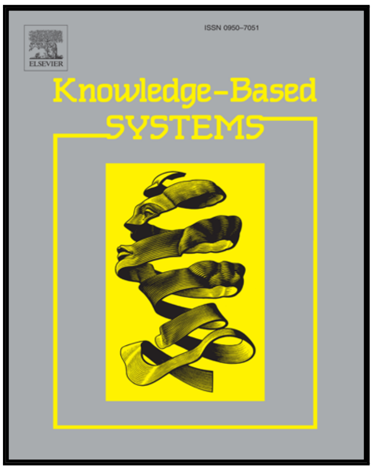


Statistics-based CRM approach via time series segmenting RFM on large scale data

Meina Song, Xuejun Zhao, Haihong E, Zhonghong Ou

PII: S0950-7051(17)30261-7
DOI: 10.1016/j.knosys.2017.05.027
Reference: KNOSYS 3925

To appear in: *Knowledge-Based Systems*

Received date: 16 November 2016
Revised date: 25 May 2017
Accepted date: 29 May 2017

Please cite this article as: Meina Song, Xuejun Zhao, Haihong E, Zhonghong Ou, Statistics-based CRM approach via time series segmenting RFM on large scale data, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.05.027

Statistics-based CRM approach via time series segmenting RFM on large scale data

Meina Song[a], Xuejun Zhao[a], Haihong E[a], Zhonghong Ou[a,*]

[a] *School of Computer Science, Beijing University of Posts and Telecommunications*
Beijing, China 100876

Abstract

Conventional customer relationship management (CRM) is typically based on RFM model, whose parameters are the recency, frequency and monetary aspects of target customers. The latest comprehensive analysis has enabled CRM to present parameters with time series. For example, researchers can account for changing trends based on an RFM model for flexible marketing strategies. Such changes might inspire telecommunication service scenarios that user value relies on long-term performance. In this study, we propose a statistic-based approach to value latent users via time series segmenting time interval of RFM in large scale data set. Apart from utilizing in Spark platform, we integrate multiple corresponding analysis (MCA) to regularize clustering results by the RFM model and extend these approaches to multiple levels. A comprehensive set of experiments, revealed interesting observations regarding the coexistence of time interval and RFM model. First, the clustering method along time interval in three dimensions of the RFM model outperforms the method along the three dimensions in each interval.Subsequently, the cooperation of RFM and MCA provides a convenient methodology for exploring CRM in large–scale data.Therefore, the RFM model with time intervals integrated with MCA in CRM are essential.

Keywords: CRM, RFM, large–scale data, MCA, time interval

1. Introduction

The main task of customer relationship management (CRM) is to value and retain users by exploring the potential relationships among users and deriving innate values of their own characteristics [1], because the characteristics interact with these relationships [2]. Characteristics are quantitative and qualitative ones; both are supposed to reflect different relationships [3]. Given the intense competition of telecommunication operators and rapid growth of telecom service data generated by smart phones, CRM for telecom service data has been a strategic initiative method for identifying high-net–worth clients and providing improved service [4].

*Corresponding author

Email addresses: mnsong@bupt.edu.cn (Meina Song), xuejunzhao@bupt.edu.cn (Xuejun Zhao), ehaihong@bupt.edu.cn (Haihong E), zhonghong.ou@bupt.edu.cn (Zhonghong Ou)

During the course of valuing users by CRM, the potential relationships among users can be divided into the external entrance evident to researchers, and the internal entrance built by their innate characteristics. Data for innate entrance is more available; thus, we explore quantitative and qualitative characteristics of internality in detail. Typically, the RFM model explores quantitative characteristics and enriches the criteria for potential relationships in CRM, because customer value can be reflected by the most recent consumption as the recency, the frequency in normal consumption, and the monetary cost of consumers in the model [5, 6, 7, 8]. Given the magnitude and complexity of telecom service data, the trend of change with time has been considered for the RFM model [5].

We investigated the quantitative characteristics in CRM to evaluate the user relationships more precisely. Considering the ever-increasing capacity of qualitative characteristics [7], we explored the innate interaction relationships among qualitative and quantitative characteristics; thus, we found that the qualitative characteristics play a significant role in CRM. For the correlativity of qualitative characteristics, researchers normally consider clustering methods with manual classification [9, 10]. Therefore, we explore a statistics-based approach to value latent users via time series segmenting time interval of RFM.

In this paper, cellphone flow in the RFM model is regarded as general currency without real records of consumption in the marketing area, with the presumption that user value detected by cellphone flow remains the same for each user. Specifically, we utilized the Spark platform to explicitly regularize telecom service data into the established data format (including quantitative and qualitative characteristics respectively). When applying the RFM model, we leverage time series analysis for proper time intervals, and integrate the clustering method (e.g., k-means) to significantly detect their respective groups. We apply multiple corresponding analysis (MCA) to regularize clustering results in the three dimensions of RFM; thus, we can discover corresponding relationships between quantitative and qualitative characteristics to enhance the research capability of CRM.

The contributions of this paper are listed as below: (1) It presents a detailed methodology for constructing time interval for the RFM model, particularly in large–scale data; Besides, we conclude that 7 a.m. is a significant watershed in the usage of cellphone flow; (2) Unlike normal clustering method that considers the swift time and fitting RFM model into every time interval, this paper contrasts clustering experiments on both dimensions of RFM and time intervals, concluding that fitting the RFM model separately in time interval guarantees precise performance. (3) MCA highlights the regularized methodology on clustering results, rather than comprehensive research on corresponding relationships between qualitative and quantitative characteristics for relationships in internality.

The rest of the paper is organized as follows: The framework and methodology are described in section 2, the experiment environment in section 3, the segmented time interval for the RFM model in section 4, the relationship detected by quantitative characteristics in section 5, the customer relationship detection in section 6, related work in section 7 and the conclusion in section 8.

2. Framework and methodology

2.1. Data framework

While exploring qualitative and quantitative relationships on telecom service data in internality, the RFM model provides quantitative relationships in three levels, whereas MCA generates qualitative ones. The telecom service data is detected by International Mobile Subscriber Identification (IMSI) per minute in one day from one base station, which includes the occurrence time, occurrence amount of cellphone flow and the qualitative perspective of functional categories, handset manufacturers, and application developers. The original data were approximately 30 GB. After IMSI diagnosis, we can collect cellphone flow information that belongs to approximately 1,000,000 users on the Spark platform. Compared with the original CRM analysis in so-called large–scale date on 10,000 users, our telecom data is more similar to large–scale data.

Guided by limits of the detected information, we presume that the cellphone flow serves the same role as currency in the marketing area when it is applied in the RFM model. In addition, qualitative characteristics have to be preprocessed before MCA is applying. Owning to the technical requirements of qualitative and quantitative data in internality, data preprocessing is necessary to divide numerical for RFM and qualitative characteristics for subsequent experiments.

For data on numerical characteristics, we can precisely collect the occurrence time and amount of cellphone flow in byte unit, whether during uploading or downloading. On the other hand, the qualitative characteristics can be collected, including the functionality, service provider, client application and user agent, where the functionality refers to the functional usage of the cellphone flow.The data format of Spark includes quantitative and qualitative data.

2.2. Spark process

In this section, we describe three features that we have added to Spark HDFS, specifically to handle challenges in ”large–scale data” environments. First, telecom service data are composed of numerical and qualitative characteristics, acquired by deploying such data on HDFS to divide the original data frame separately into each aspect; thus, the data can be immediately queried. Second, large scale processing on the time interval performs well on the regularized data format in an established data set. The leverage of time can ease the sparsity problem. We describe how time series analysis is being incorporated into arranging the quantitative data into the established time interval. Subsequently, values of three dimensions in the RFM model can be collected in each interval, and thus the vectors set as each interval in MLlib package are introduced to improve efficiency. Finally, k-means clustering in successive time duration works on vectors on disparate time durations. Based on Spark and its MLlib package, quantitative characteristics can be converted into values in three dimensions (recency, frequency and monetary) after RFM modeling and relative processing. Finally, the corresponding relationships interact well in MCA from quantitative results and qualitative characteristics. The overall process is shown in Figure 1. Thus, Spark can be a powerful tool for large–scale data.

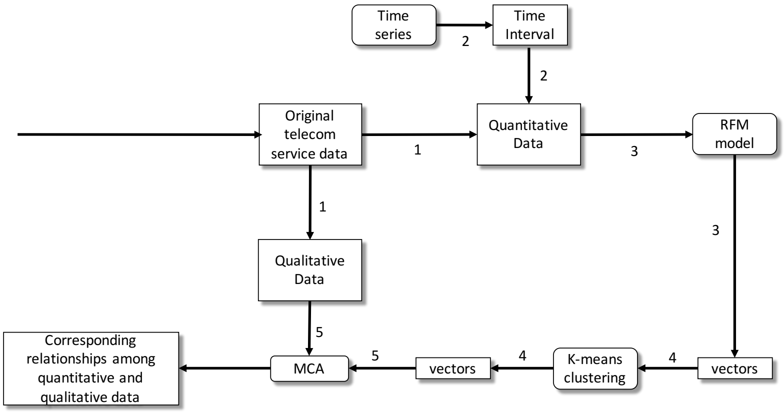


Figure 1: Phases of process planning overall. Rectangles represent present data format, circles represent methods, and figures represent the experimental sequence.

2.3. Data preprocessing

Overall, telecom service data has grown to 30 GB, with 481,905,749 rows. The cellphone flow is supposed to be indicted as Gaussian curve because of the existing large-scale data. When IMSI identifies unique endless equipment, the original data, divided into quantitative and qualitative data in the telecom service, work effectively in different experiment procedures.

During preprocessing, IMSI is the unique identification for dividing and uniting data of 1,006,149 users in externality and internality. After processing in Spark, we can collect the volume for experimental data. For the data on numerical characteristics in externality recorded per minute, similar to occurrence time and amount of cellphone flow, the information is rearranged into vectors in one hour. On the other hand, the qualitative characteristics in internality can be comprehensively collected, where categories of functionality, service provider, client application and user agent are 25,40,86,48,respectively. The data description is presented in Table 1.

2.4. Experimental procedure

In terms of the magnitude of cellphone data in a day, cellphone flow data for the RFM model is segmented in various intervals rather than the original one–minute interval. Time series analysis can provide proper segmentation for time intervals. As this approach allows us to identify the distribution in range of nuanced quantitative differences among various time intervals, experiments verify that these distributions can be decomposed into normal distributions in each interval, which are associated with a distinct mean and variance. In effect, the resulting nominal variables with RFM model traits can be presented in various intervals without violating any statistical assumptions.

In the process of adopting RFM model into various intervals, the vectors of RFM values appear; thus, we introduce k-means clustering method to unite these values

Table 1: Data on externality and internality

Data category	Parameter	Volume
Quantitative data	IMSI	1006149
	startTime	1440
	endTime	1440
	ByteUp	3264560GB
	ByteDown	3206195GB
Qualitative data	functionality	25
	service Provider	40
	client Application	86
	user agent	48

for the final outcome. Rather than clustering values on the basis of three parameters (recency, frequency and monetary) of the RFM model in each interval and gathering models that are as numerous as the amount of time intervals, we cluster values along coherent time intervals to gather three models, which are the same as the three dimensions in the RFM model. The three parameters in the RFM model separately over sequential time for persisting the RFM core meaning: user value is reflected by monetary cost, as well as embodied by frequency and recency. For better business implications, MCA can prune the unequal number of clusters in three levels of RFM model.

To study the qualitative relationships in internality, we apply MCA to avoid generating biased evaluation criteria by the RFM model and generate profound cognization for multiple innate characteristics. First, MCA can provide a comparison for ultimate values of three dimensions in the RFM model, which is a basis for scaling the numerical data into a regularized format. MCA can prune the three dimensions of the RFM model into a unified clustering number. MCA discovers corresponding relationships among qualitative characteristics and quantitative values after RFM.

Finally, we plausibly assumed that the users with quantitative characteristics can be clustered into three dimensions of the RFM model. Each dimension can be explored in MCA correspondingly with qualitative characteristics; thus, both can be mutually complimented.

3. Experimental environment

The dataset on telecom service is rapidly growing in size and complexity. Solutions are needed to harness this wealth of data by experimenting on a big data platform. Spark is a full, top–level Apache project whose key abstraction consists of resilient distributed datasets (RDDs), which are fault-tolerant collections of objects partitioned across cluster nodes that can be acted on in parallel [11]. These datasets are admitted as general big data analysis tool by open source contributors, such as Cloudera, Databricks, IBM, Intel and Map R [12]. When users create RDDs by applying operations called transformations, such as map, filter, and groupBy, to data in a stable storage system, such as the Hadoop distributed file system (HDFS). [11], the data processing of Spark satisfies the fundamental safety requirement because the output of a task in one partition can be copied from HDFS to local disks [13]. Spark is built with enough APIs and abundant optimized engines on the Hadoop platform. Spark is efficient at iterative computations and is well-suited as a general purpose batch–processing engine of parallel data [14].

Scala owns standard features of programming language, such as pattern-matching, which enables developers to use the full programming language while still making rules easy to specify; thus, we found Scala well-suited to this task [15]. Apache Spark is a popular open-source platform for large-scale data processing that is suitable for iterative machine learning tasks. In this paper we present MLlib, the open-source distributed machine learning library by Spark.

Based on the efficient implementations of large-scale machine learning algorithms on several high-level libraries, the tight integration of MLlib with Spark has the following benefits [14]. (1) When bundled with Spark, MLlib supports several languages, such as Java, Scala, Python, and R. (2) MLlib provides a high-level API that leverages Spark's rich ecosystem to simplify the development of end-to-end machine learning pipelines. [14]. (3) To the best of our knowledge, MLlib is the first production quality query optimizer built on such a language. MLlib provides efficient functionality for unsupervised study, such as k-means clustering.

Overall, Spark is implemented in Scala with built–in MLlib, which is a powerful object–oriented language with ample resources [12]. We used six nodes (five workers and one master) with the following configuration:

- Intel(R) Xeon(R) CPU E5-2620V3 @ 2400 MHz (40 virtual cores);
- 64GB RAM;
- total 40 virtual cores and 520GB virtual memory;
- 26GB capacity for .blk.gz file type once equipped with 5 executors;
- Size of 2GB per container.

Software configuration

- Ubuntu 14.04;
- jdk1.8.0_65;
- Apache Hadoop HDFS 2.6.0 with short-circuit local reads enabled;
- Apache Spark master branch (target for Spark 1.4.0 release).

4. Segmented time interval for RFM model

In this section, we conclude that the time interval for RFM model from time series analysis is the basis for detecting quantitative relationships. First, we sketched the 185 approximate tendency of cellphone flow in a given day. Subsequently, the time series analysis was introduced to segment the time interval in the RFM model on Spark. Finally, we discuss the time interval applied in the RFM model.

4.1. Data sketch

Equipped with the quantitative data identified by IMSI, we first studied the overall 190 distribution to determine the proper time interval. The overall time-flow distribution is shown in Figure 2. Each hour subset contained the mean values on 60 minutes on the given day, with occasional spikes. The cellphone flow tendency is almost hollow-like before 7.00 a.m. but exhibits a straight line after 7.00 a.m. The Gaussian distribution is based on large-scale data. Thus, we presume that the cellphone flow data obey the 195 distribution in each hour interval. By applying t-test method, we can decide whether to accept the premise of a Gaussian distribution or not. To explore argument 1 further, imagine that we have 2 sample groups labeled 1 and 2, with means μ_1 and μ_2, the variance s_1^2 and s_2^2, and the sample size N_1 and N_2. The t statistic for t-test is calculated as [16]:

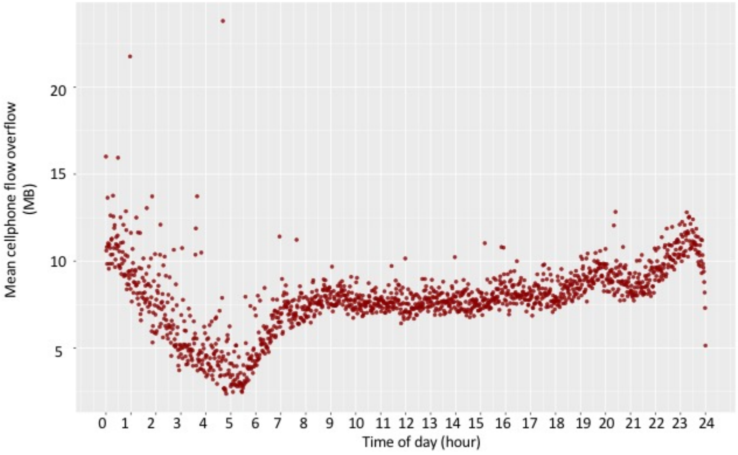


Figure 2: Overall time-flow distribution in one day.

$$t = \frac{\mu_1 - \mu_2}{s_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1}$$

200 where the pooled variance s_p^2 is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_1 - 1)s_2^2}{n_1 + n_2 - 2} \tag{2}$$

The evaluation for the hypothesis of a t-test is binary. On the one hand, the symbol 0 does not reject the null hypothesis. On the other hand, the symbol 1 can reject that hypothesis. All results are 0; thus, the presumption of a normal distribution of telecom data at each hour is satisfied. The data with nuanced variance in each hour obey a Gaussian distribution; thus, a time series can be introduced to segment the time interval in more precise ways and predict the different patterns.

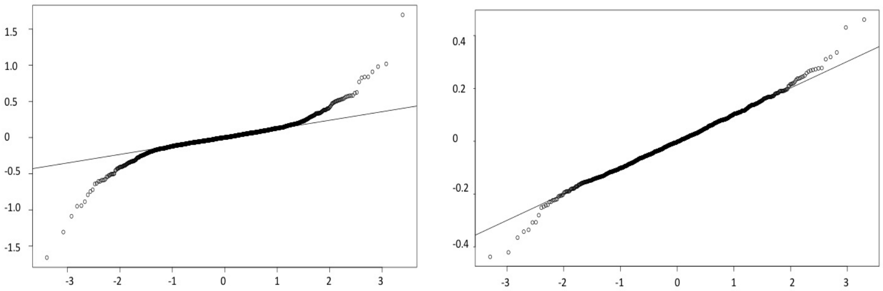
4.2. Define time interval

Given that the data in so–called 1024 intervals are too sparse, we consider rearranging and uniting the intervals. As guided by the requirement of norm distribution on rearranged time interval, the initial first seven-hour interval remains to be independent. Therefore, the remaining hour has been considered in the model procedure. First, the normality assumption of the remaining 17 hours should be tested in a QQ plot; Subsequently, the regression methods, such as normal regression and auto-Regressive and moving average (ARMA) model, are introduced for modeling. As time series analysis is introduced, we can identify user values by quantitative characteristics.

4.2.1. Normality detection

To test the normality assumption of rearranged data on the data set after 7.00 a.m., we plot standardized residuals against their normal scores on a QQ plot. If the residuals have a normal distribution, the plot should show a straight diagonal line. The figure is a scatter plot of the standardized level 1 residuals on original data set, thereby calculating for the final model including the cross-level interaction, against their normal scores. The graph indicates close conformity to normality, with no extreme outliers. Similar plots can be made for the level 2 residuals.

The poor outcomes of the QQ plot on the original data set in Figure 3(a),we consider the log (logarithm) and $diff$ (difference) methods to trim and prune the original data, thereby fitting the statistical model requirement. We test the original data, the data with log technique and data with $diff$ technique, by disobeying the normality presumption, until the data with log and $diff$ techniques appear together, whose QQ plot is presented in Figure 3(b).



(a) QQ plot with original data.

(b) QQ plot with log,diff data.

Figure 3: Comparison of QQ plot perforamce on original, log, diff and log,diff data set

4.2.2. ARMA regression

In the normal regression method, the residual standard error is 1.831 on 1,439 degrees of freedom. The multiple R-squared is 0.1788, and the adjusted R-square is 0.1782. The F-statistic is 313.4 on 1 and 1439 DF, and the p-value is below $2.2e-16$. Meanwhile, as the prior normal regression method does not present the best result , ARMA model is considered for better performance on the trimmed and processed data. First, the auto-correlation function (ACF) of sample and the bayesian information criterion (BIC) are the prior discriminator for the values of parameters: p (auto regression value), d (difference frequency for a smooth sequence), q (moving average value) in ARMA model. For the convenience and accuracy of the experiment, we extracted the final hour in the trimmed data set as the test set.

The coefficients tend to be 0 within the boundary on ACF and the gain on reality with data on lag 2 on BIC. Thus, we can define the $p = 2$, $d = 0$ and $q = 0$. After the calculation of software, σ^2 is estimated as 0.007783; the value of ACF is estimated as -2051.9, which satisfies the requirement of the t-test [17]. Furthermore, the residuals for the model fitting Gaussian distribution because the QQ plot is straight. The equation is applied in the test set, which is tested by the mean absolute error (MAE).

$$MAE = \frac{\sum_i \|x_i - bar\bar{x}_i\|}{T} \tag{3}$$

Where T is total number of test items, i is the symbol of items, x_i is the real value and $\bar{x}_i$ is the predicted value. The lower the MAE values, the better is the performance presented. Finally, the value of MAE is 1.1957, whereas that of the train set is 0.7149, thereby fitting the requirement for the model [17]. Thus, we can accept the presumption of the ARMA model, and the data set of cellphone flow with default value can be predicted in the model. The formula for the corresponding model is:

$$\begin{aligned} diff(log(y_t)) = (&0.0001 - 0.6001 diff(log(y_{t-1})) \\ &-0.3334 diff(log(y_{t-2})) + \varepsilon_t) \end{aligned} \tag{4}$$

4.3. Interval applied in RFM model

Guided by the multi-dimensional studies in the RFM model, we can preprocess the cellphone flow data in established interval: recency represents the length of time since last purchase, whereas frequency denotes the number of cellphone flow uses within a specified period. The monetary cost represents the amount of cellphone flow spent in this specified time period. Meanwhile, the tendency graph for recency, frequency and monetary levels could be gathered after data preprocessing in an established time interval.

As vectors that represent information in Figure 4, cellphone flow information has been placed in two dimensions: 8 time intervals and 3 parameters (recency, frequency and monetary) in the RFM model. Therefore, we introduced the clustering methods into the RFM model to evaluate the users and their belonging clusters.

Given the co-existing dimensions of intervals and the RFM model, we have to decide the proper dimension for clustering. On the one hand, we can cluster values along time intervals for three models. On the other hand, we can cluster values on three parameters in the RFM model to gather eight models.

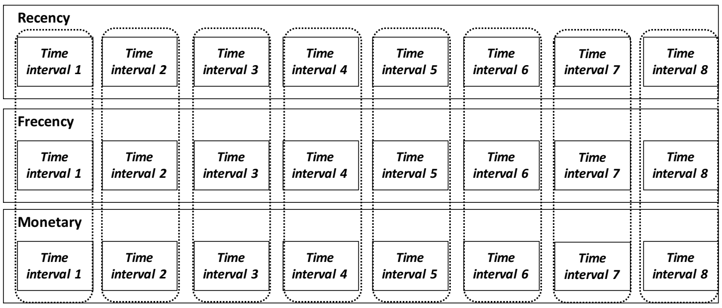


Figure 4: Format of vector on various time intervals in RFM model, horizontal lines present three dimensions of RFM model, and vertical ones present various time intervals.

5. Relationship detected by quantitative characteristics

Through the time series analysis on Spark, we can detect relationships of quantitative characteristics in internality. First, we conduct contrast experiments on the traditional RFM model and the RFM model from the time series analysis in this study. The RFM model built on various time intervals is processed on Spark; thus, we employed k-means clustering on three dimensions, namely, recency, frequency and monetary cost in the RFM model, as quantitative benchmarks for user relationship.

5.1. Experiment description

After vector building on the RFM model in various time intervals, the clustering method on these vectors has to be considered. The prior experiments clustered the recency, frequency and monetary values of the RFM model on each time interval. Thus, the three dimensions of the RFM model are rearranged by the time stamp, where the number of models equals the amount of time intervals.

Researchers disrupted the original thought of the clustering RFM model into each time interval [5]. Therefore, we applied the trend of the clustering values after the RFM model by clustering the data on each time interval based on each dimension of the RFM model. Thus, we can gather three models in three dimensions of the RFM model. From the perspective of large scale data, the defined three models outperform the number of models equal to the number of intervals.

To verify the accuracy of the time interval definition and clustering methods on the RFM model along the changing time, we conducted experiments on our presented algorithm and two contrast experiments. Our proposed algorithm A1 is clustering values along time intervals, for three models: Dimension R, Dimension F and Dimension M. The first contrast algorithm A2 clustered values on the three dimensions (recency, frequency and monetary) of the RFM model in each interval for eight models: Dimension 1, ..., Dimension 8. The last contrast algorithm A3 is conducted on the original vectors without processing by the clustering method. The comparison of algorithms A1 and A2 proves the better criteria of clustering on the combination of the RFM model and time interval whether by time interval or three dimensions of the RFM model. The comparison of algorithms A1 and A3 confirms the accuracy of introducing the clustering method into vectors.

The evaluation criteria for algorithms comparison are the average processing time, and the silhouette distance [18] of each point from the nearest cluster to effectively and vividly compare algorithms.

For each datum i in applying silhouette distance, let $a(i)$ be the average dissimilarity of i with all other data within the same cluster. We can interpret $a(i)$ as how well i is assigned to its cluster (the smaller the value, the better the assignment). We then define the average dissimilarity of i to a cluster c as the average of the distance from i to all points in c. Let $b(i)$ be the lowest average dissimilarity of i to any other cluster, of which i is not a member. The cluster with the lowest average dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i. We can define a silhouette as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

The interval for $s(i)$ in the preceding definition is $[-1, 1]$, where 1 means $b(i) >> a(i)$ and the i is well-matched in its own cluster, -1 means $b(i) << a(i)$ and the i is more appropriate in other clusters other than its own, and 0 means the datum is appropriately clustered.

5.2. Contrast experiments

First, the processing time in three compared experiments separately presents the comparison in the aspects of 2, 3, 4, 5, 6, 10, 15 and 20 groups after clustering. In Figure 4, the processing time of A1 and A2 algorithms in various dimensions stay in the same magnitude for less than 100 seconds, although the time of A1 is slightly higher than that of A2. By contrast, the processing time of the A3 algorithm is beyond 100 seconds, thereby proving the poor effect on the original data set without clustering method with no evident effect on time interval definition.

Second, the silhouette distance is the main index for identifying the experimental effects among algorithms A1, A2 and A3. We clustered groups from 2 to 20, and we presented groups from 2 to 7 as limited by the length of the article. From table 4, we have to gather the best groups for these algorithms in various dimensions. In algorithm A1, the R, F, and M dimensions perform well on clustering into 7, 5 and 7 groups; In algorithm A2, 8 dimensions perform well in 2 groups after clustering. Subsequently, algorithm A3 had worse result than A1 and A2 regardless of the number of groups.

For algorithm A2, the experiments present a clustering RFM model into two clusters in 8 time intervals, thereby performing the best clustering results. For the mathematical aspect, the clustering centers of the RFM model in each interval can be presented as the linear combination of the RFM model. As ij presents the i-th group center on j-th dimension in modeling, C_{ij} stands for the numerical combination of

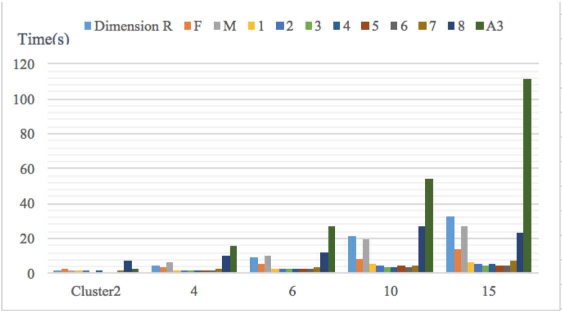


Figure 5: Experimental results of various dimensions contrasting in processing time .

Table 2: Contrast on silhouette distance in three algorithms.

Dimension	Clustering number					
	2	3	4	5	6	7
R	0.3952	0.6643	0.6055	0.5933	0.5902	**0.6821**
F	0.4543	0.6214	0.6512	**0.687**	0.5131	0.5345
M	0.5274	0.5915	0.5823	0.6294	0.6386	**0.6479**
1	**0.5922**	0.5643	0.5759	0.5644	0.5469	0.4931
2	**0.594**	0.5668	0.5627	0.5476	0.5286	0.4995
3	**0.5568**	0.486	0.5263	0.5437	0.4966	0.5088
4	**0.5465**	0.4871	0.5174	0.554	0.5302	0.4899
5	**0.543**	0.4854	0.5229	0.5527	0.5279	0.5045
6	**0.5485**	0.4939	0.5222	0.5399	0.5159	0.504
7	**0.5965**	0.5102	0.5284	0.5178	0.5109	0.4739
8	**0.6615**	0.6381	0.6413	0.6297	0.621	0.5843
A3	0.3635	0.2832	0.2341	0.2311	0.1972	0.183

RFM model, where R_{ij} is the numerical result on R dimension, F_{ij} is numerical result on F dimension and M_{ij} is numerical result on M dimension.

$$C_{ij} = R_{ij} + F_{ij} + M_{ij} \tag{6}$$

The simplified parameters in each time interval present 2^n probabilities of performance because n is the number of time intervals. Thus, we have to reject the useless probabilities and accept the right ones manually.

The bar graph in Figure 5 indicates the numerical results of cluster centers on three dimensions in the RFM model in algorithm A2. Besides, the line presents the silhouette distance of each cluster to distinguish the best performance on group 2.

Finally, we can conclude that the present algorithm A1 outperforms algorithm A3

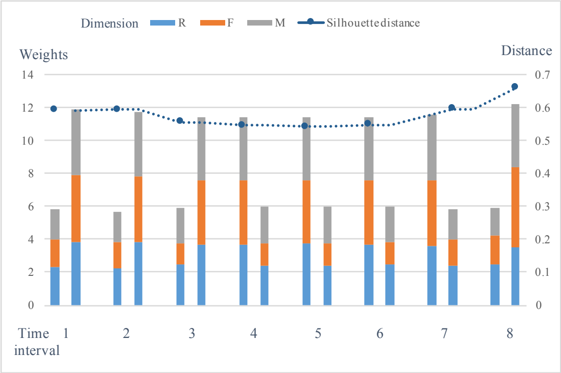


Figure 6: Performance of A2 algorithm, the bar graphs present the weights of cluster centers on three dimensions in the RFM model, and the dotted line presents silhouette distance of each cluster.

with the original vectors, and algorithm A2, thereby proving the need to define the time interval and clustering of the results of the RFM model.

5.3. RFM model

As enhanced by the user value counting on the direct material value as well as the related recency and frequency, the RFM model can also be explored. After the model building by ARMA, the stable phase from 7 a.m. is explored as one interval in conclusion, and the time before 7 a.m. can be explored in one hour.

The trimmed time interval has to be rearranged in the vectors of three dimensions, recency, frequency and monetary cost in the RFM model. Given that the RFM sets up the boundaries on each model, we can obtain the changing trends of recency, frequency and monetary models in various time intervals, which build the data trend for the time axis.

Subsequently, we gathered the processing time and silhouette distance in Figure 6 for our algorithm A1. The processing time is linear increasing unless the clustering groups approach 20, but the overall time in the three dimensions is preserved in magnitude under 100 seconds. In terms of silhouette distance, three dimensions present a wavelike tendency, particularly for dimension F. By considering the computing time consumed in approaching 20 groups, the selection of clustering into 5 or 15 groups in dimension F is the former. Besides, dimensions R and M chose 7 groups as the best classifications.

First, we cluster the recency level on 2, 3, 4, 5, 6, 7, 8, 9 clusters by silhouette clustering criteria.Subsequently, we can observe the clustering effect by dissimilarity of items in the same category and in various categories. We select 7 as the clustering amount for the silhouette criteria performing best. Second, we cluster frequency level on 2, 3, 4, 5, 6, 7, 8, 9 clusters by the silhouette clustering criteria. Subsequently, we can see the clustering effect by dissimilarity of items in the same category and in different categories. We selected 5 as the clustering amount for the silhouette criteria performing best. Finally, we cluster the monetary level on 2, 3, 4, 5, 6, 7, 8, 9 clusters

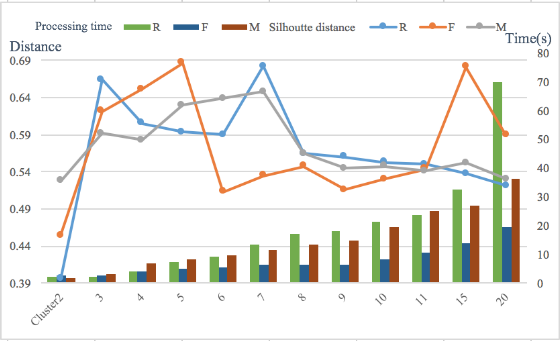


Figure 7: Performance of A1 algorithm, the bar graphs present the processing time of three dimensions of the RFM model, and the lines present performance of Silhoutte distance.

by silhouette clustering criteria. Subsequently, we can see the clustering effect by dissimilarity of items in same category and in different categories.

We selected 5 as the clustering amount for the silhouette criteria performing best. In conclusion, we take 7 clusters in the recency model, 5 clusters in the frequency model and 7 clusters in the monetary model.

The clustering process on the RFM model detects the tendency of user behavior in three numerical dimensions. Without quantitative characteristics, the scores of the aforementioned three models are insufficient to provide proper suggestions for business implications, particularly in large–scale data.

6. Corresponding relationships detected by MCA

After the experiments on Spark, which focus on the benchmarks of numerical values with the RFM model, we applied MCA to regularize the three dimensions of the RFM model into unified clustering centers, for better business implication. We explored user relationships between quantitative benchmarks and qualitative characteristics from a holistic perspective by MCA.

6.1. MCA model

Analysis the relationships of several categorical dependent variables in the large database can be realized by MCA, rather than using the analogy of regression models, which ignores the interaction effects in the same direction as the main effects [19]. Technically, MCA is obtained by applying a standard correspondence analysis on an indicator matrix (i.e., a matrix whose entries are 0 or 1) [20]. After computing the main factor with the main matrix, we gather two factors shared by multiple characteristics.

The performance of indicator matrix *CA* on nominal variables and observations will provide two sets of factor scores: one for the rows as `Dimension 1` and one for the columns as `Dimension 2`.

These factor scores are generally scaled such that their variance is equal to their corresponding eigenvalue (some versions of *CA* compute row factor scores normalized to unity).For the proximity between variables we need to distinguish two cases. First, the proximity between levels of different nominal variables means that these levels tend to appear together in the observations. Second, given that the levels of the same nominal variable cannot occur together, we need a different type of interpretation for this case. The proximity between levels means that the groups of observations associated with these two levels are themselves similar.

6.2. Regularizing three parameters in RFM model

After numerical characteristics processing after RFM model with 7, 5 and 7 groups in dimensions R, F and M, we conduct MCA on the values of three dimensions in the RFM model into regularized format, where users can be evaluated by the same criteria in various dimensions. In MCA, we concluded the inner corresponding relationships of three dimensions by preparing for the interacting relationships among quantitative and qualitative characteristics. Therefore, relationships of various dimensions in RFM are regularized in MCA and users can be precisely evaluated.

To demonstrate the interpretation of MCA on three dimensions of the RFM model, we present the corresponding results of three dimensions.

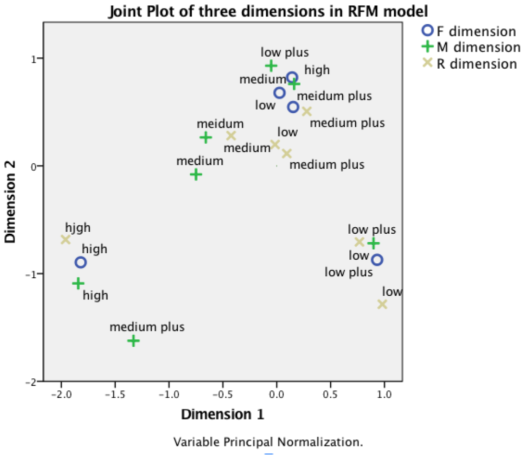


Figure 8: Relationships among video and model in the article.

From Figure 8, we can conduct the final high, medium, medium plus, low plus and low levels for recency, frequency and monetary dimensions, where medium differed from medium plus as the former presented the stable performance on medium level over time, while the latter presented a large fluctuation change around medium (the same as low and low plus).

6.3. User intents reflected by qualitative characteristics

After corresponding relationships of three dimensions in RFM model were clarified, we explored the interaction relationship among these dimensions and cellphone flow functionaries, as well as service providers, where we can reveal user intents. The user intents can be detected as the prior description about segmented users are based on demographical data [21], while the input data of this research is not so abundant.

The data description is presented in Table 3, where over 80% functionality and service provider has been presented.

Table 3: Distinct user intents on functionality and service provider with coverage=0.8

User category	Functionality	Service provider
High	Advertisement	Google AdMob
Medium	Web browsing	Baidu
Medium fluctuate	Social networking	QQ Weibo
Low	Marketplace	Taobao Youku
Low fluctuate	Video	iTunes Burstly

According to MCA, most data points are clustered around medium and medium fluctuation performance of the three dimensions from the RFM model. For service providers, high performance are AdMob and Google, both of which are advertisement companies. It is reasonable for the best advertisement company online focusing on the users with best performance, while the other advertisement companies fail to grasp the best users, leading to the blank of best performance in functionality. After that, the group using Baidu is medium , corresponding with web browsing; Users of QQ and Weibo seem to be medium fluctuate in social networking; Taobao consumers are to be low in cellphone flow, revealing corresponding relationship with marketplace. However, the video company Youku belongs to low category, while iTunes and Burstly are low fluctuate. Thus, we take a close look at video.

We conducted an MCA experiment on video-watching users with more detailed data, and gathered the relationships among video-watching behavior and the other behavior on functionality. Notably, over 95% of the users perform video viewing and web browsing, and the better their performance is, the closer the people behavior occurred in the spike time (spike time indicates the cellphone flow occurs to be extremely in the neighborhood).

Subsequently, the analysis on the video generalizes the deep use in future. To detect user behavior, users with sole customer characteristics can be aided by the conclusion of the RFM model. On the other hand, users with sole user behavior can be supported by the conclusions of multivariate statistical analysis to detect customer characteristics.

Once a person is detected with video usage, we can conclude that he/she may conduct the medium performance on the recency dimension and low performance on the frequency and monetary with 95% confidence. Meanwhile, considering that the RFM model performance on user behavior is in one separate quadrant, we can probably discover a video-addicted user.

7. Empirical basis

7.1. Empirical basis for introducing time

Owning to the previous clustering on RFM model [21], the large scale data with diverse attributes works well when demographical background is provided. However for the normal situation, the research content is unattached with any other input data. The inner attribute of this data is considered more than before, like time, to rich the outcome of model. Without this, the limited data can simply been discussed for a restricted outcome.

For customized analysis, the changing behavior along time is supposed to be detected, as user value is changing along behavior [22]. The quicker we detect the change of customer mind, the more profit we can gather from customer pocket. We can paint a clear picture of anonymous users, to reach better understanding of consumer behavior that can help improve user experience.

As the original time interval is determined by device, which is probably not suitable for objective demands of practical experiments. Thus time series analysis is allowed to revise original time interval is better for detecting proper user behavior.

7.2. Empirical basis for other situation

Despite applying in one case study, this empirical procedure is useful for other CRM researches. The time range of variables detected by the end device is reasonable in many other business areas, such as online shopping sites and video detecting system. Thus, the segmented time interval process can be easily adapted for the 'currency' in RFM model, while the real currency consuming is the commercial secret, and substitute, like cellphone flow in this situation.

The corresponding relationships approach can be applied to other areas, as long as there is a quantitative variable describing the customer that is of particular interest to the business. For instance, in the shopping website, this could be the category of purchases in such time range; in the video detecting system, it could be the facial expressions. The generality of the approach indicates that it could be interesting to develop a tool to support this kind of analysis, independently of the domain of application. In this study, various methods were applied and required to associate with each other when the data is ready for the next procedures.

8. Related work

In the telecom service scenario, researchers experiment on a large–scale data platform [23], which is beneficial in processing multiple qualitative and quantitative characteristics; thus, related work in this paper can be roughly divided into two categories. One is the general RFM model for quantitative characteristics, with the guidance of time series analysis. The other is utilizing MCA to prune the clustering results of the RFM model, specifically, discovering detailed relationships among quantitative and qualitative characteristics.

8.1. Related work on CRM for quantitative characteristics in internality

In exploring CRM in internality, the potential relationships among users can be examined. Relationships in valuing numerical characteristics, which reflect similar user performance, have been analyzed in the RFM model [5, 6, 7, 8]. The RFM model is applied well for quantitative characteristics because it not only reflects the normal value in the monetary dimension but also considers the time–related characteristics, of recency and frequency. For the RFM model, the values of the three dimensions in RFM are computed for each cluster and the customers that belonging there.

Research on CRM has evolved to its belonging detection because the data has become increasingly complex and individuals can be reflected by clusters [6, 24, 10]; thus, we consider the use of clustering methods. With the growth of large–scale data changing over time, the mature research on the RFM model normally considers time [5]. As recency, frequency and monetary coexist in the RFM model, researchers cluster the three parameters to provide the unique criteria in time intervals.

8.2. Related work on CRM for qualitative characteristics in internality

For the qualitative characteristics in internality, researchers normally cluster the characteristics by clustering methods [10, 21], without considering the corresponding relationships of characteristics in data. Sometimes, the hierarchy of characteristics can optimize the clustering method [9], and association rule with manual classification has explored multiple telecom service characteristics [10], even in a large–scale data set [21]. Besides, researchers discover unobserved heterogeneity with relationships for numerical and qualitative characteristics simultaneously [25].

However, the aforementioned explorative results cannot provide an intact hierarchy of quantitive and qualitative characteristics in corresponding relationships. As MCA provides a graph presentation of regularized results [26], we can gain competitive advantage as the multiple characteristics are analyzed one by one with less bias.

9. Conclusion

It is common to confront with practical input without demographical background, normally generating simple calculation and resulting in data waste. For such anonymous data, we concluded that the RFM model on Spark is a viable methodology for exploring relationships among users with the analysis of time series, while MCA is ready to prune results from the RFM model and build interaction relationships among multiple characteristics.

We propose a statistic-based approach to value latent users via time series segmenting time interval of RFM in a large–scale data set. Using time series analysis, We explored user relationships on coherent time, and we utilized the Spark platform to target users and discover quantitative relationships in the RFM model. Adjusted with k-means method, the clustering results on the three dimensions of the RFM model performed better than clustering on each interval. We leveraged MCA to correspond with multiple qualitative characteristics with quantitate results after the RFM model.

Currently, we are working on a prototype to demonstrate the changing tendency of user behavior. In the future, we will extend the time range of telecom service data to one week, to obtain a more comprehensive analysis results. Furthermore, we can conduct a prediction model and recommender model to improve our CRM analysis.

10. Acknowledgement

This work is supported by the National Key Project of Scientific and Technical Supporting Programs of China (Grant No. 2014BAH26F02), the Special Fund for Beijing Common Construction Project.

References

References

[1] M. Baker, M. Saren, Marketing Theory: A Student Text, SAGE Publications, 2016.
URL `https://books.google.com.sg/books?id=qhkFDAAAQBAJ`

[2] C. Chen, X. Zheng, Y. Wang, F. Hong, D. Chen, Capturing semantic correlation for item recommendation in tagging systems, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[3] F.-X. Hong, X.-L. Zheng, C.-C. Chen, Latent space regularization for recommender systems, Information Sciences 360 (2016) 202–216.

[4] N. Lu, H. Lin, J. Lu, G. Zhang, A customer churn prediction model in telecom industry using boosting, IEEE Transactions on Industrial Informatics 2 (10) (2014) 1659–1665.

[5] M. Hosseini, M. Shabani, New approach to customer segmentation based on changes in customer value, Journal of Marketing Analytics 3 (3) (2015) 110–121.

[6] R.-S. Wu, P.-H. Chou, Customer segmentation of multiple category data in e-commerce using a soft-clustering approach, Electronic Commerce Research and Applications 10 (3) (2011) 331–341.

[7] F. Hamka, H. Bouwman, M. De Reuver, M. Kroesen, Mobile customer segmentation based on smartphone measurement, Telematics and Informatics 31 (2) (2014) 220–227.

[8] L. Aburto, R. Weber, Improved supply chain management based on hybrid demand forecasts, Applied Soft Computing 7 (1) (2007) 136–144.

[9] F.-M. Hsu, L.-P. Lu, C.-M. Lin, Segmenting customers by transaction data with concept hierarchy, Expert Systems with Applications 39 (6) (2012) 6221–6228.

[10] S. Y. Sohn, Y. Kim, Searching customer patterns of mobile service using clustering and quantitative association rule, Expert systems with Applications 34 (2) (2008) 1070–1077.

[11] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, M. Franklin, S. Shenker, I. Stoica, Fast and interactive analytics over hadoop data with spark.

[12] A. G. Shoro, T. R. Soomro, Big data analysis: Apache spark perspective, Global Journal of Computer Science and Technology 15 (1).

[13] R. Xin, P. Deyhim, A. Ghodsi, X. Meng, M. Zaharia, Graysort on apache spark by databricks, GraySort Competition.

[14] X. Meng, J. Bradley, B. Yuvaz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al., Mllib: Machine learning in apache spark, JMLR 17 (34) (2016) 1–7.

[15] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, et al., Spark sql: Relational data processing in spark, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 1383–1394.

[16] G. D. Ruxton, The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test, Behavioral Ecology 17 (4) (2006) 688–690.

[17] J. D. Cryer, N. Kellet, Time series analysis, Vol. 286, Springer, 1986.

[18] H. Ling, D. W. Jacobs, Shape classification using the inner-distance, IEEE transactions on pattern analysis and machine intelligence 29 (2) (2007) 286–299.

[19] P. Wang, G. Robins, P. Pattison, E. Lazega, Exponential random graph models for multilevel networks, Social Networks 35 (1) (2013) 96–115.

[20] H. Abdi, D. Valentin, Multiple correspondence analysis, Encyclopedia of measurement and statistics (2007) 651–657.

[21] P. Q. Brito, C. Soares, S. Almeida, A. Monte, M. Byvoet, Customer segmentation in a large database of an online customized fashion business, Robotics and Computer-Integrated Manufacturing 36 (2015) 93–100.

[22] F. Kooti, K. Lerman, L. M. Aiello, M. Grbovic, N. Djuric, V. Radosavljevic, Portrait of an online shopper: Understanding and predicting consumer behavior (2016) 205–214.

[23] J. Tan, The design and implementation of big data platform for telecom operators, in: International Conference on Industrial IoT Technologies and Applications, Springer, 2016, pp. 3–11.

[24] B. Xia, B. B. Amor, H. Drira, M. Daoudi, L. Ballihi, Combining face averageness and symmetry for 3d-based gender classification, Pattern Recognition 48 (3) (2015) 746–758.

[25] G. Reikard, Predicting solar radiation at high resolutions: A comparison of time series forecasts, Solar Energy 83 (3) (2009) 342–349.

605 [26] M. Zhu, V. Kuskova, S. Wasserman, N. Contractor, Correspondence analysis of multirelational multilevel networks, in: Multilevel Network Analysis for the Social Sciences, Springer, 2016, pp. 145–172.