UNIVERSITI PUTRA MALAYSIA

STYLOMETRIC AUTHORSHIP BALANCED ATTRIBUTION PREDICTION METHOD

TAREEF KAMIL MUSTAFA

FSKTM 2011 16

STYLOMETRIC AUTHORSHIP BALANCED ATTRIBUTION PREDICTION METHOD

By

TAREEF KAMIL MUSTAFA

Thesis Submitted to the School of Graduate Studies, University Putra Malaysia,
in Fulfillment of the Requirement for the Degree of Doctor of Philosophy

October 2011

Dedicated to Professor Kamil Alshaibi
God rest his soul …

Abstract of thesis presented to the Senate of University Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

STYLOMETRIC AUTHORSHIP BALANCED ATTRIBUTION PREDICTION METHOD

By

TAREEF KAMIL MUSTAFA

October 2011

Chairman: **Norwati Mustapha, PhD**

Faculty: **Computer Science and Information Technology**

Stylometric authorship attribution is one of the important approaches in the text mining field that has received growing attention due to its delicateness. This approach concerns about analyzing texts such as novels and plays written by famous authors, trying to measure their writing style by choosing some attributes that shows uniquely belong to the author, assuming that each author has a special artistic way of writing that no other author has.

There are two major problems that tie up the progress in this field, which are the predictions accuracy results and the human expert judgment. The techniques that manage such predictions are either using the statistical attributes such as frequent words or the use of more sophisticated semantic techniques such as lexicons. Nonetheless, the results are still considerably less accurate.

In this research, we propose a new Stylometric method known as the Stylometric authorship balanced attribution (SABA) that is able to overcome these problems with higher accuracy prediction and independent from human judgments, which means that the method does not rely on the domain experts. The new method is implemented by merging three methods, which are called the computational approach, the Winnow algorithm and the Burrows-delta method. The proposed method also uses a set of more effective attributes as compared to the frequent words method. This results in higher Stylometric prediction thus far, having more alibis for author artistic writing style for authorship recognition and prediction. The effective attributes are represented by the word pair and the trio, while both are multiple words attributes.

The proposed SABA method is compared against three other methods using the computational approach, the Winnow algorithm method, and the Burrows-delta method. The results showed that the proposed method produces superior prediction accuracy and even provides a completely correct result during the final stage of the experiment.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

STAILOMETRIK PEREKAYASAAN KAEDAH TEKNIK RAMALAN PENULISAN

Oleh

TAREEF KAMIL MUSTAFA

OCT 2011

Pengerusi: **Norwati Mustapha, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Atribusi stailometrik penulisan adalah satu daripada pendekatan yang penting di dalam bidang perlombongan teks yang banyak menerima perhatian disebabkan ketelitiannya. Pendekatan ini memberi perhatian kepada penganalisaan teks seperti novel dan drama yang ditulis of penulis-penulis terkemuka, cuba untuk mengukur gaya penulisan mereka dengan memilih atribut-atribut yang secara uniknya milik seseorang penulis, dengan andaian bahawa setiap penulis mempunyai cara artistik teristimewa dalam penulisan yang tidak dimiliki oleh penulis-penulis yang lain.

Terdapat dua masalah yang besar yang mengekang perkembangan bidang ini, iaitu keputusan ketepatan ramalan dan pertimbangan daripada pakar perseorangan. Teknik-teknik yang mengurus ramalan adalah sama ada menggunakan atribut statistikal seperti perkataan kerap atau penggunaan teknik-teknik semantik yang sofistikated seperti leksikon. Walau bagaimanapun, semua keputusan masih dianggap kurang tepat.

Di dalam penyelidikan ini, kami mencadangkan sebuah algoritma stailometrik yang baharu, yang berupaya mengatasi masalah-masalah tersebut dengan menghasilkan ketepatan ramalan yang lebih tinggi dan tidak bergantung kepada pertimbangan manusia, yang mana memberi maksud bahawa algoritma tersebut tidak perlu bergantung kepada pakar bidang. Algoritma baharu ini dibangunkan dengan menggabungkan tiga kaedah, iaitu pendekatan pengkomputeran, algoritma Winnow dan kaedah Burrows-delta.

Eksperimen telah dijalankan dengan menggunakan set data daripada laman web Gutenberg, yang mengumpul banyak buku-buku kesusasteraan. Walau bagaimanapun, kajian ini mengehadkan skop kepada koleksi 50 novel daripada 5 penulis terkemuka yang telah menerbitkan karya mereka dalam Bahasa Inggeris sewaktu kurun ke-19. Dalam menjalankan eksperimen-eksperimen tersebut, 10 buah buku telah diagihkan kepada setiap penulis, yang mana 9 daripada buku-buku tersebut digunakan untuk tujuan latihan dan buku yang kesepuluh digunakan untuk tujuan pengujian.

Atribusi Stailometrik Penulisan Seimbang (SABA) yang dicadangkan dibandingkan dengan tiga lagi model yang lain, iaitu pendekatan pengkomputeran, algoritma Winnow, dan algoritma Burrow-delta. Keputusan-keputusan menunjukkan bahawa algoritma yang dicadang menghasilkan ketepatan ramalan lebih hebat malahan memberikan keputusan berketepatan penuh di dalam tahap akhir eksperimen.

ACKNOWLEDGEMENT

I would like to take this opportunity and thank my supervisor, Dr. Norwati Mustapha, for her support, guidance's, and understanding. Her comments and suggestions for further development as well as her assistance during writing this thesis are invaluable to me. Her patience, humility, tutorship, interest, teaching and research style have provided for me an exceptional opportunity to learn and become a better researcher.

I would also like to thank the committee members, Dr. Masrah Azrifah Azmi Murad and Associate Professor Dr. Md. Nasir Sulaiman for their help and valuable suggestions.

My deepest appreciation to my family for their utmost support and encouragement without which all these would not be possible, wishing health for my wife.

For the others who have directly or indirectly helped me in the completion of my work, I thank you all.

Finally, my deepest appreciation to University Putra Malaysia and beautiful Malaysia for their support, encouragement and for accepting me in their community and giving me the feeling that I am not far from home.

APPROVAL

I certify that a Thesis Examination Committee has met on 14 October 2011 to conduct the final examination of Tareef Kamil Mustafa on his thesis entitled “Stylometric Authorship Balanced Attribution Prediction Method” in accordance with the Universities and University College Act 1971 and the Constitution of the University Putra Malaysia [P.U.(A) 106] 15 March 1998. The committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Ali Mamat, PhD
Professor Assoc.
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Chairman)

Shyamala Doraisamy, PhD
Professor Assoc
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Internal Examiner)

Hamidah Ibrahim, PhD
Professor Assoc
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Internal Examiner)

Ajith Abraham, PhD
Professor
Machine Intelligence Research Labs
Scientific Network For Innovation & resea
(External Examiner)

SHAMSUDDIN SULAIMAN, PhD
Professor and Deputy Dean
School of Graduate Studies
University Putra Malaysia
Date:

This thesis was submitted to the Senate of University Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Norwati Mustapha, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Chairman)

Masrah Azrifah Azmi Murad, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

Md. Nasir Sulaiman, PhD
Associate professor
Faculty of Computer Science and Information Technology
University Putra Malaysia
(Member)

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
University Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

TAREEF KAMIL MUSTAFA

Date:

TABLE OF CONTENTS