Pattern Recognition Letters 31 (2010) 502–510

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A document clustering algorithm for discovering and describing topics

Henry Anaya-Sánchez [b,*], Aurora Pons-Porrata [a], Rafael Berlanga-Llavori [b]

[a] *Center for Pattern Recognition and Data Mining, Universidad de Oriente, Patricio Lumumba s/n, Santiago de Cuba, Cuba*

[b] *Department of Languages and Computer Systems, Universitat Jaume I, Campus del Riu Sec, Castelló, Spain*

ARTICLE INFO

Article history:
Available online 26 November 2009

Keywords:
Document clustering
Topic discovery
Topic description

ABSTRACT

In this paper, we introduce a new clustering algorithm for discovering and describing the topics comprised in a text collection. Our proposal relies on both the most probable term pairs generated from the collection and the estimation of the topic homogeneity associated to these pairs. Topics and their descriptions are generated from those term pairs whose support sets are homogeneous enough for representing collection topics. Experimental results obtained over three benchmark text collections demonstrate the effectiveness and utility of this new approach.

1. Introduction

The ever-increasing availability of textual documents has led to a growing challenge for information systems to effectively manage and retrieve the information comprised in large collections of texts according to the users' information needs. However, it is not always easy or even possible for the users to formulate such needs precisely. For example, the users may not be familiar with the vocabulary that defines the topics of their interest, or simply they may wish to get a broad summary of the collection in order to guide their searches. For this reason, there exists a great interest to develop new tools for analyzing and summarizing these collections according to their main topics.

Clustering is an unsupervised learning technique that has been widely used in the process of topic discovery from documents. Basically, clustering methods are aimed at generating document groups or clusters, each one representing a different topic. However, as pointed out in several previous works (Fung et al., 2003; Pons-Porrata et al., 2007), clustering is not enough. Firstly, the obtained clusters do not necessarily correspond to actual topics of interest. In practice, it is usual that clusters tend to merge documents from different topics. Secondly, clustering methods do not provide descriptions that summarize the clusters' contents, so that users can judge them as homogeneous and relevant.

In this context, our research focuses on two issues: (1) how to discover the topics comprised in a text collection, and (2) how to simultaneously provide a meaningful description for each topic. We consider a topic to be defined in terms of the set of collection documents concerning a particular subject or theme, whereas a description is a set of terms that can represent or distinguish the topic in the collection.

In this paper, we introduce a new clustering algorithm aimed at discovering and describing the topics comprised in a text collection. The underlying hypothesis is that topics can be identified from those highly probable term pairs generated from the document collection that are likely to represent homogeneous contents. In our method, we assume that no prior knowledge about the collection exists, and therefore no training samples are available to supervise neither the discovery nor the description processes.

Our work extends the preliminary approach introduced in (Anaya-Sánchez et al., 2008). Firstly, we provide a comprehensive formalization of the main concepts on which the method relies. Secondly, a new method for generating descriptions is presented. It provides more descriptive and suitable topic labels instead of a simple term pair. Third, we avoid the similarity threshold tuning by automatically estimating a value from the collection that produces near-optimal results. Finally, more exhaustive experiments are carried out on several benchmark text collections by comparing our proposal with the state-of-the-art methods in terms of the quality of both the discovered topics and their descriptions.

The remainder of this paper is organized as follows. Section 2 describes the main works in the literature that simultaneously concern both discovering and describing topics. Section 3 presents the topic discovery method and explains the basic concepts on which our approach relies, whereas Section 4 discusses about the time complexity of the proposal. In Section 5, we show some experimental results carried out on three benchmark text collections. Finally, Section 6 presents the conclusions and future work.

2. Related work

The issue of describing automatically detected topics was initially presented in the earlier topic discovering systems such as

[*] Corresponding author. Fax: +34 964 728435.
E-mail addresses: a1084921@alumaiouji.es (H. Anaya-Sánchez), aurora@cerpamid.co.cu (A. Pons-Porrata), berlanga@lsi.uji.es (R. Berlanga-Llavori).

0167-8655/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.patrec.2009.11.013

Scatter/Gather (Cutting et al., 1992) and Suffix Tree Clustering (STC) (Zamir et al., 1997). The former one proposes to describe detected topics with the most frequent words in their corresponding clusters. However, this may cause descriptions to contain meaningless terms that do not distinguish the topics. On the other hand, STC uses frequent word sequences in order to determine and describe document groups. Due to the space requirements of its underlying clustering algorithm, this approach is only able to deal with short documents (e.g. snippets).

Topic detection systems (TDS) were devised for large and dynamic document collections such as newswire streams. However, as these approaches mainly rely on existing document clustering techniques, they disregard the description of the detected topics. Recently, in (Pons-Porrata et al., 2007) a new TDS is proposed for detecting and describing topics by means of a cluster hierarchy. The aim of the hierarchy is to capture properly the different topic granularities. For describing topics, the *Typical Testor Theory* is applied in order to select the most discriminating frequent words of each cluster. However, this approach has several limitations. Firstly, it requires to build a hierarchy of clusters to detect the proper granularity of the topics. Secondly, descriptions are calculated once the hierarchy is built by selecting a subset of features that discriminates each cluster from the others. The calculation of those discriminating subsets uses to be computationally expensive.

More recently, several works such as FIHC (Fung et al., 2003), CFWS (Li et al., 2008) and the method proposed by Malik and Kender (2006), aim at obtaining simultaneously both the coverage of a topic and its description by means of a new clustering criterion based on the concept of *frequent term set* (i.e. a set of terms that cooccur in at least a minimum number of documents in the text collection). Under this clustering criterion, the document clusters and their descriptions are determined by the frequent term sets of the document collection. Broadly, the clusters correspond to either sets of documents that share frequent term sets or mixtures of these sets if they are similar. In this way, FIHC uses frequent word sets to construct document clusters and organize them into a topic hierarchy. The cluster descriptions are composed of the clusters' most frequent words. CFWS is a partitional method that relies on frequent word sequences to firstly reduce the high dimensionality of the documents, and then to simultaneously build and describe the clusters from the sets of documents that share the frequent sequences. In this case, the clusters' descriptions are given by the frequent word sequences used for determining the clusters. Similar to FIHC, the method proposed by Malik builds a cluster hierarchy by relying on the notion of closed interesting itemsets (Malik and Kender, 2006). One of the claims of these works is that they outperform classical document clustering algorithms such as *Bisecting K-Means* and *UPGMA* at the same time that they provide a description for the clusters relying on these term sets.

However, several issues still remain open in order to apply such algorithms. Firstly, the importance of a frequent term set only depends on the number of documents that contain it, ignoring thus the weight of each term in the documents. Secondly, it is not ensured the homogeneity of the set of documents containing the frequent term set. If we select randomly a frequent term set, it is more likely to be a language collocation or a frequent domain pattern than a true topic label. For example, pairs like {*kill*, *people*}, {*high*, *price*} and {*student*, *book*} are frequent correlations between frequent words and they are more likely to be generated than {*Spain*, *Eurocup*}, which is a topic-based term set. Indeed, there are much less topic-based term sets than non topic-based ones. The application of interesting measures (Malik and Kender, 2006) can alleviate this problem by rejecting those term sets whose terms are not highly correlated. Unfortunately, these measures cannot distinguish between a collocation and a topic-based term set.

Thirdly, approaches based on term sets produce redundant and overlapping clusters. Closed term sets have been successfully applied to reduce these redundancies, as well as pruning via interesting measures (Malik and Kender, 2006). Merging clusters is also a common strategy for reducing redundancy, mainly for 1-term sets (Fung et al., 2003; Li et al., 2008). However, even applying these strategies, the obtained clusters contain a high overlap. This produces an effect of boosting in the calculation of both micro- and macro-averaged F1 scores, which in turn hides the actual effectiveness of these approaches.

Finally, these methods need to set up a minimum support for mining frequent term sets. Determining this value is one of the most critical aspects of all these algorithms. High values for the support threshold produce a handful set of term sets, but they only cover the broadest topics (i.e. many documents will not be assigned to a topic). Instead, low support values produce either a very large set of term sets or a combinatorial explosion, mainly in large and heterogeneous document collections.

Similar to the approaches based on frequent term sets, our approach relies on highly probable term pairs generated from the collection. However, we use these pairs only as a guide to explore the possible topics of the collection. Topics and their descriptions are generated from the documents containing those term pairs deemed to be representative of a collection topic. Thus, we introduce the concept of homogeneity of a document set, which is aimed at checking if a set of documents is cohesive enough to represent a topic. On the other hand, we avoid the problem of the minimum support threshold setting by exploring term pairs from more probable to less ones until all potential topics are discovered. In this way, term pairs are generated on demand while there are documents unassigned to topics.

3. Topic discovery method

Given a document collection $\mathscr{C} = \{d_1, \ldots, d_n\}$, the proposed method aims to obtain a clustering $\mathscr{G} = \{(\delta_1, G_1), \ldots, (\delta_m, G_m)\}$, where each cluster G_i represents a collection topic ($G_i \subseteq \mathscr{C}$) and δ_i is its description.

Assuming that each topic can be generated from a term pair,[1] the clustering method relies on the probability of generating such pairs from the collection to guide the search for a "good" partition of the data. Thus, starting from the most probable pair of terms generated from the collection, its support set (i.e. the set of documents in $\mathscr{C}$ that contain both terms) is built. If this set is homogeneous in content (see Section 3.2), a cluster consisting of the set of relevant documents for the content labeled by the pair is created (see Section 3.3). The cluster description is defined by the set of all collection terms that are descriptive for the content labeled by the pair (see Section 3.4). In case that the support set is not homogeneous in content, the pair of terms is discarded. Thus, meaningless term pairs are disregarded for identifying a topic.

Once a cluster has been built, its documents are removed from the collection. Then, this process is repeated again (regarding only the remaining document collection) until either the collection is empty or no more relevant pairs exist. Finally, if there are documents not clustered yet, a singleton is created for each one considering its most probable term pair as its description. The general steps of our proposal are shown in Algorithm 1. The terminology of the algorithm is specified along the next subsections, which are devoted to explain how the highly probable term pairs are generated, when a support set is considered as homogeneous in content, and how a topic and its description are built. For clarifying

[1] Terms correspond to meaningful word lemmas directly extracted from texts. Stop words are disregarded.