Chapter 10
An Optimal Configuration of Sensitive Parameters of PSO Applied to Textual Clustering

Reda Mohamed Hamou
Dr. Moulay Tahar University of Saida, Algeria

Abdelmalek Amine
Dr. Tahar Moulay University of Saida, Algeria

Mohamed Amine Boudia
Dr. Tahar Moulay University of Saida, Algeria

Ahmed Chaouki Lokbani
Dr. Tahar Moulay University of Saida, Algeria

ABSTRACT

The clustering aims to minimize intra-class distance in the cluster and maximize extra-classes distances between clusters. The text clustering is a very hard task; it is solved generally by metaheuristic. The current literature offers two major metaheuristic approaches: neighborhood metaheuristics and population metaheuristics. In this chapter, the authors seek to find the optimal configuration of sensitive parameters of the PSO algorithm applied to textual clustering. The study will go through in dissociable steps, namely the representation and indexing textual documents, clustering by biomimetic approach, optimized by PSO, the study of parameter sensitivity of the optimization technique, and improvement of clustering. The authors will test several parameters and keep the best configurations that return the best results of clustering. They will use the most widely used evaluation measures like index of Davies and Bouldin (internal) and two external: the F-measure and entropy, which are based on recall and precision.

DOI: 10.4018/978-1-5225-5832-3.ch010

INTRODUCTION

Currently, due to the exponentially increasing amount of electronic textual information, the major problem for computer scientists is access to the content of textual information. This requires the use of more specific tools to access and siphon through the content of texts in a faster and more effective way.

Text Mining aims to develop new and effective algorithms for processing, searching, and extracting knowledge from textual and unstructured documents. One of the techniques widely used is called clustering.

Nature is a source of inspiration for researchers in various fields. These inspirations offer a natural framework to solve these problems in a flexible and adaptive way. The swarm intelligence is a field of interdisciplinary research that is relatively recent.

We are interested in studying the algorithms that are based on the specific movements of a swarm of agents to solve a problem. We chose the PSO algorithm ("particle swarm optimization") that uses a set of particles characterized by their position and velocity to optimize one or more fitness functions in a search space. This algorithm was initially proposed as a meta-heuristic for solving optimization problems.

In this paper, we use textual clustering by applying the PSO algorithm for multi-objective optimization (minimizing the intra-class distance and maximizing distances extra-class) and study the sensitivity parameters of the PSO for improvement on the quality of the textual clustering.

The study will go through in dissociable steps:

1. The representation and indexing of textual documents
2. Clustering by biomimetic approach
3. Optimized by PSO
4. Study the sensitivity parameter.

REPRESENTATION OF TEXTUAL DOCUMENTS

The machine learning algorithms cannot process directly the unstructured data: image, video, and of course, the texts written in natural language. Thus, we are obliged to pass by an indexing step.

The indexing step is simply a representation of the text as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it); this is very delicate and very important at the same time: a poor or bad representation will lead certainly to bad results.

We will represent each text as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it). In this way, we shall have a vector which represents the text and which is exploitable by machine learning algorithms at the same time. The main characteristic of the vector representation is that every language is associated with a particular dimension in the vector space. Two texts using the same textual segments are projected on identical vectors.

Several approaches for the representation of texts exist in the literature, among whom the bag-of-words representation which is the simplest and the most used, the bag-of-sentences representation, the n-gram representation which is a representation independent from the natural language and conceptual representation.