Bio-Inspired Techniques in the Clustering of Texts: Synthesis and Comparative Study

Reda Mohamed Hamou, GeCoDe Laboratory, Dr. Tahar Moulay University of Saïda, Saïda, Algeria
Hadj Ahmed Bouarara, GeCoDe Laboratory, Dr. Tahar Moulay University of Saïda, Saïda, Algeria
Abdelmalek Amine, GeCoDe Laboratory, Dr. Tahar Moulay University of Saïda, Saïda, Algeria

ABSTRACT

Today, the development of a large scale access network internet/intranet has increased the amount of textual information available online/offline, where billions of documents have been created. In the last few years, bio inspired techniques which invaded the world of text-mining such, as clustering, represents a critical problem in the digital society especially over the world of information retrieval (IR). The content of this article is a recapitulation of a set of works as a comparative study between the authors’ experiments realized by applying a set of bio-inspired techniques (social spiders(SS), 2D Cellular automata (2D-CA), 3D cellular automata (3D-CA), Artificial immune system (AIS), Particle swarm optimization (PSO)) and other techniques founded in literature (Ants Colony Optimization (ACO) and Genetic algorithms (GAs)) for solving the text clustering challenge by using the benchmark Reuter 21785. They analyse the different results in term of entropy, f-measure, execution time, and clusters number in order to find the ideal configuration (similarity measure and text representation method) for each technique. Their objectives are to improve the efficiency of text clustering systems and make decisions that can be the starting point for other researchers.

Keywords: Bio-Inspired Techniques, Similarity Measures, Text Clustering, Text Representation, Validation Tools

INTRODUCTION AND BACKGROUND

Nowadays, the information world knows a real revolution in term of digital data that touches all sectors of human activities. The development of the internet and the means of online/offline communication like Facebook, twitter, messages, and the storage of information in cloud services, make the web the largest data container where 80% of the presented data are in unstructured form (free text), that are handled by machines.

The manual treatment of textual data are very costly in time and human resources. For instance, we are confronting a lot of texts from different disciplines (medicine, IT, biological,
linguistic) and we ask a person, to understand and classify them without, any prior information and any assistance of an expert. The texts of the same area must be in the same group and the texts of different sphere in different groups. It is clearly a hard process that takes a long time.

Since the 1990s, the availability of more powerful hardware such as discs with wide capacity, faster processor, and efficient visualization tools, have turned over the birth of the mechanisms for text mining, that represents the notion of all the methods for the automatic processing of textual data available in electronic formats in order to:

- The structuration of data, in order to be a great help for the automatic decisions;
- Discover the previously unknown information;
- The extraction of, knowledge and valuable information.

One of the text mining tasks is the clustering, that is employed for automatically partitioning a set of textual documents into distinct clusters, following two conditions for the purpose to guarantee the homogeneity of clusters:

1. Minimizing the inertia intra cluster, which measures the concentration points of the cluster around the centre of gravity;
2. Maximize inertia inter cluster, which measures “the remoteness” of the clusters’ centres between them.

In the recent years, several works have been devoted with a single goal, which is to find an efficient approach for clustering. The research is nevertheless active, and the advancement around this scientific discipline is facing to various problems:

- **The choice of the ideal parameters**: Because most of clustering techniques are grounded on the representation of text, and a similarity measure, that influence the outcomes;
- **Response time**: Clustering is an NP-complete problem as proved by the number of Stirling (Gould, 1961) and the number of Bell (Bell, 1984);
- **Absence of universal technique**: In the literature every day a new paradigm invades the world of clustering, which have induced the problem that there’s no universal technique;
- **Initialisation of cluster number**: Find the clusters number initial, is a classical problem (because there is no supervisor (domain expert));
- **Data diversity**: The multiplicity of the natural language ambiguity, as biological text, medical textbook, literature text, artificial intelligence text... etc.

Our nature aged more than 5 billion years and more experiences, initiating from the most minor thing as bacteria to the complex system’s collection of human body. Meanwhile, the nature compound, the largest reserve of solutions and ideas for many kinds of problems, that human started to explore recently. In the information science community, the world celebrates the birth of new interesting domain known as bio-inspired methods, which have demonstrated their strength in different hard cases. These techniques are interesting that, reflects the natural behaviours of insects, plants and even human systems on an algorithm treatable by machine.

Our article is positioned in the framework of grouping different text clustering works through a synthesis, in the form of a comparative study between our bio-inspired techniques (automat cellular 2D in (Hamou2, 2010), cellular automat 3D in (Hamou2, 2012), immune system in (Hamou1, 2010), social spiders in (Hamou1, 2012) particle swarm optimization PSO in (Hamou,
BDD-Based Synthesis of Reversible Logic
www.igi-global.com/article/bdd-based-synthesis-reversible-logic/51676?camid=4v1a