Erasmus Mundus masters program in

Language and Communication Technology

A graph model for text analysis and text mining

Author: Thi Ngoc Quynh Do

Supervisor: Amedeo Napoli

Free University of Bozen-Bolzano, Université de Lorraine

Abstract

Automated text analysis and text mining methods have received a great deal of attention because of the remarkable increase of digital documents. Typical tasks involved in these two areas include text classification, information extrac- tion, document summarization, text pattern mining etc. Most of them are based on text representation models which are used to represent text content. The traditional text representation method, Vector Space Model, has several notice- able weak points with respect to the ability of capturing text structure and the semantic information of text content. Recently, instead of using Vector Space Model, graph-based models have emerged as alternatives to text representation model. However, it is still difficult to include semantic information into these graph-based models. In this thesis, we propose FrameNet-based Graph Model for Text (FGMT), a new graph model that contains structural and shallow semantic information of text by using FrameNet resource. Moreover, we introduce a Hy- brid model based on FGMT which is more adapted to text classification. The experiment results show a significant improvement in classification by using our models versus a typical Vector Space Model.

Keywords: text representation model, graph model, FrameNet, text analysis, text mining