An Immune-based Approach to Knowledge Acquisition

Andrés Romero Rodríguez
Universidad Nacional de Colombia
Bogotá, D.C.
caromeroro@unal.edu.co

Luis Fernando Niño
Universidad Nacional de Colombia
Bogotá, D.C.
lfninov@unal.edu.co

ABSTRACT

Keywords

1. INTRODUCTION

Today, the generation of textual information grows considerably; thus people and organizations begin working with huge amounts of data, such data can be in the form of working papers, corporate documents, e-mail and many other forms. This situation makes necesary the use of computational tools that facilitate the management of such amount of information in a reliable, secure and efficient way so that this information becomes helpful to the daily work of people instead of being a headache beacause of its difficult managenent.

Tools used curently do not solve totally the problem of information organization, because it only provides a way for information storage; such storage can be in the form of a database, bibliography management systems, e-mail organizers among others; but, even when the access to such information has been made very efficient, we still don't have adequate techniques that allow us to take advantage of the tacit knowledge contained behind the stored information to improve the user performance in the daily tasks that involve the use of such information. For this reason, the management of such amounts of information is becoming into a serious problem not only for a personal level, but also for a corporate one; because every people must fight everyday with their own information that in some cases become difficult to manage, from individual e-mail, to information management in the enterprises.

In recent years, there have been developed several techniques and methodologies focused in the adequate management of this kind of information; in these are included the text categorization and classification systems, which are designed to assign each one of the documents in the collection to a category from a specified set. Information extraction techniques allow to obtainning important information for the user in a specified topic upon a generic set of documents; the semantic web is focused in the organization of the information contained in the world wide web in a way that the knowledge contained in the documents is kept.

The problem that arises from this, is those of the information management, personal as much as corporate; so that the users do not have to worry about the storage, processing and search for the appropriate information in the adequate moment. Such search for information, is driven by the individual interests of the users, such interests can depend on the rol of the person within the company.

There are several techiques in the field of computational intelligence that can be used in order to deal with this problem, such as information extraction, documento classification, and methodologies that facilitate the representation and processing of the tacit knowledge stored into the documents. In using these techniques, there must be defined mechanisms that enable us to manage such knowledge in a way that it can be useful in the organizations and for the individuals; the main steps in order to achieve this process are:

1. Knowledge extraction from several sources
2. Representation and storage of the extracted knowledge
3. Manipulation of such knowledge and conversion into other forms useful for the users
4. Delivery of the processed knowledge to the appropriate people

2. KNOWLEDGE REPRESENTATION MODELS

Semantic information refers to the tacit knowledge stored within the documents, such knowledge is generally hidden inside the text and it is difficult to access. There are needed techniques that allow to find and represent the stored knowledge in the documents, in [12] are defined 5 principles that must be acomplished by a knowledge representation, these are:

1. A knowledge representation is a substitute, it means that inside the entity that stores the knowledge, there is a reasoning process, such as the stored knowledge is a substitute of the real world objects.

2. A knowledge representation is a set of ontologic agreements: every posible representations for an object have certain degree of error, which can lead to undesired effects in the knowledge manipulation process. For this reason, there must be reached an agreement about these representation so that such error can be avoided.

3. A knowledge representation is a fragmented theory of intelligent reasoning: this is based on the idea that knowledge is the base for the human intelligent reasoning. A representation for such knowledge won't be complete because of the limitations of the computing systems; for this reason, it is not posible a full intelligent reasoning system, instead of that, it must be fragmented.

4. It is a way for efficient computation: those representations typically offer a set of ideas that facilitates the inference process; the effectiveness of this process depends completely upon the efficiency with the stored knowledge is processed.

5. A knowledge representatin is a way for human expression: this is the way to express about the real world, the way to express concepts about the world to the machines.

With these characteristics for a knowledge representation, such knowledge must be obtained from a collection of documents containing related information about a particular domain, for this, document clustering and classification techniques must be used, also methodologies that enable the representation and processing of the knowledge stored in these documents. Typically, this information is represented using a set of keywords, semantic networks and ontologies; furthermore, it is useful to have previous information about the context of interest, posibly using public domain ontologies, as in [6].

In [3] are defined the two principal elements in the acquisition of a language that is useful in the stored knowledge management process, these are:

1. Lexicon building, this is the knowledge about the words.

2. Generatin of relationships between the words contained in that lexicon, this can be done using a semantic network or an ontology.

2.1 Ontologies

An ontology is defined as a *formal and explicit specification of a shared conceptualization* [2]; this definitions means that the ontology represents a common understanding about some domain so that it can be communicated between people as computers. The principal characteristics of an ontology are:

- Ontologies are an abstract model of some phenomenon of the world, which identifies the relevant concepts.
- The type of used concepts, as well as the restrictions that apply on them, are explicitly defined.
- The ontology must be interpreted by a computer.
- The ontology captures consensual knowledge, it is the result of collaborative effort of the domain experts.

Ontologies can be expressed using several levels of formalism, nevertheless, there are defined four categories of the more common forms of express them [5]:

1. Highly informal: written using natural language.
2. Semi-informal: Restricted and structured using natural language.
3. Semi-formal: Using a formally defined language.
4. Rigorously formal: Semi-formal, including theorems and proofs.

When creating an ontology, it must be defined the adequate formality level to write it, because the knowledge represented in the ontology must be useful for its users, and also, it must be used to support the design and development of applications that use the knowledge contained in it; most of them are systems based on agents that extracts information from the web; such agent must be able to share a common ontology to access the information they need.

Ontologies can also be classified according to the level of formalism they are written [5]:

1. Catalogue: a list of terms, no axioms, no glosses.
2. Glossed catalogue: a catalogue with natural language glosses.
3. Taxonomy: a collection of concepts with a partial order induced by inclusion.
4. Axiomated taxonomy: a taxonomy with axioms.
5. Context library: a set of axiomatized taxonomies with relations among them.

2.2 Semantic Networks

Semantic networks are way to store semantic knowledge, in the initial approach, such knowledge was based on inheritance hierarchies, later, the concept was extended to cover any type of graphic representation [8]. This networks are based in the conections of nodes that are related using some semantic characteristic. The principal components of a semantic networks, can be divided in 3 categories [1]:

- Objects
- Events
- Relations

Here, objects represents all (phisical or abstract) entities (ideas, numbers), for each object is defined its main property as the type to which it belongs. Events represents actions, also there are predicates defined thah modifie an event description adding a role. Finally, the relations connect objects of several types as well as events that involve such objects.

3. ARTIFICIAL IMMUNE SYSTEMS

The immune system is composed upon several molecules, cells and organs distributed along the body. There is no main organ that controls the function on the immune system. Their principal task is to monitor the body looking for malfunctioning cells, such cells can belong to the body or not, in this case there are strange elements that may cause diseases. The immune system is then capable to distinguish between the self and non-self into the body [4].

A set of proteins called antibodies recognize the external agents, then cells called phagocytes engulf such agents. If all these external agents remain in the same place, all the phagocytes must to do is to approach to the element that have been identified by the antibodies and destroy it [11]. The immune system has developed the ability to generate a repertoire of detectors from which, given the exposure to an antigen, there will be selected those specific that can detect such antigen in a suitable way. This system presents an almost unlimited capability to detect any chemical agent, whether it is natural or artificial [10].

Antigens are usually proteins or molecules external to the body, which are derived from patogens or malignant cells; such antigens are characterizaed by components called epitopes. In defining antigens, two main properties should be distinguished: *antigenicity*, the capacity of a given antigen to be recognized by the antigen-specific receptors expressed by T or B cells; and *immunogenicity*, the ability of ther antigen to induce an immune response [9].

An artificial immune systen can be defined as a computational system developed using ideas, theories and components taken from the natural immune system.

The natural immune system works on three different levels: phisical barriers (skin, mucosa, etc), the innate immune system and the adaptive immune system. Artificial immune systems are based upon the last level, which presents the desirable properties for a computational intelligence system such as learning and memory [12].

The main task of the immune system is to defend the body against external agents, such function can be viewed, in a general way, as the clssification of such agents in two classes: *self*, those belonging to the body; and *non-self*, those which are external or are potencially dangerous to the body. This classification process is performed using a vast collections of T cells which are capable to recognize proteins; such cells are produced by its own *learning algorithm* [13]. This learning algorithm inherent to the immune system has benn used with relative success in text classification tasks and information extraction focused in such classification [7, 13].

4. PROPOSED MODEL

5. REFERENCES

[1] J. F. Allen and A. M. Frisch. What's in a semantic network? In *ACL Proceedings, 20th Annual Meeting*, pages 19–27, 1982.

[2] V. R. Benjamins, D. Fensel, and A. Gómez-Pérez. Knowledge management through ontologies. In *PAKM*, 1998.

[3] S. Chakrabarti. *Mining the Web. Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers, 2002.

[4] L. N. de Castro and J. Timmis. Artificial Immune Systems: A Novel Approach to Pattern Recognition. In L. A. J. Corchado and C. Fyfe, editors, *Artificial Neural Networks in Pattern Recognition*, pages 67–84. University of Paisley, Jan. 2002.

[5] D. Elliman. Automatic derivation of on-line document ontologies, July 26 2001.

[6] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI*, pages 1048–1053, 2005.

[7] J. Greensmith and S. Cayzer. An artificial immune system approach to semantic document classification. Technical Report HPL-2003-141, Hewlett Packard Laboratories, July 16 2003.

[8] T. L. Griths and M. Steyvers. A probabilistic approach to semantic representation, Apr. 29 2002.

[9] R. Holtappels. Dominating immune response, immunodominance and its significance in immunity. *B.I.F. FUTURA*, 20(3), 2005.

[10] D. Izhaky and I. Pecht. What else can the immune system recognize? In *Proceedings of the National Academy of Sciences of the United States of America*, volume 95, pages 11509–11510, September 1998.

[11] S. I. Nishimura. A study of spatial formation of immune cells. *Genome Informatics*, 12:302303, 2001.

[12] A. Scime. *Web Mining: applications and techniques*. Idea Group, 2005.

[13] J. Twycross. An immune system approach to document classification. Technical Report HPL-2002-288, Hewlett Packard Laboratories, Oct. 23 2002.