UNIVERSITY OF TECHNOLOGY SYDNEY

Doctoral Thesis

Multi-Author Document Decomposition Based on Authorship

Author:
Khaled Aldebei

Principal Supervisor:
Prof. Xiangjian He

Co-Supervisor:
Dr. Wenjing Jia

Co-Supervisor:
Dr. Gengfa Fang

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Global Big Data Technologies Centre
UNIVERSITY OF TECHNOLOGY SYDNEY

January 2018

Declaration of Authorship

I, Khaled Aldebei, declare that this thesis entitled, 'Multi-Author Document Decomposition Based on Authorship' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Production Note:
Signature removed prior to publication.

Date: 16/01/2018

UNIVERSITY OF TECHNOLOGY SYDNEY

Abstract

Global Big Data Technologies Centre
UNIVERSITY OF TECHNOLOGY SYDNEY

Doctor of Philosophy

Multi-Author Document Decomposition Based on Authorship

by Khaled ALDEBEI

Decomposing a document written by more than one author into sentences based on authorship is of great significance due to the increasing demand for plagiarism detection, forensic analysis, civil law (i.e., disputed copyright issues) and intelligence issues that involves disputed anonymous documents. Among the existing studies for document decomposition, some were limited by specific languages, according to topics or restricted to a document of two authors, and their accuracies have big rooms for improvement. In this thesis, we propose novel approaches for decomposition of a multi-author document written in any language disregarding to topics, based on a Naive-Bayesian model and Hidden Markov Model (HMM). The proposed approaches of the Naive-Bayesian model aim to exploit the difference in its posterior probability to improve the performance of decomposition. Two main procedures are proposed based on Naive-Bayesian model, and they are Segment Elicitation procedure and Probability Indication Procedure. The segment elicitation procedure is proposed to form a strong labeled training dataset. The probability indication procedure is developed to improve the purity of the sentence decomposition. The proposed approaches of the HMM strive to exploit the contextual correlation hidden among sentences when determining their authorships. In this thesis, it is for the first time the sequential patterns hidden among document elements is considered for such a problem. To build and learn the HMM, a new unsupervised learning method is proposed to estimate its initial parameters. The proposed frameworks do not require the availability of any information of authors or document's context other than how many authors have contributed to writing the document. The effectiveness of the proposed algorithms is proved using benchmark datasets which are widely used for authorship analysis of documents. Furthermore, scientific papers are used to demonstrate the performance of the proposed approaches on authentic documents. Comparisons with recent state-the-art approaches are also presented to demonstrate the significance of our new ideas and the superior performance of the proposed approaches.

Acknowledgements

Sincere feelings and strongest kind words emanating from my heart go to my supervisor **Professor Xiangjian He** for his marvelous support, advice, help and practical suggestions throughout my PhD journey. I am greatly indebted to him for his excellent direction, patient instruction, steadfast encouragement and timely comments to this research throughout the past four years. I owe my research achievements to his excellent supervision. Without his support and supervision, I could not have come this far.

I would like to thank my co-supervisor, **Dr. Wenjing Jia** for her unflinching encouragement, support and her useful comments. Her valuable suggestions and guidance have been a great help and kept me moving ahead at a critical time. I would also like to thank my co-supervisor **Dr. Gengfa Fang** for his advice and assistance.

I am forever appreciate and thankful to my family for all of their love and encouragement. Special thanks to my father, Mr. Waleed Aldebei, and my mother, Mrs. Sameera Abu Alsamen, for their extraordinarily generous help during my study. Without their support, I would never have had a chance to study overseas.

I would also like to express my pure-hearted thanks to my wife, Eng. Alanoud Alnsour, for her love, understanding and patience. You have been a pillar of strength to me. I also thank my darling baby girl, Salma Aldebei, for entering into our lives and making it more fulfilling.

Finally, I would like to extend my thanks to all my colleagues, friends and the staff of the school, especially those people listed below for providing support and friendship that I needed.

- Professor Massimo Piccardi, Associate Professor Qiang Wu, Dr. Min Xu, Dr. Priydarsi Nanda, Shaukat Abedi, Sari Awwad, Ahmed Mian Jan, Minqi Li, Fairouz Hussein, Omar Alshaweesh and Muhammad Usman.

Contents

List of Figures

List of Tables

Abbreviations

AA	**A**uthorship **A**ttribution
DT	**D**ecision **T**ree
EM	**E**xpectation **M**aximisation
GMM	**G**aussian **M**ixture **M**odel
HMM	**H**idden **M**arkov **M**odel
IID	**I**ndependent and **I**dentically **D**istributed
IR	**I**nformation **R**etrieval
K-NN	**K**-**N**earest **N**eighbors
MAP	**M**aximum **A** **P**osterior
MLE	**M**aximum **L**ikelihood **E**stimation
NN	**N**eural **N**etworks
NLP	**N**atural **L**anguage **P**rocessing
NUANCE	**N**-gram **U**nsupervised **A**utomated **N**atural **C**luster **E**nsemble
PIP	**P**robability **I**ndication **P**rocedure
POS	**P**art-**O**f-**S**peech
SUDMAD	**S**equential and **U**nsupervised **D**ecomposition of a **M**ulti-**A**uthor **D**ocument
SVM	**S**upport **V**ector **M**achine
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requency
WAN	**W**ord **A**djacency **N**etwork

Publications

These are the publications resulted from this thesis.

- Khaled Aldebei, Xiangjian He, and Jie Yang. Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model. *Association for Computational Linguistics, Volume 2: Short Papers*, page 501, 2015. (**CORE A*, CCF A & ERA A**)

- Khaled Aldebei, Xiangjian He, Wenjing Jia, and Jie Yang. Unsupervised Multi-Author Document Decomposition Based on Hidden Markov Model. In *ACL (1)*, 2016a. (**CORE A*, CCF A & ERA A**)

- Khaled Aldebei, Helia Farhood, Wenjing Jia, Priyadarsi Nanda, and Xiangjian He. Sequential and Unsupervised Document Authorial Clustering Based on Hidden Markov Model. In *Trustcom/BigDataSE/ICESS, 2017 IEEE*, pages 379–385. IEEE, 2017. (**CORE A & ERA A**)

- Khaled Aldebei, Xiangjian He, Wenjing Jia, and Weichang Yeh. SUDMAD: Sequential and Unsupervised Decomposition of a Multi-Author Document Based on a Hidden Markov Model. *Journal of the Association for Information Science and Technology*, 69(2):201–214, 2018. ISSN 2330-1643. doi: 10.1002/asi.23956. URL http://dx.doi.org/10.1002/asi.23956. (**ERA A* & CCF B**)

Dedicated to My Family