faster and efficient construction of automatic ontology. Our objective is to tackle the problem faced in the field of information retrieval and classification in the current era of information over flow.

1. Introduction

Recent years has witnessed a tremendous increase in the use of world wide web and it is soon becoming an essential means of information furnishing resource [6] in diverse forms covering various domains. Searching web in its present form is however an infuriating experience for the simple fact that the data available is too superfluous [12] and non-comprehensive by machines.

This drawback has led to several research groups proposing several innovative, intelligent tools and theories of knowledge representation techniques to attain efficient information search. The most interesting solutions proposed are developing semantic web based ontologies to incorporate data understanding by machines. The objective of this approach is to intelligently represent data, enabling enhanced captured of existing information by machines.

However, some of the most discouraging drawbacks of this approach are firstly, the high cost of construction involved in building such ontologies [1,2,3,4,5] with the expertise help of domain specialists followed by the huge amount of time required for construction.

This makes it highly inaccessible to small research groups and enterprises requiring ontologies or similar methods for their research activities. Our proposal is to build an innovative semi-automated knowledge representation model where results are close to ontology in terms of efficiency, precision and recall.

The principle objective of our proposal is to enable building of semi automated ontologies with decreased time and cost for construction. To achieve this we propose our knowledge representation model called extended semantic network which forms a hybrid model between automatically constructed proximal network model and the human constructed semantic network model.

2. State of the art

One of the most basic reasons for ontology construction [1] is to facilitate sharing of common knowledge about the structural information of data among humans or electronic agents. This property of ontology in turn enables reuse and sharing of information over the web by various agents for different purposes. Ontology [2, 15] can also be seen as one of the prominent methods of knowledge representation due to its ability to represent data in a relational hierarchy it shares with the other existing data.

There are several developed tools for ontology construction and representation like protégé-2000 [4], a graphical tool for ontology editing and knowledge acquisition that can be adapted to enable conceptual modeling with new and evolving semantic web languages. Protégé-2000 has been used for many years now in the field of medicine and manufacturing.

This is a highly customizable tool as an ontology editor credited to its significant features like an extensible

knowledge model, a customizable file format for a text representation in any formal language, a customizable user interface and an extensible architecture that enables integration with other applications which makes it easily custom-tailored with several web languages.

Even if it permits easier ontology construction, the downside is its requirement of human intervention at regular levels for initial structuring of concepts for its ontology. There are several other applications like the semantic search engine called the SHOE Search.

The WWW Consortium (W3C) has developed a language for encoding knowledge on web to make it machine understandable, called the Resource Description Framework (RDF) [2]. Here it helps electronic media gather information on the data and makes it machine understandable. But however RDF itself does not define any primitives for developing ontologies.

In conjunction with the W3C the Defense Advanced Research Projects Agency (DARPA), has developed DARPA Agent Markup Language (DAML) [3] by extending RDF with more expressive constructs aimed at facilitating agent interaction on the web. This is heavily inspired by research in description logics (DL) and allows several types of concept definitions in ontologies.

The Unified Medical Language System is used in the medical domain to develop large semantic network. In the following section we introduce our approach of knowledge processing, integration and representation for information retrieval [18] problems and eventually discuss the possible solutions.

There are also several researches in the field of natural language processing techniques where conceptual vectors are used to build conceptual networks. Thematic aspects or ideas of textual segments like documents, paragraphs etc. are represented using vectors of interdependent concepts.

Lexicalized vectors have been used in information retrieval for long [21] and for meaning representation by the LSI (Latent Semantic Indexing) model [22] from latent semantic analysis (LSA) studies in psycho-linguistics. In computational linguistics, [23] proposed a formalism for the projection of the linguistic notion of semantic field in a vectorial space. This method is indeed very efficient but the main drawback here is the huge computations involved in determining the conceptual vector values [ ].

3. Our hybrid proposal – extended semantic network (esn)

The basic idea of Extended Semantic Network is to identify an efficient knowledge representation and ontology construction method to overcome the existing constraints in information retrieval and classification problems of information overflow.

To realize this goal we put our ideas into a two segment approach. The first step consists in processing large amount of textual data with the help of our mathematical models thus making our proposal of automatic ontology construction scalable. This model is termed as the recall focussed approach where the network built is mainly designed to show high recall.

The second step involves in manually constructing small semantic networks based on our designed model which in turn is derived from KL-ONE [9, 10]. This model is called the precision model. This is followed by the final process of examining carefully and efficiently the various possibilities of integrating information obtained from our mathematical model with that of the manually developed mind model.

Here, the primary idea is to develop an innovative approach obtained by combination of features from man and machine theory of concept [8], whose results can be of enormous use in the latest knowledge representation, classification, retrieval, pattern matching and ontology development research fields. In this paper we discuss and highlight the methods employed by us for information processing and integration for visualising a novel knowledge representation [7] method for automated ontology construction.

3.1. Proximal network model – Recall focussed approach

The fundamental theory of proximity is concerned with the arrangement or categorisation of entities that relate to one another often believed to favour interactive learning, knowledge creation and innovation. When a number of entities are close in proximity a relationship is implied and if entities are logically positioned; they connect to form a structural hierarchy. Our Proximal Network model is built based on this structural hierarchy, of word proximity in documents.

This approach is largely employed to enable processing of large amount of data [11, 12] in a considerably small time. Another important aspect of this approach is its ability to automatically process the input data into a network of concepts interconnected with mathematically established relations forming a recall focused approach.

The proximal network model involves three phase of processing, firstly the pre-treatment process where the documents related to the domain are analyzed in 2 stages and an output of word document matrix is obtained.

This matrix is then passed on to the intermediate process and is analyzed by the data mining and clustering algorithms namely K-means Clustering, Principle component analysis and Word association to obtain an output of word pair matrix with a value between each word pair. This value is the proximity between the words

pair in the projected space depending on their occurrence in the contents of the documents processed.

This data is further subjected to post-treatment process where partial stemming is carried on the word pair matrix depending on case based requirement.

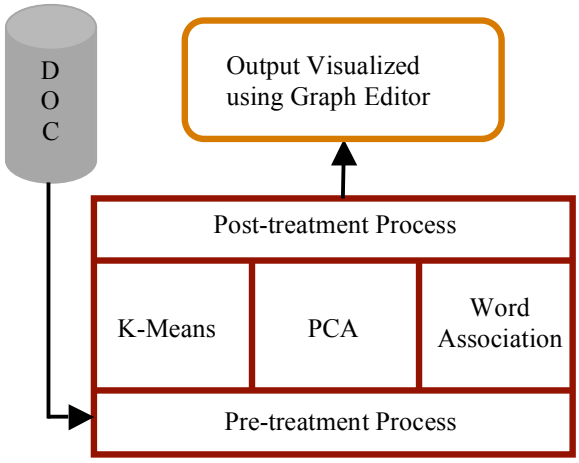

Figure1: Block diagram representing proximal network prototype

Figure2: An extract of proximal network

The output is then stored into a Mysql database and visualized using the Graph Editor, a java application developed by us for visualization and easy editing of networks. Currently, we have successfully processed around 3423 words computing their actual physical occurrence. We have been able to successfully build a proximal network of 50,000 word pair.

Currently the documents processed are related to the research activities carried out in the field

- Arabidopsis thaliana

This program is primarily concerned with the physical distance that separates words. Till date, we have successfully processed around 3423 words computing their actual physical occurrence. We have been able to successfully build a proximal network of 50,000 word pair, an extract of which is seen in figure 2. Each of these word pair is related using the value obtained from the prototype and is visualised using the simple UML link of association [10, 12].

The proximal network is however independent of the input information. This data processing method in itself can be independently used for data processing and representing knowledge in various domains irrespective of the input data. In comparison with the NLP techniques this method builds conceptual networks much faster with good recall.

For example if one needs to build a network of football players from France then it is sufficient to provide document concerning to the game and its players. The proximal network will automatically build a network on all related football players both French and international players.

Here, each node of this network is distanced based on the mathematical calculations which take mush less time (test results to be released soon) to analyse and build a network of 40,000 and above word pair. The fact that the small time taken for processing huge amounts of data makes it an important aspect in ontology construction representing multiple domain scalable.

3.2. Semantic network model – Precision focused approach

Technically a semantic network is a node- and edge-labeled directed graph, and it is frequently depicted that way. The scope of the semantic network is very broad, allowing semantic categorization of a wide range of terminology in multiple domains. Major groupings of semantic types include organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas.

The links between the semantic types provide the structure for the network and represent important relationships. Our semantic network is based on the KL-ONE model, with domain being the centre of our network which is expatiated by the domain components which in turn define concepts using the instance and inheritance relations [15, 16, 17].

We follow the scheme process [18] where minimum required information on a domain is precisely represented using the semantic relations defined above. The model is built based on the same set of documents used in proximal network and the 50 most important concepts is chosen with the help of a domain expert and is put into the semantic model.

Here each relational link used namely the compositional, instantiation and inheritance links are given a predefined unit during calculation. This model is then stored and can be visualized using graph editor.

In our semantic network prototype we reuse the documents pertaining to each field and then choose a set

of concepts representing the field in consideration. This is achieved with help of researchers of ToxNuc-E. This list of concepts pertaining to each field is furnished to the specialists who in turn rate each concept with respect to its importance in representing the field.

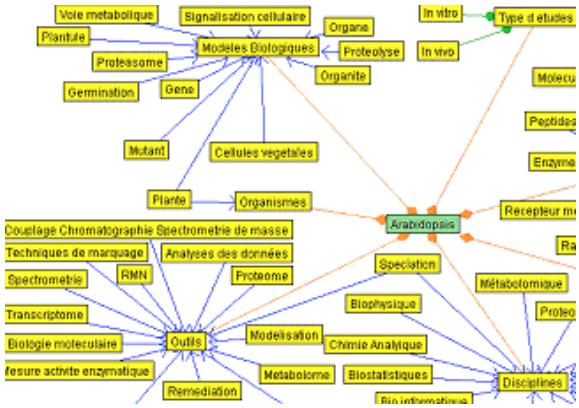

Figure 3: An extract of semantic representation of concepts for arabidopsis using graph editor

We then choose the first 50 concepts [19] most representing the field from the above list and provide it to people who were either specialists or people possessing good level of knowledge in each of these study area accompanied with our relational links.

All the links used in connecting the node is based on the UML [10] links, consisting of four different types of associative lines. The concept network thus built is based on the meaning each concept pair share, with a possibility of more than one relationship between a single pair of connected nodes.

They have been currently chosen on an experimental basis [12], after proper consideration and also analyzing the requirements of our approach. We start with our domain name representing the super class in our approach. The super class is then connected to its subclasses based on the category of the relation they share, which can be chosen from the four links we provide.

The objective here is to introduce the semantic based relation into our mathematically modeled proximal network. The network thus developed is then analyzed and merged to obtain one single semantic network for that domain. This process is repeated on different lists of concepts concerning to various domains to obtain one network for each domain. The result obtained is fed into the Mysql database along with the relational links they share.

In the above figure we see an extract of semantic network built on the domain arabidopsis with the help of experts of the domain in question. Here, the links between the concepts are used based on the relationship that the nodes share between them.

This network is then readily combined with the proximal network thus forming the extended semantic network. This network mainly works on the lines of precision. Here the position of each concept and its relation is calculated very precisely based on several data and expertise advice.

3.3. Integration of mind and mathematical models to obtain ESN

The 2 different resulting models obtained by our methods dilated in the previous sections are combined employing the simple extension methods. We retain our semantic network model as the basic architecture of our knowledge representation model. Here the common nodes between the two networks are identified by our algorithm and this is used to expand the semantic network model using results from the proximal network model.

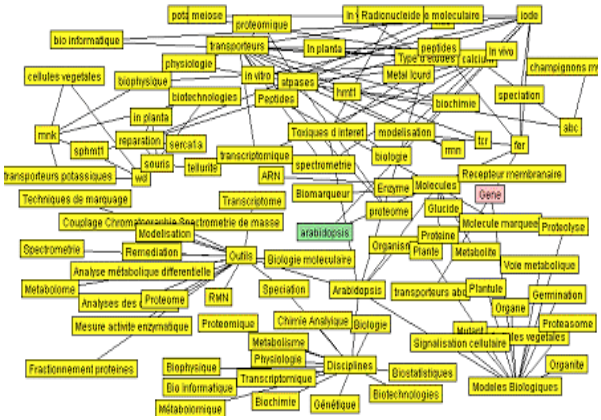

Figure4. An Extract of Extended Semantic Network Visualised using Graph Editor

The extension is defined based on the proximity value shared by the connected words. Presently we have limited this to a level of 5 extensions i.e. only the first 5 level of nodes are added from the proximal network. The directional flow of the process is restrained where the relational flow is possible only from a lower level node to the upper level node.

The figure 6 illustrates this idea where the relational flow is possible from S0, S1 node towards S2 but not vice versa. Here S0 represents the first level, S the second and so on.

Simultaneously, several other optimizing algorithms are being considered to be utilized in merging the networks to build the Extended Semantic Network. We are exploring the possibilities of using the genetic algorithms and features of neural networks to obtain an optimal result. Our present results are in the process of testing and verification.

Initial results have proved to be encouraging when verified by experts in comparison with human developed

ontology and concept networks and has been validated for providing satisfactory results.

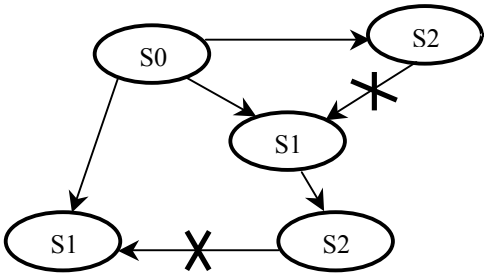

Figure5: Relational flow illustration

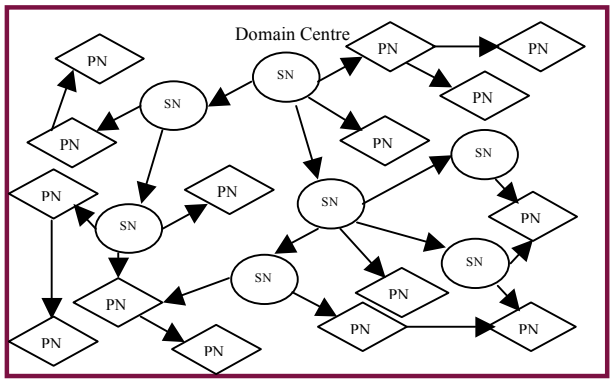

Figure6. Extended Semantic Model

The ESN prototype thus targets at initialising a new method for knowledge representation for easy ontology construction which can be employed in new generation search algorithm to facilitate information management, retrieval and sharing.

This prototype enables easy construction of conceptual networks. Unlike natural language processing techniques here no heavy computations are required. In order to develop new networks we just need a set of documents related to that particular topic.

These documents are required to be only input into the proximal network program. This automatically develops a network of nodes called the word network. This network basically contains all the different words that can be found in the input data related to the domain. Thus this forms a recall process network.

This network is then combined with the semantic network. The network being restricted to 50 nodes can be easily developed using the model provided by us. Semantic network is essentially the precision network where the nodes are placed in the network with the help of expert knowledge.

Thus when constructing the extended semantic network we just extend the precision model by adding nodes from the recall model at possible and required positions. This in turn combines the 2 networks to provide the extended semantic network.

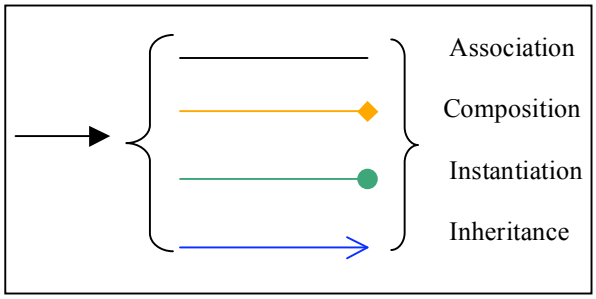

Figure 7: Links used in extended semantic network

4. Application platform

The Extended Semantic Network prototype has been developed in collaboration with the ToxNuc-E project funded by CEA (Commissariat à l'Energie Atomique). ToxNuc-E [13], is a project devoted to all the research activities carried out for controlling nuclear environmental toxicology in the living environment with several research centres like CNRS, INSERM etc involved.

It is a platform where researchers from different domains like biology, chemical, physics and nuclear, across Europe working for a common purpose, meet and exchange their views on various on-going research activities related to nuclear toxicology. The research involves in

The ToxNuc-E presently has around 660 researchers registered with their profile, background and area of research interest and are geographically displaced. Our research is applied in this platform to provide these researchers knowledge representation tool like ESN which can be easily utilized by non specialists in developing project based ontology quickly and efficiently.

5. Future work

Currently, we are experimenting on the 3 topics chosen by the researchers as the domain of major research activities. The data and the documents used in our experimental prototype of ESN are obtained from the ToxNuc-E platform. We soon intend to extend our research to all 15 research fields of ToxNuc-E.

The results of our algorithm have been subjected to testing, by human experts and have been judged to provide results very close to human constructed concept networks with reduced time of construction and very cost effective. Another important feature of ESN is its ability to customise to user needs [14] and equally providing results very close to NLP-based indexing methods

without heavy computations i.e. if a user needs specific information on specific subject it is adequate to change the input documents for the proximal network. Based on these documents the entire network is reconstructed in a time span of approximately 30 minutes.

The principle advantage of our methodology with respect to the previous work is our innovative hybrid approach of integrating machine calculations with human reasoning abilities. We use the precise, non estimated results provided by human expertise in case of semantic network and merge them with the machine calculated knowledge network from proximal results. The fact that we try to combine results from two different aspects forms one of the most interesting features of our current research.

We view our result as structured by mind and calculated by machines. One of the major drawbacks of this approach is finding the right balance for combining the concept networks of semantic model with the word network obtained from proximal model.

Our future work involves in identifying this accurate combination between the two vast methods and setting up a benchmark to measure our prototype efficiency. Our next step will be to include natural language processing techniques and lemmatises to our pre-treatment process.

Our objective is to develop an application for document classification and indexation based on the results of Extended Semantic Network. This application library is intended to be used for classification purpose in the project ToxNuc-E for better data management on the platform.

We also plan to include user modeling [14] features by monitoring the behavior; interests and research works carried out by the members of ToxNuc-E and then build a model unique to each user. This model consecutively builds a profile for each user and sequentially stores the details obtained in a database. These details can be utilized to better understand the user requirements thus helping the user in efficient data search, retrieval, management, and sharing.

Some of the major points we hope to achieve through this method of knowledge representation network are

- To make construction of semantic based concept networks cost effective by campaigning minimum human intervention. In turn reducing the construction time using mathematical models.
- To identify a good balance between mind and mathematical models to develop better knowledge representing networks with good precision and high recall.

6. Conclusion

The question on knowledge representation, management, sharing and retrieval are both fascinating and complex, essentially with the co-emergence between man and machine. This research paper presents a novel hybrid approach, specifically in the context of knowledge representation and retrieval. The proposal is to attempt at making ontology construction faster and easier. The advantages of our methodology with respect to the previous work, is our innovative approach of integrating machine calculations with human reasoning abilities thus creating a hybrid approach.

We use the precise, non estimated results provided by human expertise in case of semantic network and then merge it with the machine calculated knowledge from proximal results. The fact that we try to combine results from two different aspects forms one of the most interesting features of our current research.

We view our result as structured by mind and calculated by machines. One of the major drawbacks of this approach is finding the right balance for combining the concept networks of semantic network with the word network obtained from the proximal network.

We are also looking forward to release our results after subjected to verification and validation by several experts in the domain. We are also working towards making a beta version of our complete algorithms to be tested by various research groups during their research activities.

7. References

[1] T.R. Gruber, "Toward Principle for the design of ontologies used for Knowledge Sharing", in Proc. Of *International Workshop on Formal Ontology*, March 1993.

[2] Brickley, D. and Guha, R.V. Resource Description Framework (RDF) Schema Specification. Proposed Recommendation: *World Wide Web Consortium*, 1999.

[3] Helder, J. and McGuinness, D.L., The DARPA Agent Markup Language. *IEEE Intelligent Systems*, 2000.

[4] Natalya F. Noy, Michel Sintek, Stefan Decker, onica Crubézy, Ray W. Fergerson and Mark A. Musen, Creating Semantic web Contents With protégé 2000, Stanford University, *IEEE Intelligent Systems*, 2001.

[5] J.F Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

[6] N. Cuarino, C. Masolo, and G. Vetere, "Ontoseek: Content-based Access to the Web," *IEEE Intelligent Systems*, Volume 14, no. 3, pp. 70-80, 1999

[7] M.R Quillian,, Semantic memory. M Minsky, *Ed, Semantic Information Processing*. pp.216-270. Cambridge, Massachusetts: MIT Press, 1968.

[8] J.F Sowa, Conceptual structures: information processing in mind and machine, *Addison-Wesley Longman Publishing Co.*, Inc, Boston, MA, 1984.

[9] J Brachman, L Deborah, McGuinness, F Patel-Schneider, A Resnick Living with CLASSIC: When and How to Use a KL-ONE-Like Language, *Special issue on implemented knowledge representation and reasoning systems Pages: 108 – 113*, ACM Press, NY, USA,1991.

[10] J. Brachman, G. Schmolze, An Overview of the KL-ONE Knowledge Representation System, Cognitive Science 9(2), pp 171-216, 1985.

[11] M.E Winston, R Chaffin and D Hernnann, A taxonomy of part – Whole Relations *Cognitive Science 11*, 1987.
[12] N.J Belkin, W.B Croft, Information Filtering and Information Retrieval: Two Sides of the Same Coin?, *Communications of the ACM Vol. 35 n°12*, 1992
[13] M Ménager, Programme Toxicologie Nucléaire Environnementale : Comment fédérer et créer une communauté scientifique autour d'un enjue de société , *Intelligence Collective Partage et Redistribution des Savoirs, Nimes*, France, septembre, 2004.
[14] J Aberg & N Shahmehri, User Modelling an Aid for Human Web Assistants, User Modeling 2001: *8th International Conference*, UM 2001, Southaven, Germany, July 13-17, 2001.
[15] Natalya F. Noy and Deborah L.McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, *Ontology Tutorial*, Stanford University, Stanford, CA.
[16] Alexander maedche & Steffen Staab, "Ontology Learning for the Semantic Web", *Volume 16 IEEE Intelligent Systems*, 2001
[17] E Rosch Cognitive Representation of Semantic Categories, University of California, Berkeley, 1978
[18] Umberto Eco, Kant and the platypus, essays on language and cognition
[19] Prince, V. Lafourcade, M., Mixing semantic networks and conceptual vectors application to hyperonymy, Systems, Man and Cybernetics, Part C, IEEE Transactions on**.**
[20] Simon Polovina and John Heaton, "An Introduction to Conceptual Graphs," AI Expert, pp. 36-43, 1992.
[21] G. Salton, C. Buckley, and E. A. Fox. *Automatic query formulations in information retrieval*. Journal of the American Society for Information Science, 34(4):262-280, July 1983.
[22] Deerwester S. et S. Dumais, T. Landauer, G. Furnas, R. Harshman. Indexing by latent semantic anlysis. *In Journal of the American Society of Information science*, 1990, 416(6), pp 391-407.
[23] Jacques Chauch´e. D´etermination s´emantique en analyse structurelle : une exp´erience bas´ee sur une d´efinition de distance. *TAL Information*, 31/1, pp 17-24, 1990.