Computer Supported Data-driven Decisions for Service Personalization: A Variable-Scale Clustering Method

Ai WANG[1], Xuedong GAO[1], Mincong TANG[2]*

[1] Donlinks School of Economics and Management, University of Science and Technology Beijing, No. 30 Xueyuan Road, Beijing, 100083, China
ai.wang@uta.edu, gaoxuedong@manage.ustb.edu.cn

[2] The International Center for Informatics Research, Beijing Jiaotong University, No. 3, Shangyuancun, Haidian, Beijing, 100044, China
Mincong@bjtu.edu.cn (**Corresponding author*)

Abstract: The aim of this paper is to solve the object segmentation problem designed for service personalization in the context of individual athletic events. Focusing on certain personalized characteristics of the marathon contestants, the research puts forward a discovery method based on the variable-scale clustering (PCD-VSC). This method could be employed in order to obtain object segmentation based on scale similarity measurement. A case study is created based on a real dataset related to 59 marathon events which took place in several cities between 2017 and 2018, with a total number of 14,160 contestants. The numerical experimental results show that the PCD-VSC algorithm divides marathon runners into seventeen qualified clusters based on clear competitive and preference characteristics. Hence, this method could support the managers of marathon competitions in designing and implementing a personalized service scheme for the marathon contestants. Also, in comparison with the traditional VSC, the proposed method improves the overall accuracy and efficiency in analyzing categorical dataset with duplicate attribute values.

Keywords: Variable-scale clustering, Personalized service, Duplicate attribute values, Scale effect.

1. Introduction

As a sport that trains both physical and mental strength, city marathons have become a new way of pursuing a healthy lifestyle for people from different countries and of different ages and genders. In the past five years, the marathon has witnessed a boom in the number of host cities, registered events and participants. The success of such prestigious events like the Boston marathon, the New York marathon, the Berlin marathon, and the London marathon has turned marathon into an emerging symbol of those cities. Taking the Beijing marathon as example, the overall participation and competition level of marathon runners have been continuously improved during these five years (see Fig.1a). In 2019, the total number of registered marathon runners reached 165,704 (Marathon Running, 2019).

Marathon also drives the economic development of host cities in a direct way. According to the 2018 China road race report, the total annual consumption in 2018 China marathon reached 17.8 billion yuan, with the total consumption driven by the annual events reaching 28.8 billion yuan and the total annual industrial output reaching 74.6 billion yuan (Sports Quick Tongue, 2019). Besides, the 2018 China marathon annual work report shows that marathon competitions help host cities to stimulate the income generated by consumption on hotels, catering, shopping, tourism, transportation, etc. (see Fig.1b) (Ecological Sports, 2019).

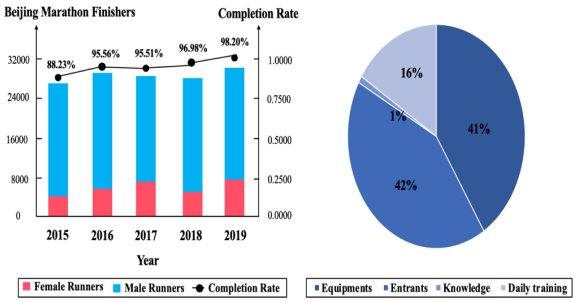

Figure 1(a). Statistics of Beijing marathon from 2015 to 2019

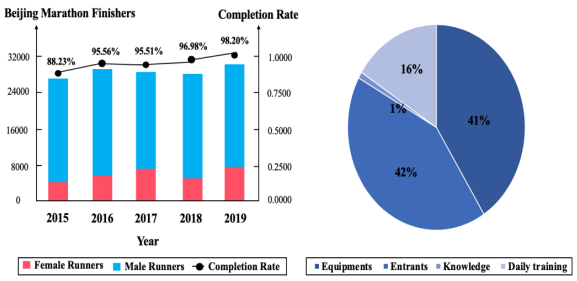

Figure 1(b). The average annual consumption feature of Chinese marathon runners in 2018

Several researchers have studied the economic benefits brought about by city marathons. (Huang and Xie, 2017) divide the economic benefits of marathon competitions into two parts: direct and indirect economic benefits.The direct economic benefits refer to the direct revenue of marathon events, including media rights, ticket revenue,

https://doi.org/10.24846/v29i1y202006

souvenir sales revenue, etc. while the indirect economic benefits refer to the benefits brought by expanding the city or the brand awareness during the period of the marathon event, and main beneficiaries are regional governments, event operators and sponsors. Yang (2019) studies the mechanism of interaction between city marathon events and local industries based on the industrial interaction theory. The results imply that marathon events and the media industry have great potential for integrated development. (Li, 2019b) focuses on the integrative development of marathon and city tourism industry, and comes to the conclusion that sports tourism which involves watching and participating in sports competitions is a significant development direction of industrial integration, which brings about new requirements for the organizing committee with regard to the quality of the provided services. Only by designing differentiated management schemes related to the individual needs of the typical consumers (runners), i.e., service personalization for city marathons, could event operators and related enterprises release their respective industrial advantages and achieve sustainable development.

Compared with the traditional service personalization in the financial sales industry (Li,2019; Davidavičienė et al., 2019), developing personalized services for the marathon industry faces more challenges on runners segmentation. On the one hand, one should identify the runners' competition characteristics. Apart from being consumers, runners are also participants in the marathon competition. Therefore, the description of their demand characteristics should not be limited to their subjective preference characteristics, their competitive characteristics should be paid more attention to, such as running capacity state and pace strategies (Elgendy & Elragal, 2016). On the other hand, the spatial and temporal characteristics related to the runners' service demand should be identified. Since different marathon events are often held in different cities and on different dates, runners' demands also change significantly following these temporal and spatial changes, that is runners' demands have (temporal and spatial) scale effects.

Therefore, this paper analyzes the data-driven service personalization problem for city marathon industry. The main contributions are as follows. Firstly, according to the concept space model (CS) of scale transformation theory, the scale similarity measurement is established. Compared with traditional similarity measurements, the scale similarity could measure the object difference not only between different attributes, but also between different observation scales under the same attribute, which overcomes the uncertainty of attribute weight assignment in previous similarity calculation. Secondly, in order to evaluate runners' competitive feature for object segmentation, the method of runners capacity index computing (RCIC) is proposed with regard to runners' personal best scores. Thirdly, a personalized characteristics discovery method based on the variable-scale clustering (PCD-VSC) is also put forward. This method yields object segmentation results with clear scale characteristics, which could help managers design a personalized service scheme. Compared to traditional variable-scale clustering method (VSC), the PCD-VSC could solve the duplicate data problem by improving the scale transformation mechanism. A real dataset of 59 city marathon events taking place between 2017 and 2018 is selected for experimental analysis, involving 14,160 runners in total. Numerical experimental results illustrate that the proposed methods are able to identify various runner clusters with scale characteristics, which directly help organizers improve the quality of services related to city marathons.

The structure of this paper is as follows. Section 2 summarizes the relevant research work, including the scale effect theory and duplicate data analysis methods. Section 3 describes the scale similarity measurements, the runners capacity index computing method (RCIC) and the personalized characteristics discovery method based on the variable-scale clustering (PCD-VSC) in detail. Experimental procedures and analysis results of real marathon datasets are discussed in section 4. Finally, section 5 sets forth the conclusion to this paper.

2. Related Works

2.1 Scale Effect

Gaining great significance in multiple research fields of natural science, scale effect describes the phenomenon through which the properties

Table 1. Example: The Multi-scale Data Model of City Marathons

Marathon	**Miles^{S0}**	*Miles^{S1}*	*Miles^{S2}*	**Climb^{S0}**	*Climb^{S1}*	*Climb^{S2}*
Race A	64.9	**[50,80)**	0	3150	**[2000,4000)**	0
Race B	77.2	**[50,80)**	0	3360	**[2000,4000)**	0
Race C	70	[50,80)	0	4935	[4000,6000)	1
Race D	93.2	[80,120)	0	4140	[4000,6000)	1
Race E	107.6	[80,120)	0	1680	[400,2000)	0
Race F	356.3	[120,+∞)	1	27390	[6000, +∞)	1

or characteristics of the observed object change once with its observation scale, especially the spatial and temporal scale (Zhang, 2017; Sethi & Karnawat, 2018). For example, (Ye, 2018) studies the scale effect of soil characteristics, and finds that the evolution law on the structural damage of yellow soil in freezing-thawing environment under different observation scales. (Wang et al., 2019) studies the scale effect of time series (like stocks and funds) on the financial field, and proposes a trend prediction algorithm based on multiple temporal scale. Experimental results using a real dataset show that the prediction accuracy of the proposed algorithm could be improved by 10% when compared with the one using only a single temporal scale. (Xu, Zhang & Liu, 2019) focuses on the spatial distribution characteristics and influence mechanism of manufacturing enterprises in the Yangtze River delta. According to a statistical analysis, the spatial distribution of manufacturing enterprises in the Yangtze River delta is unbalanced, and spatial aggregation first increases and then decreases once with the change in geographical distance, i.e., the spatial scale effect.

Scale transformation theory provides models and methods for a quantitative analysis of the scale effect of management objects, including the measurement of scale transformation rate, scale transformation strategy, scale transformation mechanism, etc. (Wang & Gao, 2019; Gao & Wang, 2019). The concept space model (CS) describes the partial order relationship between different scales of management objects under the same attribute, and is employed in order to transform the initial single-scale dataset into a multi-scale dataset (Wang & Gao, 2019b).

After interviewing the organizers of city marathons (see Section 4 for details), it has been found that a marathon event is usually characterised by four dimensions (attributes), i.e., time, location, miles and climb. Generally, the attribute time involves multiple temporal observation scales (such as month, season and year), and attribute location has multiple spatial observation scales (such as country and continent). Table 1 depicts a sample of a real city marathon dataset under two attributes namely miles and climb (see Section 4.2). In order to distinguish the difficulty level of each marathon, the organizing committee divides the basic mileage () and the basic Climb () into different gears, and formulates multiple observation scales with coarser granularity. In terms of the concept space model, the relationship between multiple observation scales under the same attribute could be clearly expressed as $Miles^{S0}\ Miles^{S1}\ Miles^{S2}$, $Climb^{S0}\ Climb^{S2}$. It can be seen that all attributes of city marathon contain multiple observation scales.

Therefore, analysing the runners' dynamic demand changes for different observation scales related to a marathon is quite crucial for object segmentation, that is the scale effect of runners' demands, so as to gain more accurate and objective runners characteristics results.

2.2 Duplicate Data Problem

The clustering analysis method in machine learning is commonly utilized to solve the customer segmentation problem for service personalization and decision support systems (Filip, 2012; Candea & Filip, 2016). It could be utilized to calculate the similarity between different customers under multiple attributes without predefined classification labels, and divide customers with more similar characteristics into the same cluster, that helps managers to develop differentiated customer management strategies (Borlea et al., 2016; Filip, 2020). Therefore, this paper approaches the cluster analysis method to study the object segmentation in the service personalization of city marathons.

According to the multi-scale data model of city marathon (see Table 1), it can be noticed that the attribute values related to basic (initial) observation scale $\{Miles^{S0}, Climb^{S0}\}$ are different for each race. However, after scaling up these two attributes namely miles and climb, the same attribute values are obtained for several races, i.e., $Miles^{S1}(RaceA) = Miles^{S1}(RaceB) = [50,80)$, $Climb^{S1}(RaceA) = Climb^{S1}(RaceB) = [2000,4000)$. The management objects *RaceA* and *RaceB* represent the duplicate values of the marathon dataset under the observation scale $\{Miles^{S1}, Climb^{S1}\}$.

However, the initial class center of the traditional clustering methods is often randomly selected (Gocken & Yaktubay, 2019). If we directly cluster the dataset with duplicate values, the duplicates might be randomly assigned to different initial clusters, which will directly increase the algorithm computational complexity.

Wu (2016) proposes a high-dimensional data clustering framework for multiple duplicate attribute values. In the beginning, all duplicate objects in the initial dataset should be identified, and eliminated from the dataset until only one representative object remains. Then, cluster analysis is carried out on the remaining data objects and the clusters are obtained as well as their cluster center. Finally, all the pre-eliminated duplicate objects are classified into matched clusters which contain the representative object, and the final clustering result is obtained. A real dataset is selected in order to verify the clustering method under different algorithm parameters. Experimental results show that the proposed method could improve clustering performance obviously under different evaluation indices.

Although the scale transformation theory has already proposed the variable-scale clustering algorithm (VSC) following the scale up transformation mechanism and scale transformation rate measurement (STR), the VSC lacks the optimal processing mechanism for duplicate value data (Wang & Gao, 2018; Wang, Gao, & Yang, 2019). This paper improves the traditional VSC by combining the duplicate value clustering method, in order to increase the efficiency of runners segmentation.

3. Personalized Characteristics Discovery Method

3.1 Scale Similarity Measurement

According to the interview results of city marathon organizers, runners' competitive characteristics are primarily related to runners' gender, endurance distance running (miles), climbing height (climb), and time, which are also affected by environmental factors during the race. Besides, in the same marathon race, male (female) runners with a shorter finish time have a higher physical capacity than other male (female) runners with a longer finish time. In different marathons, the more similar the difficulty level and the environment are, the more similar the runners' physical capacities prove to be. Thus, the capacity of different runners could be evaluated based on the finish time of the same city marathon or of similar marathon events they have participated in.

In addition, the data analysis methods and techniques are quite different depending on different data types of attribute values. Generally, there are two types of structured data attributes, i.e., continuous variables and categorical variables. The categorical variables are completely discrete and include nominal variables, binary variables and sequential variables (Zhang et al., 2017). In Section 2.1 one has already exemplified that all attributes of city marathon dataset are categorical variables. Also considering the multi-scale marathon dataset, this paper establishes the scale similarity measurement.

Definition 1 (*Scale similarity*) Let X_i, X_j represent two different objects, $A_k = \{A_k^s \mid (k = 1,2,\cdots,m) \wedge (s = 0,1,\cdots,n^k)\}$ represents the categorical multi-scale dataset, where A_k^s is the $(s+1)th$ scale of the kth attribute, m is the number of attributes, $(n^k + 1)$ is the number of scales in the kth attribute. The similarity between runner X_i and X_j is :

$$Sim^S(X_i, X_j) = \sum_{k=1}^{m} d(x_{ik}, x_{jk}) \tag{1}$$

$$d(x_{ik}, x_{jk}) = \sum_{s=0}^{n^k} \delta(x_{iks}, x_{jks}) \Big/ (n^k + 1) \tag{2}$$

$$\delta(x_{iks}, x_{jks}) = \begin{cases} 1, x_{iks} = x_{jks} \\ 0, x_{iks} \neq x_{jks} \end{cases} \tag{3}$$

where x_{ik} (x_{jk}) is the value of object X_i (X_j) under attribute A_k, and x_{iks} (x_{jks}) is the value of object X_i (X_j) under attribute scale A_k^s.

As it can be seen from definition 1, the efficiency of scale similarity calculation is relatively low due to the multiple times similarity calculation. One could even obtain a high-dimensional dataset when both the number of objects and the attributes are large. In order to efficiently identify the group of runners who have participated in similar marathon races, the method which involves fewer similarity calculation operations becomes the preferred choice.

The object-attribute space partition method is an efficient computing method for high-dimensional sparse data (Hu & Wang, 2018). After the reachable adjacency matrix is established based on the similarity relation between different objects, the strong connected domain could be directly divided by employing the subsystem judgement theorem, instead of by repeatedly calculating the similarity between objects.

Therefore, according to the object - attribute space partition method, this paper proposes an algorithm of runners capacity index computing (RCIC).

Figure 2 illustrates the basic idea of the RCIC. First, the runners' personal best record (including race time, race location, race miles, race climb, and finish time) is used to calculate the scale similarity, and divide the runners-races strong connected domain. Then, one identifies the maximum finish time $T_{\max}$ and minimum finish time $T_{\min}$ for male and female runners respectively in every strong connected domain. Finally, the runners capacity index RCI_i (see Eq.4) is calculated based on the runner's individual finish time T_i. The pseudo code of the RCIC is shown in Algorithm 1. The RCIC has advantages over the low computational complexity algorithm by calculating only once similarity of all runners, so, it can identify the set of runners who have participated in similar marathon races.

$$RCI_i = \frac{T_{\max} - t_i}{T_{\max} - T_{\min}} \times 100 \tag{4}$$

Algorithm 1 *Runners Capacity Index Computing* (D, λ)// $D = \{g,T,L,M,C,t\}$ is the initial multi-scale dataset of runners, g is the gender of runners, T is the time of marathon races L is the location of marathon races, M is the miles of marathon races, C is the climb of marathon races, t is the finish time of runners, λ is the threshold of scale similarity

1: $D^{Race} = D - \{g\} - \{t\}$
2: $A = D^{Race}.Sim^S$ // see Eq.1-3
3: **for** all $a_{ij} \in A$ **do**
4: **if** $a_{ij} \geq \lambda$ **then**
5: a_{ij}=1
6: **else**
7: a_{ij}=0
8: **end if**
9: **end for**
10: $Runners\_Races = A.SubsystemPartition$ // see (Hu & Wang, 2018)
11: **for** all $Runner_p \in Runners\_Races$ **do**
12: **if** $g(Runner_p) == female$ **then**
13: $T_{\max} = \max(t(Runners\_Races\_Female))$
14: $T_{\min} = \min(t(Runners\_Races\_Female))$
15: **else**
16: $T_{\max} = \max(t(Runners\_Races\_Male))$
17: $T_{\min} = \min(t(Runners\_Races\_Male))$
18: **end if**
19: $RCI_p = 100 \times (T_{\max} - t_i) / (T_{\max} - T_{\min})$
20: end for

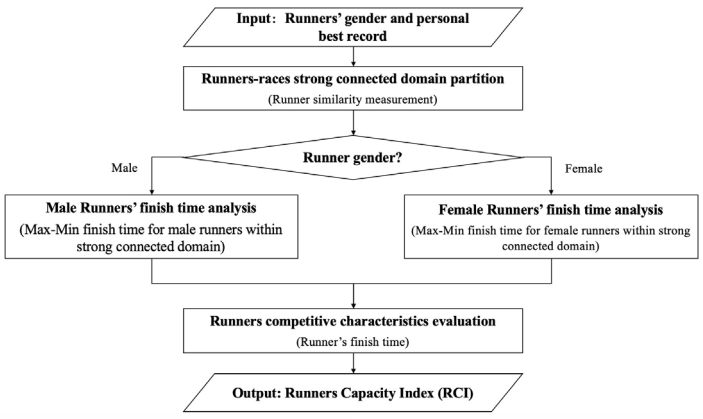

Figure 2. The algorithm framework of runners capacity index computing

3.2 Personalized characteristics discovery method based on the variable-scale clustering

The major task of service personalization for city marathons is to identify and classify runners' competitive and preference characteristics, which proves to be a multi-scale clustering analysis problem with duplicate values, as it was mentioned in Section 2. Although the scale transformation theory has already established the variable-scale clustering algorithm (VSC) based on the scaling-up transformation mechanism , the VSC always performs with a low efficiency which is caused by a random selection of the initial class center, and therefore, it lacks the optimal processing mechanism for duplicate values.

In this section, after improving the traditional VSC, one proposes the algorithm for personalized characteristics discovery based on the variable-scale clustering (PCD-VSC) (see Algorithm 2). Figure 3 illustrates the scale transformation mechanism for duplicate values. After each scale transformation, the mechanism identifies whether there are duplicate objects in the current dataset. If so, only one representative object is retained; the remaining objects are not involved in the clustering calculation process, The pseudo code of the PCD-VSC is shown in Algorithm 2.

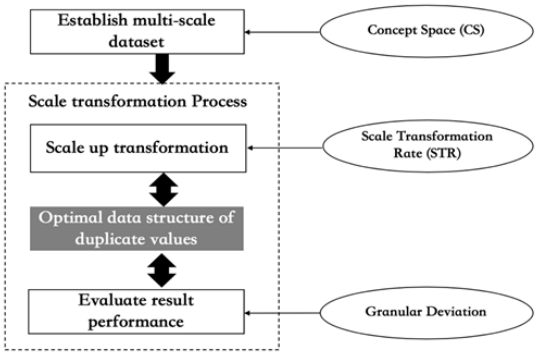

Figure 3. The scale transformation mechanism for duplicate values

Assuming that the rate of duplicate values in a n-instance dataset is $\alpha \in [0,1]$ and k is the number of clusters, the time complexity of the PCD-VSC is $O((1-\alpha/2)nkT)$, while the time complexity of traditional VSC is $O(nkt)$, where T is the iterations of the PCD-VSC, and is the iterations of the VSC. If $T = t$, the PCD-VSC reduces the time complexity by $\alpha nkt/2$ compared with the VSC, which proves the significant improvement of algorithm efficiency.

Algorithm 2 *Personalized characteristics discovery method based on the variable-scale clustering* (D, λ, S_0, k, CS) // $D = \{g, T, L, M, C, t\}$ is the initial multi-scale dataset of runners, g is the gender of runners, T is the time of marathon races, L is the location of marathon races, M is the miles of marathon races, C is the climb of marathon races, t is the finish time of runners, λ is the threshold of scale similarity, S_0 is the threshold of scale transformation rate, k is the number of clusters, CS is the concept space of RCI

$RCI = RCIComputing(D, \lambda)$ // see Algorithm 1
$D = D.MultipleScale(D, RCI, CS)$
$D^- = D.duplicates$
$D^+ = D - D^-$
$C = InitialClustering(D^+, k)$
$R_0 = \max(GrD(C_i.qualified))$ // see Wang (2019)
$D^+.delete(C_i.qualified)$
for all $C_i.qualified$ **do**
  if $C_i.qualified \cap D^- \neq \varnothing$ **then**
    $C_i.qualified = C_i.qualified \cup D^-.duplicates$
  end if
end for
$D = D^+$
for $D \neq \varnothing$ **do**
  for all $A_j \in D$ **do**
    if $STR(A_j, CH(A_j)) < S_0$ **then** // see (Wang & Gao, 2019)
      $D.update(A_j, CH(A_j))$
      $D^- = D.duplicates$
      $D^+ = D - D^-$
      break
    end if
  end for
  $C = D^+.Cluster(k - count(C_i.qualified))$
  for all $C_i \in C$ **do**
    if $GrD(C_i) < R_0$ **then**
      $D^+.delete(C_i)$
      $C_i.qualified = C_i.qualified \cup D^-.duplicates$
    end if
  end for
end for

4. Experiment results and discussion

4.1 Data collection and processing

This section focuses on an original dataset related to 59 real city marathon races, which were organized by Grit Tao Sports Development Co., Ltd. and took place between 2017 and 2018. The dataset includes the records for a total number of 14,160 runners and the host locations are distributed in 40 different cities and regions around the world (see Figure 4). Table 2 shows all variables applied for numerical experiment, including the race time (RTime), race location (RLocation), race mileage (RMiles), race climb (RClimb), and race records of all competitors (RRecords). The attributes RTime, RLocation, RMiles and RClimb all contain multiple observation scales, just like Gender, Nationality and personal best performance (PB) which belong to the runners' attributes. Similarly, the attributes PBTime, PBLocation, PBMiles and PBClimb include multiple observation scales.

4.2 Experiment process and discussion

The experimental purpose is to verify whether the RCIC and PCD-VSC could recognize the

Figure 4. The host locations of city marathons

Table 2. Summary of variables

Variable	Description
Race Attributes	
RTime	Time of city marathon races with multiple observation scales, i.e., Month, Quarter, Year
RLocation	Location of city marathon races with multiple observation scales, i.e., City/Province, Country, Continent
RMiles	Miles of city marathon races with multiple observation scales, i.e., Original miles/Level/Level plus
RClimb	Climb of city marathon races with multiple observation scales, i.e., Original climb/Level/Level plus
RRecords	Finish time (records) of all marathon runners with single scale, i.e., Minute
Runner Attributes	
Gender	Runners' gender with single scale, i.e., Male, Female
Nationality	Runners' gender with multiple observation scales, i.e., Country, Continent
Personal Best	
PBTime	Time of runners' personal best record with multiple observation scales, i.e., Month, Quarter, Year
PBLocation	Location of runners' personal best record with multiple observation scales, i.e., City/Province, Country, Continent)
PBMiles	Miles of runners' personal best record with multiple observation scales, i.e., Original miles/Level/ Level plus
PBClimb	Climb of runners' personal best record with multiple observation scales, i.e., Original climb/Level/ Level plus
PBRecord	Finish time of runners' personal best record with single scale, i.e., Minute

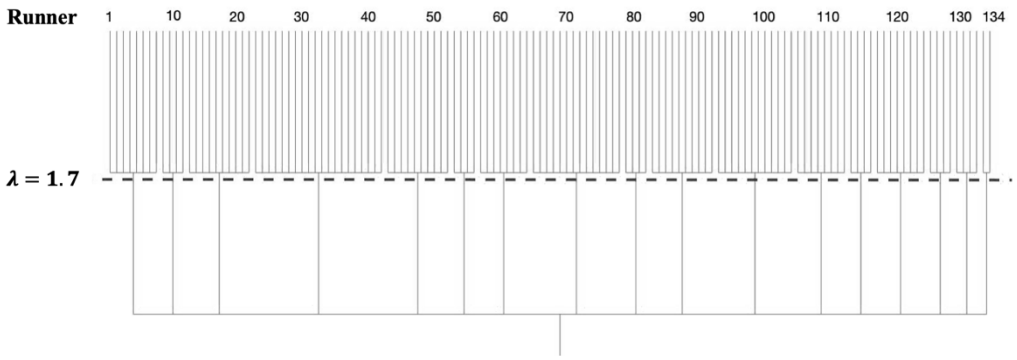

Figure 5. The runners-races strong connected domain by the RCIC

Table 3. Characteristics of Male Runners Obtained by the PCD-VSC

Iterations	#	Number of Runners	GrD	RCI	PBMiles	PBClimb	Nationality
I	1	2	**1.2504**	78.4	[50,80)	[2000,4000)	Argentina
	2	2	**1.2504**	66.5	[50,80)	[2000,4000)	Colombia
	3	6	2.0840	**89.6**	[81,120)	[2000,4000)	South Africa
II	4	**20**	3.9423	[80,100)	[81,120)	[4000,6000)	United Kingdom
	5	5	3.8063	**(0,20)**	[50,80)	[600,2000)	Australia
III	6	5	1.5754	[20,40)	[81,120)	[4000,6000)	Asia
	7	18	1.8380	[40,60)	[50,80)	[2000,4000)	Europe
	8	4	1.5754	[80,100)	[120,+)	[6000,+)	North America
	9	5	1.2603	[60,80)	[50,80)	[2000,4000)	North America

characteristics of runners and complete runners segmentation, so as to support organizing committee in the decision-making process. Thus, by combining the interview results of the organizers of marathon events, 134 runners have been randomly selected (namely, 67 male runners and 67 female runners) as experimental samples, in order to verify whether the proposed algorithms could meet the managers' requirements

Figure 5 shows the results of runners-races strong connected domain by employing the RCIC. According to all the runners' PB- and race-related information, the RCIC algorithm obtains seventeen strong connected domains for λ=1.7 , and the RCI of all runners, which provides data basis that would enable one to discover runners clusters with clear scale characteristics.

The characteristics of male runners obtained by employing the PCD-VSC algorithm are shown in Table 3. After analysing the competitive characteristics of male runners, it can be figured out that: 1) Runners in Cluster 8 have higher RCI and participate in marathon races with higher difficulty level, which would make them capable runners. On the contrary, runners in Cluster 5 have lower RCI and participate in marathon races with lower difficulty level, which would make them beginner runners. Managers could recommend marathon races and related sports information of different difficulty levels based on their competitive states; 2) Runners in Cluster 3 have the highest RCI and come from South Africa, which indicates that these runners have a great potential and deserve great attention during the marathon race.

After analyzing the spatial and temporal characteristics of male runners' service demands, it can be found that: 3) Runners in Clusters 4 and 7 are all from Europe, especially from the UK. Managers could adjust the management communication channels in a timely manner based on the runners' preferences, such as cultural habits; 4) Runners in Cluster 6 are from Asia. The marathon-related official information platform could design a language selection fouction, especially the interfaces of race regulations; 5)

Table 4. Characteristics of Female Runners Obtained by the PCD-VSC

Iterations	#	Number of Runners	GrD	RCI	PBMiles	PBClimb	Nationality
I	1	7	3.5726	59.9	**[120,+)**	**[6000,+)**	Australia
	2	3	**0.8336**	60.6	[81,120)	[4000,6000)	**Japan**
	3	3	3.3344	64.1	**[50,80)**	**[600,2000)**	South Africa
II	4	**19**	2.9998	[60,80)	[81,120)	[2000,4000)	Spain
	5	11	1.6714	**[80,100)**	[81,120)	[4000,6000)	New Zealand
	6	**13**	2.4043	(0,20)	[50,80)	[2000,4000)	France
	7	7	1.3132	[40,60)	[50,80)	[2000,4000)	**Japan**
	8	4	**0.9193**	[20,40)	[81,120)	[4000,6000)	Brazil

The granular deviation (GrD) of Clusters 1 and 2 is the smallest, which indicates that runners in each cluster are much similar based on their current observation scale. Hence, managers could recommend other runners for them when making up the network of runners.

The characteristics of female runners obtained by employing the PCD-VSC algorithm are shown in Table 4. After analysing the competitive characteristics of female runners, it can be figured out that: 1) Runners in Cluster 1 have medium RCI but participate in marathon races with higher difficulty level, which would make them radical runners. On the contrary, runners in Cluster 3 have higher RCI but participate in marathon races with lower difficulty level, that would make them potential runners. Managers could recommend marathon races and related sports information of different difficulty levels following their competitive states and preferences; 2) Runners in Cluster 5 have the highest RCI and are coming from New Zealand, which indicates that these runners have great potential and deserve great attention during the marathon race.

After analysing the spatial and temporal characteristics of female runners' service demands, it can be noticed that: 3) Runners in Clusters 4 and 6 are all from Europe, especially Spain and France. Managers could adjust the management communication channels in a timely manner based on the runners' preferences, such as cultural habits; 4) Runners in Clusters 2 and 7 are from Asia. The marathon-related official information platform could design a language selection fouction, especially Japanese interfaces of race regulations; 5) The granular deviation (GrD) of Clusters 2 and 8 is the smallest, which indicates that runners in each cluster are very similar based on their current observation scale. Hence, managers could recommend other runners for them when making up the network of runners.

5. Conclusion

With a focus the boom in city marathons, this paper presents the data-driven service personalization methods applied to the official decision support system of marathon events to improve the service quality. Firstly, the scale similarity measurement is established using the concept space model (CS) of scale transformation theory. The advantage of the scale similarity is that it could measure the object difference not only between different attributes, but also between different observation scales under the same attribute, which overcomes the uncertainty of attribute weight assignment in previous similarity calculation. Secondly, the method of runners capacity index computing (RCIC) is proposed through runners' personal best scores, which is utilized to evaluate runners' competitive feature for object segmentation. Thirdly, a personalized characteristics discovery method based on the variable-scale clustering (PCD-VSC) is also put forward. The method obtains object segmentation results with clear scale characteristics. Compared to traditional variable-scale clustering method (VSC), the PCD-VSC could solve the duplicate data problem, and help organizing committee design personalized service scheme for runners. We select a real dataset of 59 city marathon events from 2017 to 2018 for experimental analysis. Numerical experimental results illustrate that the proposed methods are able to identify various runner

clusters with scale characteristics, which meets managers' requirement in practice.

The future work will focus on a wider selection of data related to the runners' subjective preferences for runners segmentation task, so as to keep improving the service personalization level of city marathons. Also, an online service platform of marathon runners will be developed embedding our proposed methods, to further verify the algorithm efficiency.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 71272161, and in part by the China Scholarship Council.

REFERENCES

Akamani, K. & Hall, T. E. (2019). Scale and co-management outcomes: assessing the impact of collaborative forest management on community and household resilience in Ghana, Heliyon, 5(1), e01125.

Beijing Marathon. (2016). Data analysis of the 2017 Beijing marathon, Organizing Committee of Beijing Marathon (October 9). Available at: <http://www.beijing-marathon.com/html/page-12127.html>.

Beijing Marathon. (2017). Data analysis of the 2017 Beijing marathon, Sohu (October 13). Available at: <http://www.sohu.com/a/197878300_492663>.

Beijing Marathon. (2018). New Record! More than 500 people broke records for the 2018 Beijing marathon, Sina Sports (October 18). Available at :<http://www.beijing-marathon.com/html/page-12127.html>.

Borlea, I.-D., Precup, R.-E. & Dragan, F. (2016) On the architecture of a clustering platform for the analysis of big volumes of data. In IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (pp.145–150).

Candea, C. & Filip, F. G. (2016). Towards intelligent collaborative decision support platforms, Studies in Informatics and Control, 25(2), 143-152. DOI: 10.24846/v25i2y201601

Cui, F., Tang, H., Zhang, Q., Wang, B. & Dai, L. (2019). Integrating ecosystem services supply and demand into optimized management at different scales: A case study in Hulunbuir, China, Ecosystem Services, 39(1), 100984.

Davidavičienė, V., Raudeliūnienė, J. & Zubrii, M. (2019). Evaluation of Customers' Sustainable Fashion Perception, Journal of System and Management Sciences, 9(4), 50-66.

Ecological Sports. (2019). The annual report of 2018 China marathon shows that the total output of the marathon industry reached 74.6 billion yuan, Sohu (March 12). Available at: <http://www.sohu.com/a/300597739_505583>.

Elgendy, N. & Elragal, A. (2016). Big Data analytics in support of the decision-making process. Proceedia Computer Science, 100(2016), 1071–1084.

Filip, F. G. (2012). A decision-making perspective for designing and building information systems, International Journal of Computers Communications & Control, 7(2), 264-272.

Gao, X. D. & Wang, A. (2018). Variable-scale clustering. In *Proceedings of the* 8th International Conference on Logistics, Informatics and Service Sciences (pp. 221-225).

Gao, X. D. & Wang, A. (2019). Customer satisfaction analysis and management method based on enterprise network public opinion. Operations Research and Management Science, In Press.

Gocken, T. & Yaktubay, M. (2019). Comparison of different clustering algorithms via genetic algorithm for VRPTW, International Journal of Simulation Modelling, 18(4), 574-585.

Hu, Y. Y. & Wang, A. (2018). Modelling technique of the object-attribute system oriented to data mining, *International Journal of Technology and Management*, 17(3), 155-169.

Huang, R. Y. & Xie, S. Y. (2017). An analysis on the economic factors and economic benefits of the development of sports industry - a case study of marathon competition, *Journal of Guang Dong Communication Polytechnic*, 16(4), 39-42.

Li, M. (2019). A personalized service improvement scheme and implementation strategy of

agricultural and commercial bank, 3-15. Jiangxi: Jiangxi University of Finance and Economics.

Li, Y. (2019). Research on the integration and development of marathon industry and tourism industry under the new development concept, *Tourism Overview*, 10(2), 31-32.

Marathon Running. (2019). Big data released after 2019 Beijing marathon: The number of record-breaking players has reached a domestic record, Baidu (December 18). Available at: <https://baijiahao.baidu.com/s?id=1653264159611824504&wfr=spider&for=pc>.

Sethi, N. A. & Karnawat, S. N. (2018). Real Time Reporting of Inventory: An Innovation in Inventory Management, Journal of Logistics, Informatics and Service Sciences, 5(2), 1-8.

Sports Quick Tongue. (2019). Why is the current city marathon gradually becoming a calling card of a city?, Baidu (November 21). Available at : <https://baijiahao.baidu.com/s?id=1650777031055682978&wfr=spider&for=pc>.

Takashina, N., Marissa, M. L. & Baskett, L. (2016). Exploring the effect of the spatial scale of fishery management, Journal of Theoretical Biology, 390(1), 14-22.

Wang, A. & Gao, X. D. (2019). Multifunctional product marketing using social media based on the variable-scale clustering, Technical Gazette, 26(1), 193-200.

Wang, A. & Gao, X. D. (2019). Hybrid variable-scale clustering method for social media marketing on user generated instant music video, Technical Gazette, 26(3), 771-777.

Wang, A., Gao, X. D. & Yang, M. H. (2019). Hybrid variable-scale clustering method for social media marketing on user generated instant music video. In *Proceedings of the* 9th International Conference on Logistics, Informatics and Service Sciences (pp. 65-69).

Wang, J. C., Deng, Y. P., Shi, M. & Zhou, Y. F. (2019). Time series trend prediction at multiple time scales, *Journal of Computer Applications*, 4(1), 1046-1052.

Wu, S. & Fu, L. W. (2016). High-dimensional data clustering for customers with duplicate attribute values. In *Proceedings of the 6th International Conference on Logistics, Informatics and Service Sciences (LISS'2016)* (pp. 1-6). DOI: 10.1109/LISS.2016.7854441

Xu, W. X., Zhang, X. J. & Liu, C. J. (2019). Spatial distribution pattern and influencing factors of manufacturing enterprises in Yangtze River Delta: Scale effects and dynamic evolution, *Geographical Research*, 38(5), 1236-1252.

Yang, J. (2019). A study on the interaction mechanism between urban marathon and local industry in China - a case study on Wuxi international marathon, *Journal of Hubei University of Economics*, 16(12), 24-30.

Ye, W. J. (2018). Scale effects of damage to loess structure under freezing and thawing conditions. *Rock and Soil Mechanics*, 39(7), 1-9.

Zhang, D. (2017). High-speed Train Control System Big Data Analysis Based on Fuzzy RDF Model and Uncertain Reasoning, International Journal of Computers, Communications & Control, 12(4), 577-591.

Zhang, D., Sui, J. & Gong, Y. (2017). Large scale software test data generation based on collective constraint and weighted combination method, Tehnicki Vjesnik, 24(4), 1041-1050.