Cluster Analysis for Personalised Mobile Entertainment Content

Worapat Paireekreng[1], Kok Wai Wong[1]
and Chun Che Fung[1]
[1] Murdoch University, South Street, Murdoch,
6150, Perth, WA, Australia
{w.paireekreng, k.wong, l.fung}@murdoch.edu.au

Abstract. There is much attention given to emerging technologies like mobile internet because of its increasing popularity. Much research has concentrated on hardware and some have focused on personalisation in terms of content visualisation. The focus of this paper is on mobile content personalisation, seeking to understand the user groups through clustering users based on their profile. This paper focuses on the implementation of a technique known as 'Zoning-Centroid', which is the evaluation technique used to determine the appropriate number of clusters required to best cluster the given users profile. The user profile used in this paper includes mobile content usage and their demographic factors. The clustering algorithm used in this paper is k-means clustering. The results show that the proposed technique could suggest appropriate number of clusters to be used with the k-values, in order to implement for mobile entertainment content personalisation.

Keywords: mobile content user clustering, mobile content personalisation, clustering, cluster analysis.

1 Introduction

Mobile devices have improved substantially over the past few years with many added features and attributes such as internet access. Besides provide communications and mobile services, it also allows users to access a wide range of information based on personal need or context. Relatively some improvements have been performed to enhance the delivering of the mobile internet content, an area of increasing importance [21]. However, there are still problems relating to information overload and the users' behaviour on their desire contents. This creates a challenge for researchers in the domain of identifying the user's segment, and the definition of these groups according to different demographic factors and user information rankings.

In this paper, we investigate and analyse mobile internet user using clustering analysis. The technique realised in this paper could be applied in the entertainment computing domain by assisting users when accessing mobile entertainment content. The feature selection for the number of clusters and content items for each cluster is implemented for clustering with the zoning of Centroid to examine the cluster of mobile internet content user.

2 Background

2.1 Mobile Content Personalisation

Personalisation was defined as mechanisms to allow a user to adapt or produce a service to fit user's particular needs, and that after all subsequent services rendering by this service towards the user is tuned accordingly [14]. Mobile personalisation research has focused on how to facilitate the use of mobile internet. Application such as tourist guide, news update or classified information and services [7], [8] have been developed. Adaptive content which can be adjusted when the usage changed according to the context becomes important issues. It includes some researches that also looking into entertainment content and mobile games by catering different game genres for different group of users [10],[11].

Intelligent systems with machine learning and data mining play a vital role for personalisation system such as finding customer's needs[1]. Wu et al. [6] have shown that some commonly used algorithms in data mining are k-means, and SVM. Wu et al. also described k-means as a simple iterative clustering method. This is also a simple algorithm with the adaptation ability for different applications [2]. The clustering component in [16, 17] also show the mobile user clustering using demographic factors and information ranking to filter the cluster could enhance the system.

2.2 Mobile User Clustering

The research on mobile internet user can be observed from [24]. Yamakami [24] used Aging Analysis model to identify mobile internet user behavior. This model used statistical techniques to divide the users into four groups based on amount of access time. This research focused only on the frequency of the user using the mobile internet. In 2006, Okazaki [3] has includes attitudinal and demographic information for cluster analysis. It automatically determined the number of clusters which is four clusters based on Baysian Inference Criterion (BIC) techniques and TwoStep algorithm. Furthermore, [22,23] develop a formula to identify the appropriate number of clusters using a method known as 2, 3 and 4 cluster. These researches implemented k-means clustering technique to know how long they spent their time in each time zone (Always on, Morning, Daytime and Night). In another research [4], they used factors analysis related to call usage, payment behavior and additional service usage clustered by k-means and Association Rule.

K-means and determining number of clusters. K-means has been introduced by Tou and Gonzalez since 1974 [12]. K-mean can handle large data and is computational efficient with its simple implementation when compared to other techniques like hierarchical clustering [13], [24] and implemented in various areas [9], [19] such as image processing or information retrieval. Nevertheless, identifying the optimal number of cluster seems to be a problem. There are some research in identifying the number of cluster such as 'Gap Statistic' [5] which focused on well-separated cluster and uniform distribution dataset. In 2009, Muhr and Granitzer [15] proposed automatic cluster number selection by applying x-means with split and merge clustering method. They measure cluster validity with BIC and F-Score. This technique is also appropriate with known class or labeled data. Another cluster number determination technique is 'L method' [20]. However, this method did not work well with global evaluation metrics and it is unable to work with less number of clusters like 1 or 2 clusters.

Labelled clustering data. F-Score has been used in [15] to measure the quality of cluster. However, to measure the cluster analysis, prior information of the cluster is necessary. This problem is similar to Random Index (RI). Although there is research performed in using fuzzy C-means clustering with 'Induced Entropy' to evaluate the cluster, it also needs testing data for known classes such as visited and recommended web pages [18].

Cluster evaluation. Ray and Turi [19] proposed the method to evaluate the clustering techniques using k-means. It is validity ratio which is defines as

$$validity = \frac{Intra}{Inter}$$

The concept of this measurement is minimizing the sum of squared distance for intra-cluster and maximizing inter-cluster value. If the validity value is small, it can be implied that the cluster is compact compared with other k-values.

From the above, it can be observed that most of the researches have focused on the clustering in terms of mobile internet user behavior such as adoption or experience in mobile internet. Although, there were researches mentioned about number of cluster and clustering techniques, there are not used in clustering of mobile entertainment content users. In addition, with the limited computational resource user needs to get the response,. a simple algorithm should be considered for mobile computing. K-means and its acceptable k-values which are suitable for unlabelled mobile content user clusters will be carried out for this paper. This method helps to select the appropriate number of clusters with efficient computation.

3 Proposed Method and Experiment

The data source used for the experiment was obtained from the published research on the mobile internet content users in Bangkok [25]. This set of data consists of the user's content preference such as multimedia, news or information services on mobile internet. 300 randomly selected records were used as training data for clustering. There are several factors and attributes in the dataset. In this research, we have selected the key demographic factors of gender, age, income and occupation to find potential groups or clusters. These attributes were chosen in acquiring the requisite data from the mobile internet users as well as the ease of classification for further analysis.

The cluster analysis is performed using the k-means cluster technique. K-means clustering technique was selected as it provides a simple algorithm that can be used to determine cluster sizes. This allows the implementation of a clustering model at the server of the content provider in order to know the customers' characteristic and provide appropriate content to each cluster based on the cluster characteristic.

The aim of the experiment is to analyse the group based on demographic factors. The analysis should generate the appropriate number of clusters for mobile content users, leading to the identification of contents which these clusters of users will be accessing. The experiments are conducted with k-means where k=4,5,6,7 and 8 consecutively.

Zoning-Centroid. The evaluation method called '*Zoning-Centroid*' is proposed in this paper. The distance from the centre of each cluster should be used to determine the cluster's members in each cluster, and to ensure that they are appropriately distributed.

'*Zoning-Centroid*' will use the distance between centre of cluster and data to calculate the zone that this data is sought. It measures how far from the centre of this data. The zone will be divided into 5 zones. Each zone is computed from *Zone-Distance* which is derived from the difference between the maximum distance in the cluster and the minimum distance in the cluster

$$\text{ZoneDistance}_{(n,i,k)} = (\text{MaxDistance}_{(n,i,k)} - \text{MinDistance}_{(n,i,k)}) / 2^n . \qquad (\mathbf{1})$$

where, n=zone number of cluster i and k = k-values; $1 \leq n \leq 5$; $1 \leq i \leq 5$; $4 \leq k \leq 8$

Then, the '*Zone-Limit*' will be calculated from *Zone-Distance* as following

$$\text{ZoneLimit}_{(n,i,k)} = \text{ZoneLimit}_{(n-1,i,k)} + \text{ZoneDistance}_{(n,i,k)} . \qquad (\mathbf{2})$$

where, n = zone number of cluster i and k = k-values; MinDistance = n-1 for n=1.

After that, the distance of each data will be assigned to each zone according to its limits. For example, if cluster 1 and zone 1 limit is 0.924737, the data with distance below this limit will be in zone 1. In contrast, if the distance is over that limit, the data will be assigned to the subsequent zone. Figure 1 shows the concept of zoning.

We will measure the amount of cases that will fall in each zone and count the number and percentage of each zone to determine the data distribution based on '*Zoning-Centroid*'. This evaluation method will be applied in k-means between k=4 and k=8 for mobile content usage.

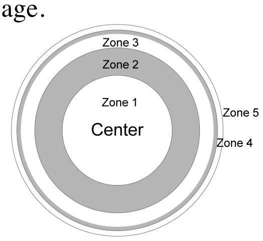

Fig. 1. The '*Zoning-Centriod*' diagram shows the data coverage for each zone

The 'Zoning-Centroid' separates each zone using exponential distance from the centre of each cluster. The main concept is based on proximity between data and centre. The first zone consumes 50% distance from the minimum to half of the maximum distance. This is because good quality cluster should contain data close to its centre as much as possible. Therefore, the first zone covers area larger than the next zone. In contrast, if the zone division implements linear zoning, the distance in each zone would be separated to be far from each other equally. As a result, the data will difficult to be separated as most data will appear to be near to its centre.

4 Experiment Results

4.1 Analysis of Cluster with k-values

The characteristics of each cluster based on demographic factors and content usage are analysed and concluded by using difference k-values as follows;

k=4. The results show that gender and age do not have any effect on clustering except cluster 4. It shows unique characteristics are teenager, low income and studying. For income and occupation, they are different in 3 clusters, therefore they are unable to be determined precisely.

k=5. At this k-value, demographic factors, income, started to show some significance and separated more precisely. In addition, age has clearer influenced on the cluster than the previous k-value. The 'teenager' is still the dominating attribute in clustering while there is no effect to clustering with gender.

k=6. The cluster of 'teenager' is maintained and gender still has no effect towards clustering. Age and occupation seem to be clearer. There are different ages in each cluster such as more than 18 years old, more than 36 years old or between 19-35 years old. In addition, income begins to be separated into less income and above average income.

k=7. There is one cluster that has the proportion equal to 5% appeared in cluster 6 and age begin to influence to clustering. Then, income is also clustered more precisely in cluster 1 and 2 by less income and more income groups. Occupation shows the groups which are free-time and employed with low income in cluster 3 and 4. It is similar to other k-values that 'teenager' cluster is separated clearly compared to other clusters.

k=8. We stop at this k value by setting up cut off point when the small cluster which proportion less than 5% appeared. The teenager group together with gender has effect in cluster 1 and 8 by division between male and female with combination of age. Furthermore, occupation also determines the group characteristic by presenting employed or having more free time. In cluster 3,4,7 which there is age between 19-35 years old, show different among cluster by occupation and income.

4.2 Number of Cluster Using 'Zoning Centroid'

As can be seen from the table 1, the cases in Zoning-Centroid for each cluster (CZCC), as expected, the percentage of cases that fall in Zone 1 is the highest percentage in each k-value. The cumulative percentage of cases between Zone 4 and Zone 5 is around 5-8%. It can be implied that 92-95% of data approximately has not fallen over to zone 3 for every k-value. In addition, it shows the highest percentage in Zone 1 followed by Zone 2 and Zone 3 which means the data for k=5 is disseminated appropriately especially in the first 2 zones. The percentage and trends of each k-values and CZCC each zone are shown in Figure 2.

To consider the cumulative of dissemination of data compared to percentile of distance from the centre of the cluster to its limit, we sum data from Zone 1 to Zone 4 which are 93.75% of the percentile of distance. This shows that the data distribution is at around 94-96%. The highest is at k=4 and it decreases slightly when k-values is increased. However, the percentage rises again when k=8. When compared with cumulative 3 zones, the result still presents trends similar to 4 zones.

K=5 shows the cumulative percentage from Zone 1 to Zone 2 at approximately 94% which is significantly higher than the other k-values. The percentage comparison can be seen from the table 2.

According to CZCC, it seems that k=5 shows the most significant results compared to other k-values based on *Zoning-Centroid* consideration with less cumulative zones (2 zones). This can be implemented to choose the appropriate number of cluster for mobile content usage.

Table 1. Number of Cases in '*Zoning-Centroid*' in Each Cluster.

Cluster	# Cases Zoning-Centroid	k=4	k=5	k=6	k=7	k=8
1	Zone 1 limit	58	22	66	25	9
	Zone 2 limit	21	21	0	2	4
	Zone 3 limit	3	0	2	0	0
	Zone 4 limit	0	1	0	0	0
	Zone 5 limit	2	1	4	3	1
2	Zone 1 limit	44	29	20	20	20
	Zone 2 limit	9	10	4	4	0
	Zone 3 limit	2	1	2	2	1
	Zone 4 limit	1	1	1	0	0
	Zone 5 limit	5	3	1	1	1
3	Zone 1 limit	57	66	33	67	43
	Zone 2 limit	20	0	2	22	25
	Zone 3 limit	4	0	0	0	7
	Zone 4 limit	2	0	0	0	0
	Zone 5 limit	3	3	1	11	4
4	Zone 1 limit	61	47	47	12	12
	Zone 2 limit	7	16	16	3	3
	Zone 3 limit	0	0	0	3	3
	Zone 4 limit	0	0	0	0	0
	Zone 5 limit	1	6	6	1	1

5	Zone 1 limit		68	43	63	63
	Zone 2 limit		4	0	6	6
	Zone 3 limit		0	1	0	0
	Zone 4 limit		0	7	0	0
	Zone 5 limit		1	3	1	1
6	Zone 1 limit			38	11	22
	Zone 2 limit			1	3	9
	Zone 3 limit			0	0	2
	Zone 4 limit			1	0	0
	Zone 5 limit			1	1	1
7	Zone 1 limit				27	19
	Zone 2 limit				9	13
	Zone 3 limit				2	0
	Zone 4 limit				0	1
	Zone 5 limit				1	2
8	Zone 1 limit					24
	Zone 2 limit					1
	Zone 3 limit					1
	Zone 4 limit					0
	Zone 5 limit					1
	Total	300	300	300	300	300
	CZCC - Zone 1	73.33%	77.33%	82.33%	75.00%	70.67%
	CZCC - Zone 2	19.00%	17.00%	7.67%	16.33%	20.33%
	CZCC - Zone 3	3.00%	0.33%	1.67%	2.33%	4.67%
	CZCC - Zone 4	1.00%	0.67%	3.00%	0.00%	0.33%
	CZCC - Zone 5	3.67%	4.67%	5.33%	6.33%	4.00%
	Total	100.00%	100.00%	100.00%	100.00%	100.00%

Table 2. The percentage sum of data dissemination in various zones.

Zone	K				
	4	5	6	7	8
Sum 4 Zones (93.75%)	96.3333%	95.3333%	94.6667%	93.6667%	96.0000%
Sum 3 Zones (87.5%)	95.3333%	94.6667%	91.6667%	93.6667%	95.6667%
Sum 2 Zones (75%)	92.3333%	94.3333%	90.0000%	91.3333%	91.0000%

* In blanket means the percentile of distance from centre to its limit

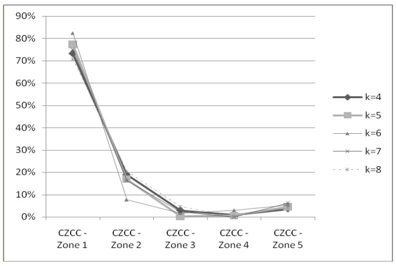

Fig. 2. Percentage of cases of ‘Zoning-Centroid’ in each zone in each cluster

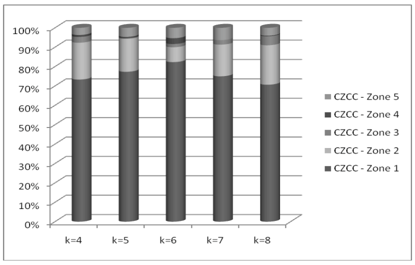

Fig. 3. Cumulative percentage of cases of ‘Zoning-Centroid’ in each zone in each cluster

4.3 Cluster Evaluation

To evaluate the quality of cluster, the method to measure the number of cluster is used and the results are shown in table 3. We use TwoStep clustering techniques to compare the results of the number of clusters. This method was used in Okazaki's research for determining the number of clusters and mobile inter adopter cluster solution [3]. However, the TwoStep algorithm with BIC (Bayesian Information Criterion) and ratio of distance measure showed that the number of auto-clustering for this mobile content usage dataset is just 2 clusters. As a result, we will ignore this measurement because the results for the clustering are unable to implement in the further stage such as customer's pattern of content usage. It is too small a number of clusters. Our method can show most numbers of cluster compared reasonably to auto-clustering with TwoStep.

The quality of clusters is then measured. As in this case, we are using unlabelled data, we will use the fundamental concept of clustering to measure the quality. The concept is based on the measurement of the minimum distance within cluster and the maximum distance between clusters. Therefore, the validity metric from Ray and Turi [19] is used. The results are presented as following:

Table 3. The validity of clustering

k	Intra	Inter	Validity
4	499.4445	3.8780	128.7895
5	430.2511	3.9670	108.4574
6	372.2088	3.2829	113.3793
7	360.6455	2.7056	133.2981
8	298.2944	2.3483	127.0281

According to the validity metric, the intra cluster value is calculated from the sum squared of distance in the cluster while the inter cluster value is selected from the minimum value of distances between cluster centers which it is desired to be maximized. The validity can imply that if the value is small, it means the cluster quality is good. From the results, it implies that for this mobile content clustering problem, the appropriate number of clusters should be 5 because its validity is the smallest compared to other k-values clustering.

Table 4. The comparison of validity and 'Zoning-Centriod' using cluster k=5

	Normal Validity	Zoning-Centriod	% of calculation reduced
Number of cases	300	283	5.6667%
Percentile of distance	100%	75%	25%

To consider the concept of 'Zoning-Centroid', it can be seen from the Table 4 that this method can reduce the number of data or cases to be calculated for finding the number of clusters and at the same time ensuring the quality of cluster by at least

5.6667%. It also decreased the percentile of distance to consider to 75% from the cluster's centre.

5 Discussions and Conclusions

This research not only recommends the optimum number of clusters for mobile internet content user groups but also provides the techniques to cluster through the use of k-means and subsequent evaluation with '*Zoning-Centroid*'. This clustering is based on demographic factors with the data provided by the users allowing both the cluster analysis to be processed easily. The 'Zoning-Centroid' can assist in determining the appropriate k-values for the number of clusters, allowing the content providers to focus on individual clusters and deliver the right content to the right group at the right time.

The results of the research potentially increase business value by determining the optimal number of clusters to be grouped for mobile content personalisation. The appropriate number of clusters is determined by the combination of a clustering technique with fundamental demographic factors.

The k-means is a simple algorithm, and therefore suitable to be used for the mobile content personalisation. The model can be built at the content provider's server and predict the user's group from incoming user's profile faster. When the content provider knows the user's characteristics, it would be easier to provide appropriate content to them quickly.

References

1. Bose, I., Mahapatra, R.K.: Business Data Mining - A Machine Learning Perspective. J. Information and Management 39, 211--225 (2001)
2. Jain, A.K.: Data Clustering: 50 Years Beyond K-Means. J. Pattern Recognition Letters (2009)
3. Okazaki, S.: What Do We Know About Mobile Internet Adopters? A Cluster Analysis. J. Information and Management 43, 127--141 (2006)
4. So Young Sohn, Kim, Y.: Searching Customer Patterns of Mobile Service Using Clustering and Quantitative Association Rule. J. Expert Systems with Applications 34, 1070--1077 (2006)
5. Tibshirani, R., Walther, G., Hastie, T.: Estimating the Number of Clusters in a Data Set via the Gap Statistic. J. Royal Statistical Society 63, 411--423 (2001)
6. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D., Steinberg, D.: Top 10 Algorithms in Data Mining. J. Knowledge and Information Systems 14, 1--37 (2008)
7. Zhang, D.: Web Content Adaptation for Mobile Handheld Devices. J. Communications of the ACM 50, 75--79 (2007)
8. Zipf, A., Jost, M.: Implementing Adaptive Mobile GI Services Based on Ontologies Examples from Pedestrian Navigation Support. J. Computers, Environment and Urban Systems 30, 784--798 (2006)

9. Gonzalez-Barron, U., Butler, F.: A Comparison of Seven Thresholding Techniques with the K-means Clustering Algorithm for Measurement of Bread-crumb Features by Digital Image Analysis. J. Food Engineering 74, 268--278 (2006)

10.Paireekreng, W., Rapeepisarn, K., Wong, K.W.: Time-Based Personalised Mobile Game Downloading. In: Z., Pan et al. (eds.) Transactions on Edutainment II. LNCS, vol. 5660, pp. 59--69. Springer, Heidelberg (2009)

11.Rapeepisarn, K., Wong K.W., Fung C. C., and Khine, M.S.: The relationship between Game Genres, Learning Techniques and Learning Styles in Educational Computer Games. In: Technologies for E-Learning and Digital Entertainment, pp.497--508. Springer-Verlag Berlin/Heidelberg, (2008)

12. Tou, J.T., Gonzalez, R.C.: Pattern Recognition Principles. Addison-Wesley, Massachusetts (1974)

13. Bose, I., Xi, C.: Exploring Business Opportunities from Mobile Service Data of Customers Using Inter-cluster Analysis. In: The IEEE International Workshop on Data Mining for Design and Marketing, Hongkong (2006)

14.Jorstad, I., Thanh, D.V., Dustdar, S.: Personalisation of Future Mobile Services. In: 9th International conference on intelligence in service delivery, Bordeaux, France (2004)

15. Muhr, M., Granitzer, M.: Automatic Cluster Number Selection using a Split and Merge K-Means Approach.In: 20th International Workshop on Database and Expert Systems Application. IEEE Computer Society, Linz, Austria (2009)

16.Paireekreng, W., Wong, K.W.: Mobile Content Personalisation Using Intelligent User Profile Approach. In: The 3rd International Conference on Knowledge Discovery and Data Mining (WKDD2010), Phuket, Thailand (2010)

17.Paireekreng, W., Wong, K.W.: Client-side Mobile User Profile for Content Management Using Data Mining Techniques. In: 8th International Symposium on Natural Language Processing, Bangkok, Thailand (2009)

18.Phatak, D.S., Mulvaney, R.: Clustering for Personalized Mobile Web Page.In: The IEEE International Conference on Fuzzy Systems (2002)

19.Ray, S., Turi, R.H.: Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation.In: the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT99), Calcutta, India (1999)

20.Salvador, S., Chan, P.: Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithm.In: 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04), Boca Raton, Florida (2004)

21.Uribe, S., Fernandez-Cedron, i., Alvarez, F., Menendez, J.M., Nunez, J.L.: Mobile TV Targeted Advertisement and Content Personalization. In: 16th International Conference on Systems, Signals and Image Processing (IWSSIP2009), Chalkida, Greece (2009)

22.Yamakami, T.: A User-Perceived Freshness Clustering Method to Identify Three Subgroups in Mobile Internet Users. In: The 2nd International Conference on Multimedia and Ubiquitous Engineering, Busan, Korea (2008)

23.Yamakami, T.: Exploratory Day-scale Behavior Assumption-Based User Clustering with the Mobile Clickstream. In: Eighth International Conference on Parallel and Distributed Computing, Application and Technologies, Adelaide, Australia (2007)

24.Yamakami, T.: Toward Understanding the Mobile Internet User Behavior: A Methodology for User Clustering with Aging Analysis. In: Fourth International Conference on Parallel and Distributed Computing, Application and Technologies, Chengdu, China (2003)

25.Paireekreng, W.: Influence Factors of Mobile Content Personalization on Mobile Device User in Bangkok. (2007)