TAXONOMY LEARNING FROM MALAY TEXTS USING ARTIFICIAL IMMUNE SYSTEM BASED CLUSTERING

MOHD ZAKREE BIN AHMAD NAZRI

UNIVERSITI TEKNOLOGI MALAYSIA

TAXONOMY LEARNING FROM MALAY TEXTS USING ARTIFICIAL IMMUNE SYSTEM BASED CLUSTERING

MOHD ZAKREE BIN AHMAD NAZRI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

APRIL 2011

DEDICATION

To almarhum ayahanda and bonda

To my wife

May this inspires our sons and daughter

Wan Nur Aqila, Muhammad and Anas

May Allah bless us all

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful. I thank to Allah for granting me strength and guidance throughout my journey to complete this study.

I could not have completed this work without the help and assistance of a large number of people and organizations.

The author has been truly blessed to have the guidance of two stellar advisors who have defined my PhD experience. My grateful thanks to Prof. Dr. Hjh. Siti Mariyam Hj. Shamsuddin for her wisdom and encouragement. Your supervisory approach has provided me the freedom to explore new ideas. I'm grateful to Prof. Madya Dr. Azuraliza Abu Bakar for her guidance and support throughout the duration of my studies. I have counted you both as mentor and friends these past years. The author would also like to thank Prof. Dr. Jon Timmis of the University of York for inspiring me to explore immune system. I am also indebted to Universiti Kebangsaan Malaysia (UKM) for funding my PhD study. Dr Pouzi Hamzah at Universiti Malaysia Terengganu and Dr. Tang Enya Kong at Universiti Sains Malaysia also deserve special thanks for their assistance in supplying the relevant tools and other research materials.

I would like to thank the Data Mining and Optimization Research Group of Universiti Kebangsaan Malaysia, particularly Prof. Dr. Abdul Razak Hamdan, Assoc. Prof. Dr. Salwani Abdullah, Tri Basuki Kurniawan, Nasser, Saif, Yahya, Salam and Hamza for providing much needed study breaks, discussions, rants, debates and most importantly as anchors for my sanity. Finally, my wife Wan Noraidah Mohammad for your intimate understanding and patience for me whilst writing up my thesis.

ABSTRACT

In taxonomy learning from texts, the extracted features that are used to describe the context of a term usually are erroneous and sparse. Various attempts to overcome data sparseness and noise have been made using clustering algorithm such as Hierarchical Agglomerative Clustering (HAC), Bisecting K-means and Guided Agglomerative Hierarchical Clustering (GAHC). However these methods suffer low recall. Therefore, the purpose of this study is to investigate the application of two hybridized artificial immune system (AIS) in taxonomy learning from Malay text and develop a Google-based Text Miner (GTM) for feature selection to reduce data sparseness. Two novel taxonomy learning algorithms have been proposed and compared with the benchmark methods (i.e., HAC, GAHC and Bisecting K-means). The first algorithm is designed through the hybridization of GAHC and Artificial Immune Network (aiNet) called GCAINT (Guided Clustering and aiNet for Taxonomy Learning). The GCAINT algorithm exploits a Hypernym Oracle (HO) to guide the hierarchical clustering process and produce better results than the benchmark methods. However, the Malay HO introduces erroneous hypernym-hyponym pairs and affects the result. Therefore, the second novel algorithm called CLOSAT (Clonal Selection Algorithm for Taxonomy Learning) is proposed by hybridizing Clonal Selection Algorithm (CLONALG) and Bisecting k-means. CLOSAT produces the best results compared to the benchmark methods and GCAINT. In order to reduce sparseness in the obtained dataset, the GTM is proposed. However, the experimental results reveal that GTM introduces too many noises into the dataset which leads to many false positives of hypernym-hyponym pairs. The effect of different combinations of affinity measurement (i.e., Hamming, Jaccard and Rand) on the performance of the developed methods was also studied. Jaccard is found better than Hamming and Rand in measuring the similarity distance between terms. In addition, the use of Particle Swarm Optimization (PSO) for automatic parameter tuning the GCAINT and CLOSAT was also proposed. Experimental results demonstrate that in most cases, PSO-tuned CLOSAT and GCAINT produce better results compared to the benchmark methods and able to reduce data sparseness and noise in the dataset.

ABSTRAK

Fitur yang diekstrak dalam pembelajaran taksonomi dari teks yang digunakan untuk menggambarkan konteks suatu perkataan lazimnya mempunyai kesalahan (hingar) dan masalah kejarangan data. Beberapa penyelidikan telah cuba mengatasi masalah kejarangan dan hingar dengan menggunakan algoritma pengelompokan seperti Pengelompokan Aglomerat Berhierarki (HAC), Pembahagi-dua K-min dan Pengelompokan Aglomerat Berhierarki Berpandu (GAHC). Walau bagaimanapun, kaedah ini mengalami masalah perolehan kembali yang rendah. Oleh itu, penyelidikan ini bertujuan untuk mengkaji penggunaan dua penghibiridan sistem imun buatan (AIS) dalam pembelajaran taksonomi dari teks Melayu dan pembangunan alat Perlombongan Teks Berasaskan Google (GTM) untuk pengekstrakan fitur bagi mengatasi masalah kejarangan data. Dua algoritma pembelajaran taksonomi dicadangkan untuk mengurangkan masalah kejarangan dan hingar dalam set data. Algoritma pertama direka dengan menghibrid GAHC dan Rangkaian Imun Buatan (aiNet) yang dinamakan GCAINT (Pengelompokan Berpandu dan aiNet untuk Pembelajaran Taksonomi). Algoritma GCAINT mengeksploitasi *Hypernym Oracle* (HO) yang memandu proses pengelompokan berhierarki untuk menghasilkan keputusan yang lebih baik berbanding kaedah lain. Namun, HO bahasa Melayu ini mengandungi perkataan sebagai hipernim atau hiponim yang salah, justru mempengaruhi kualiti taksonomi yang terbentuk. Oleh itu, kaedah kedua dicadangkan iaitu penghibridan antara Algoritma Pemilihan Klonal (CLONALG) dengan Pembahagi-dua K-min yang dinamakan CLOSAT. Keputusan CLOSAT adalah lebih baik berbanding kaedah tanda aras tersebut. Demi mengurangkan masalah kejarangan dalam set data, GTM dicadangkan. Namun, GTM menambah jumlah ralat ke dalam set data yang seterusnya mewujudkan hubungan yang salah diantara perkataan di dalam taksonomi. Pengaruh penggunaan ukuran afiniti dengan kombinasi yang berbeza (seperti Hamming, Jaccard dan Rand) terhadap prestasi kaedah cadangan turut dikaji. Jaccard didapati lebih baik berbanding Hamming dan Rand dalam mengukur afiniti diantara perkataan. Selain itu, alat penalaan parameter automatik berasaskan Pengoptimuman Partikel Secara Berkumpulan (PSO) juga dibangunkan. Keputusan kajian menunjukkan bahawa dalam kebanyakan kes, CLOSAT dan GCAINT yang ditala PSO menghasilkan keputusan yang lebih baik berbanding kaedah lain serta mengurangkan masalah kejarangan dan hingar pada set data.