Keyword Extraction Using an Artificial Immune System

Andres Romero
Computer Engineering Department
National University of Colombia
Bogota, Colombia
caromeroro@unal.edu.co

Fernando Nino
Computer Engineering Department
National University of Colombia
Bogota, Colombia
lfninov@unal.edu.co

Categories and Subject Descriptors: I.7 Document and Text Processing: Miscellaneous

General Terms: Algorithms.

Keywords: Keyword Extraction, Artificial Immune Systems, Information Theory.

1. INTRODUCTION

Extracting keywords from a single document or a document set is an important task in processes such as clustering, classification or extraction of semantic information.

The relations between the documents can be helpful in extracting keywords that represent such meaning based on the content of similar documents.

A general keyword extraction process mainly considers steps such as identifying keyword candidates, weighing each candidate, and selecting the keywords with the higher weights [3].

The most common method to select candidate keywords is n-grams, also, some preprocessing is achieved to get better keywords, including stop word removal, stemming and part of speech tagging. Once candidate keywords have been selected, a numerical weight is assigned to each word in order to determine its relative importance to the document [1].

2. KEYWORD EXTRACTION USING AIS

The set of related documents from which the keywords will be extracted are divided into several categories. The probability of finding a particular word in a document taken from a category and the probability of finding such word in a document taken from the whole document set are computed. Then the entropy of the words in the category and the total entropy of the category are calculated.

Accordingly, 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.

The immune system presented here combines two different theories. *Immune Network*: Immune network abilities to detect features are exploited and applied to detect the important words in the documents [2]. *Information Theory*: It provides a formal background to the operation of the immune network and determines how much information is provided by the antibodies to the network.

The extraction process is carried out in the following stages: (i) A document is taken from the corpus. The document is converted into a set of antigens, each representing a word contained in the document; (ii) Each antigen is presented to the immune network. Antibodies get stimulated if their words match those of the antigen; (iii) Antibodies in the network interact with one another to determine a co-stimulation between them; (iv) Antibodies with the lowest stimulus values are removed from the network, and those that remain in the network will correspond to the keywords for the specific category.

3. EXPERIMENTAL RESULTS

Some experiments were carried out to evaluate the performance of the proposed approach using the *20 Newsgroups* dataset. A set of 200 documents taken from each category were presented to the immune network. At the end of the process, the antibodies that provided the highest information were selected as the keywords for that category.

As an example, the first keywords extracted for some categories are shown next.

atheism	***windows***	***autos***	***mideast***
atheist	*window*	*drive*	*armenia*
argument	*print*	*engin*	*armenian*
atheism	*dataproduct*	*driver*	*govern*
statement	*system*	*automot*	*turk*
christian	*network*	*wheel*	*villag*

4. CONCLUSIONS

This scheme of weighing words combined with the immune network model, gives a method for keyword extraction that can be used with a small set of documents, and will continue working well as the number of documents is increased.

It is feasible to build a knowledge representation for each category with the words that are extracted from the text documents, in which the important concepts are represented by the extracted keywords.

5. REFERENCES

[1] 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*. IEEE Computer Society, 2004.

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

[3] 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.

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
ACM 978-1-59593-697-4/07/0007.