Internet Research

An online niche-market tour identification system for the travel and tourism industry

C.H. Wu G.T.S. Ho C.H.Y. Lam W.H. Ip K.L. Choy Y.K. Tse

An online niche-market tour identification system for the travel and tourism industry

C.H. Wu
School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China and
Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hunghom, Hong Kong
G.T.S. Ho, C.H.Y. Lam, W.H. Ip and K.L. Choy
Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hunghom, Hong Kong, and
Y.K. Tse
The York Management School, University of York, Heslington, York, UK

Received 16 August 2014
Revised 16 September 2014
Accepted 4 October 2014

Abstract

Purpose – The purpose of this paper is to present a novel approach for niche-market tour identification, with the objective to obtain a better segmentation of target tourists and support the design of tourism products. A proposed system, namely the Niche Tourism Identification System (NTIS) was implemented based on the proposed scheme and its functionality was showcased in a case study undertaken with a local travel agency.

Design/methodology/approach – The proposed system implements automated customer market segmentation, based on similar characteristics that can be collected from potential customers. After that, special-interest tourism-based market strategies and products can be designed for the potential customers. The market segmentation is conducted using a GA-based *k*-means clustering engine (GACE), while the parameter setting is controlled by the travel agents.

Findings – The proposed NTIS was deployed in a real-world case study which helps a local travel agency to determine the various types of niche tourism found in the existing market in Hong Kong. Its output was reviewed by experience tour planners. It was found that with the niche characteristics can be successfully revealed by summarizing the possible factors within the potential clusters in the existing database. The system performed consistently compared to human planners.

Originality/value – To the best of the authors' knowledge, although some alternative methods for segmenting travel markets have been proposed, few have provided any effective approaches for identifying existing niche markets to support online inquiry. Also, GACE has been proposed to compensate for the limitations that challenge *k*-means clustering in binding to a local optimum and for its weakness in dealing with multi-dimensional space.

Keywords Market segmentation, Information processing, Database marketing, Customer requirements, Customer characteristics, Niche tourism

Paper type Research paper

1. Introduction

The rapid advancement of the internet and computer technologies has not only increased the amount of online information available but also turned people to the web to acquire knowledge and find answers (Ho *et al.*, 2010). People's daily activities have greatly changed from "hard" information to the internet and the internet of things (IoT). There are increasing numbers of e-shopping and e-payment environments that have evolved from physical stores and banks (Hsieh *et al.*, 2013). The same trend can also be found in the global tourism industry. There is a great deal of tourism-related online information

Internet Research
Vol. 26 No. 1, 2016
pp. 167-185

1066-2243
DOI 10.1108/IntR-08-2014-0204

provided by local travel agencies, hotels and home-stay providers but the supply is less commensurate with the demand. Tourists usually have difficulty in choosing, according to their real needs, and the online information and services provided are often of no avail.

The global tourism industry has become an extremely dynamic system and it operates in a volatile environment, in which both growth and development fluctuate (Farrell and Twining-Ward, 2004). Nowadays, the boundaries between travel agencies and tour operators are fairly indistinct. They not only have to compete against local competitors but also against global companies due to emergence of integrated online travel service providers. Due to the general broadening of the public's travel experience and the huge amount of internet information available, tourists are becoming more sophisticated in their needs and preferences. They require customized products and services which suit their personalities at a reasonable price (Macleod, 2003). Therefore, the rise of niche tourism has been recognized as a response to the specific needs of customers (Novelli, 2005). Niche tourism can be characterized by multi-dimensional space which describes the characteristics of a group of tourists who share similar desires and wants (Hassan, 2000). The key to success in niche tourism significantly depends on how well the niches are defined and on the effectiveness and efficiency with which travel agencies satisfy their customers relative to their competitors. A consequence is that the use of information technologies and systems to manage information in an efficient and effective way has become vital in finding the key to success. With strategic ties between information systems and niche tourism, a competitive advantage in fulfilling customer needs can be gained.

Generally, market segmentation serves as the first step for the deployment and implementation of marketing strategies, followed by market targeting for both mass and niche marketing (Smith, 1956; Dalgic and Leeuw, 1994). Chen and Fang (2013) have stated that the use of a clustering algorithm would be an efficient method in this regard. According to data-driven approaches, the quality and types of the segments obtained, as well as their usefulness to a company, greatly rely on the selection of appropriate segmentation bases and algorithms (Wedel and Kamakura, 1998). The tourism market has been traditionally classified into four major categories, namely geographic, socio-economic and demographic, psychographic and behavioral (Mykletun *et al.*, 2001). On top of that, the benefit segmentation base, introduced by Haley (1968), has also been emphasized in some tourism research studies (Tsiotsou and Goldsmith, 2012). Although some alternative methods for segmenting travel markets have been proposed, few have provided any effective approaches for identifying existing niche markets. The industry has been keen to have computerized systems to process data, manage data, analyze data and transfer data to support decision making in travel market segmentation. To address this issue, this study proposes the use of a heuristic clustering algorithm for niche tourism market identification. In a previous study (Ho *et al.*, 2012), a robust GA-based *k*-means clustering engine (GACE) was introduced to obtain customer groupings with consideration of the data distribution and dimension quality in order to attain better resource allocation. To extend its application, an information system embedded with GACE is developed to illustrate the feasibility and capability in supporting effective marketing approaches.

This paper is divided into six main sections. Section 2 provides a quick review of the background and techniques applied in this paper. Section 3 presents the framework and algorithm of the advanced market segmentation system. In Section 4, a case study in a Hong Kong travel agency is described that illustrates the proposed methodology and implementation in an actual business environment. The results and discussion are presented in Section 5, with conclusions given in the last section.

2. Literature review

The internet has made a profound impact on the way ideas are generated and knowledge is created. The number of daily and enterprise devices that will soon be connected to the internet will be huge. Cloud, social networks, mobile devices and information are driving early opportunities in the IoT, and one has seen evidence of this already through examples such as warehouse management with IoT technologies (Lim *et al.*, 2013; Reaidy *et al.*, 2015), cloud-based manufacturing technology (Huang *et al.*, 2014), RFID-enabled healthcare applications (Ting *et al.*, 2011; Wamba *et al.*, 2013), e-learning systems (Tan *et al.*, 2014) and smart sensor networks (Chen *et al.*, 2012; Portilla *et al.*, 2014). IoT is an adoption of advanced technologies for different kinds of user-oriented applications supported by the internet backbone. The internet backbone also supports a widespread use of a variety of platforms including e-mail, blogs, forums, online communities, social networks and review sites that collect and disseminate information, hence electronic word-of-mouth communication through these platforms significantly influences consumer purchase decisions, especially in purchasing tourism products (Rong *et al.*, 2012).

"Niche tourism" is derived from "niche marketing" which originates from the niche concept in ecology (Lambkin and Day, 1989). The rise of the niche tourism marketing approach aims to respond and satisfy the needs of travelers, reflecting the growth of this specific market and its importance to the industry (Ibrahim and Gill, 2005). However, valuable market niche identification is never easy. Previous research studies concerning market segmentation in the tourism industry have been examined. Kim *et al.* (2007) suggested that provision of useful and relevant information based on customer preference was one of the important factors when selecting a travel agency. Thus, when searching for criteria, tourist preference attributes should be focussed on. There are a number of alternative segmentation factors, which include geographic characteristics (Reid, 2003), demographics (Hsu and Sung, 1997), psychographics (Hsu *et al.*, 2002) and benefits sought (Yannopoulos and Rotenberg, 1999). However, the choice of a market segmentation base varies directly with the purpose of the market study under consideration (Koc and Altinay, 2007). The choice of different bases may lead to different segments being revealed. Among all the factors of concern, it is suggested that benefit segmentation is one of the prominent segmentation bases as the benefits sought by tourists can truly reflect socio-economic buying behavior and personality (Frochot and Morrison, 2001).

In their benefit segmentation review (Frochot and Morrison, 2001), it was summarized that most of the related research had two basic components, factor combination and clustering analysis. To reduce dimensionality, principal component analysis is widely used to identify some significant dimensions by redundancy removal. In addition, Dolnicar *et al.* (2012) have proposed a biclustering method to segment vast amounts of data on the tourism industry, with their eyes on variable and grouping selection. Clustering analysis, one of the data mining techniques, is a unsupervised learning method which perform classification that entails the division of data sets or samples into subsets, so that data in the same cluster shares maximized homogeneity (Xu and Wunsch, 2009; Duan and Xu, 2012). Clustering analysis has been the dominant means of segmenting the targeted uniformed market using diverse combinations of attributes (Wedel and Kamakura, 1998), and has also been successfully adopted in outlier detection (Duan *et al.*, 2009) to generate accurate alerts. One of the most commonly used non-hierarchical methods is k-means clustering with Euclidean distance (Dolnicar, 2002; Kerschbaum, 2008), because of its ability and efficiency in handling huge amounts of data. A k-means algorithm is initiated by a user-defined parameter k and center points in an attempt to partition a given data set into k clusters. The approach minimizes the

Euclidian distance for the purpose of obtaining the least variation within each cluster, and maximizes the variation between the resulting k clusters (Dattorro, 2005). The algorithm is iterated until no records are exchanged between clusters.

However, one of the disadvantages of k-means clustering is that a local optimum usually results due to the high sensitivity to the user-defined initial cluster numbers and centers. Additionally, k-means clustering is limited to spherical shaped clusters only, as it relies on Euclidean distances from the cluster centers as a criterion function (Roberts, 1997). The existence of noise, outliers and missing values also hinder the ability of k-means clustering owing to its reliance on mean values. In an attempt to overcome this limitation, cluster validation techniques are needed for the purpose of evaluating and selecting the best performing values of the initial centers. Genetic algorithms (GA), introduced by Holland (1975), are simple optimization algorithms which search for a global optimum based on the process of repeating reproduction and evaluation in order to obtain the optimal partition of a data set. The existence of GA operators and different structures prevents the result from reaching a local optimum and improves the GA's capability and scalability (Kim and Ahn, 2008; Li *et al.*, 2011a, b; Fritzsche *et al.*, 2012; Jin *et al.*, 2014). Recently, some researchers also studied the essential characteristics of adaptation complexity in evolutionary algorithms to improve the overall performance (Wang, *et al.*, 2011).

Researchers have tried to integrate k-means clustering by assigning different roles to the algorithms in which GA is incorporated, for finding the optimal numbers of the cluster number-k. Krishna and Murty (1999) proposed a GA-based method, GKA, which uses k-means to replace the cross-over operator in an attempt to obtain the locally optimal clusters. This is accompanied by a biased mutation operator which is used to widen the search for the global optimum. A new clustering algorithm based on a genetic algorithm with gene rearrangement was proposed by Chang *et al.* (2009). Frequently, the hybrid approach of combining GA and k-means clustering algorithms is used for optimizing the initial cluster centers. Online shopping market segmentation introduced by Kim and Ahn (2008) proved that GA-based k-means clustering undisputedly improved segmentation performance when compared to other conventional clustering algorithms. Regarding multi-dimensional space, the GA is a domain-independent technique used for solving optimization problems concerning multi-attributes. Embedding an implicit degree of parallelism, GAs can therefore be applied to huge data sets and are especially suitable for large dimensional space exploration (Cordon *et al.*, 2003). More recently, integration enabled the k-means algorithm to select the optimal solution with consideration of particular dimensions for each cluster as a advanced grouping decision (Li *et al.*, 2008; Ho *et al.*, 2012).

In this paper, GACE is proposed for niche-market identification in an attempt to search for profitable and worth generating niche tourism. Its ability to select significant attributes from numerous attributes also forms the basis for its integration with the k-means clustering technique in niche tourism identification. The modified engine can enhance the performance of the k-means algorithm by making it less dependent on the pre-defined parameters, especially the randomly chosen initial cluster center, thus leading to a more reliable result in extracting shared characteristics of different niches.

3. Framework of Niche Tourism Identification System (NTIS) with GACE

In this section, the framework of the proposed system with GACE is presented in an attempt to spot niche markets in the contemporary tourism industry. As shown in Figure 1, the proposed NTIS is divided into four main parts: data collection and storage, data pre-processing, the engine – GACE and customer inquiry.

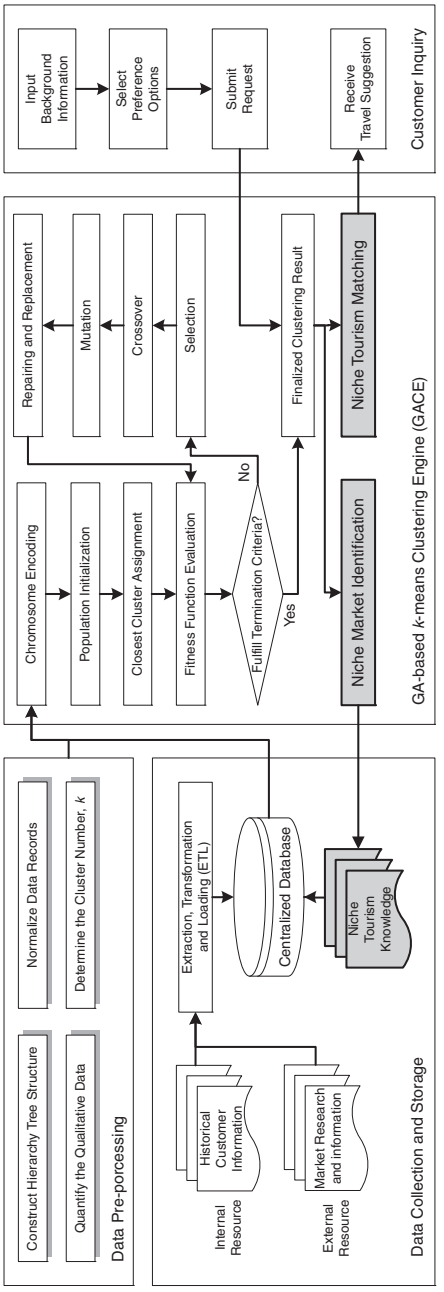

Figure 1.
Framework of the proposed system – NTIS

3.1 Data collection, storage and pre-processing

Periodically reviewing market trends and customer needs is important for success in the tourism market. Before deriving marketing strategies, it is necessary for marketers to gather information regarding the market needs. These consist of individual needs, as well as the existing offerings in the market. This can be done through market research conducted by external parties who can provide decision makers with both primary and secondary market information, hence guiding them to make critical business decisions. Extraction, transformation and loading are carried out to clean and integrate the distributed data collected from various sources. The data are then transformed to fit operational needs. This helps bring all the data together in a standard and homogenous environment before being stored in the centralized database.

Apart from the data collection and centralization, some preparatory tasks need to be done before implementing the GACE for niche-market identification. According to Pyle (1999), data pre-processing is a critical, but time-consuming and challenging task. However, it helps transform the qualitative data collected from surveys into a form suitable for valid operation of the proposed system. It consists of four tasks, which are: constructing a hierarchical tree, quantifying the data, normalizing the attribute values and proposing a pre-defined number of clusters, k.

Due to the complexity of the tourism industry, which consists of numerous sub-categories, the seven matrices affecting the purchasing decisions of travelers have been summarized. By using a tree diagram, the broad tourism industry is broken down into various details, as shown in Figure 2. The sub-categories identified in the tree diagram vary according to the major concerns of decision makers as regards market segmentation within the industry.

Since the GACE is designed for handling numerical data, it is therefore necessary to quantify qualitative attributes such as reasons for traveling, benefits sought and traveling destination. Meanwhile, data normalization is needed in order to eliminate the dominating effect of any single value in the selected attributes when calculating the Euclidean distance in clustering. Subsequently, it is necessary for decision makers to determine the number of clusters to be obtained, i.e. the value of k. The higher the value of k, the higher the computation cost, however, the benefit is that the clusters are then more specifically identified. The value of k therefore depends on the choices of the decision makers and on the resources available.

In order to apply the GACE in the next stage, chromosome encoding must be done in advance. The encoding of the chromosomes is based on the hierarchical tree diagram defined in the industry. The chromosomes encoded in this study are divided into the determining factor region and the parameter region. For categorical data, binary encoding is used, while continuous encoding is used for numerical data. The notation table for the study is shown in Table I:

Definition 1. $F_{[}x_1, x_2, \ldots, x_n]$ is a representation of all the determining factors hanging on the tree diagrams, and $P_{[}x_1, x_2, \ldots, x_n]$ is for corresponding parameters.

Definition 2. $C = \{1, 2, \ldots, N\}$ is the index set of chromosomes where N is the total number of random chromosomes generated. An individual chromosome consists of $F_{ij[}x_1, x_2, \ldots, x_n]$ and $P_{ij[}x_1, x_2, \ldots, x_n]$ in sequence, for each cluster. For those cases of $P_{ij[}x_1, x_2, \ldots, x_n]$, where $F_{ij[}x_1, x_2, \ldots, x_n]$ is equal to 0, a dummy value will still be generated. Therefore, a chromosome matrix [*RandChromo*] can be created.

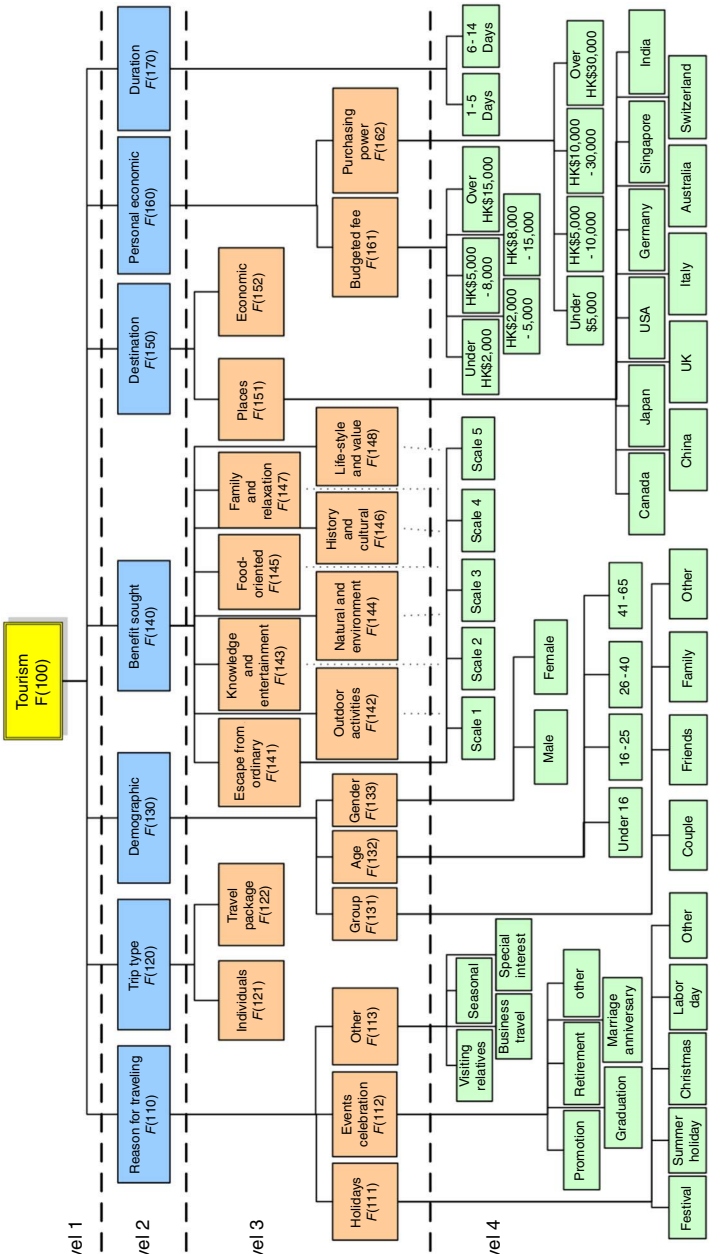

Figure 2.
Hierarchical tree diagram for the tourism industry

Symbol	Explanation
[*RandChromo*]	A $N \times 40$ matrix containing N random chromosomes, means there are N rows in the matrix
x_n	The nth attribute identified in the hierarchical tree concerning tourism industry
f	Total number of genes in the determining factor region within the tourism industry
i	The ith chromosome in the [*RandChromo*], $i \in N$
j	The jth cluster of chromosome in the [*RandChromo*], $j \in k$
k	Total number of pre-defined clusters
l	The lth raw data set collected, $l \in m$
m	Total number of raw data sets collected
p	Total number of genes in the parameter region within the tourism industry
s	Running index
C	Random chromosome in the [*RandChromo*]
F	Determinant factor region within the tourism industry
N	Population size of random chromosome
P	Parameter of corresponding determining factor within the tourism industry
$Adjust_{i,j}$	Adjustment index represents the adjustment index for cluster j of the ith random generated chromosomes

Table I.
Notation table

Unlike the random chromosome presented above, a raw data chromosome consists of parameter regions only. The data matrix is constructed corresponding to the raw data gathered from a questionnaire. There are m data records and each consists of n parameter values for the three attributes. The row vector of the raw data matrix denotes the data record entry, while the column vector denotes the parameter value, and the resulting raw data matrix is a $m \times n$ matrix.

3.2 GACE

As GACE inherits ability for optimization from the GA, it also contains the steps to mate, mutate and propagate problem solving genes to the next generation in order to generate a possible solution to a target problem. The procedure details are given in previous literature (Ho *et al.*, 2012).

3.2.1 Population initialization and fitness function computations. Initially, a random population containing a specified number of chromosomes, which represents the k cluster centers, is created. The population size, N, must be an even number for the sake of valid pairing up in the cross-over stage. The population is initialized as starting candidates, making up the parent pool.

In this study, the Euclidean distance in k-means clustering is considered as the criterion function for evaluating the fitness value of each candidate chromosome. Adjustment indices are incorporated in order to compensate for the unfairness caused by the high-dimensional space distance. The following two remarks give some ideas for defining the adjustment indices: the distance calculated using the fewest dimensions should have the largest adjustment index; since "time" is the main concern in this project, when a "time horizon" is not chosen, the optimal value of the chromosome will be large. This will act as a penalty, minimizing the corresponding selection probability. Therefore, a definition of the adjustment indices is suggested in the distance calculation:

$$Adjust_{ij} = 10\left(1 - RandChromo_{i,10(j-1)+1}\right) + \left(\sum_{s=2}^{10} RandChromo_{i,10(j-1)+s}\right)^{-1} \quad (1)$$

The data points should be assigned to the cluster with the shortest distance between all clusters. That is, the total sum of the Euclidean distance between each data point and its corresponding center of cluster should be minimized. Thus, the linear programming (LP) model is formulated to minimize the total distance:

$$\min_{x_{lj}} \sum_{j=1}^{k} \sum_{l=1}^{m} Dist_{lj} x_{lj} \tag{2}$$

Subject to:

$$\sum_{j=1}^{k} x_{lj} = 1,\ l = 1,\ \ldots,\ m, \qquad \sum_{l=1}^{m} x_{lj} = m_j,\ j = 1,\ \ldots,\ k, \qquad \sum_{j=1}^{k} m_j = m,$$

$$x_{lj} \geqslant 0,\ l = 1,\ \ldots, m;\ j = 1,\ \ldots,\ k \quad \text{and} \quad m_j \geqslant 0,\ j = 1,\ \ldots,\ k$$

Since one raw data point can only be assigned to one group, therefore the summation of x_{lj} equals 1. x_{lj} can either be 1 or 0, whereas 1 indicates that the lth raw data will be assigned to the jth cluster and vice versa. The LP model is then presented in a matrix form:

$$\min_{\vec{x}} \vec{c}^{\,T} \vec{x} \tag{3}$$

Subject to:

$A\vec{x} = \vec{b}$ and $\vec{x} \geqslant 0$ where

$$x = \begin{bmatrix} x_{1,1} \\ \vdots \\ \overline{x_{m,1}} \\ x_{1,2} \\ \vdots \\ \overline{x_{m,2}} \\ \vdots \\ x_{1,k} \\ \vdots \\ \overline{x_{m,k}} \\ m_1 \\ \vdots \\ m_k \end{bmatrix}_{(m+1)k \times 1}, \qquad c = \begin{bmatrix} Dist_{1,1} \\ \vdots \\ \overline{Dist_{m,1}} \\ Dist_{1,2} \\ \vdots \\ \overline{Dist_{m,2}} \\ \vdots \\ Dist_{1,k} \\ \vdots \\ \overline{Dist_{m,k}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(m+1)k \times 1}, \qquad b = \begin{bmatrix} 1 \\ \vdots \\ \overline{1} \\ 0 \\ \vdots \\ \overline{0} \\ m \end{bmatrix}_{(m+k+1) \times 1} \quad \text{and}$$

$$A = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{(m+k+1) \times (m+1)k}$$

3.2.2 Selection. Selection probability, $Prob_{i,1}$, is calculated for determining whether or not to select the chromomsome from the current generation using Roulette wheel selection. The culmulative probability, $Prob_{i,2}$, which is the summation of all the normalized selection probabilities, is then computed to outline the range in the Roulette wheel.

Therefore, we suggest the following way of assigning probability: let [*Prob*] be a $N \times 2$ matrix such that $Prob_{i,1}$ and $Prob_{i,2}$ represent the probability and cumulative probability assigned to the ith random chromosome respectively. A constant of value 1.01 is added to the formula to prevent the chance of getting a value of 0 in any probability:

$$Prob_{i,1} = \frac{1.01 \cdot max\left(Optimal_{1,1}, Optimal_{2,1}, \ldots, Optimal_{N,1}\right) - Optimal_{i,1}}{\sum_{s=1}^{N}\left[1.01 \cdot max\left(Optimal_{1,1}, Optimal_{2,1}, \ldots, Optimal_{N,1}\right) - Optimal_{s,1}\right]} \quad (4)$$

$$Prob_{i,2} = \sum_{s=1}^{i} Prob_{s,1} \quad (5)$$

3.2.3 Cross-over, mutation, fitness function evaluation and stopping criterion. Cross-over is a choatic process for exchanging information between two parent chromosomes for producing the next generation. Another operator, mutation, is also used for the same function as cross-over to enhance the genetic diversity, avoiding convergence of the local minimum. For the determining factor region, a gene is mutated simply by flipping its values (binary). However, for the parameter region using a continuous encoding scheme, a number within the range of the value for the parameter region will occur with equal probability. After cross-over and mutation, the fitness values of the new chromosomes are evaluated with those in the parent pool. The iterative procedure of the GA stops if the number of generations reaches the maximum generation specified. The best chromosome among the parent and mating pool provides the solution to the clustering problem. The obtained clustering result, as the first output of the GACE, is the niche-market identification which is the fundamental goal of this work.

3.3 Customer inquiry

In the proposed niche tourism system, on top of the data collection, pre-processing and computation, it is suggested that a customer interface should be built so that the niche tourism marketing can reach the target customer in a more direct and timely manner. As in the processes in the previous section, the historical customer and marketing research information has been converted into valuable information that is available, yet has never been extracted before. The niche tourism and market knowledge is then stored in a centralized database for both internal marketing purposes and external use. The centralized database is updated using appropriate marketing research approaches and also the newly mined niche tourism knowledge. Therefore, this system provides a customer inquiry function so that the niche packages derived from the knowledge can be offered for those potential travelers who fall into our resultant cluster(s). To enhance the user experience, the customer needs only to input some background information and traveling preference for the engine and database to match them up with the extracted clusters, so as to provide more suitable suggestions in terms of particular niche packages.

4. Case study

In the case study, a prototype of the NTIS with GACE is used to support the design in different niche travel packages for a large local travel agency in Hong Kong. The development environment of the prototype system is in Visual Studio 10.0 with a MATLAB engine application of the GACE. The company offers a wide variety of travel packages featuring various geographical locations. However, with increasingly aggressive competition, the company has barely maintained its market share and is planning to develop niche markets to further expand its business. In the expansion plan, their target customers are those aged from 18 to 26 who have a distinct purpose for traveling. As tourists now require customized and unique products and services, the benefits sought have becomes an increasingly determining characteristic in distinguishing travelers, according to recent tourism literature. Therefore, in this case study, the heuristic clustering algorithm is employed to reveal the niches of local young people. The goal of the study is to design one to two niche tourism packages featuring specific benefits sought by the customers, rather than on exotic geographical locations or destinations.

4.1 Preparation and setup of NTIS

Data were collected before applying the hybrid algorithms. A Web-based distribution approach was used to collect information regarding customer needs in niche tourism, and 200 valid records were obtained ($m = 200$). The collected and preprocessed customer needs information became the input data for GACE. Since the company would like to know which month(s) the niche packages should be launched, for better resource allocation, the "time horizon" attribute is the focus in this case. Therefore, a tree diagram was derived, and is shown in Figure 3. In total, ten determining factors, which include "time horizon," "budgeted fee" and eight sub-factors categorized in "benefit sought," were identified. As GA algorithms are designed to handle exact

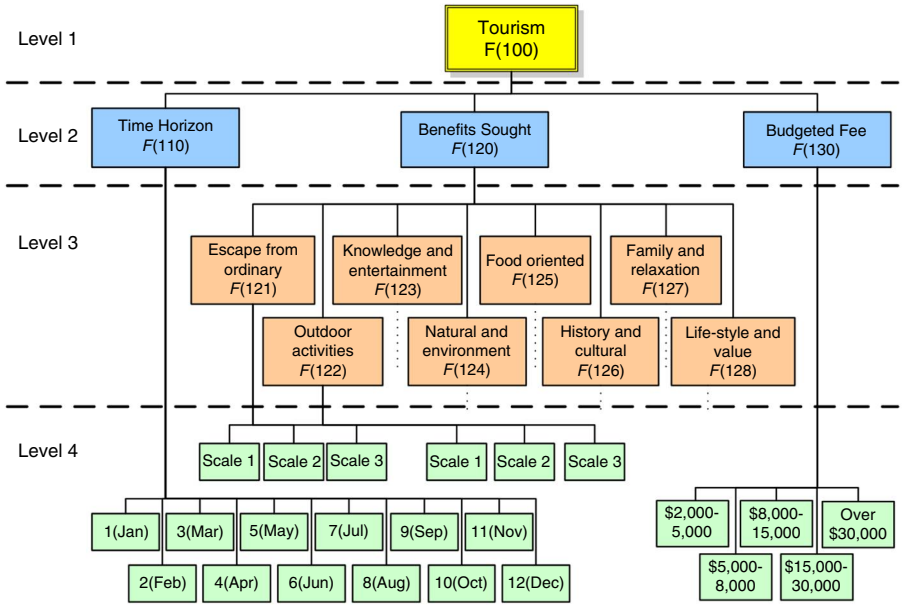

Figure 3.
Tree diagram for the case study

values rather than a value range for the attribute named "budgeted fee," a scale system was employed to represent the scale from 1 to 5 as shown in Table II.

In GACE, the GA function is undertaken together with *k*-means calculation. In each iteration of the GAs, the best *N* chromosomes with the lowest fitness values are preserved for the next iteration. The iterative process stops when the number of generations reaches the pre-defined maximum number of generations. The fitness values of the last iteration are recorded for the evaluation. Under the "benefits sought" attribute, there are eight options in total. In order to obtain a meaningful result, at least three options must be selected with their determining factor regions having a value of 1, denoting the successful selection of the options. In reality, decisions can be made by reference to the resources on hand and the feasibility of embracing several benefits in one trip. There are five input parameters in the engine: the number of clusters to obtain *k*; the population size *N*; the cross-over rate; the mutation rate; and the maximum number of generations "maxo." The parameter setting of the GACE is shown in Table III. Figure 4 shows the initial setup process of the NTIS.

Table II.
Conversion table for budgeted fee

Range of budgeted fee (HK$)	Scale
2,000-5,000	1
5,000-8,000	2
8,000-15,000	3
15,000-30,000	4
Over 30,000	5

Table III.
Parameter settings of the GACE

Settings	Parameters being examined
Number of clusters	$k=2$ or $k=3$
Population size (*N*)	16
Cross-over rate	0.8
Mutation rate	0.99
Generation (termination criterion)	250 or 500

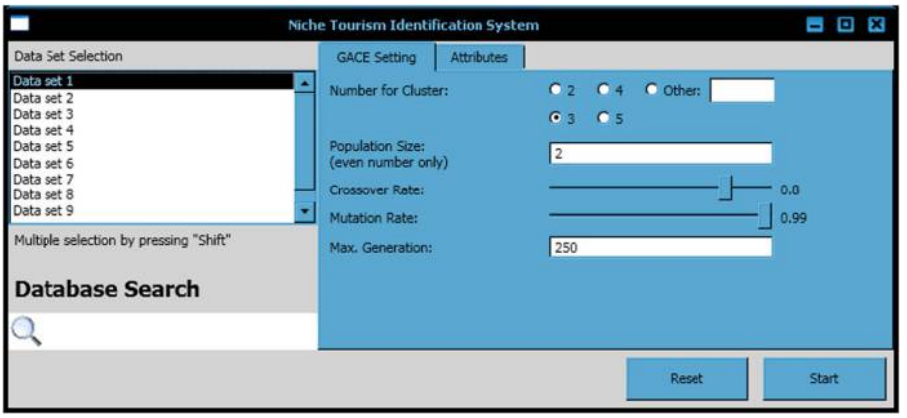

Figure 4.
User interface of NTIS for revealing the niches among customers

4.2 Results, evaluations and implications

Through the case study, the application of the NTIS is demonstrated while the performance of its GACE is also evaluated using different sets of parameters. One of the clustering results after running the GACE in the NTIS is shown in Figure 5. Summarized results of the clustering are shown in Tables IV-VII. The time taken for generating an iteration is about 1.20 seconds on average, with an initial population size (N) of 16. The variation of the fitness value of the chromosomes in the last iteration is recorded and is shown in Figure 6, for evaluation of the performance of the GACE.

From the obtained results, given above, the findings regarding the population size, number of generations and number of clusters defined can be summarized:

(1) The larger the size of the initial population in the parent pool, the more favorable the optimal values and stable outputs.

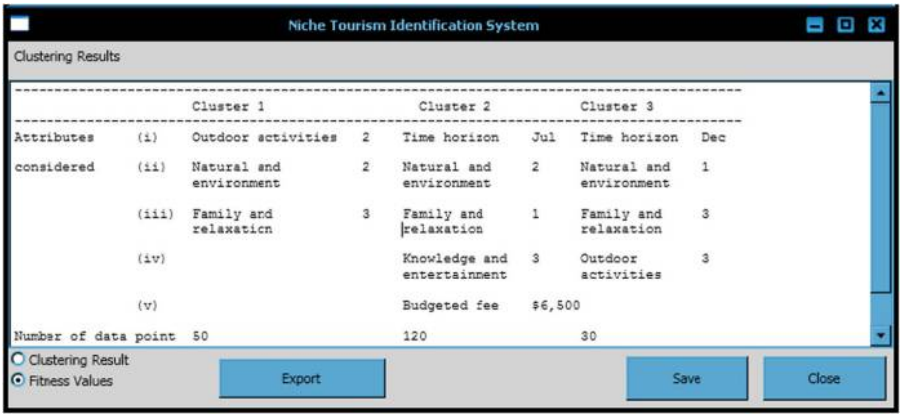

Figure 5.
A clustering result shown in the NTIS

		Cluster 1		Cluster 2	
Attributes considered	(i)	Time horizon	February	Time horizon	July
	(ii)	Food oriented	3	Natural and environment	2
	(iii)	Knowledge and entertainment	2	Knowledge and entertainment	2
	(iv)	Family and relaxation	2	History and cultural	2
Number of data points		40		160	
Total number of data		200			

Table IV.
Clustering result with $k = 2$, $maxo = 500$

		Cluster 1		Cluster 2	
Attributes considered	(i)	Time horizon	July	Food oriented	2
	(ii)	Outdoor activities	2	History and cultural	2
	(iii)	Family and relaxation	2	Knowledge and entertainment	2
	(iv)	History and cultural	2		
	(v)	Budgeted fee	$22,500		
Number of data points		170		30	
Total number of data		200			

Table V.
Clustering result with $k = 2$, $maxo = 250$

		Cluster 1		Cluster 2		Cluster 3	
Attributes considered	(i)	Time horizon	June	Natural and environment	2	Time horizon	July
	(ii)	Outdoor activities	2	Escape from ordinary	2	Knowledge and entertainment	2
	(iii)	Family and relaxation	1	History and cultural	2	Lifestyle and value	3
	(iv)	Food oriented	3			Food oriented	3
	(v)	Budgeted fee	$22,500			Budgeted fee	$11,500
Number of data points		30		55		1,153	
Total number of data		200					

Table VI.
Clustering result with $k = 3$, $maxo = 500$

		Cluster 1		Cluster 2		Cluster 3	
Attributes considered	(i)	Outdoor activities	2	Time horizon	July	Time horizon	December
	(ii)	Natural and environment	2	Natural and environment	2	Natural and environment	1
	(iii)	Family and relaxation	3	Family and relaxation	1	Family and relaxation	3
	(iv)			Knowledge and entertainment	3	Outdoor activities	3
	(v)			Budgeted fee	$6,500		
Number of data points		50		120		30	
Total number of data		200					

Table VII.
Clustering result with $k = 3$, $maxo = 250$

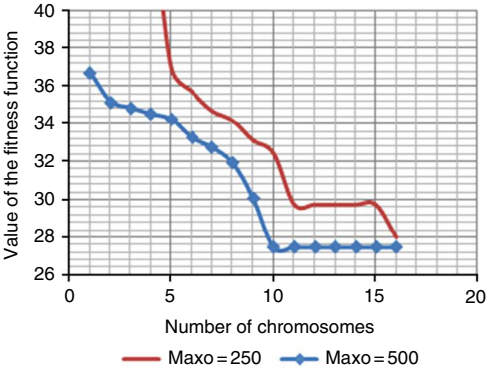

Figure 6.
Variation of fitness values of chromosomes

Note: $k = 2$

(2) The larger the value of the maximum number of generations, the smaller the fitness value, and the smaller the variation of the fitness values recorded. This implies a better result is obtained.

(3) When $k = 3$, the resulting optimal fitness value is lower than that when $k = 2$, that means a better cluster result is obtained when $k = 3$.

The clustering results act as a support which can assist the company staff of the marketing and promotion section in determining niches and subsequently deriving niche marketing strategies. They can summarize the characteristics concerning the clusters identified and hence design-specific packages for the target customers. Afterwards, the staff can compare the strategies obtained with the existing market offerings to see if any new opportunity emerges that is worth pursuing. It should be noted that the clustering results offer the staff a data-driven approach in niche tourism market identification and marketing. The clustering results should be verified to see if the results obtained are meaningful and valid for further marketing effort.

In the case study, the summarized niche, using the results obtained with parameter settings of $k = 3$ and $maxo = 500$, is about "food oriented, lifestyle and value" tours. The launching period of the tours should be in July to September with a budget approximately from eight to fifteen thousand Hong Kong dollars. With this niche identified, some tailor-made tour packages can be designed and thus such packages can be readily suggested to any interested customers in response to an online inquiry (Figure 7). Potential customers input essential background information and select their preferences, and the system provides all matched and tailor-made niche tour packages.

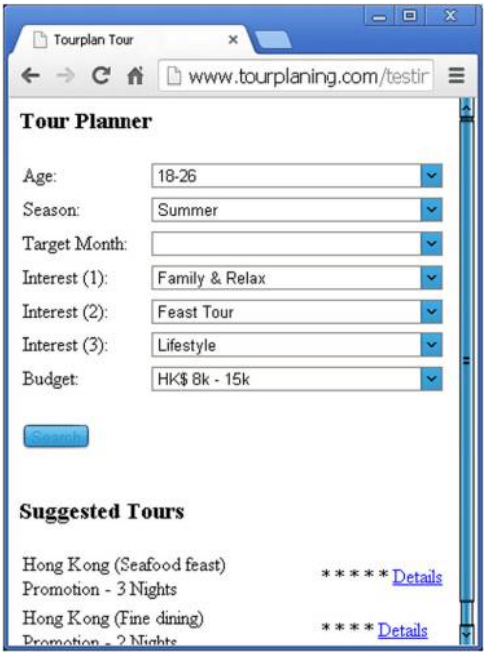

Figure 7.
Responses made to an online inquiry by the NTIS

5. Discussion

A heuristic clustering algorithm is useful in tackling problems involving optimization of computational requirements for achieving robust and fast solutions. In attempting to have a better understanding of the specific needs of the tourist, high-dimensional space segmentation is involved. This is due to the complex structure of the tourism industry in terms of the factors affecting the choice of destinations and the purpose of traveling. Therefore, the proposed GACE is suitable for global searching in such a complex situation, as it provides marketers with more insight in approaching niche opportunities in the tourism industry.

GAs operate on a heuristic basis, hence there is no guarantee that the resulting optimal solution is the "best" solution. However, heuristic methods, i.e. GA, are capable of undertaking a global search of the optimal factors within a shorter time and are more cost-effective than those algorithms which can output an exact "best" solution, such as LP. Therefore, a compromise is made between timely decision making support and a guarantee of the optimal solution. It is vital for the travel and tourism industry, because data generation, gathering, processing and application are so important for day-to-day operations (Poon, 1993).

6. Conclusions and future work

Prior to formulating market strategies and designing tourism products, customer market segmentation, which divides the whole market into groups of people with similar characteristics, should be the focus when determining potential customers for niche package tours. Therefore, the clustering approach is developed to segment and to search for niches in the market. In this paper, an online niche-market tour identification system with GACE is proposed to investigate the potential operation in niche-market identification in the tourism industry. Equipped with a global-search-capable GA, the proposed method is able to solve clustering problems involving high-dimensional space. GA-based k-means clustering is no longer confined to the limitation faced by conventional k-means clustering, rather it provides a robust solution in customer pattern segmentation. The results from the data set collected show that the larger the maximum number of generations, the better is the result obtained. In addition, the number of clusters defined can be validated by looking at the variation of the fitness values. With the proposed NTIS, travel agents and tour operators can easily create online niche travel communities or fan pages, with special interests and characteristics. The communities and fan pages can facilitate the spread of electronic word of mouth and travel experience which have influence on tourism decision making (Jalilvand and Samiei, 2012). Furthermore, the GACE can be applied to other industries and applications. For example, it can be used to explore niches in highly competitive and saturated markets, such as Hi-tech electronic products and financial investment portfolios. It is believed that the proposed GACE can lead marketers along the path to success in niche marketing.

Acknowledgments

The authors thank the Editor and Reviewers for their valuable comments and suggestions that have improved the paper's quality. The authors would also like to thank the School of Information Science and Technology of Sun Yat-sen University, the Department of Industrial and Systems Engineering of The Hong Kong Polytechnic University and The York Management School of University of York, for support in this research. This work is also partially supported by the project (G-UB97).

References

Chang, D.X., Zhang, X.D. and Zheng, C.W. (2009), "A genetic algorithm with gene rearrangement for *k*-means clustering", *Pattern Recognition*, Vol. 42 No. 7, pp. 1210-1222.

Chen, X. and Fang, Y. (2013), "Enterprise systems in financial sector – an application in precious metal trading forecasting", *Enterprise Information Systems*, Vol. 7 No. 4, pp. 558-568.

Chen, Z.Q., Ip, W.H., Wei, Y.F. and Wu, C.H. (2012), "Chaos particle swarm algorithm for energy consumption optimization in wireless sensor networks", *Sensor Letters*, Vol. 10 No. 8, pp. 1830-1835.

Cordon, O., Herrera-Viedma, E. and Lopez-Pujalte, C. (2003), "A review on the application of evolutionary computation to information retrieval", *International Journal of Approximate Reasoning*, Vol. 34 Nos 2-3, pp. 241-264.

Dalgic, T. and Leeuw, M. (1994), "Niche marketing revisited: concept, applications and some European cases", *European Journal of Marketing*, Vol. 28 No. 4, pp. 39-55.

Dattorro, J. (2005), *Convex Optimization & Euclidean Distance Geometry*, Meboo Publishing, Palo Alto, CA.

Dolnicar, S. (2002), "A review of data-driven market segmentation in tourism", *Journal of Travel & Tourism Marketing*, Vol. 12 No. 1, pp. 1-22.

Dolnicar, S., Kaiser, S., Lazarevski, K. and Leisch, F. (2012), "Biclustering: overcome data dimensionality problems in market segmentation", *Journal of Travel Research*, Vol. 51 No. 1, pp. 41-49.

Duan, L. and Xu, L.D. (2012), "Business intelligence for enterprise systems: a survey", *IEEE Transactions on Industrial Informatics*, Vol. 8 No. 3, pp. 679-687.

Duan, L., Xu, L., Liu, Y. and Lee, J. (2009), "Cluster-based outlier detection", *Annals of Operations Research*, Vol. 168 No. 1, pp. 151-168.

Farrell, B.H. and Twining-Ward, L. (2004), "Reconceptualizing tourism", *Annals of Tourism Research*, Vol. 31 No. 2, pp. 274-295.

Fritzsche, M., Kittel, K. and Blankenburg, A. (2012), "Multidisciplinary design optimization of a recurve bow based on applications of the autogenetic design theory and distributed computing", *Enterprise Information Systems*, Vol. 6 No. 3, pp. 329-343.

Frochot, I. and Morrison, A.M. (2001), "Benefit segmentation: a review of its applications to travel and tourism research", *Journal of Travel & Tourism Marketing*, Vol. 9 No. 4, pp. 21-45.

Haley, R.I. (1968), "Benefit segmentation: a decision-oriented research tool", *Journal of Marketing*, Vol. 32, pp. 30-35.

Hassan, S.S. (2000), "Determinants of market competitiveness in an environmentally sustainable tourism industry", *Journal of Travel Research*, Vol. 38 No. 3, pp. 239-245.

Ho, G.T.S., Ip, W.H., Lee, C.K.M. and Mou, W.L. (2012), "Customer grouping for better resources allocation using GA based clustering techniques", *Expert Systems with Applications*, Vol. 39 No. 2, pp. 1979-1987.

Ho, L.A., Kuo, T.H. and Lin, B. (2010), "Influence of online learning skills in cyberspace", *Internet Research*, Vol. 20 No. 1, pp. 55-71.

Holland, J.H. (1975), *Adaption in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.

Hsieh, T.C., Yang, K.C., Yang, C. and Yang, C. (2013), "Urban and rural differences: multilevel latent class analysis of online activities and e-payment behavior patterns", *Internet Research*, Vol. 23 No. 2, pp. 204-228.

Hsu, H.C.C. and Sung, S. (1997), "Travel behaviors of international students at a Midwestern university", *Journal of Travel Research*, Vol. 36 No. 1, pp. 59-65.

Hsu, H.C.C., Kang, K.S. and Wolfe, K. (2002), "Psychographic and demographic profiles of niche market leisure travelers", *Journal of Hospitality & Tourism Research*, Vol. 26 No. 1, pp. 3-22.

Huang, B., Li, C. and Tao, F. (2014), "A chaos control optimal algorithm for QoS-based service composition selection in cloud manufacturing system", *Enterprise Information Systems*, Vol. 8 No. 4, pp. 445-463.

Ibrahim, E.E. and Gill, J. (2005), "A positioning strategy for a tourist destination, based on analysis of customers' perceptions and satisfactions", *Marketing Intelligence & Planning*, Vol. 23 No. 2, pp. 172-188.

Jalilvand, M.R. and Samiei, N. (2012), "The impact of electronic word of mouth on a tourism destination choice: testing the theory of planned behavior (TPB)", *Internet Research*, Vol. 22 No. 5, pp. 591-612.

Jin, C., Li, F., Wilamowska-Korsak, M., Li, L. and Fu, L. (2014), "BSP-GA: a new genetic algorithm for system optimization and excellent schema selection", *Systems Research and Behavioral Science*, Vol. 31 No. 3, pp. 337-352.

Kerschbaum, F. (2008), "Building a privacy-preserving benchmarking enterprise system", *Enterprise Information Systems*, Vol. 2 No. 4, pp. 421-441.

Kim, D.J., Kim, W.G. and Han, J.S. (2007), "A perceptual mapping of online travel agencies and preference attributes", *Tourism Management*, Vol. 28 No. 2, pp. 591-603.

Kim, K.J. and Ahn, H. (2008), "A recommender system using GA *k*-means clustering in an online shopping market", *Expert Systems with Applications*, Vol. 34 No. 2, pp. 1200-1209.

Koc, E. and Altinay, G. (2007), "An analysis of seasonality in monthly per person tourist spending in Turkish inbound tourism from a market segmentation perspective", *Tourism Management*, Vol. 28 No. 1, pp. 227-237.

Krishna, K. and Murty, M.N. (1999), "Genetic *k*-means algorithm", *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, Vol. 29 No. 3, pp. 433-439.

Lambkin, M. and Day, G.S. (1989), "Evolutionary processes in competitive markets: beyond the product life cycle", *Journal of Marketing*, Vol. 53 No. 3, pp. 4-20.

Li, F., Xu, L.D., Jin, C. and Wang, H. (2011a), "Intelligent bionic genetic algorithm (IB-GA) and its convergence", *Expert Systems with Applications*, Vol. 38 No. 7, pp. 8804-8811.

Li, F., Xu, L.D., Jin, C. and Wang, H. (2011b), "Structure of multi-stage composite genetic algorithm (MSC-GA) and its performance", *Expert Systems with Applications*, Vol. 38 No. 7, pp. 8929-8937.

Li, M.J., Ng, M.K., Cheung, Y.M. and Huang, J.Z. (2008), "Agglomerative fuzzy *k*-means clustering algorithm with selection of number of clusters", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20 No. 11, pp. 1519-1534.

Lim, M.K., Bahr, W. and Leung, C.H. (2013), "RFID in the warehouse: a literature analysis (1995-2010) of its applications, benefits, challenges and future trends", *International Journal of Production Economics*, Vol. 145 No. 1, pp. 409-430.

Macleod, D. (2003), *Niche Tourism in Question – Interdisciplinary Perspectives on Problems and Possibilities*, Crichton Publications, Glasgow.

Mykletun, R.J., Crotts, J.C. and Mykletun, A. (2001), "Positioning an island destination in the peripheral area of the Baltics: a flexible approach to market segmentation", *Tourism Management*, Vol. 22 No. 5, pp. 493-500.

Novelli, M. (2005), *Niche Tourism: Contemporary Issues, Trends and Case*, Elsevier Butterworth-Heinemann, London.

Poon, A. (1993), *Tourism, Technology and Competitive Strategies*, CAB International, Oxon.

Portilla, J., Otero, A., Rosello, V., Valverde, J., Krasteva, Y.E., de la Torre, E. and Riesgo, T. (2014), "Wireless sensor networks: from real world to system integration – alternative hardware approaches", in Hashmi, S. (Ed.), *Comprehensive Materials Processing*, Elsevier, Waltham, MA, pp. 353-373.

Pyle, D. (1999), *Data Preparation for Data Mining*, Morgan Kaufmann, San Francisco, CA.

Reaidy, P.J., Gunasekaranb, A. and Spalanzania, A. (2015), "Bottom-up approach based on internet of things for order fulfillment in a collaborative warehousing environment", *International Journal of Production Economics*, Vol. 159, pp. 29-40. doi: 10.1016/j.ijpe.2014.02.017.

Reid, D. (2003), *Tourism, Globalization and Development: Responsible Tourism Planning*, Pluto Press, London.

Roberts, S. (1997), "Parametric and non-parametric unsupervised cluster analysis", *Pattern Recognition*, Vol. 30 No. 2, pp. 261-272.

Rong, J., Vu, H.Q., Law, R. and Li, G. (2012), "A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining", *Tourism Management*, Vol. 33 No. 4, pp. 731-740.

Smith, M. (1956), "Product differentiation and market segmentation as alternative marketing strategies", *Journal of Marketing*, Vol. 21 No. 1, pp. 3-8.

Tan, W., Chen, S., Li, J., Li, L., Wang, T. and Hu, X. (2014), "A trust model for e-learning systems", *System Research and Behavioral Science*, Vol. 31 No. 3, pp. 353-365.

Ting, J.S.L., Tsang, A.H.C., Ip, A.W.H. and Ho, G.T.S. (2011), "RF-MediSys: a radio frequency identification-based electronic medical record system for improving medical information accessibility and services at point of care", *Health Information Management Journal*, Vol. 40 No. 1, pp. 25-32.

Tsiotsou, R.H. and Goldsmith, R.E. (2012), *Strategic Marketing in Tourism Services*, Emerald Group Publication Limited, Bingley.

Wamba, S.F., Anand, A. and Carter, L. (2013), "A literature review of RFID-enabled healthcare applications and issues", *International Journal of Information Management*, Vol. 33 No. 5, pp. 875-891.

Wang, P., Zhang, J., Xu, L., Wang, H., Feng, S. and Zhu, H. (2011), "How to measure adaptation complexity in evolvable systems-a new synthetic approach of constructing fitness functions", *Expert Systems with Applications*, Vol. 38 No. 8, pp. 10414-10419.

Wedel, W. and Kamakura, W.A. (1998), *Market segmentation: Conceptual and Methodological Foundations*, Kluwer Academic Publishers, London.

Xu, R. and Wunsch, D.C. (2009), *Clustering*, John Wiley & Sons, Inc., Hoboken, NJ.

Yannopoulos, P. and Rotenberg, R. (1999), "Benefit segmentation of the near-home tourism market: the case of upper New York State", *Journal of Travel & Tourism Marketing*, Vol. 8 No. 2, pp. 41-55.

Further reading

Sheikh, R., Raghuwanshi, M. and Jaiswal, A. (2008), "Genetic algorithm based clustering: a survey", *Proceedings on Emerging Trends in Engineering and Technology (ICETET'08), Nagpur, Maharashtra, July 16-18.*

Corresponding author

Dr C.H. Wu can be contacted at: jack.wu@connect.polyu.hk