International Journal of Hybrid Information Technology
Vol. 9, No.11 (2016), pp. 79-90
http://dx.doi.org/10.14257/ijhit.2016.9.11.08

Analysis of Audience Interest and User Clustering Based on Program Tags

Fulian Yin[1], Xingyi Pan[1], Jianping Chai[1] and Wenwen Zhang[2]

[1]*Faculty of science and technology, Communication University of China, Beijing 100024, China*
[2]*Beijing Aerospace System Engineering Institute, Beijing 100076, China*
penny_pxy@hotmail.com

Abstract

This paper proposes an analysis method of user behavior to provide personalized program recommendation based on program tags in the field of broadcasting and television. Multidimensional Scaling Analysis is used to produce a quantitative description of viewing preferences. Hierarchical clustering is performed to determine the number of clusters, followed by K-means clustering to group the data according to audience interest in TV program tags. This divides the audience into groups with similar viewing preferences.

Keywords*: television program tag; Audience Interest of TV program tags; clustering algorithm; viewing preference*

1. Introduction

With the rapid increase of television programs as well as the growing popularity of digital televisions, audience is overloaded with new television programs coming out every day. A lot of Chinese television program classification systems are proposed, however, all of them focus differently, and fail to be very comprehensive [1]. Furthermore, each system adopts a single standard, which makes it overly rigid and unable to capture the details in television programs. There has been some research done on multiple attributes of television programs [2-3]. In [4], a Delayering Tagging System of Television Programs is proposed, which describes the television program contents in a multidimensional way. On the other hand, with growing popularity of digital televisions and the growing diversify in content of television programs; analysis of audience preference is becoming more and more important to each part of the television industry. For viewers, analysis of their preferences helps providing more programs that probably interest them. To the program producers and distributers, identifying different viewing preference groups helps direct program production, procurement and advertising delivery. For television and broadcasting regulators, analysis of audience preferences allows to grasp the current trend of television and broadcasting.

Audience preferences are determined by mathmatically analyzing user behavior data. The result is a quantitive description of the degree of user preference for a particular subject. User behavior data includes explicit data and implicit data. The explicit data is suitable for the analysis of audience preference. The explicit data is used in calculation of the similarity between users. For instance, user rankings for programs and videos [5-7] and video-on-demand data as well as video tags [8] are used to calculate user similarities in television program recommendation systems. Both explicit data and implicit data such as viewing duration, together with a TV ontology containing other sematic attributes of TV programs, are used to calculate

ISSN: 1738-9968 IJHIT

Degree Of Interest (DOI) which quantifies user preferences [9-10]. Implicit data, including user behavior and viewing records, can also be analyzed to obtain explicit user preferences. User behaviors are classified into three categories according to different purposes and utilized to compute user preferences [11]. User viewing records in DVB are stored as user profiles, and analyzed by The Ant Colony System to divide users into different interest groups [12]. For sports video analysis, three models including TV scenario, Internet scenario and attributes, are used to build a user preference model. A method of user preference analysis on viewing records is proposed in the TV scenario model [13]. A method of computing user preferences based on context features and a measure of influence of context value are proposed [14]. Due to large number of programs, it is not feasible to use user-program binary data for user preference analysis. In one research, rating data are seperated into two portions, and used in a bias model to capture users‘affinity to items [15]. In another research, program features are taken into consideration instead of programs to capture user interest, to decrease the dimensions number [16]. Based on existing technology, this paper proposes a method of calculating Audience Interest, quantifying audience preferences in television programs, and performs user clustering of Audience Interest using K-means cluster algorithm. It resolves the problem of excessive calculation consumption caused by fulsome programs when using user-programmed data. The result shows this method can cature audience prefernces and recognizes crowds with different viewing preferences.

2. Audience Interest based on Program Tags

2.1. Delayering Tagging System of Television Programs

Most traditional classification systems are treelike structured.[4] These classification systems overly focus on program category, while paying little attention to capture single program’s multiple features. As a result of the rapid increase of television programs, classification systems are getting more complicated and classification criteria becoming more specific. To solve this problem, Delayering Tagging System of Television Programs is propused in [4]. This system describes program contents in multiple dimensions. By using Delayering Tagging System of Television Programs in this paper, audience preference are captured and analyzed in a novel way. The structure of this system is shown in Figure 1.

2.2. Audience Interest of Program Tags

A computing method Audience Interest of TV Program Tags (AIT) is proposed in this paper, quantifying user preferences of program tags and describing the distribution of user tastes. AIT, short of Audience Interest of TV Program Tags, is computed from viewing records. Features and types of TV programs are used

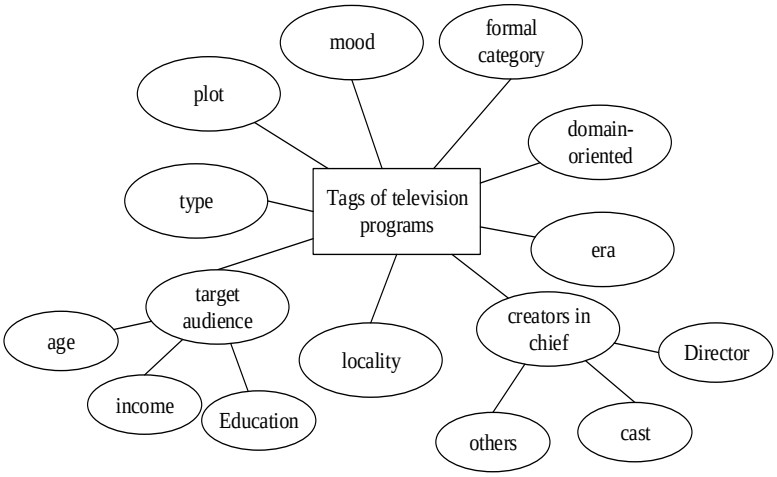


Figure 1. Delayering Tagging System of Television Programs

instead of programs, so AIT is transformed from the relationships between users and programs to the one between users and program tags. The relationship between users and program tags can capture user preferences generally. Audience Interest qualifies user preferences of program tags and can spot audience crowds with different tastes, providing solid foundation for researches of recognizing and clustering of audience viewing tastes. Furthermore, AIT also solves the problem of dimension superabundance of user-program interest calculation in aforementioned researches of audience tastes.

2.2.1. Viewing Indicators

Viewing indicators are used to reflect the viewing effect and the spread of TV programs. The ratings are the most common used viewing indicators. The ratings represent the percentage of audience (or households) who watch a certain TV program during a certain period of time account for the overall audience (or households) Liu [16]. Due to different measurements engaged by the domestic institutions during the rating investigations, different sample households and different computing methods, the results of the ratings are different. In this paper, the ratings (Rtg%) are computed from the viewing duration, and the formula is shown in (1).

$$Rtg\% = \left(\sum_{i=1}^{n} time_i \times w_i\right) / \left(T \times N\right) \times 100\% \quad (1)$$

In (1), $time_i$ and w_i denote the ith viewing duration and the weight respectively. T denotes the total duration of this period, and N denotes the total number of viewers.

2.2.2. Audience Interest of TV Programs Tags

This paper first proposes the concept of Audience Interest of TV Programs Tags, and quantifies user preferences of different TV program tags. This method solves the problem of dimension superabundance of user-program data. The concept of Audience Interest of TV Programs Tags is formed based on the concept of the ratings. Specific components are described as follows. They are User Program Weight (UPW), frequency-considered User Program Weight (FUPW), frequency factor and Audience Interest of TV program tags (AIT).

UPW (User Program Weight) is a quantified value of user preference of a single TV program. In UPW, a single user‘s viewing duration of a single TV program is calculated to measure the viewing proportion of a certain program for the certain user.The formula of UPW is shown in (2).

$$UPW_{ij} = \sum_{j \in P_i} \left(t_{ij} / Time_j \right) \quad (2)$$

In (2), j denotes the TV program j and t_{ij} denotes the ith user's viewing duration of TV program j. $Time_j$ denotes the total duration of TV program j. And P_i denotes all TV programs viewed by the ith user during a certain period of time. Every value of *UPW* represents a user's viewing proportion of a TV program. To avoid inclining TV programs of long duration, the ratio of viewing duration and total duration of the viewed program is chosen rather than the viewing duration.

FUPW (Frequency-considered User Program Weight) takes frequency factor into consider based on (2), which means FUPW is weighted UPW by frequency factor. The formula of FUPW is shown in (3).

$$FUPW_{ij} = UPW_{ij} \times freq_{ij} \quad (3)$$

The formula of $freq_{ij}$ is shown in (4).

$$freq_{ij} = d_{ij} / \mathrm{D} \quad (4)$$

If a user watches the same TV program times, it is more convincing that the user preference is stronger than he/she watches a TV program only once. So FUPW is calculated as weighted UPW.

AIT (Audience Interest of TV Program Tags) transforms FUPW to a more general description of user preferences according to the relation between TV program tags and TV programs in Delayering Tagging System of TV programs. The formula of AIT is shown in (5).

$$AIT_{ik} = \sum_{k \in Tag_j} FUPW_{ij} \quad (5)$$

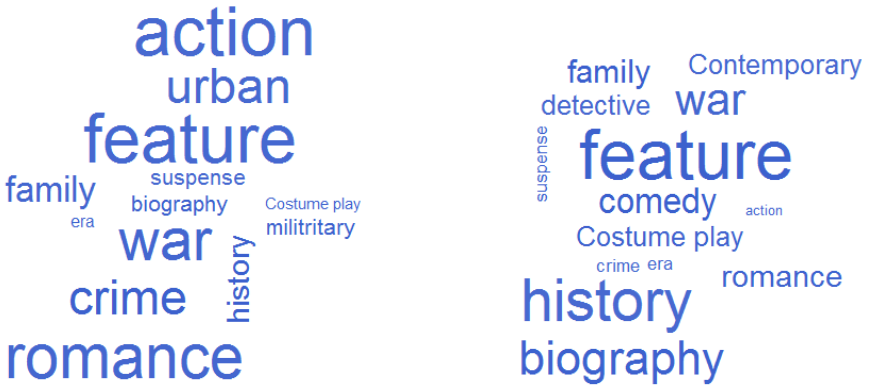
In (5), $FUPW_{ij}$ denotes the ith user's Frequency-considered User Program Weight of program j. Tag_j is a set of TV program tags of program j. AIT is calculated by summing every user's FUPW values of all TV program tags and transforms the user-program relationship to user-programtag relationship.

3. Analysis of Audience Interest

3.1. Personal Tag Cloud of Audience Interest

In this paper, a visualization method of AIT presented in a tag cloud is proposed. Tag clouds are visualized tag set, where tags differs from each other in a visuable way, and these differences are determined by their own weight values. In this paper, the value of *AIT* is selected as the visuable weight. The sizes of tags indicates different value of *AIT*. Figure 2. shows two users' tag clouds of personal Audience Interest of TV Program Tags. From Figure 2. it can be seen that the sizes of *feature*, *romance*, *action* and *war* on the left side are much bigger than others, indicating User A has strong perferences for romance stories and action plays. From the tag cloud on the right side, it can be see that *feature*, *history*, *biography*, *comedy* and *costume* play are in much bigger sizes than others, indicating that User B prefers to history and biography stories, costumed TV series and comedy than others.

3.2. Audience Multidimensional Scaling based on AIT



(a) User A (b) User B

Figure 2. Personal Tag Cloud of Audience Interest of TV Program Tags of Two Users

Multidimensional Scaling (MDS) is a multivariate statistical analysis method. In this paper, MDS is used to visualize the distribution of user preferences. User preferences dimensions are reducted and user preferences are projected into a two-dimensional space. MDS analysis for user viewing preference calculates user distances based on AIT matrix, and vividly displays the distribution in low-dimensional space such as two-dimensional one and three-dimensional one after dimension reduction. In the low-dimensional visualization, every dot represents viewing preference of one user, and the relative position between every two dots indicates the distribution of user viewing perferences, from which the relationships of user viewing perference is inferred. The fundamental of MDS model for AIT is discribed as follows.

The distance between every two users is calculated from AIT matrix. The formula is shown in (6).

$$dist_{ij} = \{ \sum_{k=1}^{n} (\mathrm{AIT}_{ik} - \mathrm{AIT}_{jk})^2 \}^{1/2} \quad (6)$$

In (6), $dist_{ij}$ denotes the distance between ith user and jth user, computed as euclidean distance of n tags. Relative positions of user distances and the whole picture is obtained from MDS analysis.

In Figure 3, every dot represents a user. The horizontal axis as well as the vertical axis is of no specific meanings but two-dimensional axes of user viewing preference. From Figure 3, it is inferred that most users aggregate on the middle left edge in the figure, the other dots distribute radially to the right side in three directions, upper, horizontal and lower respectively, and the distribution becomes gradually sparse along the directions. It is inferred that, most user viewing preferences are quite concentrated, and others distribute in three radial directions. These users have viewing preferences obviously different from most users, and these users have some strong personality viewing preferences. There are several scattered dots away from the gathering center. These correspondent users shows great difference from the masses and are far away from each other. It is inferred that these correspondent users have quite personal and strong viewing preferences.

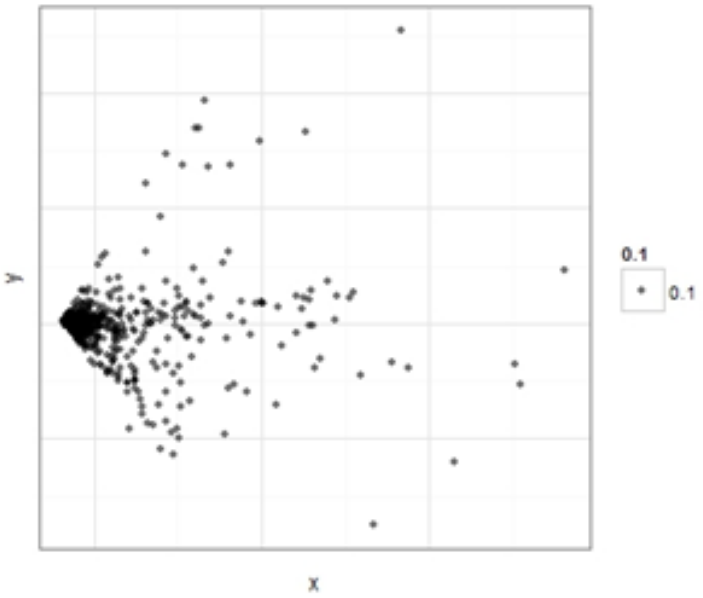


Figure 3. Multidimensional Scaling Analysis of Audience Interest of TV Program Tags

4. Audience Crowd Clustering based on AIT

4.1. Audience Clustering

Audience clustering is to divide audience into several groups of similar users, and the best division is where the similarity of all users in the same group maximizes and the similarity between each crowd minimizes. In broadcasting and television, audience clustering study is still blank. In this paper, hierarchical clustering is performed to determine the number of clusters, followed by K-means clustering algorithm on AIT data and finally group audience with different viewing preferences.

4.2. Hierarchical clustering of Audience Interest of TV Program Tags

Hierarchical clustering is a method to build a hierarchy of clusters from the AIT data set. The hierarchical clustering used in this paper is a bottom up approach, also called the agglomerative type. Firstly, every user is taken as a crowd, and then these crowds are agglomerated according to their distances between each other until all users are agglomerated as one crowd.

There is a hierarchical clustering dendrogram shows the hierarchy of clusters, shown in Figure 4. In Figure 4, every leaf node represents a user, and the vertical axis represents cluster height, that is the value of the criterion associated with the clustering method used in this algorithm. It is inferred that all users can be divided into two branches. The cluster heights of the sub-branches in the left branch are much higher than the right ones and the cluster is much smaller, indicating greater differences within these groups of smaller sizes. The sub-branches in the right branch, conversely, have smaller cluster heights and are in larger sizes, indicating the correspondent crowds are of larger sizes and fewer within diffrences. In this paper, number ‘6‘ is selected as the number of clusters in K-means clustering according to the distribution of AIT data.

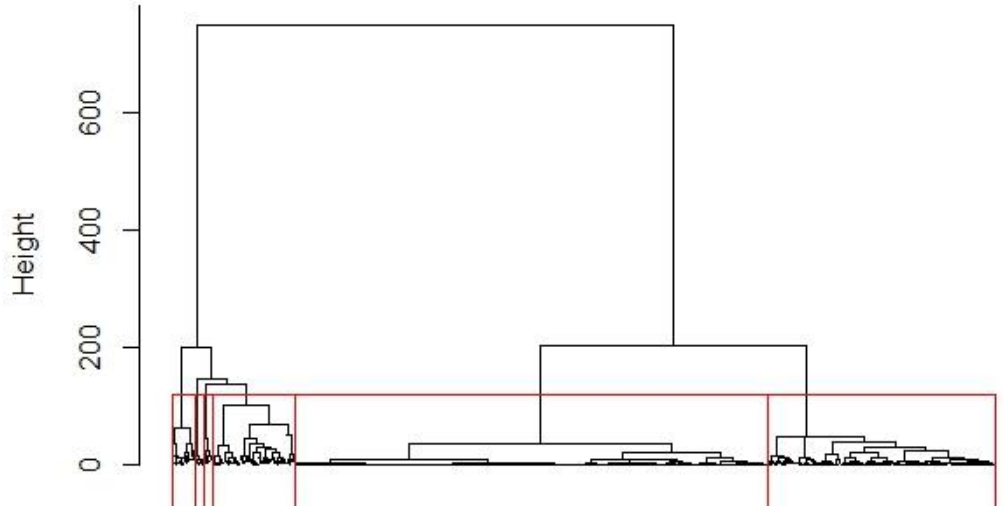


Figure 4. Hierarchical Clustering Dendrogram of Audience Crowd Clustering Based on AIT

4.3. K-Means Clustering of Audience Interest of TV Program Tag

K-means clustering is an iterative clustering algorithm which divides users into specific *k* clusters [18].

A data set of n users is given in K-means clustering algorithm, and *k* groups are obtained after clustering. Every group has specific viewing preferences, and $k \le n$. In other words, the users are divided into *k* groups and the result also meets two standards as follows:

- There is more than one user in each crowd.
- Every user must belong to only one crowd.

The number of clusters is pre-specific. The error value of the objective function is gradually reduced by interactive calculation until the objective function value converges. Finally the clustering result is obtained.

The process of K-means clustering algorithm is shown as follows:

(1) Select k users from AIT data set D randonly as initial cluster centers.
(2) repeat
(3) for every user *i* in data set *D* do
(4) Compute the distance between i and k cluster centers.
(5) Assign user *i* to its nearest cluster (the one of shortest distance)
(6) end for
(7) Calculate the mean value of all users in each cluster and take the mean value as the center of the new cluster.
(8) until the centers of k clusters do not change any more

4.4. Simulation and Performance Analysis

First of all, it is important to choose a proper value of k in K-means clustering. In this paper, the within cluster sum of squares are chosen to estimate how much differences within clusters during k changing from 1 to 20.

Figure 5 shows the within cluster sum of squares from *k*=1 to *k*=20. In Figure 5, the gradually decreasing trend of within cluster sum of squares shows that the larger k is, the smaller the sum of squares are. It is seen that the decrease from *k*=1 to *k*=5 is the

sharpest, and the decrasing trend gradually levels off from k=6 to larger numbers. In this paper, we select 6 as k.

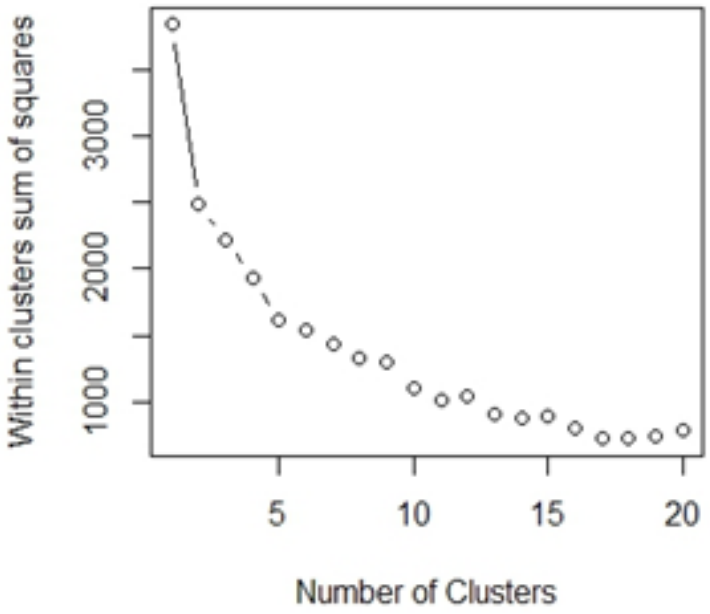


Figure 5. Within Crowd Sum Of Square ($1 \le k \le 20$)

In this paper, 6 is selected as k in K-means clustering algorithm. The result of K-means clustering algorithm for AIT is shown in Figure 6. In Figure 6, every number represents a user in data set D, and the number is the correspondent cluster which the user belongs to. The horizontal and vertical coordinates are coordinate projections as the result of discriminant coordinates calculation.

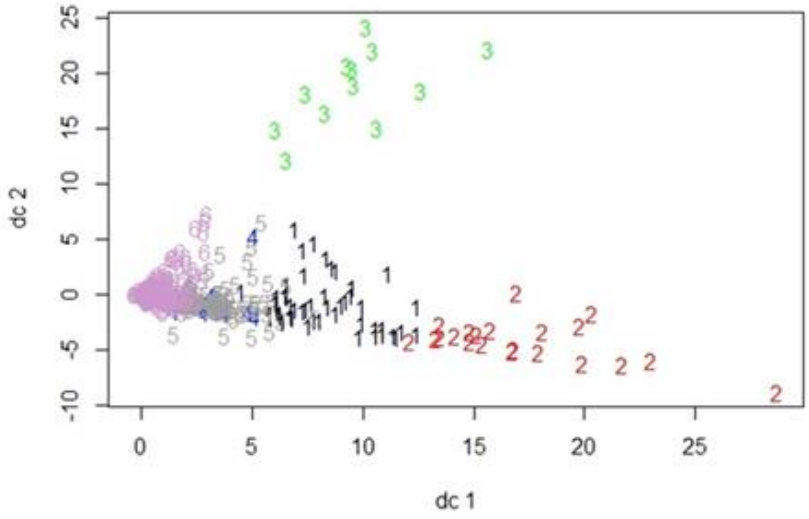


Figure 6. K-Means Clustering of Audience Groups Based on AIT

The cluster result is shown in Table 1. Except one crowd of no obvious preferences, the other 5 crowds show 5 different viewing preferences. The 5 viewing preferences are summarized as follows:

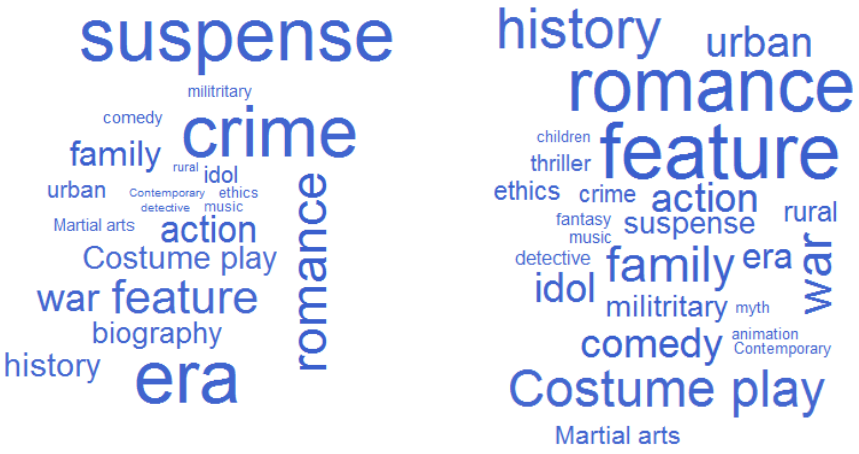
Table 1. Cluster Centers in K-means Clustering of Audience Crowd Based on AIT

k=6，and AIT of center>0.01

Cluster	1	2	3	4	5	6
romance	2.050274	4.385637	4.082295	0.739732	0.70368	0.121073
action	0.255133	0.59926	0.096215	0.336682	0.33662	0.030891
Costume play	0.624346	0.642407	1.030577	0.307315	0.315599	0.051472
family	0.922538	2.582059	0.150492	0.214538	0.21117	0.045505
feature	2.457939	4.906053	0.632882	0.725608	1.098066	0.132222
history	0.441955	0.900611	0.14942	0.37865	0.723145	0.062886
era	0.242969	0.748403	0.0753	2.860816	0.129971	0.021664
idol	0.285334	0.083336	2.86727	0.146829	0.065387	0.029634
comedy	0.742793	0.69478	0.193102	0.037797	0.133863	0.031884
war	0.279541	0.869104	0.103053	0.445419	0.694019	0.049935
biography	0.444729	0.367309	0.142951	0.334307	0.335524	0.03438
urban	0.308006	0.240035	2.849105	0.165599	0.11984	0.031982
country	0.168654	0.246913	0.001408	0.003765	0.040214	0.009321
suspense	0.178878	0.532169	0.103506	2.856234	0.104779	0.013643

TV dramas tagged with *costume play* and *history*, TV dramas tagged with *history* and *war*, TV dramas tagged with *idol* and *urban*, TV dramas tagged with *romance* and *family* and TV dramas tagged with *crime, suspense* and *era*.

Tag clouds are used to visualize the result of K-means clustering by presenting the AIT values of correspondent tags in different sizes. Tag clouds of Crowd 4 and Crowd 6 are shown in Figure 7. The biggest tags of Crowd 4 are *crime, suspense* and *era*, as is shown on the left side. As for Crowd 6, *feature* and *romance* are the biggest tags in tag cloud on the left side. It is inferred that users in Crowd 4 show a strong preferences for TV dramas tagged with *crime, suspense* and *era*, and users in Crowd 6 prefer to TV dramas tagged with *romance, feature, history* and *costume play*.



(a) Crowd 4 (b) Crowd 6

Figure 7. Two Tag Clouds of Clustering Result

5. Conclusion

This paper proposes Audience Interest of TV Program Tags based on TV program tags. Audience Interest of TV Program Tags overcomes the problem of dimension superabundance in user-program interest calculation. It records user viewing preferences from a more general perspective like program types, and analyzes the viewing data multi-dimensionally, providing analytical data for audience analysis and audience positioning research. Visualized distribution of Audience Interest is shown by MDS analysis. Audience with different viewing preferences are grouped by K-means clustering algorithm for Audience Interest. This results also grasp the current trend of television and broadcasting the effect of macro analysis of viewing trend.

Acknowledgements

This work was carried out by X. Pan, F. Yin, J. Chai and W. Zhang. We gratefully acknowledge the kind support from the State Administration of Radio, Film and Television and the Communication University of China, Beijing, China.

References

[1] H. Zhang, "A systematic classification of Chinese TV programs", Commun. Univ. China Press, Beijing, **(2007)**
[2] Y. Liu, Z. Xia, Y. Li and Z. Yang, "Multi-dimensional combination' of TV program Classification and Coding Design", Modern Commun. (J. Commun. Univ. China), vol. 1, no. 22, **(2003)**.
[3] C. Chen and W. Sun, "Media • People • Modernization", Social and Science Press, Beijing, **(1997)**.
[4] X. Pan, F. Yin and J. Chai, "Delayering Tagging of Television Programs and Association Rule Mining", Proceedings of IEEE 17th International Conference on Computational Science and Engineering, Chengdu, China, **(2014)**.
[5] A. Martinez, J. Arias and A. Vilas, "What's on TV tonight? An efficient and effective personalized recommender system of TV programs", IEEE Trans. Consumer Electron., vol. 55, no. 286, **(2009)**.
[6] H.-J. Kwon and K.-S. Hong, "Personalized smart TV program recommender based on collaborative filtering and a novel similarity method", IEEE Trans. Consumer Electron., vol. 57, no. 1416, **(2011)**.
[7] S. Lee, D. Lee and S. Lee, "Personalized DTV program recommendation system under a cloud computing environment", IEEE Trans. Consumer Electron., vol. 56, no. 1034, **(2010)**.
[8] J. Park, S.-J. Lee, S.-J. Lee, K. Kim, B.-S. Chung and Y.-K. Lee, "Online video recommendation through tag-cloud aggregation", IEEE MultiMedia, vol. 18, no. 78, **(2011)**.

[9] Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. López-Nores, J. García-Duque, A. Fernández-Vilas and R. P. Díaz-Redondo, "Exploiting synergies between semantic reasoning and personalization strategies in intelligent recommender systems: A case study", J. Syst. Software, vol. 812371, **(2008)**.
[10] R. Sotelo, Y. Blanco-Fernández, M. López-Nores, A. Gil and J. J. Pazos-Arias, "TV program recommendation for groups based on muldimensional TV-anytime classifications", IEEE Trans. Consumer Electron., vol. 55, no. 248, **(2009)**.
[11] H. Shin, M. Lee and E. Y. Kim, "Personalized digital TV content recommendation with integration of user behavior profiling and multimodal content rating", IEEE Trans. Consumer Electron., vol. 55, no. 1417, **(2009)**.
[12] Y.-C. Chen, H.-C. Huang and Y.-M. Huang, "Community-based program recommendation for the next generation electronic program guide", IEEE Trans. Consumer Electron., vol. 55, no. 707, **(2009)**.
[13] F. Sanchez, M. Alduan, F. Alvarez, J. Menendez and O. Baez, "Recommender system for sport videos based on user audiovisual consumption", IEEE Trans. Multimedia, vol. 14, no. 1546, **(2012)**.
[14] S. Song, H. Moustafa and H. Afifi, "Advanced IPTV services personalization through context-aware content recommendation", IEEE Trans. Multimedia, vol. 14, no. 1528, **(2012)**.
[15] G. Dror, N. Koenigstein and Y. Koren, "Web-scale media recommendation systems", Proc. IEEE, vol. 100, no. 2722, **(2012)**.
[16] M. Krstic and M. Bjelica, "Context-Aware Personalized Program Guide Based on Neural Network", IEEE Trans. Consumer Electron., vol. 58, no. 1301, **(2012)**.
[17] Y. Liu, "Television ratings parsing: investigation, analysis and application", Commun. Univ. China Press, Beijing, **(2000)**.
[18] Algorithm Realization of Improved K-Means ClusteringAlgorithm in Credit Analysis of Policy holders J. Harbin Univ. Sci. and Tech., vol. 14, no. 116, **(2009)**.

