Customers' Needs for Digital Terrestrial Television Broadcasting: an Analysis of Weblog Data

Junichi Kato
Department of Media and Communication Studies, Tsukuba International University, Ibaraki, Japan 108-0023
(E-mail: junichikato01@gmail.com)

Abstract The diffusion of CHIDEJI is a national level event. To clarify customers' needs for CHIDEJI is clue for cultivating national size markets. So the aim of this paper is as follows. First, we need to look into customers' needs by exploratory methods. Second, consumers forget their needs about CHIDEJI and distort them. So I have to use text written at that time as data. I clarify whether some words are attractive for other customers or not. By using text mining, I answered these questions. To first and second questions, we can improve its' attractiveness by preparing the programs about IT, drama, and disasters. Finally, it improves the attractiveness of CHIDEJI to lend itself to TV programs on the political and social issues.
Key words CHIDEJI; Weblog Data; Text mining; Marketing

1 Introduction

The diffusion of CHIDEJI, which is Japanese abbreviation of terrestrial digital TV broadcasting receiver, is a national level event. Figure 1 shows progress of diffusion of CHIDEJI[1].

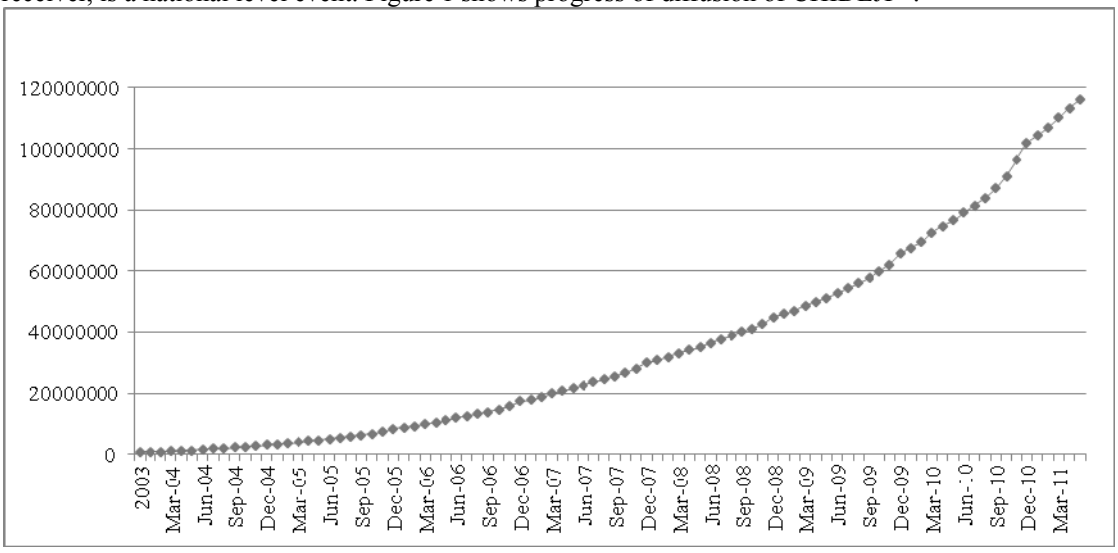

Figure 1 The Domestic Shipment of Terrestrial Digital TVBroadcast Receiver

What kind of needs do Japanese customers have in this huge phenomenon? To clarify this kind of needs is a help for cultivating a national size markets. The aim of this paper is the following three. First, almost all researches about CHIDEJI are limited to technological aspects. There are few researches to discuss customers' needs. I could find only reports of private companies, for example goo search and japan.internet.com[2]. Researchers try to clarify customers' needs by using questionnaires in these reports. However, the method of questionnaires requires to restrict and to fix items of questionnaires in advance. At first step of researches, we need to look into customers' needs by exploratory methods.

Second, there is a possibility that consumers, who adopt their TV to terrestrial digital broadcasting at an early stage, forget their needs about CHIDEJI and distort them. So I have to use text written at that time as data. From the above two points, I execute data mining by using weblog as data.

Finally, according to Rogers (2003), the process of diffusion includes critical mass [3].
After going beyond that point, diffusion is accelerated. CHIDEJI was diffused at national level. We can image that some customers who are very interested in CHIDEJI attract and explain about the great aspects of CHIDEJI to other customers. As the result of this type of communication, the diffusion rate

gets closer to critical mass. But we cannot detect whether this type of communication is before or after critical mass, because the aim of this paper is not time series analysis. I clarify whether some words are attractive for other customers or not, by using weblog analysis. I discuss the three above research questions in the following data-mining.

2 Procedure of Web Text Mining

This research employs the weblog contents' mining procedure which Kato and Ishikawa (2011) proposed [4]. This procedure is constituted by five steps. First step is to retrieve data from weblog texts. I decide the target keyword which expresses markets we want to analyze. I gather all weblog articles of goo blog which its authors, who have used the target keyword not less than one time, have written.

Second step is to choose similar words with target keyword as product keywords. Third step is to select words which are characterized of weblog authors as author keywords.

Fourth step is to categorize authors by two keywords which are selected in step 2 and 3. In this step, if authors frequently use product keywords in their weblog articles, then I assume that they are strongly interested in the markets. I call them loyal authors. On the other hand, if authors less use product keywords in their weblog articles, then I call them long tail authors. I take this difference into account and categorize authors.

Lastly, fifth step is a labeling. It is for easy interpretation that I put labels on the categories. Those labels express specifications of loyal and long tail authors. I use those labels for understanding customers' needs. In the following sections, I clarify customers' needs about CHIDEJI by the above steps.

3 Analysis

3.1 Results

In the first step, I retrieved data as follows. Target keyword was CHIDEJI. I got 1000 weblog articles by using goo blog search. I gathered all weblog articles which those authors wrote. The number of all authors was 525. When I could not get weblog articles over 99% of all articles, I eliminated those authors from the following analysis. As the result, the number of authors was 485, the number of articles was 652886, and the number of words was 430065.

Time periods of weblog data was from 2001/4/29/ to 2011/6/30. I extracted text data from all weblog articles and analyzed those text data by morphological analysis. Then I used words, which I got by this morphological analysis, as data in the following procedure.

In the second step, I selected product keywords. Product keywords were defined as similar words with target keyword CHIDEJI. I analyzed every weblog articles as one unit, not authors as one unit. I employed words, which were similar with CHIDEJI over threshold level, as product keywords. This similarity was measured by cosine measure. I decided that threshold level of similarity was 0.00342455145158, and employed 10000 words as product keywords. Cumulative relative similarity score of these 10000 words occupied almost 80 %.

In the third step, I chose personal keywords from words. First I gathered weblog articles by each author. Next, I employed words, which one author frequently used but others did not use very much, as personal keyword. Personal keywords were defined as characterizing authors over threshold level. This threshold level was 452.612757602 and the number of personal keywords was 10000.

The fourth step was nested analysis. This nested analysis meant the categorization of each category. First I gathered weblog articles by each weblog author, and then categorized authors by using product keywords in step 2. Through this categorization, I clustered authors who used similar words with product keywords. The method of clustering was SOM (Self Organization Map). I got four groups by this clustering.

Next I calculated the ratio of index of significance of product keywords by each author and then calculated mean value of upper 25 % by each group. I assumed that the group of maximum value of this average was a group which had high level loyalty for this product. So I defined authors of this group as loyal authors. The number of loyal authors was 58. On the other hand, the group of minimum average score was long tail authors. Those long tail authors came from long tail of Anderson (2008) [5]. The number of long tail authors was 298.

I categorized authors of every group, who were categorized by product keywords, by personal keywords. This categorization created four groups of 58 loyal authors and of 298 long tail authors. I categorized authors and segmented markets so far. Hereafter, I interpreted these clusters and clarified

customers' needs in detail. However, I just finished categorization of authors and could not interpret these clusters. So I labeled four groups for easy interpretation.

Fifth step was a labeling. Four groups of loyal authors included 31020 words (10 authors), 125848 words (27 authors), 22122 words (8 authors), and 40463 words (13 authors) respectively. Four groups of long tail authors included 98167 words (69 authors), 145713 words (116 authors), 111254 words (73 authors), and 69622 words (40 authors) respectively. I combined authors by each group and calculated p-value of chi-squared for each word [6]. Groups of loyal authors have 153555 words. Groups of long tail authors have 256798 words. I selected 1 % of those words in loyal and long tail groups on the basis of p-value of chi-squared. I used 1536 words for labeling of loyal authors and 2568 words for labeling of long tail authors.

3.2 Discussion

From 1536 and 2568 words, I labeled loyal and long tail authors. First, loyal authors had three labels. First of all, I gathered and labeled TV related words, because these words were directly related to CHIDEJI. First label was related to TV. This label included words which were CATV, DVD, ANIME, CHANNERU (channel), TEREBI (television), TEREBIKYOKU (TV broadcasting company), NUHSU (news), GASITU (image quality), BANGUMI (TV program), WOWOW, AQUOS (Japanese TV brand name), MINPOU (private broadcasting companies), and etc.

From these words, consumers, who were interested in CHIDEJI, were not only interested in hardware of TV (for example, AQUOS, GASITU), but also interested in software of TV (for example, ANIME, BANGUMI, WOWOW). So it was not only important to develop hardware, but also important to develop the charming software.

Second label was related to PC, home electronics, and information technology. This label included words which were Amazon, Panasonic, Vista, Windows, iPhone, iPod, yahoo, CPU, UIRUSUBASUTAH (brand name of anti-virus software), ANDOROIDO (android), ADSL, KIHBOHDO (keyboard), Monster TV (brand name of digital TV tuner for PC) and etc.

This label included home electronic. So there is no doubt that this label included home electronics' makers (for example, Panasonic), but I could find the PC and IT of words (for example, Monster TV) that were not directly related to CHIDEJI, too. These words were related to tools to watch TV program of CHIDEJI from PC. From these words, I understood that authors were interested in watching TV not only from TV but also from PC.

Final label was related to political and social issues. This label had words which were GYOUSEI (public administration), KOUSAI (high court), GOUHOU (legal), KUNI (country), KOKKAI (diet), SANIN (House of Councillors), SHYUSHO (prime minister), SHYOUSHIKA (declining birth rate), SHINBUN (newspaper), NENKIN (pension), and etc.

From these political and social related words, authors were not only interested in ANIME, but also interested in TV program of political and social issues.

On the other hand, long tail authors had three labels. First label was related to TV. This label included words which were yomiuri, TORANSFOHMA (Transformer), NIMO (nemo), HARUHI, BYONHON (Korean actor, Byung Hun), KAMENRAIDAH (Kaman rider), SAIGOU, SENSYUURAKU (final day of a sumo tournament), YATAROU, TAKECHI, SATSUGUN (troops of Satsuma), SATSUMA, TAKASUGI, YOUDOU, CHOUSYU and etc. Especially, long tail authors used words related to RYOUMADEN. RYOUMADEN was TV drama that succeeded last year. This drama dealt with the Meiji restoration. From these words, long tail authors were not only interested in political and social issues, but also interested in TV drama of historical events.

Second label was related to PC, home electronics, and information technology. This label included words which were CANON, OLYMPUS, goo, AKIBA, BROGU RANKINGU (blog ranking), YAHOO, and etc.

Final label was related to political and social issues. This label had words which were KANBOU (the secretariat), KANRYOU (bureaucrat), GIIN (Diet member), GIKAI (assembly), KYOUSANTOU (Communist Party), KEIKI (market), KEIZAI (economy), GYOUSEI (public administration), JIEITAI (self-defense forces), DAISHINSAI (great earthquake), CHIJI (governor), AMAKUDARI, TOUKYOUDENRYOKU (Tokyo electric company), FUTENMA, HOUSHASEN (radiationradial rays), HOUSHANOU (radioactivity), JISHIN (earthquake), TOYOTA JIDOUSHA (Toyota motor company), SESHIUMU (cesium), GENSHIRYOKU (nuclear energy), GENPATSU (nuclear power plant), etc. When I compared these long tail authors' words with loyal authors' words, long tail authors used words which were related to the 2011 Tohoku earthquake and tsunami.

Because long tail authors used a various words, I could not express them by a few words. Long tail

authors, who recently warmed to CHIDEJI, used words related to the 2011 Tohoku earthquake and tsunami. According to words of Rogers, they would be categorized into laggards rather than innovators.

4 Conclusions

This paper had the following two aims. The first was (1) to analyze customers' needs by an exploratory method and (2) to analyze customers' raw voice of those days. From words of loyal and long tail authors, I understood that we could improve CHIDEJI's attractiveness by preparing the TV programs related to IT, drama (for example, RYOMADEN), and disasters (the 2011 Tohoku earthquake and tsunami).

The second object was to find out clues to make the new market takeoff by analyzing weblog of authors who were interested in CHIDEJI. According to words of loyal authors, it improved the attractiveness of CHIDEJI to lend itself to TV programs focused on the political and social issues. As a result, not only the hardware (TV itself) but also the software (TV program) was important for making the new market takeoff.

Though I clarified customers' needs from contents of massive weblog in this research, I reckoned that if I used other data source along with weblog data I could propose more useful marketing insights which could not be shown by only weblog data in future research.

References

[1] http://www.jeita.or.jp/japanese/stat/digital/2011/index.htm
[2] http://research.goo.ne.jp/database/data/000067/
[3] Rogers EM. Diffusion of Innovation, 5th Edition: Free Press, 2003
[4] Kato J. & Ishikawa M. Semi-Automatic Procedure for Market Segmentation by Using Massive Weblog Data[A]. The 2011Spring National Conference of Operations Research Society of Japan, 2011: 104-105 (In Japanese)
[5] Anderson C. The Long Tail Revised and Updated Edition[M]. Hyperion, 2008
[6] Jin M. Statistical Analysis for Text Data[M]. Iwanami Shoten, 2009 (In Japanese)