SELF-ORGANIZIN MAP CLUSTERING METHOD FOR THE ANALYSIS OF E-LEARNING ACTIVITIES

MUSA WAKIL BARA

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Award of Degree of Masters of Computer Science

Faculty of Computing

Universiti Teknologi Malaysia

JANUARY, 2017

I dedicated this work to my late father Mallam Wakil Jampai Bara.

ACKNOWLEDGEMENT

All praise is due to Allah the Beneficent, the Merciful who creates man and honor him with knowledge. May the Peace and Blessings of Allah be upon His Noble Prophet Muhammad ﷺ who commands his followers to seek for knowledge.

I would like to appreciate the management of UTM in general and Faculty of Computing in particular for the provision of qualified staff and enough & up-to-date research materials.

I sincerely and faithfully acknowledge the effort of my supervisor for her tireless and uncountable support who taught me the research from afresh. Dr. Bahiah Binti Haj Ahmad my supervisor, the backbone of my research I say a big thank you.

I also acknowledge effort and fatherly advice of Eng. Hussaini Abacha, the rector of Mai Idriss Alooma Polytechnic Geidam Yobe state and the entire management of MIAPOLY for its support. I also thank the body of TETFund for their intervention which is the financial backbone of my studies.

The appreciations also go to Dr. Hassan S. Jibrin for his maximum support right from my application stage up to completion stage. I would like to acknowledge supports of Mallam Maidala B. Bularafa (DFS) Alh. Muh’d Jiddah and Mal. Umar M. Tela for looking after my family while on my study.

I appreciate the supports of my friends at home, at my institution and course mates for their courageous advices.

ABSTRACT

Students' interactions with e-learning vary according to their behaviours which in turn, yield different effects to their academic performance. Some students participate in all online activities while some students participate partially based on their learning behaviours. It is therefore important for the lecturers to know the behaviours of their students. But this cannot be done manually due to the unstructured raw data in students' log file. Understanding individual student's learning behaviour is tedious. To solve the problem, data mining approach is required to extract valuable information from the huge raw data. This research investigated the performance of Self-organizing Map (SOM) to analyze students' e-learning activities with the aim to identify clusters of students who use the e-learning environment in similar ways from the log files of their actions as input. A study on Meaningful Learning Characteristics and its significance on students' leaning behaviors were carried out using multiple regression analysis. Then SOM clustering technique was used to group the students into three clusters where each cluster contains students who interact with the E-learning in similar ways. Behaviors of students in each cluster were analyzed and their effects on their learning success were discovered. The analysis shows that students in Cluster1 have the highest number of interactions with the e-learning (Very Active), and having the highest final score mean of 91.12%. Students in Cluster2 have less number of interactions than that of Cluster1 and have final score mean of 75.65%. Finally, students Cluster3 have least number of interactions than the remaining clusters with final score means is 36.57%. The research shows that, students who participate more in Forum activities emerged the overall in learning success, while students with lowest records on interactions have lowest performance. The research can be used for early identification of low learners to improve their mode of interactions with e-learning.

ABSTRAK

Interaksi pelajar dalam e-pembelajaran berbeza mengikut tingkah laku mereka dan seterusnya, menghasilkan kesan yang berbeza kepada prestasi akademik mereka. Sesetengah pelajar mengambil bahagian dalam semua aktiviti yang disediakan dalam talian, manakala sesetengah pelajar hanya terlibat dalam sebahagian aktiviti sahaja berdasarkan tingkah laku pembelajaran mereka. Oleh itu, adalah penting bagi pensyarah untuk mengetahui tingkah laku pelajar mereka. Walaubagaimanapun, ianya tidak boleh dilakukan secara manual kerana data mentah dalam fail log pelajar adalah tidak berstruktur. Memahami tingkah laku pembelajaran setiap individu pelajar adalah membosankan. Untuk menyelesaikan masalah ini, pendekatan perlombongan data diperlukan untuk mendapatkan maklumat berharga daripada data mentah yang besar. Kajian ini menyiasat prestasi kaedah Self Organising Map (SOM) untuk menganalisis aktiviti e-pembelajaran pelajar dengan tujuan untuk mengenal pasti kelompok pelajar yang menggunakan persekitaran e-pembelajaran dengan cara yang sama. Satu kajian mengenai ciri-ciri pembelajaran yang bermakna dan kepentingannya ke atas kelakuan pelajar telah dijalankan dengan menggunakan analisis regresi berganda. Teknik SOM telah digunakan untuk mengumpulkan pelajar kepada tiga kelompok di mana setiap kelompok mengandungi pelajar yang berinteraksi dengan e-pembelajaran dalam cara yang sama. Kelakuan pelajar dalam setiap kelompok telah dianalisa dan kesannya terhadap kejayaan pembelajaran mereka ditemui. Analisa menunjukkan bahawa kelompok1 mengandungi pelajar yang mempunyai bilangan tertinggi interaksi dengan e-pembelajaran (Sangat Aktif), dan yang mempunyai min skor akhir tertinggi 91.12%. Kelompok2 pula mengandungi pelajar yang mempunyai bilangan interaksi yang kurang berbanding kelompok1 tetapi mempunyai min skor akhir 70.65%. Akhir sekali, kelompok3 mengandungi pelajar yang mempunyai sekurang-kurangnya beberapa interaksi daripada kelompok lain tetapi skor akhir mereka adalah 36.57%. Berdasarkan tindakan, kelompok1 disebut sebagai 'Sangat Aktif' manakala kelompok2 adalah 'Aktif' dan kelompok3 disebut sebagai 'kurang-Aktif. Walaupun berdasarkan kejayaan tentu, kelompok1 adalah mengandungi Pelajar Tinggi, Kelompok2 mengandungi Pelajar Rendah manakala kelompok3 mengandungi pelajar sederhana. Hasil kajian ini dapat membantu pensyarah untuk mengetahui bahawa pelajar mempunyai tingkah laku pembelajaran yang berbeza dan setiap tingkah laku membawa hasil pembelajaran yang berbeza.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACI	-	Active, Constructive, Intentional
FCM	-	Fuzzy C-Means
LMS	-	Learning Management System
Moodle	-	Modular Object-Oriented Dynamic Learning Environment
MRA	-	Multiple Regression Analysis
SOM	-	Self-Organizing Map
UTM	-	Universiti Teknologi Malaysia

CHAPTER 1

INTRODUCTION

1.1 Overview

The advancement in information technology has enhanced the effectiveness of web-based education (e-learning) system. The e-learning system allows students from anywhere and at any time to carryout different learning activities such as reading, presentations, chatting and assignment. These activities usually take palace via a platform called Learning Management System (LMS), a platform that utilizes various technologies mostly on the internet to provide access to the courses and facilitates communications between students and tutors and/or among the students. The LMS provides a large volume of heterogeneous data for use by the students and the tutors as well. The learning materials are enough to meet the requirements to achieve the learning objectives and accommodate student needs. The learning management system also contains information about students' personal information, learning styles & behaviors and usability preferences (Dráždilová et al., 2010), these information are stored in log files.

Even though the learning management system contains such huge amount of data, it however, does not provide a useful means of knowledge discovery about the data in the system's log files, thus making it challengeable to manually analyze the data (Alias et al., 2013).

. Another factor is the growth of data in the LMS, making the data unstructured, thus making it difficult to extract the desired information in the best or average case time (Félix et al., 2007). When considering the fact that learners are from different background such as educational, cultural and psychological, their learning styles and behaviors may also vary (Technology & Benha, 2010) & (Myszkorowski & Zakrzewska, 2013). The success of e-learning depends greatly on the learner's positive interest toward e-learning. To make the e-learning more effective, the e-learning environment must be adjusted to meet individual/student needs, but this is very difficult in terms of cost, time and storage management. A better way to solve the problems is to divide (classify) the students according to groups of certain similarities such as learning styles and behaviors. To achieve the process, a data mining approach is required.

Data mining has various definitions by researchers, but its meaning and goal remained the same. It is the process of analyzing data from different perspectives and summarizing it into useful information, it can also be defined as the iterative and interactive process of discovering valid, novel, useful and understandable knowledge (patterns, models, rules, etc.) in massive databases. In e-learning, data mining (also called educational data mining) is an emerging discipline, concerned with development of methods for exploring the uniqueness of data that come from educational database, and use those methods to better understand learners and their learning settings (Efrati et al., 2014). These methods include classification, clustering, prediction and visualization. The process and phases of data mining in e-learning will be discussed in chapter three (research methodology) of this research. A data mining technique that groups students into certain similarities is called clustering.

Clustering is an unsupervised supervised partitioning of objects or patterns into groups of similar objects (clusters). More precisely, clustering technique groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups (Rokach & Maimon, 2010). To achieve the best result of clustering, different computational intelligence methods such as Artificial Neural Networks, Fuzzy Logic, Decision Tree, K-means, Self-Organizing Map (SOM), etc. are used (Drážkilová et al., 2010).

Factors that affect students' interactions with e-learning include their learning behaviors. The learning behaviors of students are defined by meaningful learning characteristics. The meaningful learning characteristics show how students behave upon interaction with e-learning. However, students' learning behaviors cannot be identified directly from the log files. Data mining techniques can be used to extract the desired information from log files and use the information to identify students' learning behavior. Lack of knowledge about students' learning behavior may risk the students in their learning outcome. This is because the lecturer may not identify a certain student who has weakness in participating in some e-learning activities such as forum discussion or wiki (Tai et al., 2014), (Akcapynar et al., 2014) and (Alias et al., 2013).

1.2 Problem Background

The performance of students in e-learning environment depends on how meaningful the student learning behavior is. In other words, the success factor of e-learning is greatly dependent on meaningful learning characteristics that define student's way of interaction with the learning environment. Some students may have similarities in terms of learning behavior and vary with other students. Different learning behaviors yield different learning performance. Some learning behaviors can lead students to success while some can lead students to failure. Understanding individual learning behavior is tedious especially when they are many. Grouping such students according to their behaviors is a good solution that attracts the attention of many researchers such as (Zakrzewska, 2008) and (Zakrzewska, 2012). The idea of grouping students according to behaviors in interaction with the e-learning is to identify students that have weak learning behavior and improve the system to motivate them, it also help the administrators to adjust the e-learning environment such that students in certain group would be treated easily.

Researchers used different methods for grouping students according how they interact with the e-learning. The problem of clustering students is associated with the choice of appropriate features/attributes that reflect their learning characteristics and the right algorithm (tool) to implement the analysis.

For example Fuzzy techniques was used by (Technology & Benha, 2010) to cluster students based on their behaviors into different categories, the author used five different attributes on the actions of students on the e-learning. The attributes include the learners' visit on the web based on On/Off campus, Day/Night hours, Class/Lab vs Non-class days, Number of hits and Number of Class-Notes download each attribute is weighted according to its impact on reflecting students' behavior. However, these attributes cannot justify student's learning behavior because it is based on time rather than actions.

In their work, (Alias et al., 2013) analyzed students' action and behavior while using Moodle e-learning system in order to identify the students behaviour. The authors used attributes such as Course_View, Assignment_View, Assignment_Submit, Assignment_I, Assignment_G and apply SOM clustering technique to group the students. They however did not state the significance of the meaningful learning characteristics on student learning behavior. Artificial Neural Network was also used by (Bernad et al., 2015) to identify students' learning styles and its relation to learning effectiveness. The approach was evaluated with data from 75 computer science undergraduate students including their behavior data in university courses and their results on the Index of Learning Style (ILS) questionnaire. (Octavia et al., 2015) used C-means clustering technique to analyze students' e-learning usage based on meaningful learning characteristics, active, authentic, cooperate, collaborative and intentional. The attributes used in the research are the 21 activities found in Moodle e-learning environment. The activities include assignment view, assignment submit, blogs, journal, chat, forum, etc.

SOM clustering method was applied by (Akcapynar et al., 2014) to analyze data obtained from 74 undergraduate students who took Computer Hardware course, the researchers use ten (10) attributes of each student for the clustering. The attributes include login count, assignment count, total time spent on discussion forum and so on. The result was made by using MATLAB software. Three different clusters were found and are termed as higher learners (cluster1), medium learners (cluster2) and low learners (cluster3). The researchers only grouped the students but did not discuss on the effects of students' behavior of each cluster to their learning success.

Based on the literature, some limitations were identified such as identifying the significance of meaningful learning characteristics on student's learning behavior and identifying the effects of students' learning behavior in each cluster on their learning success. These problems are formulated in Figure 1.1.

E-learning is the transfer of knowledge through digital signals using computers and other digital devices.

Learning Management System (LMS)

Provides access to the e-learning resources and facilitates communication between students and lecturers and/or among community of students stores all activities of students in log files. Examples of LMS are Moodle, FSLM, ATutor, and Axis.

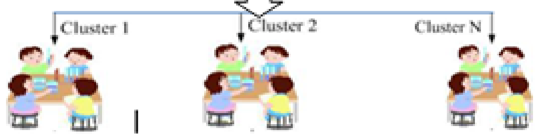
E-learning Activities

Issues

- The LMS contains huge raw data
- LMS does not provide means to discover vital information about students' learning behaviors;
- Different behaviors of students have different effects on their academic performance

Solutions

- ✓ Preprocess the raw data for analysis
- ✓ Choose the best data mining tool for the analysis
- ✓ Group the students into clusters of students who interact with the e-learning in similar ways



Analyze the behaviors of students in each cluster and identify its effects on their academic performance

Figure 1.1 Problem Formulation

1.3 Problem Statement

Unlike traditional face-to-face learning system in which the learning activities take place via direct physical contact, the e-learning system supplements and eases the activities of learning through electronic devices such as the computer. The learning materials are uploaded online and the students can access them at anytime from anywhere. Most activities such as course viewing and course downloading, assignment submission, discussion forum and assessments are done online via a platform called learning management system. However, students' interactions with the LMS vary according to their learning behaviors. Some students are highly active; some are medium active while some are low active. Students also vary prioritizing the e-learning activities. Some students pay more attention on attendance and compulsory assignment while some will give priority to most of the activities such as participating in discussion forums, participation in class exercise and searching for additional learning material other than the ones produced by the lecturers.

In order to improve the e-learning such that the aforementioned variation among the students will be minimized, the following issues are put into consideration.

1. What are the significances of meaningful learning characteristics on students' learning behaviors upon interaction with e-learning environment?
2. Does Self-Organizing Map (SOM) capable of grouping the students into clusters of students with similar ways of interaction with the e-learning environment?
3. What are the effects of students' behaviors from each cluster on their learning performance?

This research is concerned with students' involvement in e-learning. It studies and analyzes the effects of students' behaviors to their learning performance upon carrying out the e-learning activities. The research used data obtained from log files of Computer Science undergraduate students taking a Data Structure course at

Faculty of Computing, University Teknologi Malaysia. The research is aimed to investigate the performance of SOM clustering technique to obtain clusters of students who have similar was of interacting with the e-learning environment, and analyze the effects of their behaviors on their learning performance. Students who have high grades are termed as highly active while those with fewer grades are termed as low active. The proposed model will improve the existing model by identifying the low active students and convert them to become highly active.

1.4 Dissertation Aim

The main aim of this research is to investigate the performance of Self Organizing Maps (SOM) algorithm in grouping e-learning students into clusters of students who have similar ways of carrying out their activities in e-learning environment, and to identify the effects of students' learning behavior in relation to their academic performance from each Cluster in order to identify High-Active and Low-Active students; and to improve the Low-Active students to become High-Active.

1.5 Objectives

This dissertation is intended to achieve the following objectives:

1. To study and analyze the significance of meaningful learning characteristics on student's learning behavior upon their interaction with e-learning.
2. To investigate the performance of Self-Organizing Map (SOM) clustering technique in grouping students into clusters of students with similar ways of interaction with the e-learning.

3. To correlate the effects of students' learning behavior in relation to their academic performance in order to improve the comprehensibility of e-learning environment.

1.6 Research Scopes

The scopes of this dissertation are described as below:

1) The research focuses on students' behavior upon interaction with the e-learning environment as the input data;
2) Multiple regression analysis was done in SPSS
3) The dataset is obtained from log files of 22 Computer Science undergraduate students taking a Data Structure course from Faculty of Computing, Universiti Teknologi Malaysia;
4) The Self-organizing Map (SOM) phases are done using MATLAB.

1.7 Significant of the Research

This research is aimed to describe students' involvement in e-learning. It demonstrates how data mining methods are implemented to analyze the activities of students as they interact with the e-learning environment. The output of the research is the clusters of students who use the e-learning environment in the same manner. The research result is useful to lecturers by providing feedback about the learning behavior of students and how the behavior affects the student's academic performance. It can also be used to trace student who is at risk of failing a particular course, this can be done if the lecturer discovers that a new student's learning behavior resembles that of students who previously failed the course. The research can also help lecturers to identify and improve the structure of a particular course or topic within a course that the students find it difficult to comprehend. Therefore, the overall achievement of this research is to improve the comprehensibility of e-learning environment by understanding the behaviors of the students.

REFERENCES

Ahmad, N. B., Ishak, M. K., Alias, U. F., & Mohamad, N. (2015). An Approach for E-Learning Data Analytics using SOM Clustering. *International Journal of Advances in Soft Computing & Its Applications, 7*(3).

Akcapynar, G., Altun, A., & Cosgun, E. (2014). Investigating Students' Interaction Profile in an Online Learning Environment with Clustering. 109-111. doi: 10.1109/icalt.2014.40

Alias, U. F., Ahmad, N. B., & Hasan, S. (2006). STUDENT BEHAVIOR ANALYSIS USING SELF-ORGANIZING MAP CLUSTERING TECHNIQUE. *learning, 22*, 27.

Amandu, G. M., Muliira, J. K., & Fronda, D. C. (2013). Using moodle e-learning platform to foster student self-directed learning: Experiences with utilization of the software in undergraduate nursing courses in a Middle Eastern university. *Procedia-Social and Behavioral Sciences, 93*, 677-683.

Ariffin, N. H. M., Abd Rahman, H., Alias, N. A., & Sardi, J. (2014). *A survey on factors affecting the utilization of a Learning Management System in a Malaysian higher education.* Paper presented at the e-Learning, e-Management and e-Services (IC3e), 2014 IEEE Conference on.

Bapu, G. K., Deshmukh, M. P. R., Ashok, M. B., Shamrao, S. P., & Tanaji, S. G. Clustering Moodle Data As a Tool For Profiling Students.

Bernard, J., Chang, T.-W., Popescu, E., & Graf, S. (2015). Using Artificial Neural Networks to Identify Learning Styles. *9112*, 541-544. doi: 10.1007/978-3-319-19773-9_57

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. 11-15. doi: 10.1145/2567574.2567604

Bovo, A., Sanchez, S., Héguy, O., & Duthen, Y. (2013). *Clustering moodle data as a tool for profiling students.* Paper presented at the e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on.

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems *Evolution of teaching and learning paradigms in intelligent environment* (pp. 183-221): Springer.

Chattopadhyay, S., Pratihar, D. K., & Sarkar, S. C. D. (2012). A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms. *Computing and Informatics, 30*(4), 701-720.

Chen, Y., Qin, B., Liu, T., Liu, Y., & Li, S. (2010). The Comparison of SOM and K-means for Text Clustering. *Computer and Information Science, 3*(2), 268.

Dráždilová, P., Obadi, G., Slaninová, K., Al-Dubaee, S., Martinovič, J., & Snášel, V. (2010). Computational intelligence methods for data analysis and mining of elearning activities *Computational intelligence for technology enhanced learning* (pp. 195-224): Springer.

Efrati, V., Limongelli, C., & Sciarrone, F. (2014). A Data Mining Approach to the Analysis of Students' Learning Styles in an e-Learning Community: A Case Study *Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge* (pp. 289-300): Springer.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. New York: Springer.

Ghouila, A., Yahia, S. B., Malouche, D., Jmel, H., Laouini, D., Guerfali, F. Z., & Abdelhak, S. (2009). Application of Multi-SOM clustering approach to macrophage gene expression analysis. *Infect Genet Evol, 9*(3), 328-336. doi: 10.1016/j.meegid.2008.09.009

Jiang, Y. H. (2014). The Application Research of Fuzzy C Means Clustering Algorithm in Middle School Students' Network Learning Media Preference. *Applied Mechanics and Materials, 644-650*, 2051-2054. doi: 10.4028/www.scientific.net/AMM.644-650.2051

Kardan, S., & Conati, C. (2010, June). A framework for capturing distinguishing user interaction behaviors in novel interfaces. In *Educational Data Mining 2011*.

Kularbphettong, K., & Tongsiri, C. (2012). Mining Educational Data to Analyze the Student Motivation Behavior. *World Academy of Science, Engineering and Technology, 68*(2012), 1256-1259.

Mavroeidis, D., & Marchiori, E. (2011). A novel stability based feature selection framework for k-means clustering *Machine Learning and Knowledge Discovery in Databases* (pp. 421-436): Springer.

Miche, Y., Akusok, A., Veganzones, D., Björk, K.-M., Séverin, E., du Jardin, P., . . . Lendasse, A. (2015). SOM-ELM—Self-Organized Clustering using ELM. *Neurocomputing, 165*, 238-254.

Myszkorowski, K., & Zakrzewska, D. (2013). Using fuzzy logic for recommending groups in e-learning systems *Computational Collective Intelligence. Technologies and Applications* (pp. 671-680): Springer.

Niu, K., Niu, Z., Liu, D., Zhao, X., & Gu, P. (2014). A Personalized User Evaluation Model for Web-Based Learning Systems. 210-216. doi: 10.1109/ictai.2014.39

Octaviani, D., Othman, M. S., Yusof, N., & Pranolo, A. (2015). APPLIED CLUSTERING ANALYSIS FOR GROUPING BEHAVIOUR OF E-LEARNING USAGE BASED ON MEANINGFUL LEARNING CHARACTERISTICS. *Jurnal Teknologi, 76*(1).

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review, 10*(1).

Premlatha, K., & Geetha, T. (2015). Learning content design and learner adaptation for adaptive e-learning environment: a survey. *Artificial Intelligence Review, 44*(4), 443-465.

Ratnapala, I., Ragel, R., & Deegalla, S. (2014). *Students behavioural analysis in an online learning environment using data mining.* Paper presented at the Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Sabitha, A. S., Mehrotra, D., & Bansal, A. (2015). Delivery of learning knowledge objects using fuzzy clustering. *Education and Information Technologies*. doi: 10.1007/s10639-015-9385-5

Sisovic, S., Matetic, M., & Bakaric, M. B. (2015). Mining student data to assess the impact of moodle activities and prior knowledge on programming course success. 366-373. doi: 10.1145/2812428.2812459

Tai, D. W.-S., Wu, H.-J., & Li, P.-H. (2007). *A hybrid system: neural network with data mining in an e-learning environment.* Paper presented at the Knowledge-Based Intelligent Information and Engineering Systems.

Technology, E. E., & Benha, T. (2010). Evaluation of E-Learners Behaviour using Different Fuzzy Clustering Models : A Comparative Study, *7*(2), 131–140.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). A Learning Analytics Approach to Correlate the Academic Achievements of Students with Interaction Data from an Educational Simulator *Design for Teaching and Learning in a Networked World* (pp. 352-366): Springer.

Yunianta, A., Yusof, N., Othman, M. S., & Octaviani, D. (2012). Analysis and categorization of e-learning activities based on meaningful learning characteristics. *World Academy Of Science, Engineering And Technology: Singapore*, 723-728.

Zakrzewska, D. (2008). Cluster analysis for users' modeling in intelligent E-learning systems. *New Frontiers in Applied Artificial Intelligence*, 209-214.

Zakrzewska, D. (2009). Cluster analysis in personalized e-learning systems *Intelligent Systems for Knowledge Management* (pp. 229-250): Springer.