Detecting the Changing Points of Multiple-Regression Model on the Basis of the Relations between Audiences' Rating and the Matching between Needs and Contents

KATO Junichi [1)], NINOMIYA Shoji [2)]

1) Tsukuba International University, 2) Osaka University of Economics

【Abstract】The aim of this paper is to show you the procedure about the following four points. First, we explore audiences' needs by using an exploratory method. Second, we build multiple-regression model to explain audiences' ratings by the degrees of matching between audiences' needs and contents of drama. Third, we measure the degrees of matching by an experimental method. Fourth, we detect the points of structure change of this model. We propose the following procedure about the above four points. About the first point, we clarify audiences' needs by using the procedure of blog text mining (KIP) shown in Kato and Ishikawa (2011). Second, we refer to Rust and Alpert model and build the logistic multiple-regression model to explain audience ratings by the degrees of matching. Third, we use experimental methods that subjects answer questionnaire about drama after watching its DVD. Fourth, we detect the changing points of regression model by using Stepwise Chow Test proposed by Ninomiya (1977). We set forth the research direction to build multiple-regression model that explains audience ratings by the degrees of matching between contents and needs, and to detect the structure changing points. These are contributions of this research.

【Keywords】Marketing, Creation of Markets, Audience Ratings

1. Introduction

Marketers make much of the matching between providers' services and customers' needs. Markets are not made from only services of service providers or only customers' needs. Marketers think the matching between these both sides creates markets.

In this research, we propose the procedure to detect the points of creation of markets on the basis of matching between services of service providers and customers' needs. We need to gather knowledge in multiple disciplines to solve a problem like this. We call this a knowledge co-creation in this paper. From this point of view, the problem of this research is an object of knowledge co-creation.

Hereafter, we apply our research interests to the television audiences' ratings and summarize our prospective contribution of this research. An audience's rating is used as an important index for understanding that the TV program is supported by latent TV viewers or not.

This judgment varies depending on the contents of the TV program. At the same time, audiences' ratings vary depending on the audiences' needs from viewpoint of the time series. As the result, when we evaluate the changing of the audience ratings, we need to take the following two points into considerations. The first point is the varying of the contents of TV program (providers' side) and the other point is the varying of the needs of audiences (customers' side).

Especially, in the case of a long term TV program like a drama, we think audiences' needs change with time. If we fail to understand the change of audiences' needs from the viewpoint of times series, we confuse contents of drama with needs of audiences. We need to build an evaluation model which includes producers (contents of TV drama) and audiences (needs) sides for using audiences' ratings as appropriate indexes

We propose the following four procedures in this research. First, we clarify customers' needs from massive blog texts. Second, we build a model that includes producers and audiences sides. Third, we collect real data that are related to matching between contents of TV program and audiences' needs by using experimental methods. Fourth, we detect the changing points of the model by using Stepwise Chow Test. The main content of this paper is to propose the procedure like this.

2. Outline of the procedure

In this research we propose to clarify customers' needs from blog texts data by using KIP (KIP: Kato &

Ishikawa Procedure) which shown in Kato & Ishikawa (2011). Blog texts are written by customers. This means blog texts are raw voices of customers. We explain the audiences' ratings by the matching between customer needs, which we extract from customers' raw voices, and contents of a drama. We build a multiple regression model like this.

In this procedure, the multiple explanatory variables in a multiple regression model grasp the changing of audience ratings. Additionally, we contain audience ratings and the matching (of contents of TV drama and audiences' needs) into one regression model and unify them. Then if we apply Stepwise Chow Test to a multiple regression model, we can detect the points of the structure changing of a multiple regression model. The explanatory variables of its model are the matching between contents of TV drama and needs of audiences and the explained variable of its model is an audience's rating.

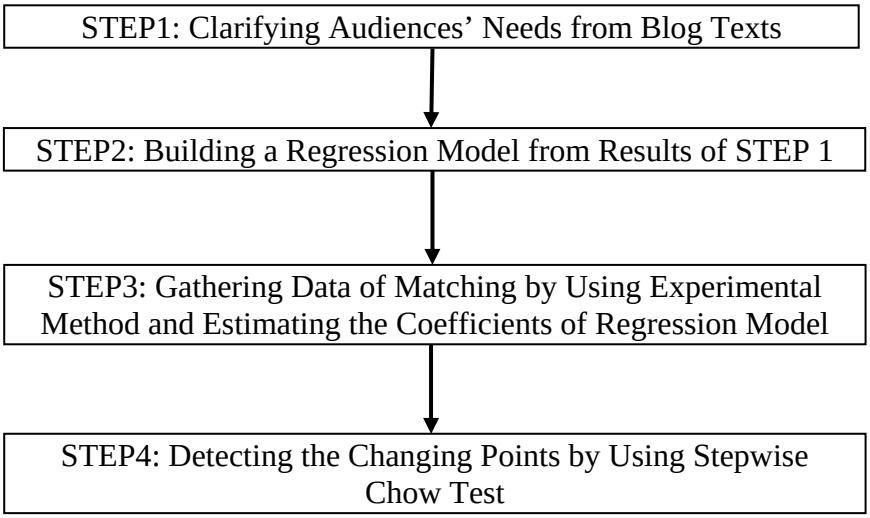

Fig 1: Outline of the Procedure

Figure 1 shows the procedure of this research. STEP 1 of figure 1 is to clarify audiences' needs of all time periods of TV drama by using KIP shown in Kato and Ishikawa (2011). STEP 2 of figure 1 is to build a multiple regression model. Explanatory variables of this model are the degrees of matching between needs we clarified and contents of TV drama. An explained variable of that model is an audience's rating. The existing marketing researches proposed some models for forecasting the audiences' ratings. We build our model on the basis of the existing models. STEP 3 in figure 1 is to gather data of degrees of matching between needs of audiences and contents of TV drama by using an experimental method. We estimate coefficients of a multiple regression model by using these data. Lastly, STEP 4 in figure 1 is to detect the structure changing points by using Stepwise Chow Test.

3. Procedure

3.1 Outline of KIP (Kato and Ishikawa's Procedure)

In this section, we focus on STEP 1 in figure 1 and explain the procedure to clarify customers' needs form blog texts as the following 6 steps. The outline of the following 6 steps is shown in figure 2 in the next page. We show you each step in figure 2.

STEP1 in figure 2 is a data collection. We decide one word which expresses the market we would like to analyze. This keyword is named as target keyword. We gather all authors who used target keyword in their blogs not less than one time. Then we retrieve all blogs of all authors we gathered. Texts of these blogs are divided into words (mainly noun) and the frequency of these words is used as fundamental information in the following analysis.

STEP 2 and STEP 3 in figure 2 are keywords selection. First, we calculate tf・idf values by using the frequency as data. The tf・idf values are important indexes which are often used in text mining. A tf means a term frequency and an idf means an inverted document frequency. We calculate a tf・idf value to each word. Through these calculations, we grasp the relations between words and blog texts by tf・idf values.

We choose words as criteria from words we collected in STEP 1 for clustering of blog authors. The standard of selection of these words is the degree of similarity with the target keyword beyond threshold level. We call these words product keywords. We calculate degrees of similarity by using a cosine measure. Raw data of this calculation are tf・idf values.

Additionally, we choose a second criterion for clustering of blog authors. We select words from all words which are gathered in the STEP 1 in figure 2. These selected words characterize blog authors beyond threshold level. The degrees of characterization are calculated from tf・idf values. These values show us the relations between words and blog authors. We call these words personal keywords. These product and personal keywords are two criteria for clustering of authors.

STEP 4 and STEP 5 in figure 2 are steps of author clustering. We categorize authors by using product keywords and personal keywords. We do not need to execute both STEP 4 and STEP 5 in figure 2.

From viewpoint of marketers, we cluster authors by degrees of customers' loyalty to the product we would like to analyze, because we can easily interpret a practical implication from the result. So in many cases, we execute STEP 5 in figure 2.

We categorize authors by using product keywords. Next, we select loyal authors and longtail authors from the clustered authors. We calculate the relative percentages of frequency of product keywords. On the basis of this result, we divide all authors into high level loyalty authors (loyal authors) or low level loyalty authors (longtail authors). Finally, we segment the selected authors by personal keywords.

In STEP 6 in figure 2, we put labels on loyal authors and longtail authors and clarify the customers' needs. This is procedure for clarifying customers' needs, Kato and Ishikawa (2011) proposed, by using blog texts as data. However, labeling is not easy task. So in Kato (2012) and Kato et al. (2013), we put labels on principal axes by using principal components analysis (PCA). STEP 6 in figure 2 is a little complicated and we explain this step in next section in detail.

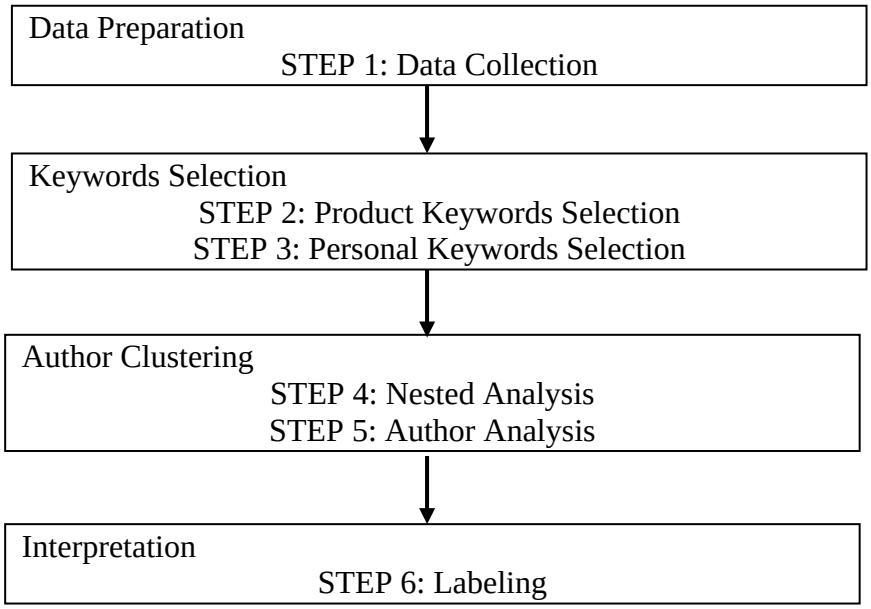

Fig 2: Outline of KIP

3.2 PCA (Principal Component Analysis) and labeling

In this research, we employ the following procedure. First, we make the loyal authors and words matrix and the longtail authors and words matrix. Elements of these matrixes are frequency of words. Second, we calculate chi-squared values and its p-values.

Third, we sort all words by using chi-squared values and p-values which we calculated. These values correspond to each word respectively. We sort all words in ascending order on the basis of p-values. And next, we sort all words in descending order on the basis of chi-squared values. The first criterion is a p-value and the second criterion is a chi-squared value. These words are sorted by degrees of characterizing loyal authors or longtail authors. Finally, we decrease all words to 0.1 percentages of all words by using chi-squared values and

p-values. By this reduction, we can get words that are more characteristic than other words for blog authors.

Though we decrease words to 0.1 percentages of all words, there are still a lot of words. We execute principal component analysis for these words, calculate principal axes for four groups of authors, and put labels on these axes. We use chi-squared values which correspond to each authors' cluster as data for principal component analysis.

First step of principal component analysis is to extract principal axes. We extract axes which explain markets that consist of four authors' clusters. The number of principal axes is decided by calculating of cumulative coefficients of determination. We employ number of principal axes until 80 percentages of cumulative coefficients of determination.

Second step of principal component analysis is a choice of words which are grounds to put labels on principal axes. We choose these words on the basis of eigenvectors. An eigenvector is calculated to each word by principal component analysis. These eigenvectors show size of coefficients to each word.

So we can choose appropriate words to put labels on principal axes by paying attention to the large eigenvectors' words. We employ the top 50 words to put labels on principal axes, because it is difficult to put labels on principal axes from very large number of words. As above, we put labels on loyal and longtail authors separately.

3.3 Regression model and experimental data

In this section, we focus on STEP 2 and STEP 3 in figure 1 and we explain a model building and experimental methods. We build our regression model on the basis of Rust and Alpert Model. Rust and Alpert (1984) is an article that shows the model explaining audience ratings to be called Rust and Alpert Model later. This model expresses audience ratings by a relative utility to be provided from a specific program. This utility is explained by two factors. They are (1) socio-demographic characteristics and a type of TV program and (2) flow states.

Shachar and Emerson (2000) revised Rust and Alpert Model from the following three points. The utility of audience was revised by (1) matching between performers and audiences' socio-demographic characteristics, (2) matching between performers and (un)observable characters, and (3) switching costs.

In this research, we change the next points of these previous articles. First, we change socio-demographic characteristics to customers' needs. In market segmentation, when we cannot get data about customers' needs, we use socio-demographic characteristics. In this research, we clarified customers' needs before a model building. Therefore, we use customers' needs instead of socio-demographic characteristics.

Second, we do not include variables in conjunction with the race in consideration of the fact of our country. Third, Rust and Alpert (1984) and Shachar and Emerson (2000) are micro models on the basis of a personal utility. But we build macro model by using audiences' ratings. Fourth, we change variables of performers of TV drama to variables of characters of TV drama. As the result, the audiences' ratings about the specific TV program are expressed as the next multiple logistic regression model.

$$\ln\left(\frac{R}{1-R}\right) = a_0 + a_1 C + a_2 P + a_3 S + \epsilon \cdots\cdots\cdots \quad \text{Eq. (1)}$$

Within equation (1), R is an audience rating, C is a needs' matching with characters, P is a needs' matching with contents of a program, and S is a switching cost.

We can estimate coefficients of equation (1) by real data. We can get real data by an experimental method and can get audience ratings from the web site of Video Research Inc. An experimental method is as follows. Subjects watch all episodes of DVD of specific drama. After watching each episode, each subject answers a questionnaire that is related to audiences' needs which are clarified by KIP.

These answers show us the degrees of matching between contents of drama and needs of audiences. As the above, we can get data which show us the degrees of matching. By using data of audience ratings and data of the degree of matching, we estimate coefficients of a regression model and execute Stepwise Chow Test.

3.4. Stepwise Chow Test

In this section, we focus on STEP 4 in figure 1. When the equality of the coefficient of a regression model is rejected within all sample period of the given data, it is called the structural change of the model like this with change of coefficients. Any existence of structural change means that the performance of a model has deteriorated statistically. Generally the officially approval test method about the structural change of a model is based on the method of Chow (1960). This test is applied for "the specific time point of being given a priori." When the structural change is found statistically, it will be considered as the division point of dividing the given data into

two homogeneous data groups. It can be considered that this division point is “the turning point of structural change” in the field of economy, management and society.

Stepwise Chow Test is a method without using a priori information to estimate both “the time point” and “the number” of structural change simultaneously. The idea of this test is as follows. It is supposed that any time point has the possibility of structural change because it is unknown at which time point the structural change happened for all sample period of the given data. Then, under the assumption that the structural change has happened at every time point for the sample period, F of Chow statistic value is calculated at the every division point, after total sample term (N) is divided into two parts; that is, period I and period II, where $N=n_1 + n_2$, n_1 and n_2 are divided sample number of period I and period II respectively. Chow test is continuously executed respectively. The detail procedure is as follows.

(1) Firstly, total sample data is separated into two parts, period I of $n_1=1$ and period II of $n_2=N-1$, to calculate F value of Chow. Secondly, next F value is calculated with period I of $n_1=2$ and period II of $n_2=N-2$. Thirdly, F value with period I of $n_1=3$ and period II of $n_2=N-3$ and so on. Finally, the last F value is calculated with period I of $n_1= N-1$ and period II of $n_2=1$. In this way, the each calculation of Chow’s F is repeated continuously while moving the time point of division. After all, F values of N-1 pieces will be obtained

(2) On condition that no F value exceeds the significant level, it is concluded that no structural change has happen. Then the data is judged to be homogeneous. However, there is a significant F, it is judged for the division point to be a turning point of structural change and for the data to be composed of two heterogeneous data group. And it progresses to the procedure of (3). Moreover, when some F values become significant continuously, named as "transition period" of the structural changes which frequently happens, maximum value of F among them is selected to be a turning point of structural change. The reason is that F value is larger, so that the evidence which rejects a hypothesis is stronger. And it advances to the procedure of (3). In Stepwise Chow Test, the above-mentioned procedure (1) and (2) is called “the first step”.

(3) The point estimated to be a turning point through “the first step” is considered to be a division point of the data, and then the data is separated by dividing into two at that time. And, above-mentioned procedure of (1) and (2) are repeated for each divided data respectively. This procedure is called “the second step”.

(4) If the turning point is found through the second step, it is separated again into two parts as a different data, and procedure (1) and (2) are repeated. This procedure is called “the third step”.

(5) Such procedure is repeated until a significant F value is not calculated. Each procedure of the repetition is named one by one as “the fourth step”, “the fifth step” and so on.

(6) The estimated turning points of structural changes are mostly fixed in the stage where the procedure (5) ended. It is possible as a result that plural number of maximum significant Fs, namely of the turning points is found throughout the all steps. In such case, all the data groups which are combined on basis of the every estimated turning points are taken up, and procedure (1) and (2) are repeated.

In Stepwise Chow Test, the procedures from (1) to (5) are called “the main step”, and a procedure (6) is called “the sub-step”. The estimation by “The main step” has experientially effective most as it is. “The sub-step” may reinforce the “the main step” result and may play the role which sometimes corrects the “the main step” result. By completing the above procedures, both “the time of a turning point” and “the number of turning points” of the structural change can be estimated simultaneously.

4. Conclusion

Based on the notion that we grasp creation of markets by matching between providers’ services and customers’ needs, we propose the concrete procedures. We can summarize these procedures as the following four points.

First, we clarify audiences’ needs by using the procedure of blog text mining (KIP) shown in Kato and Ishikawa (2011). Second, we refer to Rust and Alpert model and build the multiple logistic regression model to explain audiences’ ratings by the degrees of matching.

Third, we use experimental methods that subjects answer a questionnaire about a drama after watching its DVD. Fourth, we detect the changing points of a regression model by using Stepwise Chow Test proposed by Ninomiya (1977).

We set forth the research direction to build multiple-regression model that explains audiences’ ratings by the degrees of matching between contents and needs, and to detect the structure changing points. These are contributions of this research.

Notes

(1) We applied the procedure we show you in this paper to an empirical research. However, the result of this empirical research was not satisfactory. So we do not explain the results of an empirical research in detail in the body of this paper. Hereafter, we summarize the results in this note. We take “RYOUMADEN” as the concrete example for our empirical research. First, we clarify customers’

needs about TV drama RYOUMADEN by using blog texts as data. We retrieved the following data. Our data are 687 blog authors, 698, 307 blog entries, and 1, 258, 055 words. The time periods we retrieved data are from March 9, 2004 to October, 18, 2011. We employ 20,536 words as product keywords and 20, 536 as personal keywords. We clustered blog authors by using these product and personal keywords as criteria. Therefore, loyal authors are 60 and longtail authors are 36. We narrow down words which characterize these authors. As the result, 475 words characterize loyal authors and 104 words characterize longtail authors. We use these words for principal component analysis. From the result of PCA, two axes (fashion and outlier) grasp loyal authors and two axes (Kansai region and policy and economy).explain longtail authors. We interpret customers' needs by using these two axes. We use the result of these interpretations for making a questionnaire. We gather data from 20 subjects who attended this experiment about the matching between contents of drama and needs of customers. These subjects watch all episodes of DVD of RYOUMADEN. After watching each episode of DVD, each subject answers the questionnaire we prepare on the basis of the result of KIP. Questions in the questionnaire are needs' matching about character of the drama, needs' matching about contents of the drama, and a switching cost. The data of audiences' ratings are got from websites of Video research Inc. These data are average audiences' ratings of Kanto region. We used the above data for analyzing and executing Stepwise Chow Test. I program for executing Stepwise Chow Test by using the statistical environment R (https://sites.google.com/site/junichikatopapers/home/programs). Anyone can access and use this program only for academics. We detect the structure changes between 25 and 26 episodes from the results of Stepwise Chow Test. However, the null hypothesis "coefficients of multiple regression model = 0" was not rejected. Therefore, values of correlations between explanatory variables are high and this result of the point of structure change is not reliable from statistical viewpoint. We do not explain the result of empirical research in the body of this paper. This research is supported by the individual research fund in 2012 of Osaka University of Economics.

Reference

Back number: Weekly high television ratings program 10 4 Drama [Kanto Region] (http://www.videor.co.jp/data/ratedata/backnum/2010/index.htm) [2012, February 13]

Chow, G. C. (1960), "Test of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, Vol.28, No.3, pp.591-605.

Kato, Junichi. (2012), "Exploring Keywords to Create Tourism Markets by Using Blog Text Mining", *Collected Papers for Presentation In the 49th Annual Meeting of the Japan Section of the RSAI (Annual Meeting of The Japan section of the Regional Science Association International)*, 6 pages.

Kato, Junichi., Mamoru Imanishi & Saburo Saito. (2013), "Exploring Customers' Needs for Kyushu Shinkansen By Using Blog Text Data", *International Winter Conference on Business and Economics Research*, CD-ROM pp.1-18.

Kato, Junichi., & Masahiro Ishikawa. (2011), "Semi-Automatic Procedure for Market Segmentation by Using Massive Weblog Data", *The 2011 Spring National Conference of Operations Research Society of Japan*, pp. 104-105 頁。

Ninomiya, Shoji. (1977), "Stepwise Chow Test", *The Economic Studies Quarterly*, Vol.28, No.1, pp.50-60.

Programs, Junichi KATO PAPERS (https://sites.google.com/site/junichikatopapers/home/programs) [2012, November, 6]

Rust, Ronald, and Mark L. Alpert (1984), "An Audience Flow Model of Television Viewing Choice Model", *Marketing Science*, Vol.3, No.2, pp.113-124.

Shachar, Ron and John W. Emerson (2000), "Cast Demographics, Unobserved Segments, and Heterogeneous Switching Costs in a Television Viewing Choice Model," *Journal of Marketing Research*, Vol. 37, No. 2, pp. 173-186

Contact information
Address: 6-20-1, Manabe, Tsuchiura, Ibaraki, Japan
Name: Junichi KATO
E-mail：junichikato01@gmail.com