Non-disjoint grouping of text documents based Word Sequence Kernel

Chiheb-Eddine Ben N'Cir*, Afef Zenned**,Nadia Essoussi***

*LARODEC, ISGT, University of Tunis
chiheb.benncir@isg.rnu.tn
**LARODEC, ISGT, University of Tunis
afef.zenned@gmail.com
***LARODEC, ISGT, University of Tunis
nadia.essoussi@isg.rnu.tn

Abstract. This paper deals with two issues in text clustering which are the detection of non disjoint groups and the representation of textual data. In fact, a text document can discuss several themes and then, it must belong to several groups. The learning algorithm must be able to produce non disjoint clusters and assigns documents to several clusters. The second issue concerns the data representation. Textual data are often represented as a bag of features such as terms, phrases or concepts. This representation of text avoids correlation between terms and doesn't give importance to the order of words in the text. We propose a non supervised learning method able to detect overlapping groups in text document by considering text as a sequence of words and using the Word Sequence Kernel as similarity measure. The experiments show that the proposed method outperforms existing overlapping methods using the bag of word representation in terms of clustering accuracy and detect more relevant groups in textual documents.

1 Introduction

Text clustering is an important application within the information Retrieval field (IR). It aims to group similar documents in the same cluster, while dissimilar documents must belong to different clusters without using any predefined categories. This definition can be a crucial issue in many real life applications of text clustering where a document needs to be assigned to more than one group. This issue arises naturally because a document can discuss several topics and can belong to several themes. For example, a newspaper article concerning the participation of a soccer in the release of an action film can be grouped with both of the categories Sports and Movies.

Many clustering methods have been proposed to solve the problem of the detection of non disjoint groups in data. This kind of application is refereed as overlapping clustering (Diday, 1984), (Fellows et al., 2011). Our works concerns the detection of groups based k-means algorithm. Existing overlapping methods, when applied to text document clustering (Cleuziou,