Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 325–334

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Semi-Supervised Clustering Algorithms for Grouping Scientific Articles

Diego Vallejo-Huanga[1], Paulina Morillo[2], and Cèsar Ferri[3]

[1] Universidad Politécnica Salesiana, Department of Computer Science, Quito, Ecuador
dvallejoh@ups.edu.ec
[2] Universidad Politécnica Salesiana, Research Group IDEIAGEOCA, Quito, Ecuador
pmorillo@ups.edu.ec
[3] Universitat Politècnica de València, DSIC, València, Spain
cferri@dsic.upv.es

Abstract

Creating sessions in scientific conferences consists in grouping papers with common topics taking into account the size restrictions imposed by the conference schedule. Therefore, this problem can be considered as semi-supervised clustering of documents based on their content. This paper aims to propose modifications in traditional clustering algorithms to incorporate size constraints in each cluster. Specifically, two new algorithms are proposed to semi-supervised clustering, based on: binary integer linear programming with cannot-link constraints and a variation of the K-Medoids algorithm, respectively. The applicability of the proposed semi-supervised clustering methods is illustrated by addressing the problem of automatic configuration of conference schedules by clustering articles by similarity. We include experiments, applying the new techniques, over real conferences datasets: ICMLA-2014, AAAI-2013 and AAAI-2014. The results of these experiments show that the new methods are able to solve practical and real problems.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Clustering with constraints, Size constraint, K-Medoids, Linear programming

1 Introduction

Machine learning is defined as a subfield of artificial intelligence (AI) that addresses the study and construction of models capable of learning from the data [22]. Unsupervised learning is a machine learning methodology whose task is to induce a function that presents hidden structure from unlabelled data. Clustering is an example task of unsupervised learning. Cluster analysis has the objective of dividing data objects into groups, so that objects within the same group are very similar to each other and different from objects in other groups [24].

In many cases the data or problems, per se, have certain implicit restrictions, which traditional clustering algorithms do not take advantage of. At present, certain restrictions of size and relations of belonging of objects to the clusters have been incorporated into the clustering process, which have

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.206

demonstrated that the performance of the algorithms proposed for the solution of this type of problems increases significantly [26]. In clustering with size constraints, the cluster size refers to the total number of objects in each cluster [25].

Document clustering is defined as the division of a documents collection into groups according to their content [14]. Document clustering has been applied to many fields of study, such as: information retrieval, topic detection and content tracking, all of them are intrinsically related to language [4]. For the systematic treatment of language, in this paper, natural language processing (NLP) techniques are used to characterize the documents that are intended to be grouped.

A scientific paper is a written report describing original research results and generally published in journals or scientific conferences. One of the main drawbacks that arise when organizing the sessions of a conference is the large number of topics addressed by the documents presented, which are disseminated in different areas of knowledge and structures that, a priori, seem to have no relationship. In addition, the problem becomes much more complex if the times that are allocated for each session in a conference are limited, so assigning the number of papers to be exposed in a session is restricted by a specific amount. This scenario can be categorized as a problem of document clustering with size constraints.

This work addresses the problem of the automatic generation of conference schedules by using clustering techniques oriented to the grouping of documents with size constraints. When grouping scientific documents (papers), we need to take into account similarities between some features of the papers, e.g.: abstract, title, keywords, corpus, etc. We can consider these similarities by a basic weighted averaging of the individual similarities. However, in this mixing step, we loose the property of representing the documents in an Eucludian space. Many of the existing clustering methods need to represent the instances in an Euclidean space and therefore they cannot be directly applicable for this problem. In this paper we present two new semi-supervised clustering algorithms with size constraints that are able to solve the proposed problem: CSCLP - Clustering algorithm with Size Constraints and Linear Programming, and K-MedoidsSC - K-Medoids algorithm with Size Constraints. These algorithms can group elements taking into account size constraints of the target clusters. Additionally, we only need to have a distance or dissimilarity matrix between the elements to be clustered (i.e. the algorithms do not require an Euclidean space to work).

This paper is organised as follows. Section 2 presents the previous work related to clustering algorithms with size constraints. The formalisation of the two new clustering algorithms is described in Section 3. Section 4 includes simulation results and experiments for the validation of the proposed algorithms: in the first instance on multivariate benchmarking datasets and subsequently with documentary datasets, which will represent conference papers in machine learning area. The holistic methodology proposed for document clustering is also presented in this section. Finally, concluding remarks and future work are presented in Section 5.

2 Previous work

A first approximation of clustering algorithms with size constraints is presented in [13], where the goal is to find equal sized clusters as well as clusters of different sizes, through Fuzzy C-means algorithm (K-Means variation) and Lagrange multipliers. There are other many proposes that have focused on the modification of classical partition algorithms (such as K-Means) for the incorporation of size constraints, for instance: [20] and [8]. In [10] the authors propose a constraint programming formulation of some of the most famous clustering methods: K-medoids (does not use the dissimilarity matrix as input), DBSCAN and Label Propagation.

The article [26] introduces an algorithm that takes as a starting point the K-Means or Metric Pairwise Constrained K-Means (MPCK-Means) algorithms, to transform the size constraint problem into an inte-

Download English Version:

https://daneshyari.com/en/article/4960963

Download Persian Version:

https://daneshyari.com/article/4960963

Daneshyari.com