MAXIMUM TOTAL ATTRIBUTE RELATIVE OF SOFT-SET THEORY FOR DEFINITE CATEGORICAL DATA CLUSTERING

RABEI RAMLI

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

Lihat Sebelah

MAXIMUM TOTAL ATTRIBUTE RELATIVE OF SOFT-SET THEORY FOR EFFICIENT CATEGORICAL DATA CLUSTERING

RABIEI MAMAT

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

STATUS CONFIRMATION FOR DOCTORAL THESIS

MAXIMUM TOTAL ATTRIBUTE RELATIVE OF SOFT-SET THEORY FOR EFFICIENT CATEGORICAL DATA CLUSTERING

ACADEMIC SESSION : 2013/2014

I, **RABIEI MAMAT,** agree to allow this Doctoral Thesis to be kept at the Library under the following terms:

1. This Doctoral Thesis is the property of the Universiti Tun Hussein Onn Malaysia.
2. The library has the right to make copies for educational purposes only.
3. The library is allowed to make copies of this report for educational exchange between higher educational institutions.
4. ** Please Mark (√)

☐	CONFIDENTIAL	(Contains information of high security or of great importance to Malaysia as STIPULATED under the OFFICIAL SECRET ACT 1972)
☐	RESTRICTED	(Contains restricted information as determined by the Organization/institution where research was conducted)
☐	FREE ACCESS	

Approved by,

(WRITER'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address:

LOT 16221, KAMPUNG TANJUNG
21020 BATU RAKIT
KUALA TERENGGANU
TERENGGANU DARUL IMAN

Date : _____

Date: _____

This thesis has been examined on date .....

And is sufficient in fulfilling the scope and quality for the purpose of awarding the Degree of Doctor of Philosophy.

Chairperson:

PROF. DR. HJ. MOHD IDRUS BIN HJ. MOHD MASIRIN

Faculty of Civil and Environmental Engineering

Tun Hussein Onn University of Malaysia

Examiners:

PROF. DR. ABDUL RAZAK BIN HAMDAN

Faculty of Technology and Information Science

National University of Malaysia

ASSOC. PROF. DR. HJH. ROZAIDA BINTI GHAZALI

Faculty of Computer Science and Information Technology

Tun Hussein Onn Universiti of Malaysia

MAXIMUM TOTAL ATTRIBUTE RELATIVE OF SOFT-SET THEORY FOR EFFICIENT CATEGORICAL DATA CLUSTERING

RABIEI MAMAT

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

MARCH 2014

I hereby declare that the work in this project report is my own except for quotations and summaries which have been duly acknowledged

Student : .....

RABIEI MAMAT

Date : .....

Supervisor : .....

PROFESSOR DR. MUSTAFA MAT DERIS

ACKNOWLEDGEMENT

In the name of Allah, The Most Beneficent, The Most Merciful

All praises be to Allah, the Lord of the universe. May Allah bestow His mercy and grace upon His most beloved prophet Muhammad PBUH, his family and his friends. My deepest gratitude to the grace of Allah SWT as with his bounty and mercy, then I can successfully completed this PhD thesis. I would like to take this opportunity to acknowledge the guidance and cooperation of my supervisors, examiners, colleagues, friends and family members during this study period.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Dr. Mustafa Mat Deris, for his continuous support, patience, encouragement and motivation, enthusiasm and knowledge. Also the highest appreciation goes to my second supervisor Dr. Tutut Herawan from Universiti Malaya, for guiding me through his meticulousness, accuracy, dedication to work and insistence on perfection, and by his questions and comments. I could not have imagined having a better second supervisor like him. For both of you, may Allah repay all of your kindness, Insha'Allah.

In this opportunity I also would like to extend my heartfelt thanks to the examiner of this thesis Professor Dr. Abdul Razak Hamdan and Associate Professor Dr. Rozaida Ghazali, without their constructive comments this thesis will not provide any benefits to anyone.

I also would like to thanks to my parents Gayah Jusoh, Hamzah Mohamed and Noriah Ibrahim, my lovely wife Norhasimah Hamzah, my sweet childrens Azrie Najmuddin, Azriena Najiha, Nafiesa and Mohammad Naimzafran for their prayers, love and encouragement. Thanks for everybody who contributed to this achievement in a direct or indirect way.

ABSTRACT

Clustering a set of categorical data into a homogenous class is a fundamental operation in data mining. A number of clustering algorithms have been proposed and have made an important contribution to the issues of clustering especially related to the categorical data. Unfortunately, most of the clustering techniques are not designed to address the issues of uncertainties inherent in the categorical data. However, handling the data uncertainty is not an easy task. One method of handling the data uncertainty in categorical data clustering is by identifying the partition attribute in the information system. But, with this approach, the computational cost is still a major issue and the resulting clusters is still dubious. Thus, in this thesis, the concept of attribute relative which is based on the theory of soft-set is discussed and consequently introduces an alternative technique to the partition attribute selection approach for the used in the categorical data clustering. A technique which called Maximum Total Attribute Relative (MTAR) is able to determine the partition attribute of the categorical information system at the category level without compromising the computational cost and at the same time enhance the legitimacy of the resulting clusters. Experiments on sixteen (16) UCI-MLR benchmark datasets demonstrate the potentials of MTAR to achieved lower computational time with the improvements up to 90% as compared to TR, MMR, MDA and NSS. Experiments also show the objects in the clusters produced by MTAR technique has obvious similarities and the generated clusters also have better objects coverage simultaneously increased the cluster validity up to 23% in term of entropy as compared to MDA.

ABSTRAK

Pengklusteran satu set data berkategori ke dalam kelas homogen adalah operasi asas dalam perlombongan data. Beberapa algoritma pengklusteran data telah dicadangkan dan telah memberi sumbangan yang besar kepada isu-isu pengklusteran terutamanya yang berkaitan dengan data berkategori. Malangnya, kebanyakan teknik pengklusteran tidak direka untuk menangani isu-isu ketidakpastian yang wujud dalam data berkategori. Walau bagaimanapun, pengendalian ketidakpastian data bukanlah satu tugas yang mudah. Salah satu kaedah pengendalian ketidapastian data dalam pengklusteran data berkategori ialah dengan mengenalpasti atribut partisi dalam sistem maklumat. Tetapi melalui pendekatan ini, kos komputasi masih menjadi isu utama dan kluster-kluster yang dihasilkan masih diragui kesahihannya. Justeru itu, di dalam tesis ini, konsep relatif atribut yang berasaskan kepada teori set-lembut dibincangkan dan seterusnya memperkenalkan satu teknik alternatif kepada pendekatan pemilihan atribut partisi untuk digunakan dalam pengklusteran data berkategori. Teknik yang dipanggil sebagai Jumlah Atribut Relatif Maksimum (MTAR) dapat menentukan atribut partisi sesebuah sistem maklumat berkategori diperingkat kategori tanpa menjejaskan kos komputasi dan pada masa yang sama meningkatkan kesahihan kluster-kluster yang dihasilkan. Uji kaji ke atas enam belas (16) set data penanda aras UCI-MLR menunjukkan potensi MTAR untuk mencapai masa komputasi yang lebih rendah dengan penambahbaikkan sehingga 90% berbanding dengan TR, MMR, MDA dan NSS. Eksperimen juga menunjukkan objek di dalam kluster-kluster yang dihasilkan oleh teknik MTAR mempunyai persamaan yang jelas dan kluster-kluster yang dihasilkan juga mempunyai litupan objek yang lebih baik dan pada masa yang sama meningkatkan kesahihan kluster sehingga 23% berasaskan entropy berbanding dengan MDA.