Word embedding-based techniques for text clustering and topic modelling

With application in the healthcare domain

Sattar Seifollahi

Principal Supervisor: *Prof. Massimo Piccardi*

A thesis presented for the degree of

Doctor of Philosophy

School of Electrical and Data Engineering

University of Technology Sydney

September, 2019

Word embedding-based techniques for text clustering and topic modelling

Sattar Seifollahi

Abstract

In the field of text analytics, document clustering and topic modelling are two widely-used tools for many applications. Document clustering aims to automatically organize similar documents into groups, which is crucial for document organization, browsing, summarization, classification and retrieval. Topic modelling refers to unsupervised models that automatically discover the main topics of a collection of documents. In topic modelling, the topics are simply represented as probability distributions over the words in the collection (the different probabilities reveal what topic is at stake). In turn, each document is represented as a distribution over the topics. Such distributions can also be seen as low-dimensional representations of the documents that can be used for information retrieval, document summarization and classification. Document clustering and topic modelling are highly correlated and can mutually benefit from each other.

Many document clustering algorithms exist, including the classic *k*-means. In this thesis, we have developed three new algorithms: 1) a maximum-margin clustering approach which was originally proposed for general data, but can also suit text clustering, 2) a modified global *k*-means algorithm for text clustering which is able to improve the local minima and find a deeper local solution for clustering document collections in a limited amount of time, and 3) a taxonomy-augmented algorithm which addresses two main drawbacks of the so-called "bag-of-words" (BoW) models, namely, the curse of dimensionality and the dismissal of word ordering. Our main emphasis is on high accuracy and effectiveness within the bounds of limited memory consumption.

Although great effort has been devoted to topic modelling to date, a limitation of many topic models such as latent Dirichlet allocation is that they do not take the words' relations explicitly into account. Our contribution has been two-fold. We have developed a topic

model which captures how words are topically related. The model is presented as a semi-supervised Markov chain topic model in which topics are assigned to individual words based on how each word is topically connected to the previous one in the collection. We have combined topic modelling and clustering to propose a new algorithm that benefits from both.

This research was industry-driven, focusing on projects from the Transport Accident Commission (TAC), a major accident compensation agency of the Victorian Government in Australia. It has received full ethics approval from the UTS Human Research Ethics Committee. The results presented in this thesis do not allow reidentifying any person involved in the services.

Dedication

To my family, Sona and Selin, and my parents.

Declaration

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Sattar Seifollahi declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 20/1/2020

Acknowledgements

I would like to acknowledge my great appreciation for my principal supervisor Professor Massimo Piccardi for providing me with the necessary guidance throughout my PhD research. The valuable comments and suggestions provided by him were immensely helpful to me in successfully writing the journal and conference papers on which this thesis is based. It has been an honor and a pleasure working with him. Thanks to my associate supervisors, Associate Professor Adil Bagirov, Associate Professor Federico Girosi and Dr Ehsan Zare Borzeshi for supporting me for this thesis and for helpful discussions. I would like also to thank the Transport Accident Commission (TAC) managers, an accident compensation agency of the Victorian Government in Australia, for their support and for providing the data, and the Capital Markets Cooperative Research Centre (CMCRC) and the School of Electrical and Data Engineering of University of Technology Sydney for providing great support and financial assistance throughout my candidature. This research has received ethics approval from UTS (UTS HREC REF NO. ETH16-0968). Finally, I would like to give special thanks to my family members, Sona and Selin, for their great support and giving me huge energy to work on my thesis, and thanks to all the friends and other staff members of UTS and Federation University Australia who have helped me in many different ways during my PhD study.

Contents

List of Figures

List of Tables