
**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 Giroi 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

1	Introduction	15
1.1	Motivation	15
1.2	Thesis objectives	17
1.3	Publications	19
1.4	Thesis outline	19
1.5	Common notations and symbols	20
2	Literature Review	21
2.1	Cluster analysis	22
2.2	Document representations for clustering	24
2.2.1	Methods based on term frequency	25
2.2.2	Ontology based techniques	26
2.2.3	Methods based on word embeddings	27
2.3	The k -means algorithm and its variants	29
2.3.1	The k -means algorithm	30
2.3.2	The k -means++ algorithm	31
2.3.3	Spherical k -means algorithm	32
2.3.4	Global k -means algorithm	33
2.3.5	Modified global k -means algorithm	34
2.4	Maximum margin clustering	36

2.5	Topic modelling	39
2.5.1	Latent semantic analysis	42
2.5.2	Non-negative matrix factorization	42
2.5.3	Latent Dirichlet allocation	43
2.6	Deep learning and beyond	44
3	Algorithms for Documents Clustering	45
3.1	A maximum margin clustering algorithm	45
3.1.1	Formulation of the clustering problem	46
3.1.2	Initial clusters	47
3.1.3	Post-processing of initial clusters	47
3.1.4	The proposed algorithm for clustering large-scale data	48
3.1.5	Convergence of the proposed algorithm	50
3.1.6	Experiments	50
3.2	An incremental algorithm for clustering document collections	61
3.2.1	Problem formulation	62
3.2.2	Initialisation of the cluster centers	63
3.2.3	An incremental clustering algorithm and its implementation	66
3.2.4	Experimental results	66
3.3	Conclusion	73
4	Taxonomy-Augmented Features for Text Analytics	75
4.1	Hierarchy of word clusters	75
4.2	Taxonomy-augmented features given the hierarchy of word clusters	77
4.3	Taxonomy-augmented features given a set of words	78
4.4	Taxonomy-augmented features for document clustering	80
4.5	Taxonomy-augmented features for document classification	83

4.5.1	Semantic analysis	87
4.6	Conclusion	88
5	Topic Modelling	89
5.1	A semi-supervised Markov topic model	89
5.1.1	Initial process	90
5.1.2	Generative model	92
5.1.3	Experiments	93
5.2	Cluster based topic learning	96
5.2.1	Topic-word learning	97
5.2.2	The document-topic matrix	98
5.2.3	Evaluation of the proposed method for document classification	99
5.3	Conclusion	102
6	Conclusion	104

List of Figures

Figure 2.1	Main machine learning approaches.	21
Figure 2.2	Difference between hard clustering (left) and soft clustering (right).	22
Figure 2.3	The two models of Wor2Vec.	28
Figure 2.4	The main steps in topic modelling.	40
Figure 2.5	Another view on topic modelling.	40
Figure 3.1	Objective function values for SAMMC and KM with varying number of clusters (subset of four datasets).	59
Figure 3.2	Objective function values for SAMMC and KM with varying number of iterations (subset of four datasets); notations “SAMMC k ” and “KM k ” stand for SAMMC and KM with K clusters.	60
Figure 3.3	Variation of the objective function value over 20 initial solutions. “SAMMC-lower value” and “SAMMC-upper value” are the plot of the average function value of SAMMC minus and plus the standard deviation, respectively; likewise, “km-lower value” and “km-upper value” are the corresponding values for KM.	60
Figure 3.4	Objective function values by removing specific steps from the proposed algorithm (ablation analysis). “SAMMC-pert” and “SAMMC-gmm” are the function values by removing Steps 4. and 3. from SAMMC algorithm, respectively; “km” and “SAMMC” stand for KM and SAMMC algorithms.	61
Figure 3.5	Cluster validity (Dunn) index for datasets 1 – 6	71

Figure 3.6	Cluster validity (Davies Bouldin) for datasets 1 – 6	72
Figure 4.1	Three layers of the hierarchy of word clusters.	76
Figure 4.2	The process of extracting features via the hierarchy of word clusters.	77
Figure 4.3	The process of extracting features via the hierarchy of word clusters and a set of predefined words.	79
Figure 4.4	Connectivity and silhouette measures of all models for the PCalls dataset.	82
Figure 4.5	Connectivity and silhouette measures of all models for the WebKB dataset.	82
Figure 4.6	Connectivity and silhouette measures of all models for the Reuters dataset.	83
Figure 4.7	Average accuracy from 10-fold cross-validation. The horizontal axis maps the classifier and the coloured bars represent the feature vectors.	86
Figure 4.8	Average of the absolute deviations (AVDEV) of accuracies from their mean. Horizontal axis shows classification methods and vertical axis is the AVDEV value.	86
Figure 4.9	Evolution of the chosen features in the phone calls of two randomly-selected clients. The semantic scores are computed by Algorithm 13.	87
Figure 5.1	Graphical models of (a) LDA model and (b) the proposed model, SHMTM. The number of prior topics in SHMTM is K_0 ($K_0 \geq K$) and the vocabulary size is V_0 ($V_0 \leq V$).	90
Figure 5.2	A sample set of words and their relations in the transition matrix. The relations have been created using 200 prior topics on the phonecalls data set.	94

Figure 5.3 PMI score with the phonecalls data set for the SHMTM model using 1) the top words of prior topics (SHMTM topwords) and 2) all the words in the collection (SHMTM all words), and for the LDA model with Gibbs sampling using 3) the top words only (LDA Gibbs topwords) and 4) all the words in the collection (LDA Gibbs all words) 94

Figure 5.4 PMI score with the filenotes data set for the SHMTM model using 1) the top words of prior topics (SHMTM topwords) and 2) all the words in the collection (SHMTM all words), and for the LDA model with Gibbs sampling using 3) the top words only (LDA Gibbs topwords) and 4) all the words in the collection (LDA Gibbs all words) 95

Figure 5.5 Overall process of the cluster-based topic learning. 97

Figure 5.6 Accuracy and standard deviation (%) of models for D1 using xgb classification. The bar is for the accuracy and the line with dots shows the standard deviation. 101

Figure 5.7 Accuracy and standard deviation (%) of models for D2 using xgb classification. The bar is for the accuracy and the line with dots shows the standard deviation. 101

Figure 5.8 Accuracy and standard deviation (%) of models for D3 using xgb classification. The bar is for the accuracy and the line with dots shows the standard deviation. 101

Figure 5.9 Accuracy and standard deviation (%) of models for D4 using xgb classification. The bar is for the accuracy and the line with dots shows the standard deviation. 102

List of Tables

Table 1.1	Table of the main notations used in this thesis.	20
Table 3.1	Dataset summary.	52
Table 3.2	Objective function values for the SAMMC algorithm over datasets 1-8	55
Table 3.3	Objective function values for the SAMMC algorithm over datasets 9-16	56
Table 3.4	Clustering errors obtained with the compared algorithms over datasets 1-8; for compactness of notation, E_1, E_2, E_3, E_4, E_5 and E_6 are the errors obtained using KM, MBKM, DPGMM, FCM, MMC and SAMMC, re- spectively	57
Table 3.5	Clustering errors obtained with the compared algorithms over datasets 9-16; for compactness of notation, E_1, E_2, E_3, E_4, E_5 and E_6 are the er- rors obtained using KM, MBKM, DPGMM, FCM, MMC and SAMMC, respectively	58
Table 3.6	Dataset summary.	67
Table 3.7	The function value and relative errors	69
Table 3.8	The function value and relative errors	70
Table 3.9	The optimal number of clusters and cumulative CPU time for com- puting up to 100 clusters. K_1^* and K_2^* stand for the optimal number of clusters using SKM and SMGKM, respectively, and t_1 and t_2 for the time	73

LIST OF TABLES

Table 4.1	Dataset summary.	81
Table 5.1	Top words from four topics for SHMTM using only the top words or the whole dictionary with the phonecalls data set	95
Table 5.2	Top words from four topics for LDA using only the top words or the whole dictionary with the phonecalls data set	96
Table 5.3	Data set summary.	100