Knowledge-Based Systems 56 (2014) 141–155

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Beyond cluster labeling: Semantic interpretation of clusters' contents using a graph representation

François Role *, Mohamed Nadif

LIPADE, Université Paris Descartes, 45, rue des saints-pères, 75006 Paris, France

ARTICLE INFO

Article history:
Received 11 February 2013
Received in revised form 6 November 2013
Accepted 7 November 2013
Available online 20 November 2013

Keywords:
Cluster labeling
Clustering
Exploratory data analysis
Visualization
Network analysis

ABSTRACT

Efficient clustering algorithms have been developed to automatically group documents into subgroups (clusters). Once clustering has been performed, an important additional step is to help users make sense of the obtained clusters. Existing methods address this issue by assigning to each cluster a flat list of descriptive terms (labels) that are extracted from the documents, most often using statistical techniques borrowed from the field of feature selection or reduction.

A limitation of these unstructured descriptions of clusters' contents is that they do not account for the meaningful relationships between the terms. In contrast, we propose a graph representation, which makes the clusters easier to interpret by putting the descriptive terms in context, and by performing some simple network analysis. Our experiments reveal that the proposed method allows for a deeper level of interpretation, both when the clusters under study are homogeneous and when they are heterogeneous. In addition, evaluation procedures presented in the paper show that the graph-based representation of each cluster, while being very synthetic, still quite faithfully reflects the original content of the cluster.

1. Introduction

Users are increasingly overwhelmed with huge amounts of textual data that they do not have enough time to fully analyze and digest. To help them organize this large body of information, clustering algorithms have been developed that automatically group similar documents. While clustering is an invaluable tool to organize datasets into regions of interest, another crucial step is to help users understand what the obtained clusters are about. Several cluster labeling algorithms have been proposed for this purpose. Most of them consist in extracting words or short phrases that are deemed to best represent the topics of the clusters. However, while useful, this keyword-based labeling approach does not take into account the complex relations that may exist between terms in a cluster or between terms from different clusters. In fact, in order to provide an useful interpretation, a system must investigate these relations to detect that within a cluster a set of different terms are interrelated and concern the same topic, or that there is some degree of overlapping between clusters, although they may not have any terms in common. The main contribution of this paper is therefore to propose methods that go beyond flat labeling by allowing users to explore the relationships between the terms in the clusters and thus better interpret the clusters' contents.

Our solution to this challenge can be briefly summarized as follows. After clustering a set of documents represented as standard tf-idf vectors, we compute a centroid vector for each cluster, and use the terms associated with the main components of this vector as roots of a graph describing the corresponding cluster. This graph is a network of related terms, which is built from a term-term similarity matrix directly derived from the corpus (Fig. 1). The structure of the network can then provide both visual clues and numerical indicators that help users to gain a better sense of the context in which the descriptive terms are used.

We evaluate our method by measuring its ability to faithfully reflect the clusters' contents. In fact, depending on the performance of the clustering tool and hence the coherence of the obtained clusters, more or less easy to interpret descriptions can be derived. In any case, the system must not distort the original. Section 5 presents an evaluation method which we tested on several datasets.

The plan of the paper is as follows. Section 2 discusses related work. In Section 3 we introduce our technique for building a synthetic, graph-based representation of each cluster. Section 4 contains a running example and gives the results of some experiments we performed on three datasets commonly used in document clustering. Section 5 proposes a method to evaluate if the synthetic representation faithfully reflects the clusters' contents. Finally, Section 6 concludes and presents directions for future research.

* Corresponding author. Tel.: +33 143451276.
E-mail address: francois.role@parisdescartes.fr (F. Role).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.11.005

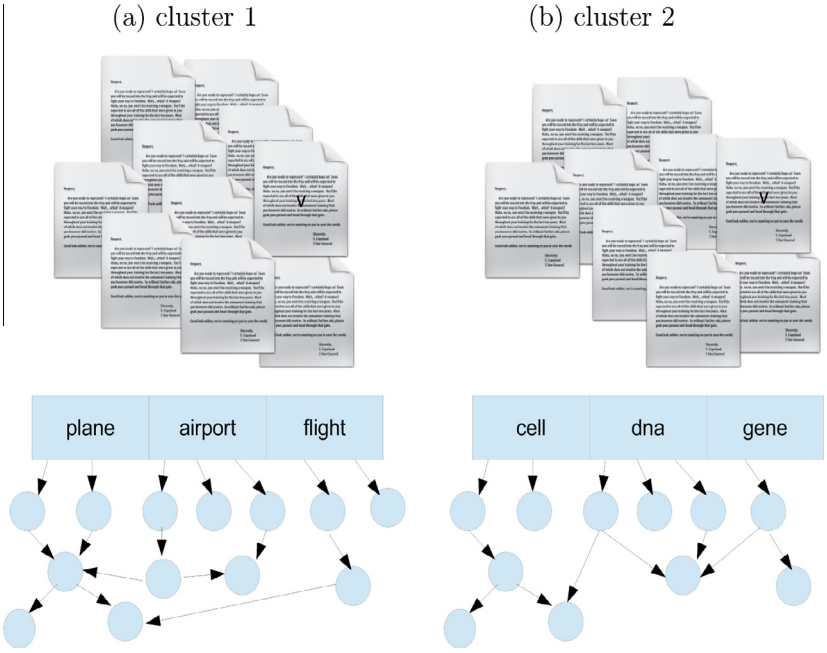

Fig. 1. Each cluster is represented by a graph. The roots of the graph (rectangles) are the words that correspond to the top tf-idf components of the centroid vector for a given cluster. The other vertices (circles) are words that can be reached by traversing the term similarity graph computed from the documents in the dataset.

2. Related work

Once a clustering method has been tried out on a data set, the obtained clusters need to be analyzed. Interpreting clustering results usually involves two different aspects: numerically assessing the results generated by the clustering method, and on the other hand trying to make sense of the obtained clusters. When class labels are available clustering quality can be numerically assessed by comparing the clustering with the gold standard classification using measures such as purity, normalized mutual information, rand index, and micro-averaged recall and precision. When class labels are not available, clustering quality can be assessed by measuring how well separated the clusters are and how compact they are. The most common measures that are used in this setting include the silhouette index [1], the Dunn's index and the Davies Bouldin index [2].

Besides these numerical quality indices, a common approach to help understand what clusters are about is to label them with a list of terms or short phrases deemed to be good descriptors of their contents. Several cluster labeling algorithms have been proposed for this purpose. The first family of techniques compute a label for a cluster by looking only at the contents of this cluster. In this approach, the most naive method is to select as labels for a cluster the most frequent terms in this cluster. A more advanced solution is to select as labels the most weighted terms in the cluster's centroid, or alternatively the titles of the documents that are closest to the centroid. Some authors have also proposed to use sentences or frequent phrases as cluster labels. For example, LINGO [3] discovers frequent phrases using suffix arrays and then retains as cluster descriptors those that can be matched to the abstract concepts obtained by decomposing the original term-document matrix using *Singular Value Decomposition* (SVD); see for instance [4]. As another example, Thirunarayan et al. [5] present a system to select cluster labels for sets of news documents obtained using queries involving companies and events, such as 'Oracle acquires Sun'. All the sentences of the documents are abstracted as a set of stems, and the system chooses labels among the sentences that contain phrasal references to the entities and events of interest. A problem with the previous methods is that they tend to favor terms that are frequent only in the cluster over terms that are frequent in the rest of the collection.

To overcome this bias, one can use feature selection methods to select terms that best characterize one cluster in contrast to other clusters. The goal here is to identify the terms that are the best indicators for cluster membership, which is often done by computing the mutual information between terms and clusters or by applying a chisquare test [6]. Finally, when clusters are obtained using methods such as *Latent Dirichlet Allocation* (LDA) [7] or *Non-negative Matrix Factorization* (NMF) [8,9], yet another possibility is to label each cluster (topic) with a list of words ordered by their salience in that cluster [10]. While interesting, all the above described methods sometimes fail to yield suitable labels. A distinct line of research has therefore investigated whether the use of external resources may help produce more meaningful labels for end-users. Following the work by Syed et al. [11], Carmen et al. [12] so propose a two-step process. In the first phase, given a cluster, the system selects a set of terms that maximize the Jensen–Shannon divergence between the cluster and the entire collection. In a second stage, a query involving these terms is issued against Wikipedia to find related Wikipedia page titles and category names. Both the original terms and the phrases derived from Wikipedia are considered as potential candidates for serving as cluster labels. In the same spirit, Tseng [13] and Bouras and Tsogkas [14] first extract candidate terms and then try to map these terms to Wordnet hypernyms in view of producing generic labels which do not necessarily exist in the clustered documents.

Whether or not they rely on external resources, a common feature of the above described approaches is that they produce a flat list of descriptive labels. In contrast, we propose a more structured representation of the clusters' contents. Being graph-based, this representation lends itself to the efficient techniques developed in the field of data analysis. In fact, it is now widely recognized that integrating graph visualization and statistical methods helps users discover important features in a dataset and thus greatly facilitates sense-making [15].