An Immune-based Approach to Knowledge Extraction

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

In this paper a methodology for keywords extraction from text documents is presented; the idea is to combine a mathematical background on information theory and a biological foundation of the immune system, to find the words that provide the most information to the categories in which the documents are grouped. Some experiments using only the information theory approach show that the methodoogy is able to find important words. This extracted keywords will be useful to build a knowledge representation of the texts.

Keywords

Keyword extraction, information retrieval, knowledge representation

1. INTRODUCTION

Nowadays, the generation of textual information have grown 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 raises the necessity of using 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 databases, bibliography management systems, e-mail organizers and many other formats; that tools help us to represent the explicit knowledge contained in the texts, but there are other type of knowedge known as the tacit knowledge, that is, the knowledge behind the text, that related to the meaning of the phrases and paragraphs. Such tools make the access to such information very efficient, but 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.

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 time period. The search for the relevant 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, document 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

In this paper we are focused on the problems related to knowledge extraction and representation from a set of related text documents. Section 2 explores the metholodogies

used to represent knowledge extracted from text documents, it includes ontologies and semantic networks models. In section 3 are presented the *Artificial Immune Systems* starting from a biological perspective to define an algorithm suitable for the task we are interested in. The porposed model to solve the problem of knowledge information is presented in section 4. Section 5 shows some preliminar experiments made using the artificial immune sysytem, Finally there are some conclusions in section 6.

2. KNOWLEDGE REPRESENTATION MODELS

Semantic information refers to the tacit knowledge stored within the documents, such knowledge is generally not explicit 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 [14] 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 [7].

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 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 [9]. 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.

This networks are very useful because it provides information about the concepts and the relationships between them, such relationships can be of several types, like composition, inheritance, etc. The common feature is that it always represents a semantic connection between many interesting concepts.

2.2 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 [6]:

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 [6]:

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.

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 functions of the immune system. An important task accomplished by the immune system is the monitoring of 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. One of the roles of the immune system is that of distinguish between the self and non-self into the body [5].

The immune system works in three different layers: physical barriers (like the skin), the innate immune system and the adaptive immune system. Most of the artificial immune system models that have been devoloped are besd in the last layer, which presents the desirable properties for a computational intelligence system like learning and memory [14].

Antigens are usually proteins or molecules external to the body, which are derived from patogens or malignant cells; such antigens are characterized by regions 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 [10].

The adaptive immune response is composed of the cellular and the humoral responses [4]:

1. **Cellular Immune Response:** A cell infected by a virus, can dagrade such virus and transport sections of its proteins to the membrane to present it; that is called an *Antigen Presenting Cell*, Helper T cells can detect the proteins that are being presented and are activated. Helper T cells are available to generate identical copies of itself. Such activated T cells circulate through the body destroying that cells that infected cells.
2. **Humoral Immune Response:** This response is initiated by cells called macrophages, which engulf antigens, such as bacteria and viruses, and process it to put in its cellular membrane and present them. Again, T cells detect such antigens that are being presented and are activated; activated T cells are cloned producing identical copies. Later, T cells help B cells to diferentiate into antibody producing cells (plasma cells). A B cell that finds an antigen seen previously that activated a T cell, engulf such antigen and transport it to its membrane to present it. Is a T cell detects the antigens presented by B cells, such T cell helps the B cell to produce copies that will differentiate into plasma cells. Plasma cells produce identical antibodies (also called immunoglobulins) that are specific to the antigen recognized by the B cell. That recently generated antibodies are capable to recognize the antigens to make easier the task accomplished by the phagocites. If all of this antigens remain in the same place, all the phagocites must do is approach to the antigen identified by the antibodies and destroy it [13].

The biological immune system have developed the ability of generate a set of detectors which, while are exposed to an antigen, are selected those that recognize in an adequate way such antigen. This system presents an almost unlimited capacity to detect any chemical agent, whatever it is, natural or artificial [11].

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

One of the several functions of the immune system is to defend the body against external agents, such function can be viewed, in general terms, 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* [15]. 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 [8, 15].

4. PROPOSED MODEL

The idea is to extract the words that provides the most information into related text documents, to achieve this goal, an artificial immune network is developed using ideas taken from the biological immune system and a mathematical background, this techniques in conjunction will find and

extract such words known as *keywords*.

The immune network take as input a set of documents grouped into categories of related documents, for each category, a set of keywords is generated. Such keywords will be useful in posterior tasks of document clssification or knowledge representation.

4.1 Mathematical Background

The documents from which the keywords are extracted, are divided into several categories; in each category, there are some related documents that are the base for the keywords extraction process as shown in figure 1.

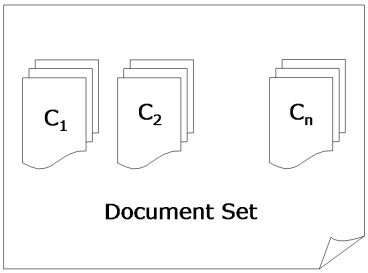

Figure 1: Documents divided into categories

Let's define some variables from such document set:

- $P_{c_i}(w)$: Probability of finding the word w in a document taken from the category c_i
- $P(w)$: Probability of finding the word w in a document taken from the whole document set.

From this definitions, we can calculate some information useful in the process of keywords extraction:

Let $E_{c_i}(w)$ be the entropy of the word w into the category c_i, that is

$$E_{c_i}(w) = -P_{c_i}(w) * \log_2[P_{c_i}(w)]$$

E_{c_i} will be the total entropy of the category c_i:

$$E_{c_i} = \sum_{w \in c_i} E_{c_i}(w) = -\sum_{w \in c_i} P_{c_i}(w) * \log_2[P_{c_i}(w)]$$

Let $E(w)$ be the total entropy of the word w into the whole document set, that is

$$E(w) = -P(w) * \log_2[P(w)]$$

E_{total} will be the total entropy of the document set:

$$E_{total} = \sum_{w} E(w) = -\sum_{w} P(w) * \log_2[P(w)]$$

Finally, let $I_{c_i}(w)$ the amount of information provided for the word w to the category c_i, and $I(w)$ the amount of information provided for the word w to the whole document set:

$$I_{c_i}(w) = E_{c_i} - E_{c_i|w}$$

Where $E_{c_i|w}$ is the conditional entropy of the category c_i given the word w, and

$$I(w) = E_{total} - E_{total|w}$$

Where $E_{total|w}$ is the conditional entropy of the document set given the word w.

The words that are of interest here, are those which provide a great amount of information to the whole document set, but a low information gain to the category in which is contained. That means, the word is useful to discriminate between categories, but inside the category is a common word that provides low information.

4.2 Biological Foundations

An Immune network will be trained with the words contained into the documents in order to extract the appropiate keywords using the folowing rules:

1. Each document is converted into a set of antigens; each antigen represents a word into the document and contain information about the category of the document and the frequency of the word into the document.
2. The antigens are presented to the immune network, each antigen must be detected by an antibody.
3. An antibody represents a word and also contains information about the frequency of that word in all the categories in the document set.
4. The interaction between the immune network is modeled in order to supress and stimulate the adequate antibodies; such interaction reflects the mathematical background presented in the later subsection. The idea is to stimulate those antibodies that provides great amount of information to the whole document set, but a lower amount into the categories.

5. PRELIMINARY EXPERIMENTS

The preliminary experimets was run using the 20 Newsgropus dataset, this dataset contains 20 categories and almost 500 training documents per category.

The experiments made are divided in two phases; the first phase consist of the calculation of the information each word provide to each category. The second phase will consider the interactions between the antibodies that detect the words contained into the documents.

The experiments showed here, consider only 5 categories and 100 training documents per category. The categories are:

1. alt.atheism
2. rec.autos
3. sci.electronics

4. comp.os.ms-windows.misc
5. talk.politics.mideast

The words that provides the most information gain to each categories as showed in tables 1 to 5:

Word	Information Gain
atheist	0.474311424745864
god	0.447382858120202
christian	0.347742779780387
moral	0.340678674923835
atheism	0.340113218272995

Table 1: First 5 keywords for alt.atheism

Word	Information Gain
car	0.564732218566477
wheel	0.291705829038465
truck	0.260315446698788
com	0.256058250377163
drive	0.255089244372633

Table 2: First 5 keywords for rec.autos

Word	Information Gain
electron	0.230044299128662
host	0.209919905811148
electr	0.209723874081105
circuit	0.206420997830274
data	0.200044500132098

Table 3: First 5 keywords for sci.electronics

6. CONCLUSIONS

From the preliminary experiments, it is clear that using a simple scheme to give importance to the words into the text documents, we can obtain words that represents in a good way the content of the document. This words can be used in a classification task, in which, given a document, by identifying the words it contains, it is easy to determine the category to which the document belongs.

The achieved experiments only consider the mathematical background of information theory [12] and still dont take into account the biological foundations of the immune system. The next step is to include those ideas of information theory into the stimulation and supression functions of the antibodies into the immune network.

With the words that are extarcted from the text documents, it will be possible ton build a knowledge representation of each category, in which the important concepts are represented by the keywords extracted.

7. 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.

Word	Information Gain
window	0.752500416388495
ms	0.747942508552121
organ	0.726826982948376
line	0.701681747828291
from	0.701272803086351

Table 4: First 5 keywords for comp.os.ms-windows.misc

Word	Information Gain
govern	0.347864431885075
against	0.315356207346521
turkish	0.304862338655776
armenia	0.293135721951149
until	0.287594412760493

Table 5: First 5 keywords for talk.politics.mideast

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

[4] G. M. Cooper and R. E. Hausman. *The Cell: A Molecular Approach*. Sinauer Associates, 2 edition, 1997.

[5] 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.

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

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

[8] 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.

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

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

[11] 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.

[12] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

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

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

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