ROUGH SETS THEORY FOR TRAVEL DEMAND ANALYSIS IN MALAYSIA

WONG JENN HWEE

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

OCTOBER 2008

ABSTRACT

This study integrates the rough sets theory into tourism demand analysis. Originated from the area of Artificial Intelligence, the rough sets theory was introduced to disclose important structures and to classify objects. The Rough Sets methodology provides definitions and methods for finding which attributes separates one class or classification from another. Based on this theory can propose a formal framework for the automated transformation of data into knowledge. This makes the rough sets approach a useful classification and pattern recognition technique. This study introduces a new rough sets approach for deriving rules from information table of tourist in Malaysia. The induced rules were able to forecast change in demand with certain accuracy.

ABSTRAK

Kajian ini menggabungkan *Teori Set Kasar* di dalam mandala perlancongan di Malaysia. Konsep ini merupakan salah satu cabang teknik di dalam bidang Kepintaran Buatan. Konsep *Teori Set Kasar* dipersembahkan untuk mengenalpasti kepentingan struktur dan pengelasan data bagi ojeck yang berkaitan. Kaedah ini menyediakan takrifan dan tatacara untuk mencari ciri-ciri yang berbeza di dalam satu kelas yang berkaitan. Kajian ini juga mencadangkan satu struktur piawai untuk menjelmakan data input ke dalam bentuk pengetahuan yang bermakna. Hasil kajian ini adalah satu set peraturan daripada jadual maklumat perlancongan Malaysia. Peraturan-peraturan yang diperolehi itu boleh digunapakai untuk meramal pola kedatangan pelancong ke Malaysia samada kedatangan yang bertambah atau berkurang. Analisa peramalan ini dilaksanakan dengan mencari kekuatan peraturan berdasarkan kepada pengukuran kepentingan peraturan, panjang peraturan dan cakupan peraturan.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

In an economy, such as Malaysia, in which a significant part of export revenues are due to foreign tourism, it is important for policymakers to understand the sensitivity of foreign tourism demand with respect to its main determinants. One of the key and preliminary elements in the planning process is to study the demand for tourist arrivals in terms of both volume and determinants. At the macro level, accurate forecasting results can help a destination predict the contributions and consequences of visitors to the local economy, culture, and environment. As well, the government bodies can project public revenues from tourism, and ensure that appropriate capacity and infrastructures can be maintained [21]. At the micro level, practitioners can use accurate forecasting results to set up operational requirements while investors can study project feasibility [21]. As a result, accurate planning can minimize if not totally avoid the economic loss due to either excessive or inadequate supply.

An increased number of papers have studied tourism demand forecasting, however, these past studies have predominantly applied statistical or econometric techniques to measure and predict the future market performance [22] in terms of the number of tourist arrivals in a specific destination. Econometric forecasting techniques are very much highly exploited in empirical studies but only margin improvement could be attained with a substantial amount of efforts because the development of such techniques might have reached a plateau for the time being. Because of this situation, academic researchers have attempted to incorporate data mining techniques to tourism demand forecasting and have achieved some ground breaking outcomes [23].

In general, data mining refers to the process of discovering useful patterns, correlations, and rules, which are previously unknown, by filtering through a large amount of data stored in some repositories [24]. The central idea for data mining is to perform an automatic or semi-automatic identification to determine the hidden relationships and patterns which are beyond humans' manual capacity. In a business environment, data mining would be useful for managers to analyze and explore market opportunities and threats, and particularly those inherent in growing or declining markets.

This study presents an approach that applies rough sets theory to form a model for tourism demand Malaysia. Rough sets theory was introduced by Pawlak in 1982. Its methodology provides definitions and methods for finding which attributes separates one class or classification from another and based on this theory one can propose a formal framework for the automated transformation of data into knowledge. Since inconsistencies are allowed and membership in a set does not have to be absolute, the potential for handling noise gracefully is big. Besides, rough sets approach is based on data-mining techniques to discover knowledge.

Rough Sets are efficient and useful tools in the field of knowledge discovery to generate discriminant and characteristic rules and fairly useful clever technique that has been applied to the many domain and is used for the discovery of data dependencies, evaluates the importance of attributes, discovers the patterns of data, reduces all redundant objects and attributes, and seeks the minimum subset of attributes.

Unlike some soft computing technique, rough set analysis do not required external parameters and only use the information presented in the known data. It does not need membership functions and prior parameter settings. It can extract knowledge from the data itself by means of indiscernibility relations and generally needs fewer calculations compare to fuzzy set theory. The attribute reduction algorithm removes redundant information or features and selects a feature subset that has the same discernibility as the original set of features. The selected features can describe the decision as well as the original whole features set, leading to better prediction accuracy.

From the travel demand analysis point of view, this aims at identifying subsets of the most vital attributes influencing the tourist arrival. The chosen subsets are then engaged within a decision rule generation process, creating descriptive rules for the classification task, which may potentially reveal profound knowledge. These decision rules are more useful for experts or policymakers to analyze and gain understanding into the problem at hand. Decision rules extracted by rough set algorithms are concise and valuable, which can be benefit to the experts by enlightening some knowledge hidden in the data.

A rough set is a formal approximation of a crisp set which is *conventional set*, in terms of a pair of sets which give the *lower and the upper approximation* of the original set. The lower and upper approximation sets themselves are crisp sets in the standard version of rough set theory [35]. Rough set technique consists of discretization process, reduct generating, rules derivation and classification. There are many discretization algorithms inside rough set technique to discretize the continuous valued attributes. One of them is *Equal Frequency Binning discretization*.

The *reducts* is the subset of attributes in the information system which are more important in knowledge represented in the equivalence class structure than other attributes. The subset of attributes can fully characterize the knowledge in the data by itself. The reduct of an information system is not unique. There are many subsets of attributes which preserve the equivalence class structure expressed in the information system. The set of attributes which is common to all reducts is called the core. The core is the set of attributes which is possessed by every valid reduct, and therefore consists of attributes which cannot be removed from the information system without causing collapse of the equivalence class structure. The cores are as the set of necessary attributes for the category structure to be represented. Reduct with minimum cardinality is also needed. The reduct with minimal cardinality is the reduct with minimal length. Then the *rules* derivations are based on the generated reducts. The measurements of the significant rules are based on the support of the rules generated, the length of the rules and the rule important measure (RIM) which is elaborated in the next chapter.

The rough sets approach has been found successful in pattern recognition and object classification in medical and financial fields (Slowinski & Zopounidis, 1995; Tanaka & Maeda, 1998). The theory has been incorporated into tourism and hospitality research by Law and Au (1998, 2000), and Au and Law (2000).

However, no work has ever linked with rough sets theory in modelling and forecasting Malaysia tourism demand analysis. Hence, this study is an attempt to forecast the travel demand in Malaysia and the impact of advertisement broadcasted by the media with the theme of *TAHUN MELAWAT MALAYSIA.*

1.2 Problem Background

Normally the tourism data often grow very large so that human inspection and interpretation of the data is not feasible. There is a gap between data generation and data understanding as a result. So, tools and techniques that can assist in extracting unknown interesting patterns buried in the data would be useful to help bridge this gap.

To understand the relationship between tourist arrivals and their determining factors, most of the existing studies focuses on tourism demand forecasting apply economic models that use mathematical functions, which require many statistical assumptions and limitations [2]. However, the models do not provide the sufficient predictive ability when it comes to problems involving interactions among many interdependent variables with unknown probability distribution. In other mean, those models are unable to perform consistently well in situations where the exogenous variables correlate with each other, and when distributions of the samples of variables do not meet the required independent and identical distribution (iid) condition [1]. Econometric forecasting techniques are very much highly exploited in empirical studies but only margin improvement could be attained with a substantial amount of efforts because the development of such techniques might have reached a plateau for the time being. In the context of tourism, Law [4] stated that one of the intrinsic problems that managers have is the large amount of raw data carried in the industry, and these data are basically not comprehensible to the non-technical practitioners.

Articles on tourism demand modelling incorporate up-to-date developments in econometric methodology have reached conflicting conclusions in terms of the methods that generate the most accurate forecasts. For example, Kulendran & King (1997) and Kulendran & Witt (2001) found that economic models were still outperformed by simple univariate time series models. By contrast, Kim and Song (1998) and Song, Romilly and Liu (2000) found that the forecasting performance of econometric models was superior to simple time series models.

Three main reasons conflicting results may arise. First, due to different methodologies [25], the performance of econometric models is sensitive with. Therefore, the Johansen co-integration technique [26] used by Kulendran and King (1997) and Kulendran and Witt (2001) may well lead to different conclusions than the Engle–Granger two-stage approach [27] used by Kim and Song (1998) and Song et al. (2000). Second, different data frequencies may lead to different conclusions. For instance, Kim Song (1998) and Song *et al.* (2000) used annual data, whereas Kulendran and King (1997) and Kulendran and Witt (2001) used quarterly data. It may well be that annual data have fewer unit roots and fewer co-integrating vectors than the same series at quarterly frequency, and different co-integrating relationships usually lead to different Error Correction Models (ECMs). Third, econometric studies of tourism demand generally assume that the structure of the model used for forecasting is constant over time. For example, the parameters of the model remain unchanged over the sample period. This assumption may be too restrictive, and result in time series models out-performing econometric models.

In view of the growing importance of data mining in tourism demand analysis, various published algorithms have been applied to forecast tourism demand in tourism research journals. The most commonly used algorithms are neural networks, Bayesian classifier, genetic algorithms, and fuzzy time-series theory. The intelligent techniques such as neural networks and fuzzy theory are based on assumptions for knowledge about dependencies, probability distributions and large number of experiments [28]. It cannot derive conclusions from incomplete knowledge or manage inconsistent information like tourism dataset. Fuzzy theory need to convert the numerical rules to the table rules form and produce long operation to have the result. On the other hand, rule-based classification process associated with neural network is not easy to explain as rules that are meaningful to the user. Moreover, in the neural networks, more robust features are required to improve the performance [29]. Meanwhile, the genetic algorithm development is in highly cost because having the mutation and crossover operation in it.

On the other hand, Support vector machine (SVM) has been found useful in handling classification tasks in case of the high dimensionality and sparsity of data points and has been among as a popular approach to efficiently treating the tourism data structure. As compared with neural network based method, L–J approach with combined kernel functions was observed to have a better performance. In addition, L–J method has the advantage on the basis of a single training run and is easier to compute for feature selection as compared with other SVM based methods. Although the approach of SVM with kernel function is useful for classification, however the computation speed is relatively slow when the kernel functions are complicated. Instead, its performance must be improved especially for complex data [35]. This is particularly important for people who want to obtain a high level of accuracy in advanced areas.

The discussions of the previous studies mentioned above are summarized in Table 1.1.

Table 1.1: Summary of Artificial Intelligent Technique in Tourism Demand Analysis.

No	Technique	Description
1	Artificial Neural Network	- Based on assumptions knowledge about dependencies, probability distributions and large number of experiments. - Cannot derive conclusions from incomplete knowledge or manage inconsistent information. - Not easy to explain as rules that are meaningful to the user. - More robust features are required to improve performance.
2	SVM	- Computation speed is relatively slow when the kernel functions are complicated.
3	Fuzzy Time Series	- Based on assumptions knowledge about dependencies, probability distributions and large number of experiments. - Cannot derive conclusions from incomplete knowledge or manage inconsistent information. - Need to convert numerical rules to the table rules form. - Produce long operation to have the result.
4	Genetic Algorithm	- Highly cost because having the mutation and crossover operation in it.

1.3 Problem Statement

Statistical assumption for analysis demand tourism cannot predict the outcome when various variable correlate with each others. Thus, using rough sets concept to analyse the tourism demand will be a better way compare to statistic assumption.

1.4 Project Aim

The aim of the study is to exploit rough sets mechanism into the travel demand analysis. It presents the results analysis of data sets, and demonstrates how rough set theory can be applied in tourism demand analysis.

1.5 Objective

The objective of this research is

i. To model the Malaysia travel demand analysis using Rough sets theory.
ii. To produce travel demand analysis information table.
iii. To evaluate the effectiveness of Rough sets theory in travel demand analysis.

1.6 Project Scope

The main focus of this project is to model the Malaysia travel demand analysis using Rough Sets theory. The scopes for this project are as follows:

i. Malaysia travel demand analysis data.
ii. A comparison will be done on multi regression in rough sets.

1.7 Conclusion

This report is organized as follows: Chapter 2 give the literature reviews about the methodology of the research and concepts of the rough sets theory. Chapter 3 discussed the Research methodology. Chapter 4 depicts the empirical results and model performance depicts the empirical results and model performance and Chap 5 is concluded with discussions on implications.

References

1. Carey Goh*, Rob Law. *Incorporating the rough sets theory into travel demand analysis.* School of Hotel and Tourism Management, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. (2002).
2. Frechtling, D.C.: *Forecasting Tourism Demand: Methods and Strategies.* Butterworth-Heinemann, Oxford (2001).
3. Law, R., Goh, C., Pine, R. *Modeling tourism demand: a decision rules based approach.* School of Hotel & Tourism Management, The Hong Kong. (2000).
4. Law, R.*Hospitality data mining myths*. FIU Hospitality Review 16(1), 59–66 (1998).
5. Pawlak, Z. *Rough Sets, Theoretical Aspects of Reasoning about Data*, *Kluwer Academic, Dordrecht*; (1991).
6. Holte R.C (Machine Learning., vol. 11, pp. 63–91, 1993) Very simple classification rules perform.
7. Alagar V S, Bergler *S,* Dong F Q eds. *Incompleteness and Uncertainty in Information Systems*. London: Springer-Verlag; 1994.
8. Wang Guoyong. *Rough Sets Theory And Knowledge Acquisition*. Xi'an: Xi'an Jiaotong University Press; 2001.
9. Nguyen, H.S. *Discretization Problem for Rough Sets Methods*. Proc. Of First Intern. Conf. on Rough Sets and Current Trend in Computing (RSCTC'98), Warsaw, Poland; 1998. 545-552.
10. Xiangyang Wang, A., Jie Yang, A., Xiaolong Teng, A., Weijun Xia, B., Richard Jensen, C. *Feature Selection Based On Rough Sets And Particle Swarm Optimization*. Pattern Recognition Lett. 28; 2007. 459-471
11. Janusz, A., Starzyk, J., Nelson, D.E., Sturtz, K. *A Mathematical Foundation For Improved Reduct Generation In Information Systems.* Knowledge Informat. Syst. 2; 2000. 131–146.

12. Skowron, A., Rauszer, C. *The Discernibility Matrices And Functions In Information Systems*. In: Slowinski, R. (Ed.), Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Dordrecht; 1992. 311–362.

13. Wroblewski, J. *Finding Minimal Reducts Using Genetic Algorithms*. In: Proc. Second Annual Join Conf. on Information Sciences, Wrightsville Beach, NC; 1995. 186–189.

14. Bazan, J., Nguyen, H.S., Nguyen, S.H., Synak, P., Wroblewski, J. *Rough Set Algorithms In Classification Problem*. In: Polkowski, L., Tsumoto, S., Lin, T.Y. (Eds.), Rough Set Methods and Applications. Physica-Verlag, Heidelberg, New York; 2000. 49–88

15. Indranil Bose. *Deciding The Financial Health Of Dot-Coms Using Rough Sets.* School of Business, University of Hong Kong; 2006.

16. http://en.wikipedia.org/wiki/Gross_Domestic_Product

17. http://en.wikipedia.org/wiki/Consumer_price_index

18. http://en.wikipedia.org/wiki/Population

19. Pawlak Zdislaw. *Rough Set Approach To Knowledge-Based Decision Support*, Eumpeon Journal of Operolional Reseerch. Na.99; 1997. 48-57.

20. Jerzy W. Grzymala- Busse*, Introduction to Rough Set Theory and Applications.*University of Kansas, Lawrence, Polish Academy of Sciences, 01- 237 Warsaw, Poland.

21. Witt, S.F., and Witt, C.A.: *Forecasting tourism demand: A review of empirical research. International Journal of Forecasting 11(3),* 1995, 447–475.

22. Law, R.: Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. Tourism Management 21, 331–340 (2000)

23. Delen, D., Sirakaya, E.: Determining the efficacy of Data-mining methods in predicting gaming ballot outcomes. Journal of Hospitality & Tourism Research 30(3), 313–332 (2006)

24. Chen, L.D.: A review of the Data Mining literature. In: Proceedings of the Hong Kong International Computer Conference ’97, pp. 23–31 (1997)

25. Clements, M. P., & Hendry, D. F. (1998). *Forecasting economic* time *series*. Cambridge: Cambridge University Press.

26 Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12, 231–254.

27 Engle, R. F., & Granger, C.W. J. (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica*, 55, 251–276.

28. C.Z. Ye, J. Yang, D.Y. Geng, Y. Zhou, N.Y. Chen. *Fuzzy Rules to Predict Degree Of Malignancy In Brain Glioma.* Med. Biol. Comput. Eng. 40 (2); 2002. 145–152.

29. A.E. Hassanien. *Fuzzy Rough Sets Hybrid Scheme For Breast Cancer Detection.* Quantitative Methods and Information Systems Department, Image and Vision Computing 25; 2007. 172 –183.

30. P.K. Simpson. *Fuzzy Min-Max Neural Networks. Part 1. Classification.* IEEE Trans. Neural Networks 3; 1992. 776–786.

31. P.K. Simpson. *Fuzzy Min-Max Neural Networks. Part 2. Clustering.* IEEE Trans. Fuzzy Syst. 1; 1993. 32–45.

32. J.R. Quinlan. *Induction Of Decision Trees.* Mach. Learn. 1; 1986. 81–106.

33. J.M. Zurada. *Introduction to Artificial Neural Systems.* West Publishing Co., New York; 1992.

34. W. Andrew. *Statistical Pattern Recognition.* Oxford University Press Inc., Oxford, 1999.

35 Su, C. -T., & Yang, C.-H. *Feature Selection For The Svm: An Application To Hypertension Diagnosis.* Expert Systems with Applications; 2006.

36. Haiyan Song, Gang Li. *Tourism demand modeling and forecasting—A review of recent research, Science Direct, 2007*

37. McIntosh, R.W., Goeldner, C.R., Ritchie, J.R.B.: *Tourism: Principles, Practices, Philosophies*. John Wiley & Sons, New York, 1995

38 Law, R.: *Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting.* Tourism Management 21, 331–340, 2000.

39. Delen, D., Sirakaya, E.: *Determining the efficacy of Data-mining methods in predicting gaming ballot outcomes*. Journal of Hospitality & Tourism Research 30(3), 313–332, 2006.

40. Uysal, M., El Roubi, S.E.: *Artificial neural networks versus multiple regression in tourism demand analysis*. Journal of Travel Research 38, 111–118, 1999.

41. Law, R., Au, N.: *A neural network model to forecast Japanese demand for travel to Hong Kong*. Tourism Management 20, 89–97, 1999

42. Cho, V.: *A comparison of three different approaches to tourist arrival forecasting.* Tourism Management 24, 323–330, 2003.

43 Wang, C.H.: *Predicting tourism demand using fuzzy time series and hybrid grey theory*. Tourism Management 25, 367–374, 2004.

44. Pai, P.F., Hong, W.C.: *An improved neural network model in forecasting arrivals.* Annals of Tourism Research 32(4), 1138–1141, 2005

45. Azuraliza Abu Bakar, Siti Mariyam Shamsuddin.: *Rough Set for Data Mining.* Soft Computing Research Group, University Technology Malaysia, page 2, 2007.

46. Aleksander Ohr, Jan Komorowski, Andrzej Skowron, Piotr Synak: *Rosetta Part 1 System Overview.* Knowledge System Group, Norwagian University of Science and Technology, Norway. page 15, 1998.

47. Haiyan Song, Kevin K.F Wong, Kaye K.S Chon.: *Modelling and Forecasting The Demand of Hong Kong Tourism.* Hospitality Management 22, page 438, 2003.

48. Kok, Y. P. *Rough Set for Predicting the Kuala Lumpur Stock Exchange Composite Index Returns*. Faculty of Science & Information System, Universiti Teknologi Malaysia; 2003.