An Artificial Immune System Based on Information Theory for Keyword Extraction from Text Documents

Sistema Inmune Artificial Basado en Teoría de la Información para la Extracción de Palabras Clave de Documentos de Texto

Andrés Romero, Ing., Fernando Niño, PhD.
Laboratorio de Investigación en Sistemas Inteligentes
Universidad Nacional de Colombia
Bogotá, Colombia Sede Bogotá
caromeroro@unal.edu.co, lfninov@unal.edu.co

Recibido para revisión 26 de Marzo de 2007, aceptado 15 de Junio de 2007, versión final 31 de julio de 2007

***Resumen*—En este artículo se presenta un modelo para la extracción de palabras clave; el cual extiende los conceptos básicos usados en dicha tarea, con el fin de proporcionar un marco teórico formal que permita determinar la importancia de las palabras clave para los documentos. El modelo propuesto combina un sistema inmune artificial con un fundamento matemático basado en la teoría de la información; este nuevo modelo tiene la ventaja de no requerir ningún conocimiento del dominio, así como el uso de un diccionario o cualquier información previa acerca del contenido de los documentos. El resultado final es un conjunto de palabras clave para cada categoría en el conjunto de datos usado.**

***Palabras Clave*—Extracción de Palabras Clave, Sistemas Artificiales Inmunes, Teoría de la Información.**

***Abstract*—This paper presents a model for keyword extraction, extending the basic concepts commonly used in this task, in order to get a formal background that allows determining the importance of the keywords to the documents. The proposed model combines an artificial immune system with a mathematical background based on information theory; this new model has the advantage that does not need any domain knowledge, neither the use of a stopword list or any previous information about the content of the documents. The final result is a set of keywords for each category into the corpus used.**

***Keywords*—Keyword Extraction, Artificial Immune Systems, Information Theory.**

I. Introduction

Recently, the generation of textual information has grown considerably; thus people and organizations work with huge amounts of data, such data can be in the form of working papers, corporate documents, e-mail among others. This situation raises the need of using computational tools to facilite the management of such amount of information in a reliable, secure and efficient way.

Many of the documents generated everyday are related to one another in some way. This relation between documents means that different documents, using different words, can have similar meaning, or they focus on similar subjects. This relation can be helpful in extracting keywords that somehow represent such meaning based on the contents of similar documents.

Obtaining keywords from a set of related documents, some tasks relevant to document processing, such as document clustering, document classification and information extraction can be improved.

An interesting application of the extracted keywords from the documents is to help in the task of search documents by topic instead of by textual matching, which means that documents can be searched using information provided by the keywords about it, instead of matching the query with the words within the document.

However most widely used techniques to extract keywords from documents have good performance and provide useful keywords, they have some requirements that in many cases are extremely difficult to fulfill. Such requirements are, commonly, the following:

- Previous Domain Knowledge.
- Part of Speech Tagging.
- Stop Word list.
- Supervised Learning.

The model proposed in this paper avoids the need to fulfill such requirements because it does not require any previous knowledge. The model is only based on the words contained in the documents and relationships between documents. In other words, there is no need to tag the words, because it has been designed to find important words and also to detect the words that do not provide much information. For this reason it is not necessary to have a stop word list or to use supervised learning because the information is only extracted from the documents.

This paper is organized as follows: in section 2, some techniques commonly used to extract keywords are summarized. Section 3 describes the main aspects of the techniques used for keyword extraction. A brief description of the immune system and some immune concepts used in the proposed model are shown in section 4. In section 5, the proposed immune based approach is detailed; specifically, a mathematical background based on information theory and how these concepts are included in the operation of the artificial immune system are presented. In addition, some experiments were carried out using some of the categories of the *20 Newsgroups* dataset, which are described in section 6. Finally, section 7 presents the conclusions of the work and shows some directions about further work to continue it.

II. Keyword Extraction Process

Keyword extraction can be viewed as a particular case of a more general process called feature extraction, in which the features can be extracted from a set of elements that contain related information about a particular domain. In order to achieve this process, techniques such as clustering and classification have been used. Some methodologies to represent and process the features stored in such elements have been proposed too. This information has been usually represented as a set of keywords, semantic networks and ontologies. In this work, the representation of such information will be in the form of keywords, therefore, in general, it is helpful to have previous information about the context [8].

Keywords and keyphrases are usually manually selected; in many contexts, authors are who assign such keywords to their own documents. This approach works well when every document has a set of keywords or keyphrases; but, in practice, this is not reached because a huge amount of information could not have an adequate structure. This is why such keywords must be automatically extracted from the document content; to achieve this, there are two main approaches [23]:

- **Assign keywords**: Phrases that best describe the document are selected from a controlled dictionary. In the training phase, a set of documents are associated to each phrase in the dictionary, then a classifier is constructed for each phrase. Each new document is processed by all the classifiers and the adequate phrases are assigned to the document. The keywords selected in the training phase are the only ones that can be assigned.
- **Extract keywords:** This approach does not consider a dictionary, instead, keywords are selected automatically from the text. Information extraction and lexical processing techniques are used to extract phrases with high probability to characterize the document. Training documents are used only to adjust the parameters of the extraction algorithm.

In [2], a general process to keyword or feature extraction is divided into two main phases:

1. Construction of a lexicon.
2. Generation of relationships between the words.

This general process is usually detailed in the state of the art methodologies for keyword extraction, and they mainly consider the following steps:

1. Identifying candidates to be keywords.
2. Weighting each candidate.
3. Selecting the keywords with the highest weights.

A. *Identifying Candidates*

The most common method used to select candidate keywords is known as *n-grams*, in which a set of *n* consecutive words are considered as a single term. Common values for *n* are usually *1*, *2* or *3* (higher values are not considered).

In adition to simply select the *words* or *n-grams* from the documents, some preprocessing is commonly achieved to get better keywords; such preprocessing includes the following steps:

- Stop word removal.
- Word stemming.
- Part of speech tagging.
- Deleting proper nouns.

B. *Weighing Candidates*

Once the candidate keywords have been selected, a numerical weight is assigned to each one of the words in order to determine its relative importance to the document or category. The most commonly used techniques to weight the words are:

- *tf* × *idf*: This technique attempts to assign a score or rank to each word. It is based on frequency of occurrence of a word in the document, and on the inverse of the frequency of occurrence of the word in the whole document set. Usually this measure is normalized to the interval [0, 1].

- *z-score*: Consists of assigning a score to each word, in this case, this score is based in the number of occurrences of the words in the document, the average occurrences in all the documents and the variance. This measure is similar to the standarization of a normal random variable.

In general, these measures try to assign a higher value to words that appear frequently in a document, but that do not appear so often in the rest of the corpus.

III. Related Work

Most of the work in keyword extraction follows the process described in the previous section. Many of the approaches focus on a specific domain and involve previous knowledge of that domain. Only a few techniques are domain independent and do not require any previous information. A summary of some techniques commonly used in keyword extraction is presented next.

In [14] two techniques for keywords extraction were compared on a biological domain. In such work, the methods used to evaluate the importance of the words to be considered as keywords are *z-score* and *tf* × *idf*, they also used some previous information to improve the process: they used a stemming algorithm and filtered the words using a stop word list.

The *tf* × *idf* method is used in [18] in addition to a bayesian classifier, the idea is to evaluate each word and then classify it as belonging to a class named *keyword* or a class *non-keyword*. They also used some preprocessing to improve the process, such as stemming, stop word removal, and semantic tagging of the candidate keywords. A technique called *named entity recognition* in the specific domain of the documents was used too. Finally, to evaluate the relevance of the selected keywords a dictionary is used.

A keyphrase extraction based on the naive Bayes learning scheme known as KEA was presented in [6] and detailed in [23], the general process follows these steps:

- Candidate keywords are selected from the text.
- Stop words are removed.
- Each keyword is ranked using *tf* × *idf*, which measures how specific a keyword or keyphrase is to a given document.
- *tf* × *idf* is discretized to apply a bayesian classifier
- Based on a bayesian learning system, each word is assigned a probability of being a keyword, and those with higher probabilities are selected as keywords.

The training process is achieved using a set of documents with their keywords previously defined.

In [15] a technique to achieve keyword extraction with some variants from the traditional methods is considered. They extracted the keywords from a single document without using a corpus, the method used there is *tf* × *idf* improved with word co-occurrence, that is, there are some words that are frequent, and they find the co-occurrences of the remaining words with such frequent terms.

Another method to keyword extraction is shown in [13]; here, the documents are represented using a vector of features based on word co-occurrence. Then the documents are grouped to identify such important words.

A different approach was used in [21], where a genetic algorithm to perform automatic keyword extraction in a supervised environment was developed.

In [20] a method for keyword extraction based on stopwords, *tf* × *idf* and bigrams is presented; besides, the extracted keywords are used to cluster webpages.

A technique to generate aditional features to the ones contained into the documents is presented in [8], such features form a set of keywords, which are used along with the words contained in the document to process them. This feature generation process is based on a domain specific knowledge, which is represented as ontologies containing hundreds of concepts. The feature generator analyzes and maps the documents into concepts belonging to the ontologies.

A technique to feature extraction known as *Word Clusters* is presented in [1]. This technique tries to cluster related words in the category they represent. To achieve this, the *Information Bottleneck* algorithm is used, which generates compact representations that improve the document processing. The generated clusters represent the main features of the document and also show an implicit relation between concepts.

A method for document categorization and the simultaneous generation of keywords for each category is presented in [7]. This method is based on K-Means, and each cluster is represented as a set of keywords.

In this algorithm, each cluster has a set of features and a weight, from which the keywords for each category will be selected. Weights are adjusted in each iteration and at the end those weights determine the most relevant keywords.

This approach has two main advantages:

- Generated clusters have a semantic meaning.
- From the keywords in each cluster, it is possible to generate a description of the cluster.

IV. Artificial Immune Systems

The natural immune system consists of molecules, cells and organs distributed throughout 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 that case there are strange elements that may cause diseases. One of the roles of the immune system is to distinguish self from non-self in the body [5].

All body cells carry molecular markers on their surface that enable them to be identified as *self* by the immune system cells. This cell marking molecules are called the *major histocompatibility complex (MHC).* Some proteins in the MHC are altered when the cell is infected by a virus. These molecules alert others cell in the immune system to begin an immune response.

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 developed are based on the last layer, which presents the desirable properties for a computational intelligence system like learning and memory [19].

A. *Cell Interactions*

1) *Antigen Detection*

Antigens are usually proteins or external molecules to the body, which are derived from pathogens 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 the antigen to induce an immune response [11].

Antigens are either free-floating in the body, in this case, the antigens are detected by B lymphocytes; or antigens can be expressed as part of an infected cell. Antigens can also be engulfed by macrophages, which digest the antigen and present it to other immune cells.

2) *Immune Responses*

a) *Innate Immune Response*

It refers to the part of the immune system with which we are born, it does not change or adapt to specific pathogens. It provides a rapid first line of defense, to keep early infection in check; this response involves roaming cells, such as macrophages, that detect and engulf extracellular molecules and materials, clearing the system of both debris and pathogens [10].

b) *Adaptive Immune Response*

The adaptive immune response is composed of the cellular and the humoral responses [3]:

1. **Cellular Immune Response:** A cell infected by a virus can degrade 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 throughout the body destroying 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 differentiate 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. If 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 of recognizing the antigens to make easier the task accomplished by the phagocytes. If all of these antigens remain in the same place, all the phagocytes must do is approach to the antigen identified by the antibodies and destroy it [16].

The biological immune system has developed the ability of generating a set of detectors which, when exposed to an antigen, those that recognize the antigen in an adequate way are selected. This system presents an almost unlimited capacity to detect any chemical agent, whatever it is, natural or artificial [12].

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 main functions of the immune system is to defend the body from external dangerous agents, such function can be viewed, in general terms, as the classification of such agents as belonging to one of two classes: *self*, those belonging to the body; and *non-self*, those which are external or are potentially dangerous to the body. This classification process is performed using a vast collections of T cells which are capable of recognizing proteins; such cells are produced by its own *learning algorithm [22]*. This learning algorithm inherent to the immune system has been used with relative success in text classification tasks and information extraction focused in such classification [9] [22].

Artificial Immune System has demonstrated to be useful in the task of feature extraction. Such feature extraction process can be achieved from multiple sources, such as images [4], email [17] and general text [22]. This feature extraction process can be efficiently used to obtain important features from text documents, such features extracted from the texts can represent the knowledge contained in the documents. There are some work on text processing, feature extraction from text documents [9], [22] and spam identification [17], [19] using Artificial Immune Systems.

V. KEYWORD EXTRACTION USING AIS

The methods presented in the previous section solve in a good way a specific task in keyword extraction; this is true because each method is performed on a particular domain, using previous knowledge of the problem such as dictionaries, stop word lists, part of speech tagging or any information that facilitates the process.

The idea of using an artificial immune system to perform this task comes from the abilities of the natural immune system, one of its functions is to protect the body, which can be viewed, in a general way, as the classification of entities into two classes, *self*, those belonging to the body and *nonself*, those external to the body and potentially dangerous. This classification is performed using a great number of T cells that are able to recognize proteins, this T cells have been produced using its own *learning algorithm* [22]. Some computational models have been developed based on such learning algorithm inherent to the immune system, which have been successfully used in text classification and extraction of semantic information [9], [22].

Artificial immune systems have demonstrated to be useful in the process of feature extraction from several sources, such as images [4], email [17] and general text [22]. Therefore, in this work an immune-based approach is proposed to obtain the most relevant terms (*keywords*) contained in text documents.

The artificial immune system developed in this work is based on some concepts in two theories, which are independent, but mixed they give a solid background to perform keyword extraction:

- *Immune Network Theory*: The idea is to exploit the abilities of the immune system to detect features and to apply it to detect important words from the documents.
- *Information Theory*: Gives a formal background to the operation of the immune network by measuring the amount of information that antibodies contribute to the immune network.

A. Mathematical Background

The basis for the operation of the immune network is a solid mathematical foundation, based on information theory where the idea behind is to determine how much information is provided by each word to the category and to the corpus.

The documents from which the keywords will be extracted, are divided into several categories; specifically, the keyword extraction process will work on a set of related documents as depicted in Figure 1.

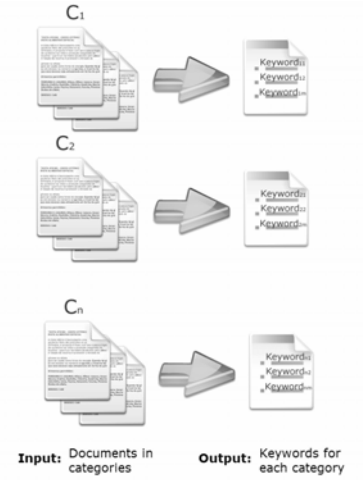

Figure 1. Keywords extracted from each category

Given a particular document set, let us define the following variables:

- $P_{c_i}(w)$: 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 these definitions, some information useful in keyword extraction process can be computed as follows:

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

$$E_{c_i}(w) = -P_{c_i}(w)\log\left[P_{c_i}(w)\right] \qquad (1)$$

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

$$E_{c_i} = \sum_{w\in c_i} E_{c_i}(w) = -\sum_{w\in c_i} P_{c_i}(w)\log\left[P_{c_i}(w)\right] \qquad (2)$$

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

$$E(w) = -P(w)\log\left[P(w)\right] \qquad (3)$$

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

$$E_{total} = \sum_{w} E(w) = -\sum_{w} P(w)\log\left[P(w)\right] \qquad (4)$$

Finally, let $I_{c_i}(w)$ be the amount of information provided by the word *w* to the category c_i:

$$I_{c_i}(w) = E_{c_i} - E_{c_i|w} \qquad (5)$$

where $E_{c_i|w}$ is the conditional entropy of the category c_i given the word *w*:

$$E_{c_i|w} = -\sum_{j\in c_i} P\left(w_j\middle|w\right)\log\left[P\left(w_j\middle|w\right)\right] \qquad (6)$$

and $I(w)$ is the amount of information provided by the word *w* to the whole document set:

$$I(w) = E_{total} - E_{total|w} \qquad (7)$$

where $E_{total|w}$ is the conditional entropy of the document set given the word *w*:

$$E_{total|w} = -\sum_{j} P\left(w_j\middle|w\right)\log\left[P\left(w_j\middle|w\right)\right] \qquad (8)$$

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

B. The Immune Network

An immune network was developed to detect important words in a document; the analogy with the natural immune system is that a set of antibodies will be able to detect the antigens presented to the system. In this case the antigens will correspond to the words contained in the documents; the antibodies also represent words and detect those antigens corresponding to the same word. The antibodies are then evaluated to determine whether they will live or die. At the end, the keywords for each category will be selected from the surviving antibodies.

1) *Immune Network Used in Keyword Extraction*

a) Antigens

Antigens are the entities that will be detected by the immune network; they store information about the words and the categories in which the words appear.

Each processed document is converted into a set of antigens which will be presented to the network, as shown in Figure 2.

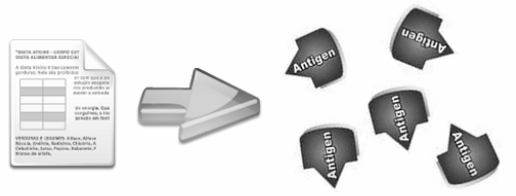

Figure 2. Antigens from each document

b) Antibodies

They store information about the word and the probability of finding that word in documents from each category. Antibodies will detect the antigens and will be stimulated in the categories which the antigen is found.

c) Interaction between antibodies

Each pair of antibodies into the network present a stimulus which represents the conditional probability for the words represented by the antibodies in the categories. When two antibodies detect antigens coming from the same document, they are co-stimulated (figure \ref{fig:img4}).

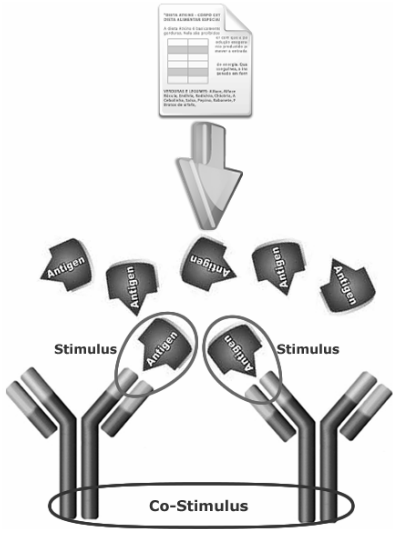

Figure 3. Co-stimulation of antibodies

d) Network Metadynamics

The network is regulated to control the number of antibodies removing those that represent words that provide the lowest amount of information to all the categories.

The amount of information provided by each word is calculated and those with the highest values are candidate words to be selected as keywords for each category. Antibodies with the lowest values represent useless words or those that are very common in all the categories and are not useful to discriminate between categories.

2) *The Algorithm*

The proposed keyword extraction algorithm takes as input a set of documents divided into categories, each document has the information about the category in which it is contained. One document is taken randomly from the dataset each time, this document is processed in the network; after presenting a number of documents, the interactions between the antibodies are calculated and those that provide the lowest information are removed and posterior occurrences of the words that they represent are filtered. At the end of the process, the antibodies that provide the highest information to each category are selected to be the keywords for that category. The performed process is as follows:

1. Each document in the corpus is converted in a set of antigens, each antigen represents a different word in the document and contains the category to which the document belongs; words that appear more than once in the document are converted in only one antigen, this is because only one antibody will detect them.
2. Each antigen is presented to the network, and the antibody which detects the word contained in the antigen is selected. An antibody detects an antigen if the word contained in the antibody is the same as the word contained in the antigen. If there is not any antibody that detects the antigen, then a new antibody is created with the word contained in the antigen.
3. The antibody that detects the antigen being presented is stimulated, this is, its information about the probabilities to find that word is updated for the category to which the antigen belongs, and the probability to the whole document set is also updated.
4. Each pair of antigens from the document are presented to obtain the corresponding antibodies that detect them. Such antibodies are co-stimulated, this co-stimulus represents the joint probability of the words in the document.
5. After presenting some documents, the network is regulated to suppress the antibodies that do not provide useful information. This is carried out by considering the co-stimulus of each pair of antibodies, and those with lower stimulation values are deleted.

The proposed keyword extraction algorithm is summarized in *Algorithm 1*:

Algorithm 1 Training

```
for each document $k$ in training set do
  convert document into antigens
  for each antigen $ag_i$ from document $k$ do
    present the antigen to all the antibodies
    find the antibody $ab$ with the highest affinity to $ag_i$
    if $ab$ is nothing then
      create antibody $ab$ with the word from $ag_i$
    end if
    stimulate antibody $ab$, i.e. update the frequency of the word in the corresponding category
  end for
  for each pair of antigens $ag_i$ and $ag_j$ from document $k$ do
    find antibodies $ab_i$ and $ab_j$ that detect the antigens
    co-stimulate $ab_i$ and $ab_j$
  end for
  if suppress then
    for each antibody $ab$ do
      calculate the co-stimulation with the rest of the antibodies, that is, find the information provided by the antibody to each category
    end for
    for each category do
      mark the antibodies with the highest information provided as keywords
      delete the antibodies with the lowest information provided
    end for
  end if
end for
```

VI. Experimental Results

Some experiments were carried out to validate the performance of the proposed algorithm. Particularly, some word extraction experiments were performed on the *20 Newsgroups* dataset. This dataset consists of 20 categories and about 500 documents in each category.

To achieve this goal, only 5 from the 20 categories in the dataset were used for the first experiments. In each category, only 200 documents were considered, in an incremental fashion; this means that in the first stage, only 50 documents per category were processed. Then the interactions between the antibodies were calculated and the least stimulated antibodies are removed, i.e. those that provide the lowest information are removed and posterior occurrences of the words that they represent will be filtered. After that, 50 more documents were processed following the same rules and so on until 200 documents per category.

Those antibodies with lower stimulus are deleted from the network and they conform a first barrier, that means that further occurrences of the words contained in such antibodies will be filtered in a first stage, and those words will not be processed by the immune network. Words represented by the antibodies in such barrier can be used as a stop word list, because they do not provide any useful information for any category in the corpus.

The categories used are:

alt.atheism
comp.os.ms-windows.misc
rec.autos
sci.electronics
talk.politics.mideast

These 1000 documents contain about 8200 different words, which are used to generate the same number of antibodies in the initial network. However, many of the antibodies corresponding to such words are subsequently removed from the network due to the interactions between the antibodies. The set of removed words are not considered in posterior documents and are used to form a stopword list, which is another output of the immune network besides the keyword list.

For each category are considered, at the end of the process, the first 10 keywords, i.e., those that provide the higher information to the category as shown in Table 1.

Table 1. Keywords extracted for the 5 categories

Atheism	*Windows*	*Autos*	*Electronics*	*Mideast*
atheist	*window*	*drive*	*electron*	*armenia*
argument	*print*	*engine*	*voltag*	*armenian*
atheism	*dataproduct*	*driver*	*devic*	*govern*
statement	*system*	*automot*	*volt*	*turk*
christian	*network*	*wheel*	*phone*	*villag*
word	*microsoft*	*dealer*	*resistor*	*land*
belief	*program*	*ford*	*transform*	*soviet*
exist	*file*	*mile*	*frequenc*	*territori*
moral	*card*	*tire*	*power*	*against*
bibl	*softwar*	*owner*	*panel*	*israel*

The keywords shown in Table 1 represent important concepts for each category. It is difficult to measure how good a keyword is to a category, this process of assigning a goodness value to each keyword can be achieved with the help of an expert in the particular domain. But with a look to the keywords, it is clear that the keywords extracted represent in a good way the categories.

VII. Conclusions

A method for keyword extraction based on the immune system was developed. The proposed method considers a mathematical background that defines in a formal way interaction between the antibodies in the artificial immune network. Such interaction results in a keyword extraction process that exhibits good performance and produced good sets of keywords.

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

This word weighting scheme combined with the immune network model gives an on-line method for keyword extraction that can be used on a small set of documents, and will still work well as the number of documents increases.

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

In addition to the main product of the immune network, antibody interaction and the information theory process not only helps to determine the keyword list, but also a stopword list is constructed from the antibodies with lowest stimulus which are deleted from the network.

An important advantage of the proposed model is that it does not need any previous information about the content of the documents. In fact, the only information needed about the documents are the categories in which they are grouped, but no additional information about each category is needed; this is reflected in the following consequences:

- **Language independence:** Because the words are extracted directly from the documents and the process is based on those words, such documents may be written in any language. It is important to notice that besides the documents may be in any language, it is necessary that all of them are in the same language for the process to work correctly.
- **No stopword list required:** As stated before, the process does not need a stop word list to filter the documents since it generates such list from the interactions between antibodies. As this list is generated dynamically, it can be used as soon as it is produced to filter additional documents being processed by the immune network.

Starting from this approach, the next step will be to consider *n-grams*, which are sets of consecutive words contained in the documents, these *n-grams* can provide more information than single words because they consider relationships between different words and there is an implicit count of co-occurrences that can lead to good *keyphrases*.

References

[1] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3:1183–1208, 2003.

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

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

[4] G. Cserey, A. Falus, W. Porod, and T. Roska. An artificial immune system for visual applications with CNN-UM. *In ISCAS (3)*, pages 17–20, 2004.

[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] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In D. Thomas, editor, *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99-Vol2),* pages 668–673, S.F., July 31–Aug. 6 1999. Morgan Kaufmann Publishers.

[7] H. Frigui and O. Nasraoui. Simultaneous categorization of text documents and identification of cluster-dependent keywords, Apr. 07 2002.

[8] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1048–1053, Edinburgh, Scotand, Aug. 2005.

[9] 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.

[10] S. Hofmeyr. An interpretative introduction to the immune system. *Design Principles for Immune Systems and Other Distributed Autonomous Systems, Oxford University Press,* pages 3–28, 2001.

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

[12] 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.

[13] N. Kang, C. Domeniconi, and D. Barbará. Categorization and keyword identification of unlabeled documents. In *ICDM*, pages 677–680. IEEE Computer Society, 2005.

[14] Y. Liu, B. J. Ciliax, K. Borges, V. Dasigi, A. Ram, S. B. Navathe, and R. Dingledine. Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering. In *CSB*, pages 394–404. IEEE Computer Society, 2004.

[15] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169, 2004.

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

[17] T. Oda and T. White. Developing an immunity to spam. In *Genetic and Evolutionary Computation - GECCO 2003*. *Genetic and Evolutionary Computation Conference, Chicago, IL, USA. Lecture Notes in Computer Science, Vol. 2723, Springer*, pages 231–242, 2003.

[18] N. F. Samatova, B. Park, R. Krishnamurthy, R. Munavalli, C. Symons, D. J. Buttler, T. Cottom, T. J. Critchlow, and T. Slezak. Information extraction from unstructured text for the biodefense knowledge center. In *Research & Development Partnerships in Homeland Security, Boston, MA, United States*, 2005.

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

[20] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Using keyword extraction for web site clustering. In *WSE*, pages 41–48. IEEE Computer Society, 2003.

[21] P. D. Turney. Learning algorithms for keyphrase extraction. Inf. *Retr*, 2(4):303–336, 2000.

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

[23] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. *CoRR*, cs.DL/9902007, 1999.

Andrés Romero: Ingeniero de Sistemas de la Universidad Nacional de Colombia, estudiante de la Maestría en Ingeniería de Sistemas y Computación. Sus áreas de trabajo actuales incluyen Sistemas Inmunes Artificiales, Computación Evolutiva, Extracción y Representación de Conocimiento.

Fernando Niño: Ingeniero de Sistemas; MSc. en Matemáticas; MSc. Computer Science; PhD. Computer Science. Actualmente vinculado al Departamento de Ingeniería de Sistemas e Industrial de la Universidad Nacional de Colombia. Sus áreas de trabajo actuales incluyen Bioinformática, Sistemas Inmunes Artificiales y Computación Bioinspirada.