Information Sciences 265 (2014) 36–49

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An efficient Particle Swarm Optimization approach to cluster short texts

Leticia Cagnina [a,*], Marcelo Errecalde [a], Diego Ingaramo [a], Paolo Rosso [b]

[a] LIDIC (Research Group), Universidad Nacional de San Luis, Argentina
[b] Natural Language Engineering Lab. – ELiRF, DSIC, Universitat Politècnica de València, Spain

ARTICLE INFO

Article history:
Received 5 February 2013
Received in revised form 2 October 2013
Accepted 14 December 2013
Available online 22 December 2013

Keywords:
Clustering
Short-text corpora
Particle Swarm Optimization

ABSTRACT

Short texts such as evaluations of commercial products, news, FAQ's and scientific abstracts are important resources on the Web due to the constant requirements of people to use this on line information in real life. In this context, the clustering of short texts is a significant analysis task and a discrete Particle Swarm Optimization (PSO) algorithm named CLUDIPSO has recently shown a promising performance in this type of problems. CLUDIPSO obtained high quality results with small corpora although, with larger corpora, a significant deterioration of performance was observed. This article presents CLUDIPSO*, an improved version of CLUDIPSO, which includes a different representation of particles, a more efficient evaluation of the function to be optimized and some modifications in the mutation operator. Experimental results with corpora containing scientific abstracts, news and short legal documents obtained from the Web, show that CLUDIPSO* is an effective clustering method for short-text corpora of small and medium size.

1. Introduction

In recent years, document clustering has become a fundamental process in many tasks as enhancing the results returned by search engines, text mining, unsupervised text organization and information retrieval. In many of these domains, the clustering task has involved documents and content available on the Web. This interest in the use of clustering techniques in these cases, can be appreciated in past events such as the *The Spock Challenge* competition, but also in more recent events related to benchmarking activities on *Web People Search*.[1]

In this context, much of the useful information to be processed is taken from Web repositories whose documents are, frequently, short texts with a few tens or hundreds words, such as scientific abstracts, news and short technical and legal documents. For instance, in most digital libraries and on line repositories users have usually free access to abstracts of scientific papers but not to their full texts. Organizing that huge volume of short texts is an important challenge, as it has been observed in many works on clustering of scientific abstracts [1,10,34].

Several techniques have been developed to solve clustering problems and those based on the Swarm Intelligence (SI) paradigm seem to be specially attractive because of their robust performance [4,30,31,50].[2] In those cases where clustering techniques are applied to corpora containing *very short* documents, further difficulties are introduced due to the low frequencies

* Corresponding author. Tel.: +54 2664425622.
E-mail addresses: lcagnina@unsl.edu.ar (L. Cagnina), merreca@unsl.edu.ar (M. Errecalde), prosso@dsic.upv.es (P. Rosso).

[1] http://nlp.uned.es/weps/, tasks 1 and 2 on clustering information about people on the Web and clustering company tweets.

[2] In the present work, we focus on clustering methods that adequately match the manual classification criteria of human experts (or "ground truth"). This degree of correspondence is usually determined with *external validity measures* (EVM), like the entropy or the *F*-measure. In this context, the expressions "good performance", or "good quality" refer to those cases where the resulting groups show good values for some EVM (*F*-measure in our case).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.12.010

of the document terms. In this type of domains, an interesting SI algorithm named Particle Swarm Optimization (PSO) [41], has been successfully used [5,24].

In this article, we extend a preliminary proposal of a discrete PSO algorithm named CLUDIPSO [5]. For that, we present a detailed analysis and a discussion of the results obtained with different short-text corpora. The study clearly shows that the performance of the algorithm deteriorates as the number of documents to be clustered increases. This is mainly due to the particular particle representation utilized to describe the obtained clusterings. To deal with this problem, we present a modified version of CLUDIPSO named CLUDIPSO* which incorporates modifications aimed at improving the algorithm's performance. These modifications include a new representation of particles to reduce their dimensionality, a more efficient evaluation of the function to be optimized i.e. the Silhouette coefficient (as aftereffect of the previous modification) and some changes to the mutation operator.

The experimental work with CLUDIPSO* considers short-text corpora containing documents available on the Web (scientific abstracts, news and short legal documents) that significantly differ in number of documents, number of terms per document, number of groups and vocabulary overlapping, among others. The results are compared with those obtained by other three representative clustering algorithms: K-Means [33], K-MajorClust [25] and CHAMELEON [29].

The remainder of the paper is organized as follows. Section 2 gives a brief description of previous interesting works on clustering of short texts. Section 3 presents some general considerations about the perspective of considering clustering as an optimization problem, describing the cluster validity measure used as objective function to be optimized. Section 4 describes the previous version of the algorithm (CLUDIPSO) and Section 5 proposes the new improved version (CLUDIPSO*). In Section 6 some general features of the corpora used in the experiments are presented. The experimental setup, the analysis of the results obtained from the empirical study and the computational complexity analysis is provided in Section 7. Finally, some general conclusions are drawn and possible future works are discussed in Section 8.

2. Related works

Clustering of (short) texts is an active research field which can be analyzed from different point of views such as efficiency, effectiveness, difficulty of the task and diversity of the approaches to address it. In this section, we first consider some efficient text and short-text clustering methods that have recently obtained good quality results. Then, the difficulties of document clustering in general and short-text clustering in particular are analyzed considering the limitations of common methods to reflect real semantics. Next, some recent approaches that attempt to overcome these limitations are described. Finally, a few approaches based on the main technique used in our proposal (PSO) are briefly explained.

The development of algorithms that produce good quality groupings efficiently is a very relevant issue in text clustering. For instance, in HSCLUST [15], a pure Harmony Search algorithm adapted for clustering tasks is combined with K-Means in three different ways: replacing the refining stage of HSCLUST with K-Means, running K-Means after each iteration of HSCLUST and using the result in the next iteration of the algorithm, and improving the clustering obtained with HSCLUST with a one-step K-Means. These combinations allow to improve the quality of the clusters of documents in an efficient way. Another technique clustering which is simple, fast and effective is through the recursive propagation of information (messages) between documents. This idea is proposed in the *Affinity Propagation* method [19]. In that algorithm, the clusters are represented by a subset of *exemplars* (chosen randomly at first) and iteratively the method finds high quality exemplars (refining the clusters) and the corresponding clusters emerge.

Short-text clustering poses data sparseness and instability problems that directly affect the quality of the obtained results. An alternative to improve those results is by incorporating internal and external semantics to a clustering method. In [23] the authors proposed a framework with these characteristics arguing that internal semantics provides a deep understanding of texts through the use of a three-level hierarchical view while that external semantics incorporates concepts derived from multiple resources as Wikipedia and WordNet. The aspect of quality in the results of short text clustering is also studied in [42]. This study considers several clustering algorithms and different similarity measures: Cosine Similarity (CS), Latent Semantic Analysis (LSA), Short Text Vector Space Model (SVSM) and Kullback Leiber Distance (KLD). Three Hierarchical Agglomerative Clustering (HC) algorithms were analyzed: Single Link HC, Complete Link HC and Average Link HC; also the Spectral Clustering (SPEC). The results allow to conclude that the considered metrics do not always represent the correct quality of the clusters.

These difficulties that document clustering poses to current standard methods have also been analyzed from a cognitive science perspective. Text clustering methods heavily rely on a similarity measure that is supposed to adequately reflect the semantic "closeness" between documents. For instance, standard methods as the *vector space model* (VSM) (the method used in the present work) represent documents in a high-dimensional space, in which each dimension of the space corresponds to a word in the document collection. Here, the underlying metaphor consists in using *spatial proximity*, frequently determined by geometric measures like the cosine similarity, for *semantic proximity* [36]. This approach, the same as other simple approaches based on co-occurrences of words [48], will have serious problems to catch some subtle semantics that human beings use in speech and writing.

Cognitive sciences have provided a lot of interesting examples of why standard approaches can have trouble extracting real "semantic" information from texts alone [16–18,32]. French argues in [16] against the approaches that separate representations of reality from the process of manipulating them. His hypothesis is that every thing might be eventually seen as