Text Clustering using Distances Combination by Social Bees:
Towards 3D Visualisation Aspect

Hadj Ahmed Bouarara, GeCode Laboratory, Tahar Moulay University of Saida Algeria, Saida, Algeria
Reda Mohamed Hamou, Department of Computer Science, Tahar Moulay University of Saida, Algeria, Saida, Algeria
Abdelmalek Amine, Tahar Moulay University of Saida Algeria, Saida, Algeria

ABSTRACT

Recently, the researchers proved that 90% of the information existed on the web, were presented in unstructured format (text free). The automatic text classification (clustering), has become a crucial challenge in the computer science community, where most of the classical techniques, have known different problems in terms of time execution, multiplicity of data (marketing, biology, economics), and the initialization of cluster number. Nowadays, the bio-inspired paradigm, has known a genuine success in several sectors and particularly in the world of data-mining. The content of this work, is a novel approach called distances combination by social bees (DC-SB) for text clustering, composed of four steps: Pre-processing using different methods of texts representation (bag of words and n-gram characters) and the weighting TF-IDF, for the construction of the vectors; Bees' artificial life, the authors have imitated the functioning of social bees using three artificial worker bees(cleaner, guardian and forager) where each one of them is characterized by a distance measure different to others generated from the artificial queen (centroid) of the cluster (hive); Clustering using the concept of filtering where each filter is controlled by an artificial worker, and a document must pass three different obstacles to be added to the cluster. For the experiments they use the benchmark Reuters 21578 and a variety of validation tools (execution time f-measure and entropy) with a variation of parameters (threshold, distance measures combination and texts representation). The authors have compared their results with the performances of other methods existed in literature (Cellular Automata 2D, Artificial Immune System (AIS) and Artificial Social Spiders (ASS)), the conclusion obtained prove that the approach can solve the text clustering problem; finally, the visualization step, which provides a 3D navigation of the results obtained by the mean of a global and detailed view of the hive and the apiary, using the functionality of zooming and rotation.

Keywords: 3D navigation, Entropy, F-measure, Filtering, Social Bees, Text Clustering

DOI: 10.4018/IJIRR.2014070103

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION AND PROBLEMATIC

The information universe, is now enduring a big revolution that involves all sectors globally and in dissimilar fields. Today’s, the web is the greatest accumulation of data, especially with the evolution of the communication means Internet / Intranet, have given a birth to a new concepts like, big data that reflects the large amount of unstructured data (textual documents) available online / offline. The digital society was enriched every day with a new substance, which makes it difficult to manage. For this reason we need to develop tools that help us to find within a reasonable time the desired information, performing certain tasks in our place, and facilitate our life.

In the past 20 years and with the development of computers, visualization tools and the instruments for automatic processing of information, as data mining, applied to extract the valuable information from a large volumes of data. We attempted in our proposed work, to consider that we have a huge quantity of textual documents, and we ask a person to classify them according to its domain, without any external help. However, this person has, no cognitive background about these documents. This process requires that the classifier must read all the documents in order to get the links between them. This kind of costly problem represents the virtual image of our work in the machine called clustering, in which the aim is to treat a set of textual documents and arrange them in homogenous classes of reflections where the documents of the same class must be similar, and the ones of different classes must be as dissimilar as possible.

Many works had been done over this area and several systems had seen the light, based on classical techniques that are faced with multiple obstacles:

- The pick of the distance measure criterion
- The selection of the texts representation method
- The initialisation of the cluster number
- Execution time caused by the number of documents existed

The current scientific world, was considerably built up with the inaugural appearance of novel concepts and prototypes. Actually, for each encountered problem, we must observe the nature; it may already have the same problem, where it had found solutions, long years ago. The bio-mimicry, consists to copy the living by getting advantages from solutions and innovations made by nature. In this paper, we will imitate the lifestyle of social bees, in order to introduce a new artificial model called Distances Combination by Social Bees (DC-SB) to solve the problem of text clustering that represents a topicality challenge in the scientific middle. Our problematic, is placed in the intersection of several subjects as shown in Figure 1.

STATE OF THE ART

The automatic classification (clustering) has attracted considerable attention from research and industry, various documents have been published in the subject, and many commercial systems and software are being developed. They provide a very important meaning in modern life where all texts clustering systems follow the same process: i) text representation, ii) construction of distance matrix, iii) modelling, iii) evaluation of results (Buhmann, 2003).

Classical Clustering Techniques (CCTs)

The CCTs can be classified into two large families:

1. **Tree Classification Family TCF**: The TCF is a recursive family that we cannot alter the source program in running. It establishes a tree of clusters, based on the movement of granularity. We can distinguish two types of approaches following the principle of this family: agglomerative (bottom) and down (divisive) (MacQueen, 1967).
Related Content

A Review of Answering Queries over Ontologies Based on Databases
[www.igi-global.com/chapter/a-review-of-answering-queries-over-ontologies-based-on-databases/138711?camid=4v1a](www.igi-global.com/chapter/a-review-of-answering-queries-over-ontologies-based-on-databases/138711?camid=4v1a)

Tweet Sentiment Analysis with Latent Dirichlet Allocation
[www.igi-global.com/article/tweet-sentiment-analysis-with-latent-dirichlet-allocation/127002?camid=4v1a](www.igi-global.com/article/tweet-sentiment-analysis-with-latent-dirichlet-allocation/127002?camid=4v1a)

Information Retrieval from Deep Web Based on Visual Query Interpretation
[www.igi-global.com/article/information-retrieval-from-deep-web-based-on-visual-query-interpretation/90441?camid=4v1a](www.igi-global.com/article/information-retrieval-from-deep-web-based-on-visual-query-interpretation/90441?camid=4v1a)