Study of Sensitive Parameters of PSO: Application to Clustering of Texts

Reda Mohamed Hamou, Department of Mathematics and Computer Science, Dr. Moulay Tahar University of Saïda, Saïda, Algeria
Abdelmalek Amine, Department of Mathematics and Computer Science, Dr. Moulay Tahar University of Saïda, Saïda, Algeria
Ahmed Chaouki Lokbani, Department of Mathematics and Computer Science, Dr. Moulay Tahar University of Saïda, Saïda, Algeria

ABSTRACT

In this paper, the authors study the parameter sensitivity of the technique of particles warm optimization (PSO) for the clustering of data, in particular the text. They experienced the PSO parameters by varying within a range of research and we noted the best result of clustering based on three measures of assessment, internal, which is the index of Davies and Bouldin and two external based on recall and precision that are the F-measure and entropy. Every time they finished an experimentation of a parameter, it is fixed to its optimal value for the next experiment parameters. The results showed a high sensitivity of some parameters on the result of clustering.

Keyword: Clustering, Indexation, Particle Swarm Optimization, Reuters 21578, Text Mining

INTRODUCTION AND PROBLEMATIC

Data mining has expanded significantly since the mid-90s and since then the development of new efficient algorithms for processing a large number of data. Textual data digitized constitute up to 80% of the daily flow of information stored in business. The amount of information available in the world explodes. In one study (Frawley, 1991), researchers estimate that it doubles every 20 months. In the same time, data accessible via the Internet quadruple (Lawrence, 1998). This is called text mining (text mining). Hence, a need for tools to manage, organize and analyze in depth and in a reasonable time this mass of data is required. These tools are especially needed in areas such as customer management, analysis of the results of investigations, intelligence, health and science, etc...

One of the techniques widely used in data mining, including text mining is the clus-
Classification and unsupervised in particular, can contribute to the problematic mentioned.

Biology has always been a source of inspiration for researchers in various fields. These researchers are based on properties of living things to offer original buildings, with amazing abilities and able to solve difficult problems where more traditional approaches encounter difficulties. These inspirations offer a natural frame work to solve these problems in a flexible and adaptive way by allowing from accessing a wider space of solutions, and offering a range of features such as: adaptation, tolerance, distributed computing, and generalization. This is the broad field of biomimetic computer.

In this area many researchers are interested in how animals move in groups. No individual control others but never the less of forms and complex behaviors can appear during these moves. This is the swarm intelligence which is a field of interdisciplinary research, relatively recent, it has been growing in popularity these days. This intelligence is defined as: “Any attempt to design algorithms or devices for solving combinatorial problems inspired the collective behavior of social insect colonies and other animal societies” and includes many algorithms-based population agents. We are interested in the study of algorithms that use more specific movements of a swarm of agents to solve a problem. This is the PSO algorithms (“particle swarm optimization”) using a set of particles characterized by their position and velocity to maximize or minimize a function in a search space. Interactions take place between the particles in order to obtain efficient global behavior. These algorithms initially proposed as a meta-heuristic for solving optimization problems. Clustering is a combinatorial problem, this meta-heuristic is used in this study to optimize operational research problem in this case minimizing intra-class distance and maximizing distances extra-classes in a pattern of clustering data. In the field of extraction and data mining for many applications requesting the use of a classification algorithm (Fayyad, 1996) in order to achieve multiple objectives. In this paper, we apply the PSO optimization technique to achieve a double objective (minimizing the intra-class distance and maximizing distances extra-class) and study the sensitivity of parameters of the PSO for improvement on the quality of clustering data including text.

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 by opting for a judicious choice resulting from the study.

REPRESENTATION OF TEXTUAL DOCUMENTS

The texts in natural language cannot be directly interpreted by a classifier or by classification algorithms from which the need for a mathematical representation of the text such that we can perform analytical processing thereon, all maintaining maximum semantics. The representation, which is generally used, is the use of a vector space as target representation space. The main feature of the vector representation is that each linguistic unit is associated with a specific dimension in the vector space. Two texts using the same textual segments are therefore projected onto identical vectors.

Several approaches for the representation of texts exist in the literature among which there are the bag of words representation, which is the simplest approach and widely used, the representation “bag of phrases” representation by lexical roots and of course the representation by n-grams which is independent representation of natural language (Shannon, 1948).

In general, the n-gram is defined as a sequence of n consecutive characters. The principle is that n-gram for a chain of k characters surrounded by blanks; we generate (k + 1) n-grams. An example of cutting for the word “porte” in 2-gram is the following: “_porte_” _p, po, or, rt, te e_.

Once we extracted all n-gram by a document, we define a list of n-grams sorted in descending order of their frequency of occurrence(Hamou, 2012).These methods are
Concept and Definition of Complexity
www.igi-global.com/chapter/concept-definition-complexity/24185?camid=4v1a

Technology Enhanced Language Learning in Virtual Worlds
www.igi-global.com/chapter/technology-enhanced-language-learning-virtual/56087?camid=4v1a