Hybrid Method for Knowledge Processing Integration and Representation

Reena Shetty, Pierre-Michel Riccio, Joël Quinqueton

Extended Semantic Network for Knowledge Representation

Reena T. N. Shetty	Pierre-Michel Riccio	Joël Quinqueton
EMA, Fontenblu, Paris	*Researcher, LGI2P, Nimes*	*Professor-LIRMM, 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 66 38 70 48*

Abstract

The proposition Extended Semantic Network is an innovative tool for Knowledge Representation and Ontology construction, which not only infers meanings but looks for sets of associations between nodes as opposed to the present method of keyword association. Our objective here is to achieve semi-supervised knowledge representation technique with good accuracy and minimum human intervention, using the heuristically developed information processing and integration methods. This research is realised by obtaining a technical co-operation between mathematical and mind models to harvest their collective intelligence.

1. Introduction

The past few years has witnessed tremendous upsurge in information availability in the electronic form, attributed to the ever mounting use of the World Wide Web (WWW). For many people, the World Wide Web has become an essential means of providing and searching for information leading to large amount of data accumulation. Searching web in its present form is however an infuriating experience for the fact that the data available is both superfluous and diverse in form. Web users end up finding huge number of answers to their simple queries, consequentially investing more time in analyzing the output results due to its immenseness. Yet many results here turn out to be irrelevant and one can find some of the more interesting links left out from the result set.

One of the principal explanations for such unsatisfactory condition is the reason that majority of the existing data resources in its present form are designed for human comprehension. When using these data with machines, it becomes highly infeasible to obtain good results without human interventions at regular levels. So, one of the major challenges faced by the users as providers and consumers of web era is to imagine intelligent tools and theories in knowledge representation and processing for making the present data, machine understandable.

Several researches has been carried out in this direction and some of the most interesting solutions proposed are the semantic web based ontology to incorporate data understanding by machines. The objective here is to intelligently represent data, enabling machines to better understand and enhance capture of existing information. Here the main emphasis is given to the thought for constructing meaning related concept networks [16] for knowledge representation. Eventually the idea is to direct machines in providing output results of high quality with minimum or no human intervention.

In recent years the development of ontology [1, 7] is gaining attention from various research groups across the globe. There are several definitions of ontology purely contingent on the application or task it is intended for. Ontology is one of the well established knowledge representation methods; on a formal ground ontology defines the common vocabulary for scientists who need to share information on a field or domain. One has seen in the past years that various research groups have been devotedly experimenting semantic related [16] ontology aimed at making web languages machine understandable.

Given the practical and theoretical importance of ontology development, it is not surprising to find a large number of enthusiastic and committed research groups in this field. Here, the section 2 of this paper highlights some of the researches carried out in the field and attempt to give a brief review of these studies. The section 3 introduces our approach to this problem and exemplifies our algorithm continued with our results and prospective. Finally the paper ends with our conclusion and drawbacks on our approach and acknowledgement to one and all concerned.

2. Related work

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 main means of knowledge representation through its ability to represent data with respect to semantic relation 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 structuring the concepts for its ontology.

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.

There are several other applications like the semantic search engine called the SHOE Search. 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, representation and integration for information retrieval [18] problems and eventually discuss the possible solutions.

3. Hybrid approach - 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. To realize this we put our ideas into practice via a two phase approach. The first phase consists in processing large amount of textual information using mathematical models to make our proposal scalable. The second phase consists in examining carefully and efficiently the various possibilities of integrating information obtained from our mathematical model with that of the manually developed mind model.

The first phase of our proposal is carried out by realising a network of words mathematically computed using different statistical and clustering algorithms. Thus creating a proximal network computationally developed, depending essentially on word proximity in documents. The second phase is ensured by a heuristically developed method of network extension using the outputs from the mathematical approach. This is achieved by considering the manually developed semantic mind model as the entry point of our concept network.

Here, the primary idea is to develop a innovative approach obtained by combining the features of 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 used by us for information processing and integration aimed at visualising a novel method for knowledge representation [5] and ontology construction.

3.1. Proximal network for efficient data processing

Proximity is the ability of a person or thing to tell when it is near an object, or when something is near it. This sense keeps us from running into things and also can be used to measure the distance from one object to another object. The simplest proximity calculations can be used to calculate distance between entities thus avoiding a person from things he can hit. Proximity between entities is often believed to favour interactive learning, knowledge creation and innovation. The basic theory of proximity is concerned with the arrangement or categorisation of entities that relate to one another. 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. This concept is largely used in medical fields to describe human anatomy with respect to positioning of organs.

Our Proximal Network Prototype model is built based on this structural hierarchy, of word proximity in

documents [12]. This approach is mainly employed to enable processing of large amount of data [8] 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.

For building this prototype we systematically employ three phases for identifying our data to build the final network. We first start with a set of documents related to 3 major fields out of the 15 fields in the nuclear environmental toxicology domain, furnished by the project ToxNuc-E. The documents obtained are first converted into simple txt format using an external converter and is later fed into our first stage called the pre-treatment process. This process is carried out in different approach which is not dealt with in this paper. Here, in this process the input document is processed in several stages and an output of word frequency matrix is created with rows representing the words and columns representing the document name.

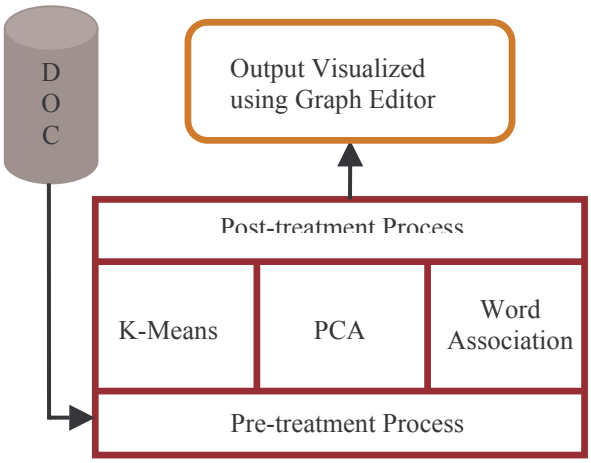


Figure1: Block diagram representing Proximal Network prototype

This word document matrix is then passed on as input to the 3 algorithms called the K-means, Principal component analysis and Word association. The algorithms have been modified accordingly to our processing criteria by adding certain a additional computations. Java has been used as the programming language for these, where each algorithm provides an output in the form of a word pair matrix with the mathematical values representing the relational weight between the word pair. The outputs from all the algorithms are then combined using the simple mean calculation and thus a single value for each word pair is estimated.

The output is then subjected to the post-treatment process where partial stemming is carried out with an objective of not losing the important information during the stemming process. The output from the previous step is then stored into a Mysql database. This data can be later visualised using the Graph Editor, a java application developed for visualisation and easy editing of networks

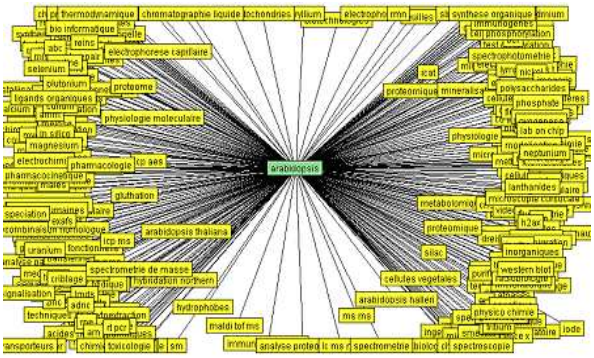


Figure2: An Extract of Proximal Network

Currently the documents processed are relating to the research activities carried out in the chosen 3 fields namely

- Arabidopsis thaliana,
- Alteration and reparation of DNA and
- Methodology and speciation

This program is primarily concerned with the physical distance that separates words in a double dimensional space. 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, 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].

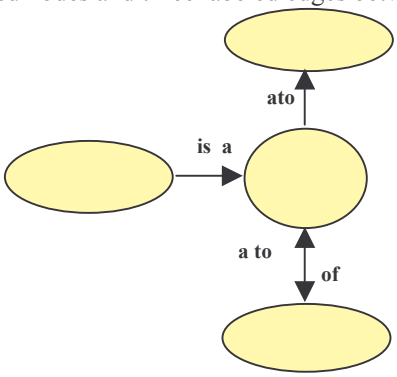
This data processing method in itself can be independently used for processing and representation data in various domains. The fact that the small time taken for processing huge amounts of data makes it an important aspect in ontology construction representing multiple domain scalable.

3.2. Semantic network prototype

Semantic Network [7, 6, 17] is basically a labelled, directed graph permitting the use of generic rules, inheritance, and object-oriented programming [11]. It is often used as a form of knowledge representation. It is a directed graph consisting of vertices, which represent concepts and edges, representing semantic relations between the concepts. The most recent language to express semantic networks is KL-ONE [9].

There can exist labeled nodes and a single labeled edge relationship between Semantic nodes. Further, there can be more than one relationship between a single pair of

connected words: for instance the relationship is not necessarily symmetrical and there can exist relationship between the nodes through other indirect paths. Below is a fragment of a conventional semantic net, showing 4 labeled nodes and three labeled edges between them.



Technically a semantic network is a node- and edge-labelled directed graph, and it is frequently depicted that way. The scope of the semantic network is broad, allowing for the 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.

In our semantic network prototype we reuse the documents pertaining to each field and then choose a set of 50 concepts most representing the field. This is achieved with help of specialists and researches of ToxNuc-E. We list a set of concepts pertaining to each field and then provide it to the specialists who in turn rate each concept with respect to its importance in representing the field.

We then choose the first 50 concepts [19] most representing the field 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 as shown in figure 4. 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 analysing 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 four links represent the simple UML links of association, composition, instantiation and inheritance.

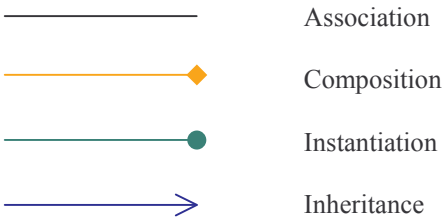


Figure 4: Links used in Semantic Network

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

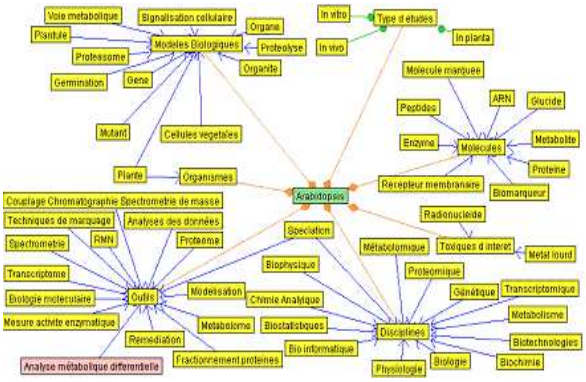


Figure 5: An Extract of Semantic representation of Concepts for arabidopsis using Graph Editor

The figure 5 shows a semantic network developed on the project arabidopsis. This is our first version of semantic network which is now being upgraded with new values and other relational links based on results.

3.3 Integration of mind and mathematical model to obtain ESN

The 2 different result models obtained by our methods elaborated in the previous sections are combined with simple extension methods. We call this the hair extension method. Here we start with semantic network retaining all its nodes as the starting core of our network. This network

is then expatiated by adding nodes from proximal network.

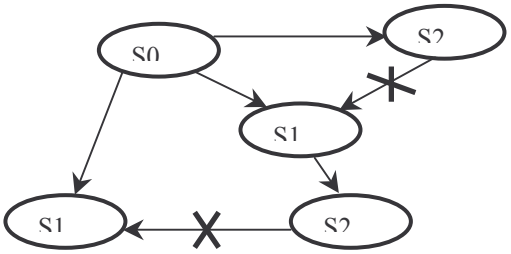


Figure 6: Relational flow illustration

Here we identify the common nodes between the two models and then add on the rest of the network node related to this node in the proximal network. Presently we have limited this to a level of 5 extensions i.e. only the next 5 level of nodes are added from the proximal network. We also use the process 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.

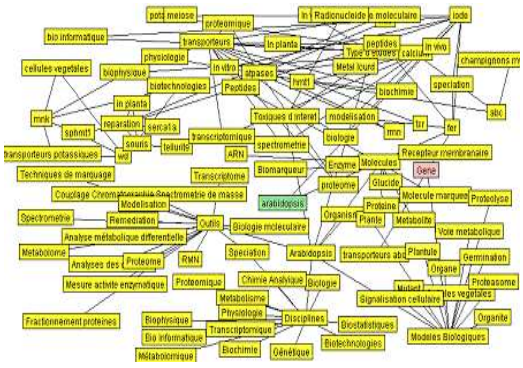


Figure 7: An Extract of Extended Semantic Network Visualised using Graph Editor

Simultaneously, several other optimising algorithms are being considered to be utilised 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 have been verified by experts in comparison with human developed ontology and concept networks and validated for providing satisfactory results.

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.

3.3.1. Application on environmental nuclear toxicology. 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 several research centres like CNRS, INSERM etc linked with CEA. It is a platform where researchers from different domains like biology, chemical, physics and nuclear, working for a common purpose, meet and exchange their views on various on-going research activities. Related to nuclear toxicology.

The ToxNuc-E presently has around 660 researchers registered with their profile, background and area of research interest. The objective of our research is to assist these researchers to achieve better knowledge representation and to support for easy information retrieval from the vast data base of information. Currently we are experimenting on the 3 topics or domain chosen by the researchers as the domain of major research activities. All the data and the documents used in our experimental prototype of ESN are obtained from the ToxNuc-E platform. We will soon extended our research to all 15 research fields of ToxNuc-E.

3.3. Results and Future Work

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. It has also proved to take much less time for construction and very cost effective. We are on the conclusion that the results are exceedingly encouraging in terms of accuracy. Our next step will be to include natural language processing techniques like stemming 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 modelling [14] features by monitoring the behaviour; 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.

4. 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 collaborative working method, 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.

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. Our future work would be to identify this accurate combination between the two vast methods and setting up a benchmark to measure our prototype efficiency.

5. Acknowledgement

We would like to use this opportunity to thank all the researchers of ToxNuc-E for providing us with their data and expertise. We would also like to thank the reviewers of this paper for their useful comments.

6. 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] Rational Corporation: UML Notation Guide 2, 2000.

[11] M.E Winston, R Chaffin and D Hernnann, A taxonomy of part – Whole Relations *Cognitive Science 11*, 1987.

[12] S.A. Mahé, P.M. Riccio et S. Vailliès: des elements pour un modèle: la lutte des classes! *Revue Génie Logiciel, n°58*, Paris, septembre 2001.

[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] 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

[19] R Davis, B.G Buchanann, Meta-Level knowledge: Overview and applications, *IJCAI, ACM SIGIR, n° 5*, Cambridge, 1984.