Abstract

The main topic of this doctoral dissertation is the extraction of valuable information associated to text documents using different graph-based representations. **The core contributions** relies on **the novel representation of text documents as graphs** using lexical and syntactical elements of the language and **the employment of this kind of representations, to solve text classification tasks or to find text patterns related to a topic** in an innovative way.

The **first part** (Chapters one to three) of this dissertation describes the role of graph-based representations on a text mining process to solve or understand distinct text problems using non traditional techniques considering the state of the art progress.

The **second part** (Chapters four to six) of this dissertation explores the use of graphs and specifically enriched/non-enriched co-occurrence graphs on different text classification tasks, such as authorship or sentiment analysis, where most of the time, texts documents are structured, labeled and created ad hoc. The goal of this part, is to obtain the author, sentiment or demographic aspects of text documents using two novell supervised learning approaches. The first one, based on the extraction of text features from a graph using a vector representation with traditional classification algorithms (like support vector machines) and the second one, based on calculating the similarity between graphs and the use of heuristics.

The **third part** (Chapter seven) of this dissertation focuses on the use of co-occurrence and user-interaction graphs to analyze and extract meaningful information on large amounts of unstructured data in the context of an explicit topic. The idea, is to extract valuable textual patterns without

considering a specific classification task. These patterns could be used by experts on a field to understand the topic in a better way considering the growth of textual data and the use of big data and data science techniques. In this part, the combination of a statistical and a graph mining approach is proposed to extract different kinds of significant patterns.

Finally, the t**heoretical and practical implications associated to the use of graphs** are discussed in the **last part** of this dissertation (Chapter eight). Particularly, the implementation and use of this structures as a good alternative to represent text documents is highlighted, keeping in mind that graphs can map different linguistic aspects of text documents into a richer data structure, which otherwise could not be used in an easy and integrated manner. Additionaly, the goal for this part is to show how graphs are used to represent text documents in a practical manner, independently of the text topic or classification task and how this kind of representations are a valuable asset to extract information that other representations can not find due to their simplicity.