EXTENDED SEMANTIC NETWORK FOR KNOWLEDGE REPRESENTATION

An Hybrid Approach

Reena T. N. Shetty, Pierre-Michel Riccio, Joël Quinqueton
Doctorate-EMA Paris, Assistant-professor-LGI2P Nimes, Professor- LIRMM Montpellier

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. The objective here is to achieve semi-supervised knowledge representation technique with good accuracy and minimum human intervention. This is realized by obtaining a technical co-operation between mathematical and mind models to harvest their collective intelligence.

Key words: Extended Semantic Network, Artificial Intelligence, Collective Intelligence, Proximal Network, Semantic Network, User Modeling, Knowledge, Knowledge Representation & Management, Information Retrieval .

1. INTRODUCTION

The past few years has witnessed tremendous upsurge in data 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 since the data available is surplus and in diverse forms. 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

Please use the following format when citing this chapter:

Shetty, R.T.N., Riccio, P.-M., Quinqueton, J., 2006, in IFIP International Federation for Information Processing, Volume 228, Intelligent Information Processing III, eds. Z. Shi, Shimohara K., Feng D., (Boston: Springer), pp. 135–144.

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 [17] 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 [2, 8] 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 [17] ontology aimed at making web languages machine understandable.

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 [17, 25] 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 [5], 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 customisable tool as an ontology editor credited to its significant features like an extensible knowledge model, a customisable file format for a text representation in any formal language, a customisable 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 of 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) [3]. 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 Defence Advanced Research Projects Agency (DARPA), has developed DARPA Agent Markup Language (DAML) [4] 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 to this problem of knowledge representation, management and information retrieval [19] and eventually discuss the possible solutions.

3. HYBRID APPROACH- EXTENDED SEMANTIC NETWORK(ESN)

3.1 General View

Extended Semantic Network is data representing network resulting from the collaboration involving two networks, one automatically constructed proximal network and the second manually constructed semantic network. Here, the primary idea is to develop a modern approach combining the features of man and machine theory of concept [9], which can be of enormous use in the latest knowledge representation, classification, pattern matching and ontology development fields. We propose to visualize a novel method for knowledge representation [6] partly based on mind modeling and partly on the mathematical method.

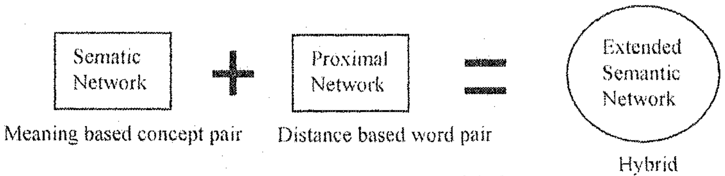

Figure1. Schematic Representation of ENS

In ESN we endeavor to develop a network of concepts based on human constructed semantic network projected as the main central part of the network which is later subjected to elaboration utilizing the statistical data obtained by our mathematical models based on the data clustering and mining algorithms. This generates a new approach for knowledge representation which can later be used for optimising, information search and classification procedures, and enabling easy and fast information retrieval. The ESN forms a hybrid structure [22] by inheriting the features of both the source networks; computed differently and independently, making it a robust and an optimal approach.

Our proposal is to construct a network of concepts similar to ontology but using a method where minimal human intervention is required. We call this a semi supervised network of concepts representing certain qualities of an ontology which later is expatiated by adding the information obtained from the mathematically elaborated proximal network. Our assumption here is that this method will produce the same output as any traditional ontology but will greatly decrease the construction time, attributed to its mathematical modeled extension. 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
- To minimise time invested in construction by introducing mathematical models without loosing on quality.
- To identify a good balance between mind and mathematical models to develop better knowledge representing networks with good precision and high recall.

3.1.1 Semantic network

Semantic Network [8] is basically a labelled, directed graph permitting the use of generic rules, inheritance, and object-oriented programming [9]. 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 [10].

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 labelled nodes and three labelled 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.

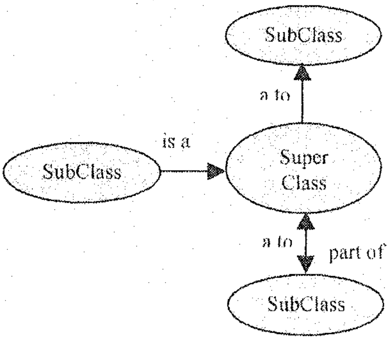

Figure 2 . Multi-labeled Semantic Relation

In our semantic network prototype all concept relations are built based on the meaning each concept pair share, with a possibility of more than one relationship between a single pair of connected nodes. All the links used in connecting the node is based on the UML [11] links, consisting of four different types of associative lines as shown below.

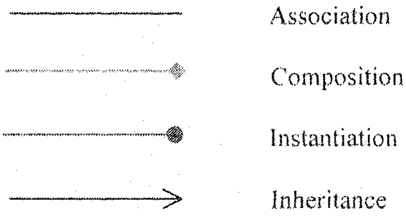

Figure 3. Links used in Semantic Network

They have been currently chosen on an experimental basis [12], after considering and analysing the requirements of our approach. We start with our domain name representing the super class in our approach. The super class is 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.

3.1.2 Proximal network

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. The Proximal Network Prototype model is built based on this structural hierarchy, of word proximity in documents [13]. Here proximity is calculated purely considering the physical distance of its occurrence at a given instance. We use UML link of association to connect words or nodes proximally closer.

Results obtained from the semantic network are considered as the centre of our network on which the ESN network will be constructed. We extend the results of semantic network by adding on the results obtained by the proximal network thus making it an Extended Semantic Network. The demonstrable prototype of ESN has been developed based on the data of ToxNuc-E project [14].

3.2 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[14], is a project devoted to all the research activities carried out in Biological, Chemical and Nuclear domains in several research centres 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 nuclear toxicology related on-going research activities.

The ToxNuc-E [14] 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 15 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.

3.2.1 Semantic network prototype

Our semantic network prototype is developed grounded on the view of a set of specialist representing each of the chosen domains from the project ToxNuc-E. To begin with, we choose a set of 50 concepts pertaining to the preferred domain of research. We then consult people who are either specialists or people possessing good level of knowledge in each of these areas of study.

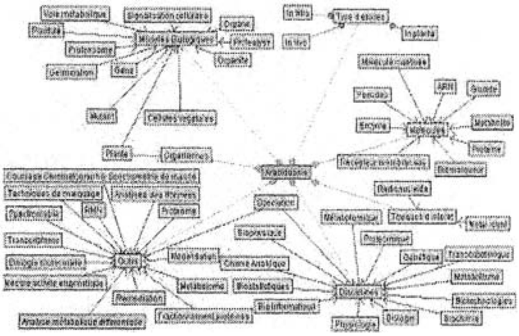

Figure 4. An Extract of Semantic representation of Concepts for arabidopsis using Graph Editor

These people are provided with the concept list on which they are requested to develop a semantic network depending on their individual view point. The network thus developed is 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 semantic network is then stored into the MySql database and visualized using graph editor - a java application developed by us and used for facilitating construction of networks and also for editing purpose.

3.2.2 Proximal network prototype

The documents relating to the numerous research activities being carried out in the chosen 15 field of nuclear toxicology in plants and animals forms our data. These documents are subjected to a pre-treatment process to obtain a matrix of words and documents as rows and columns respectively. Here java is used as the programming language and all the data used are stored in the MySql database.

This program is primarily concerned with the physical distance that separates words in a given space. Currently, we have successfully processed around 3423 words to calculate the physical distance between them, using various mathematical algorithms. The result obtained here is then fed into the graph editor for graphical visualisation. This helps in displaying the results from the program along with a value calculated for every word pair – in this case 50,000 word pair, forming the proximal

Network. These results are then stored into the database which is later used to combine with the results of the semantic network.

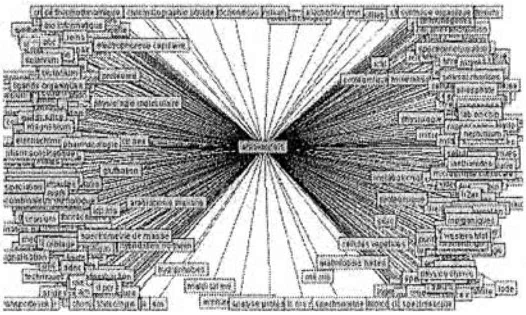

Figure5. An Extract from Proximal Network

At present, the 2 different results are combined with simple extension methods. 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.

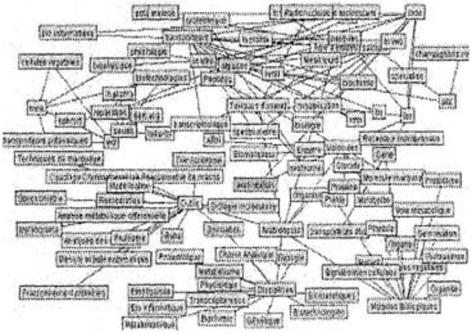

Figure6. An Extract of Extended Semantic Network

Our preliminary results have been verified by experts in comparison with human developed ontology and concept networks and have been validated for providing satisfactory results. We are now working on live data from ToxNuc-E to develop an ESN network to be later compared with classical ontology and rated by domain experts for attaining our benchmark.

3.3 Advantages 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 also on the conclusion that the results are exceedingly customisable depending on the user's domain of interest. 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 [15] 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 into 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.

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 attempts to present an hybrid knowledge representation approach accurate as ontologies but faster and easy to construct. The advantages of our methodology with respect to the previous work, is our innovative approach of combining 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.

ACKNOWLEDGEMENTS

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.

REFERENCES

[1] Natalya F. Noy and Deborah L.McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, Stanford University, Stanford, CA.

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

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

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

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

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

[7] J Voss, P Danowski, B Chapter - citebase.eprints.org, 2005.

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

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

[10] J Brachman, L Deborah, McGuinness, F Patel-Schneider, A Resnick Living with CLASSIC: When and How to Use a KL-ONE-Like Language, 1991. [11] Rational Corporation: UML Notation Guide 2, 2000.

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

[13] 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.

[14] 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.

[15] 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.

[16] E Reingold, J Nightingale, "Artificial Intelligence".

[17] Alexander maedche & Steffen Staab, "Ontology Learning for the Semantic Web".

[18] J Link-Pezet., P Glize, C Régis,, A cognitive approach to intelligent databases. On line, London: Learned Information, 1992.

[19] 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

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

[21] A Maedche, S Staab, Representation & Learning, IEEE Intelligent Systems, 2001.

[22] E Rosch and B. Mervis, Family Resemblances: Studies in the Internal Structure of Categories, University of California, Berkeley, 1989

[23] E Rosch Cognitive Representation of Semantic Categories, University of California, Berkeley, 1978

[24] E Rosch "Cognitive Reference Points", University of California, Berkeley, 1978.

[25] 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