1K2-3

Comparison of clustering methods for large scale TV viewing data

Yaling Tao[*1] Yoshiaki Mizuoka[*1] Kouta Nakata[*1] Ryohei Orihara[*1]

[*1] Knowledge Media Laboratory, Corporate Research and Development Center, Toshiba Corporation

This paper presents and compares two television-viewing pattern extracting methods based on clustering techniques. A set of viewing patterns is extracted from over 50,000 devices' log data and visualized in a novel way. Television viewing habits such as the duration each channel is viewed, the channel-changing behaviors can be seen visually. Furthermore, two metrics reflecting the number of extracted viewing patterns and channel-changing patterns are defined to evaluate the pattern extracting methods. Our experimental results show that hierarchical agglomerative clustering (HAC) outperforms K-means in extracting an interesting viewing pattern.

1. Introduction

Television industry is experiencing a revolution with the rapid development of multimedia and network communication technologies. More abundant television contents and richer consumption experiences are provided to users. For example, they can time-shift their viewing, watch their favorite shows repeatedly, or switch to the Internet at any time. The viewing habits of users are becoming more diverse, and thus it is more difficult to capture users' viewing patterns. Television advertisers and service providers face challenges in attracting users in the incredibly competitive market. Efficient delivery of advertising and contents plays a key role in cutting cost and promoting competitiveness for them. In order to accomplish that, how to extract the target consumer group from the large community of users becomes an issue to be solved in television industry.

This work aims to discover television viewing patterns from large scale log data collected from televisions. The log data record information that when and which channel users have viewed. In this work, we focus on daily habits of viewing television. Users select programs to watch according to their preferences, but in some cases they habitually watch programs such as certain morning news. Such daily habits may be difficult to find since they are not based on program's popularity or strong preferences but rather on unintentional customs. It is desirable to find such patterns because, for example, commercials that require viewers' full attention are unlikely to be effective during the unintentional viewing.

We introduce data mining approaches to discover viewing patterns of the time, the day of the week and channels. A more detailed description of log data is presented in [Mizuoka 2017]. we compare two television-viewing pattern extracting methods based on clustering techniques. A set of viewing patterns is extracted from over 50,000 devices' raw log data and visualized in a novel way. Television viewing habits such as the duration each channel is viewed, the channel-changing behaviors can be seen visually. Furthermore, two metrics reflecting the number of extracted viewing patterns and channel-changing patterns are defined to evaluate the pattern extracting methods.

The rest of this paper is organized as follows. Section 2 gives previous works related to television viewing data. Section 3 describes the television viewing data and Section 4 presents several clustering algorithms and techniques used in this work. Section 5 illustrates the experimental results. Finally, Section 6 concludes this work with a summary and gives the future work.

Contact: Yaling Tao, R & D Center, Toshiba, 1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan yaling1.tao@toshiba.co.jp

2. Related work

Although a number of works have studied on television data analysis and recommendation systems, there has been a few works for extracting viewing patterns by utilizing clustering techniques, for example, [Spangler 2003], [Chang 2012], [Kurapati 2002], [Chaney 2013], [Xu 2012], etc.

Most of those works derive viewing patterns by using demographic information of users, such as age, gender, and hobby. As we know, however, the user information is unreliable and very little can be available due to privacy limitations. [Xu 2012] deals with this problem by representing user preferences based on the categories of programs they watched. As stated in Section 1, we focus on discovering habitual viewing patterns which may contain intentional and unintentional viewing. Predicting users' preferences is out of the scope in this work.

Furthermore, to the best of our knowledge, no previous work uses clustering techniques other than K-means. The effectiveness of clustering techniques for television data also has not been discussed in any other work.

3. Dataset

The dataset used in this work are logs collected from televisions in Tokyo, from October 1 to 31, 2016. Each row of dataset represents each television as a 2689-dimensinal feature vector that captures the viewing time information of 8 terrestrial channels in Japan. The viewing time is counted in seconds in every 30-minute interval and represented as the mean value by day. The structure of each vector is shown in Figure 1.

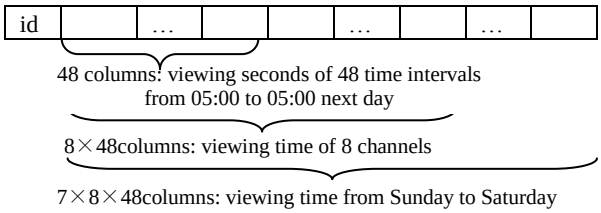

Figure 1 Data structure

Although the log data counted in seconds include rich and detailed information like the way of viewing, their high dimensionality brings more challenges for data analysis.

Considering the viewing time is the key information of log data, we plot the distribution of the number of televisions over the total viewing time in a histogram shown in Figure 2. From this Figure, we see that over 10% televisions are used less than 10 hours during this month. These data can be eliminated as noises.

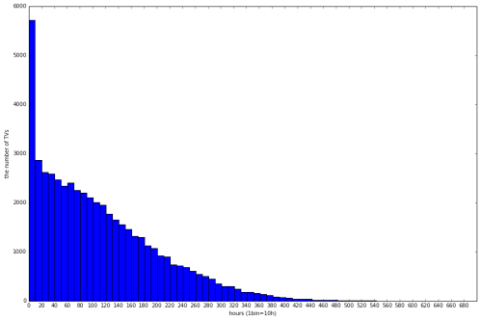

Figure 2 Distribution of the number of televisions over the total viewing time

4. Clustering techniques

4.1 Two methods

We use two of the most studied clustering methods, K-means and Hierarchical agglomerative clustering (HAC) in this work. K-means initially selects K random points as the centroids of the clusters. Then it reassigns all points to their closest centroids and recalculates the new centroid of each newly assembled cluster. This process is repeated until the centroids do not change.

HAC starts with singleton clusters and recursively merges two the closest clusters to one parent cluster until the K clusters have been built.

In our work, the Euclidean distance metric is used to measure the similarity of two clusters. Both HAC and K-means receive the target number of clusters K as a parameter and minimize the within-cluster sum of squares. Only the K-means proceeds with a partitioning way while the HAC with an agglomerative way.

4.2 Cluster visualization

Considering the goal of our work is to extract television viewing patterns, we design a method to identify a pattern from clustering results. In order to do that, we first present each cluster in a novel way by which the feature of the cluster can be read visually.

Figure 3 shows an example of clusters. The visualized cluster illustrates information including the cluster ID, the number of televisions gathering in this cluster, and the popularity of channels. The figure consists of 7×8×48 grid units. The horizontal axis shows 8 channels and 7 days a week while the vertical axis shows 48 time slots of a day from 05:00 to 05:00 (next day). Each grid unit is colored by a value between 0 and 1. This value reflects the proportion of viewing time in a time interval of 30 minutes. All clusters generated by clustering methods can be presented in this way for further analysis.

From this example of visualized cluster, we can read that about 100 users often watch channel 2 and 4 on weekdays. Channel- changing happens at 9:00 and 18:00. It can be possible that the behavior reflects users' preference to the program in channel 4 at 9:00 and one in channel 2 at 18:00. However, this explanation is unlikely because the both timings are the middle of the show in channel 4. Probably, the users chose channel 4 at 9:00 as a harmless background video after their favorite shows in channel 2 ended.

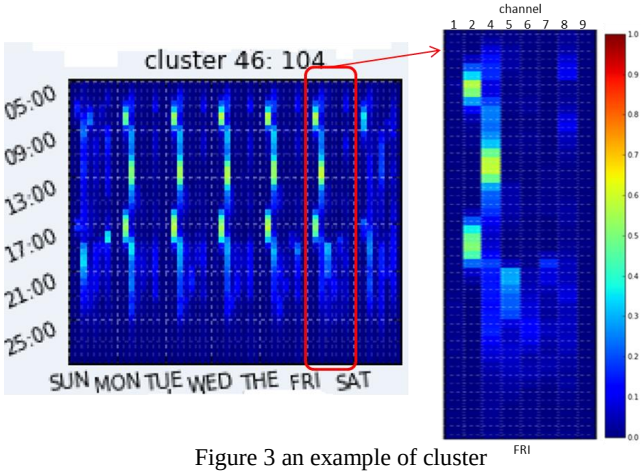

Figure 3 an example of cluster

4.3 Pattern identification

From those visualized clusters, we intuitively think a colorful cluster is a viewing pattern. This way can be quantified and so a cluster is formally identified as a *viewing pattern* if only it satisfies:

1) The number of televisions gathering in the cluster reaches an acceptable threshold, denoted by *th_size*.
2) There exist one or more grid units colored by setting values greater than a threshold, denoted by *th_time*.

In addition, we can also read the channel-changing information from some visualized clusters (e.g. Figure 3). A viewing pattern can be identified as a *channel-changing pattern* if only it has obvious channel-changing, i.e., there exist two grid units in columns of a day satisfy:

1) Color setting values of both units are greater than *th_time*.
2) These two units are adjacent in vertical direction but not in horizontal direction.

The channel-changing patterns are often difficult to find and exactly what we want because they could be examples of the unintentional viewing described in Section 1.

4.4 Evaluation of clustering

Evaluation indexes such as Dunn's Index [Dunn 1973], DB's Index [Davies 1979], have been often used in a wide range of applications. In this work, on the basis of our cluster visualization, we define two intuitive metrics, named *Extracting Index* and *Changing Index*, to evaluate a clustering method. The extracting Index is defined as the ratio of identified viewing patterns in the all clusters. For example, when we obtain 100 clusters by using a clustering method, if 70 of them pass the pattern deification described in Section 4.3, we say the extracting Index of this clustering method is 0.7. The closer this Index value is to 1, the clustering method is effective.

Similarly, we define the *Changing Index* as the ratio of channel-changing patterns.

5. Experimental results

In this section, experiments are executed by using K-means and HAC. The results are compared and evaluated with the Extracting Index and Changing Index defined in Section 4. On the basis of the data distribution shown in Figure 2, we select data that the total viewing time is not less than 10 hours as our target dataset.

We run K-means and HAC to generate 100 clusters and compare their results. As the results of K-means strongly depend on the initial centroids, K-means is run with 10 different centroids seeds and the best output is used as the final result. The maximum number of iterations of a single run is set to 300. Thresholds *th_size* and *th_time* are set to 100 and 0.3, respectively. This means we want to find out clusters in which not less than 100 users gathered and the ratio of viewing time in some timeslots is not less than 0.3.

The results are listed in Table 1. From the results we see that HAC extracted less viewing patterns but more channel-changing patterns than K-means. As HAC calculates distances between any two clusters and merges two the closest clusters to a new cluster in each step, the closer two clusters are, the earlier two clusters merge to a new one. This way yields the smallest within-cluster distance in each step, this is exactly consistent with our objective that minimizes the within-cluster sum of squares. On the other hand, K-means randomly selects K points as the centroids at the first step and then recursively reassigns each point to an appropriate cluster. The reassignment process aims at minimizing the within-cluster sum of squares but not considering all points globally. This leads to K-means overlooks some unusual patterns. For example, if K similar points which may generate an unusual cluster (e.g. channel-changing pattern) are unfortunately divided in the K initial clusters, it is difficult to gather again in the following reassignment steps.

Table 1 Results of two clustering methods

	Extracting Index	Changing Index	Time (h:mm:ss)
K-means	0.78	0.09	0:38:19
HAC	0.72	0.14	9:51:21

Furthermore, it is not obvious to specify an appropriate K for the K-means to carry out a given task. For example, if we want to find more channel-changing patterns we may need to set K to be greater than 100. We have found the threshold by try and error in our experiments. That means we have to run the K-means many times to select an appropriate K. This also is a time-consuming task. The execution time shown in Table 1 is the result without considering the time on determining the K.

On the other hand, after running the HAC once, we can investigate a lot of clustering candidates by examining the dendrogram. The dendrogram is independent of the number of clusters. The HAC is preferable for an exploratory data analysis such as channel-changing patterns extraction. Furthermore, some parallel algorithms can be considered to alleviate the HACs' time-consuming nature. This will be investigated in our future work.

From the results, we also find out that channel-changing patterns take a small portion (less than 20%) of the all clusters, for both K-means and HAC. One reason is that there are only really a few channel-changing patterns due to the diversity of users' viewing habits. As mentioned in Section 4.3, channel-changing patterns are exactly we want to discover. HAC has shown its potential for the channel-changing patterns extraction. It is worth effort in this direction.

6. Conclusions

In this work, we present and compare two television-viewing pattern extracting methods. Television viewing habits such as the duration each channel is viewed, the channel-changing behaviors can be seen visually. Furthermore, two metrics reflecting the number of extracted viewing patterns and channel-channel changing patterns are defined to evaluate the pattern extracting methods. Our experimental results show that HAC outperforms K-means in extracting channel-changing patterns while K-means can obtain more viewing patterns than HAC.

Some issues such as the time complexity of HAC, the appropriate number of clusters, and the effectiveness of HAC for more complex logs will be investigated in our future work.

References

[Spangler 2003] W. E. Spangler, M. Gal-Or, J. H. May: Using Data Mining to profile TV viewers, Communications of the ACM, vol. 46, No. 12, 2003.

[Chang 2012] R. M. Chang, R. J. Kauffman, I. Son: Consumer Micro-Behavior and TV Viewership Patterns: Data Analytics for the Two-Way Set-Top Box, Inter. Conf. on Electronic Commerce'12, August 6-8, Singapore, 2012.

[Kurapati 2002] K. Kurapati, S. Gutta: Instant Personalization via Clustering TV Viewing Patterns, IASTED's ASC, 2002.

[Chaney 2013] A. J.B. Chaney, M. Gartrell, J. M. Hofman: Mining Large-scale TV Group Viewing Patterns for Group Recommendation, Microsoft Research, Technical Report, MSR-TR-2013-114, 2013.

[Xu 2012] M. Xu, S. Berkovsky, I. Koprinska: Time Dependency in TV Viewer Clustering, UMAP Workshops, vol. 872, 2012.

[Dunn 1973] J. C. Dunn: A Fuzzy Relative of the ISODATA Process and Its Use in detecting compact Well-Separated Clusters. Journal of Cybemetics. Vol. 3, No. 33, pp. 32-57, 1973.

[Davies 1979] D. L. Davies and D. W. Bouldin: Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, No. 2, pp. 95-104, 1979.

[Mizuoka 2017] Mizuoka et al.,: Analysis of Television Viewing Pattern with Clustering Large-Scale Log Data, JSAI2017.