Analysis and Clustering of Movie Genres

Hasan Bulut and Serdar Korukoglu

Abstract— Most of the movies blend a genre with other genres; that is movie directors combine elements from different genres with each other. A movie may blend the love-oriented plot of the romance genre with Western or Science Fiction. Hence a movie may belong to several genres. A movie is also related with some keywords to describe the contents of the movie. These keywords are usually used during search to bring a movie according to user's interest. In this paper, we establish genre keyword sets from movie keywords and use these keyword sets to analyze the proximity of genres with each other. In this study, we use movie data from The Internet Movie Database (IMDB). Genres are classified using hierarchical clustering algorithm and principal component factor analysis (PCFA). The study shows us which genres are mostly used together in a movie. We show that results of the two analyses support each other.

Index Terms— Hierarchical clustering, Internet Movie Database, movie genres, principal component factor analysis

1 Introduction

Movies are part of almost everybody's life to fulfill his/her entertainment needs and hence constitute a large portion of the entertainment industry. Several websites host movie metadata and provide users the facility to search and find movies of his/her interest. The Internet Movie Database (IMDB) is a popular site cataloging almost every movie ever made. It is an excellent online database to find detailed information about movies, TV series and videos. The IMDB contains information such as genre, keywords, year, language, user ratings and many other features related to those movies and videos. However, IMDB contains huge amount of those data that should be analyzed by researchers.

Movies are classified into a number of genres to help users to direct their search to some specific categories. However, most of the movies blend a genre with other genres; that is movie directors combine elements from different genres with each other. A movie may blend the love-oriented plot of the romance genre with Western or Science Fiction. Hence a movie may belong to several genres.

[1] and [2] classify movies into four genres based on the basis of computable visual cues; Comedies, Action, Dramas or Horror films. [3] presents a method of movie genre categorization based on scene classification of movie trailers. Similar to the work presented in [1], they also classify movies into four categories Comedies, Action, Dramas or Horror. [4] characterizes the measurable traits of the musical scores utilized in Comedies, Action, Dramas and Horror movie genres to determine the feature categories carrying the most valuable information distinguishing them in a broad sense.

[5] gives a global overview of the entire movie and actor space. It visualizes all movies as well as major co-actor relationships. [6] presents a case study for the visualization and analysis of large and complex temporal multivariate networks derived from the Internet Movie Database (IMDB).

[7] uses data mining techniques to analyze the factors contributing to the rating of a movie. [8] predicts movie grosses using regression and k-nearest neighbor models on IMDB data and news data. [9] combines recommender systems with information search tools for better search and browsing. They use a collaborative filtering algorithm to generate personal item authorities for each user and combine them with item proximities.

[10] develops models and algorithms for predicting the helpfulness of reviews. The study aims to find the most helpful reviews residing among the large amount of low quality reviews.

In this paper, we present a novel approach that analyzes the proximity of movie genres with each other and discover the most related genre pairs and triples. In order to achieve this, we have established genre keyword sets from movie keywords. Genres may have common keywords. We have classified genres using hierarchical clustering algorithm. Also we have compared it with the principal component factor analysis results among genres. The study shows us which genres are mostly used together. We have shown that results of the two analyses are close to each other.

The remainder of the paper is organized as follows: Section 2 introduces data used from The Internet Movie Database (IMDB). Section 3 explains how genre keyword sets are constructed from movie keyword sets. Section 4, first presents the keyword distributions of genres. Then, hierarchical clustering method is applied on the movie data and the closest genre pairs and triples are discovered. Principal component factor analysis is performed on the same data and two analyses are compared with each other. Finally, in Section 5, we conclude that the results of the two analyses support each other.

- *H. Bulut is with the Computer Engineering Department, Ege University, Bornova, Izmir 35100 Turkey*
- *S. Korukoglu is with the Computer Engineering Department, Ege University, Bornova, Izmir 35100 Turkey*

2 THE INTERNET MOVIE DATABASE (IMDB)

The Internet Movie Database (IMDB) is an excellent online database to find detailed information about movies, TV series and videos. The IMDB contains information such as genre, keywords, year, language, user ratings and many other features related to those movies and videos. Undoubtedly, the vast amount of IMDB data contains much valuable information which needs to be researched.

The IMDB is available in a number of inconsistently structured text files, which is laid out to be human-readable, not machine-readable. The format of the data makes it difficult to directly use the source data for information extraction. Hence, the raw data needs to be preprocessed or transformed into another suitable format.

The IMDB data contains 49 separate text files. The common factor linking the information in these files is the title of the movie. The production year in parenthesis is appended to the title of the movie to account for different versions of the title. Some of the titles may include letters TV, V or VG in parenthesis as well to indicate that the title is a TV series, video or video game respectively, i.e. Sand (2001) and Sand (2001) (V). Also, if there are multiple movies with the same title in the same year, roman numbers are appended to the year, i.e. Sand (2010/I) and Sand (2010/II).

Each file provides information related to a separate feature such as genres are given in genre.list while keywords are given in keywords.list file. The convention used in these files are as follows: genre.list file provides genre information of titles in <title>|<genre> format, keywords.list file provides keyword information of titles in <title>|< keyword > format, etc. on each seperate line. If there are more than one genre or keyword for a title, then the title is repeated for that genre or keyword, i.e. <titleA>|<genre1>, <titleA>|<genre2>, etc.

However, files may contain some text at the beginning and the data may sometime comprise some errors. Also, the spacing between titles and related information is not same and not all values were available for each line. The non-standardized structure of the files requires parsing them in different ways.

For our research purposes we have used the following files: movies.list, genres.list, keywords.list and language.list. As their names indicate, movies.list file contains <title>|< year> pair, genres.list file contains <title>|< genre> pair, keywords.list file contains <title>|< keyword> pair and language.list file contains <title>|< language> pair on every line. After processing these files, titles are linked to their genres, keywords and language.

3 CONSTRUCTING GENRE KEYWORD SETS

Movie data contains a number of keywords to describe to movie. A movie may belong to several genres. Therefore, the movie keywords are included into genre keyword sets that are specified for that movie. For instance, let's consider the following movies with related keyword and genre sets:

Movie m_1
$Keywords_{m1}$= {k_1, k_2, k_3, k_4, k_5}
$Genres_{m1}$= {g_1, g_2, g_3}
Movie m_2
$Keywords_{m2}$= {k_2, k_3, k_7}
$Genres_{m2}$= {g_1, g_2}
Movie m_3
$Keywords_{m3}$= {k_1, k_6, k_8}
$Genres_{m3}$= {g_2, g_3}

For the above example, we establish the genre keyword sets as follows:

Step 1: Combine all keywords from movies which contain the genre in its list.

$Keywords_{g1}$= {k_1, k_2, k_2, k_3, k_3, k_4, k_5, k_7}
$Keywords_{g2}$= {k_1, k_1, k_2, k_2, k_3, k_3, k_4, k_5, k_6, k_7, k_8}
$Keywords_{g3}$= {k_1, k_1, k_2, k_3, k_4, k_5, k_6, k_8}

Step 2: Associate a weight with each keyword in the keyword set. Weight values are obtained by normalizing the total weight of the keyword set to 1. Keyword weight is assigned a value proportion the number of keyword repetitions within the keyword set.

$Keywords_{g1}$ = {<k_1, 0.125>, <k_2, 0.25>, <k_3, 0.25>, <k_4, 0.125>, <k_5, 0.125>, <k_7, 0.125>}

$Keywords_{g2}$ = {<k_1, 0.1818>, <k_2, 0.1818>, <k_3, 0.1818>, <k_4, 0.0909>, <k_5, 0.0909>, <k_6, 0.0909>, <k_7, 0.0909>, <k_8, 0.0909>}

$Keywords_{g3}$ = {<k_1, 0.25>, <k_2, 0.125>, <k_3, 0.125>, <k_4, 0.125>, <k_5, 0.125>, <k_6, 0.125>, <k_8, 0.125>}

Step 3: After combining all movie keywords for all genres a matrix is formed where rows represent keywords and columns represent genres. For the above example the matrix is shown in TABLE 1.

TABLE 1
KEYWORD-GENRE MATRIX FOR THE EXAMPLE

	g_1	g_2	g_3
k_1	0.125	0.1818	0.25
k_2	0.25	0.1818	0.125
k_3	0.25	0.1818	0.125
k_4	0.125	0.0909	0.125
k_5	0.125	0.0909	0.125
k_6	0.0	0.0909	0.125
k_7	0.125	0.0909	0.0
k_8	0.0	0.0909	0.125

4 ANALYSIS OF IMDB DATA

We have chosen movies with English language titles between years 2006 and 2010, a five year period, which makes a total of 48483 titles, with 27 genres and 19561 keywords. The distribution of the number of titles per genre is shown in Fig. 1. In Fig. 1, the genres are sorted in decreasing order of number of movies they have. The first ten genres (Short, Drama, Adult, Comedy, Documentary, Thriller, Horror, Action, Romance and Crime) constitute almost %80 of the movies.

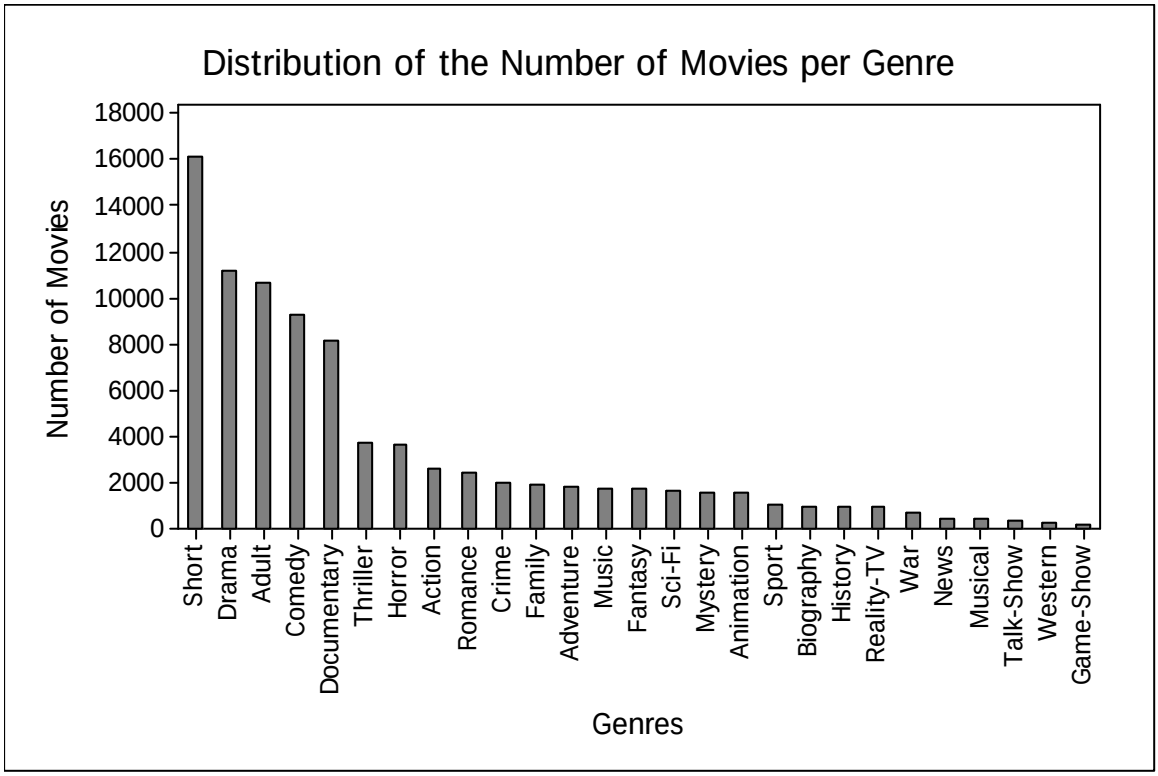

Fig. 1. Distribution of the number of movies per genre

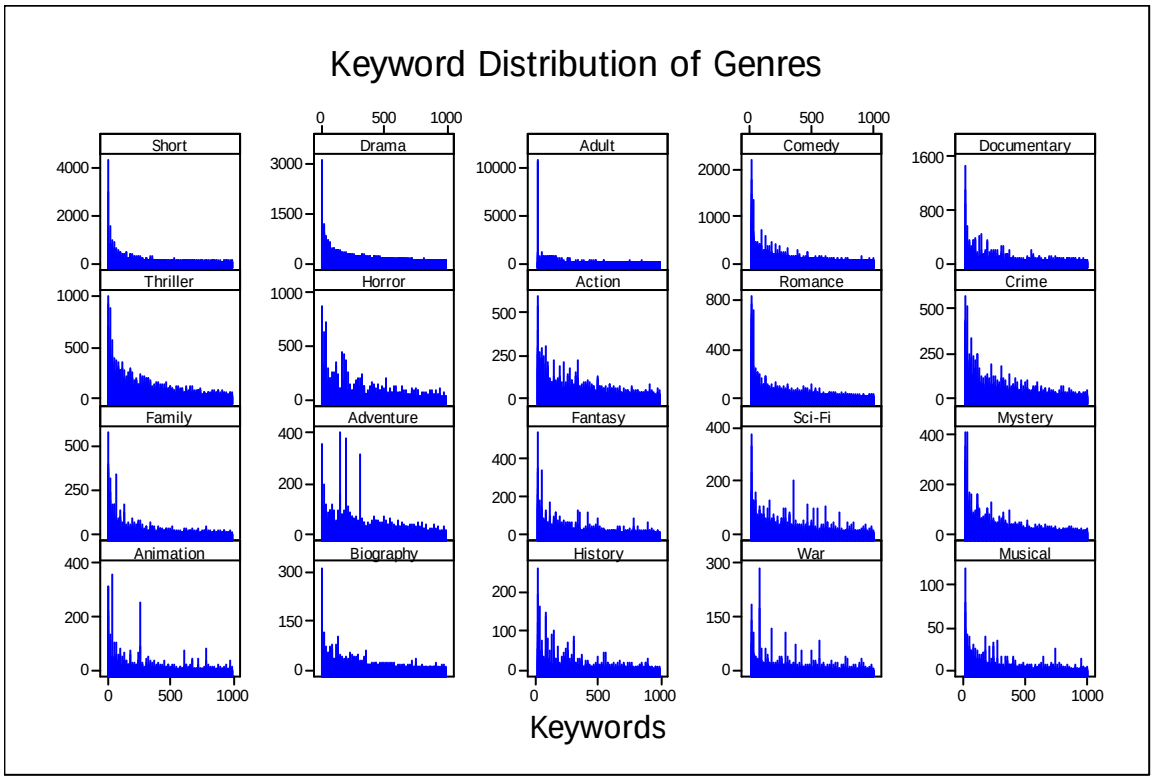

Fig. 2. Keyword distribution of genres

Keywords are sorted in decreasing order and the corresponding values (number of keywords) are plotted for 20 genres. The first 1000-keyword distributions for genres are given in Fig. 2. As shown in section 3, the frequency of a keyword may differ from genre to genre. It is observed that the distributions are positively skewed.

4.1 The Distance and the Pearson Correlation Coefficient Values between Genres

The distance and the Pearson correlation coefficient values between genres are computed, after determining the keyword weight values for genres. The result is given in Table 2. The lower left triangle shows the distance between genres and the upper right triangle shows the Pearson Correlation coefficients. All of the values in Table 2 are significant at least at 1% level ($p<0.01$). As seen from Table 2, the smallest correlation value is between News and Adult which is 0.018. The largest correlation is between Thriller and Mystery, which is 0.967.

There are 9 pairs which correlation coefficient values are above 0.9: Drama-Comedy, Drama-Romance, Documentary-Biography, Thriller-Horror, Thriller-Action, Thriller-Crime, Thriller-Mystery, Crime-Mystery and Reality-TV – Game-Show. There are also 49 genre pairs which correlation coefficient values are between 0.8 and 0.9. Adult has the least correlation with all other genres. The highest correlation value of Adult is 0.186, with Romance and the lowest correlation value of Adult is 0.018, with News. Most of the genres are highly inter-correlated with each others and creates a methodological problem for analysis of the data.

Correlation coefficients between genre pairs are also given as matrix plot in Fig. 3. Matrix plot provides visualization of genre correlations with each other. As seen from Fig. 3, the variations between genre pairs display different features for each pair**.** **S**mooth linear patterns are easily identified between some genre pairs having high correlation values. It is also interesting that the graph patterns of Adult with other genres are quite different than remaining genre pairs.

TABLE 2
DISTANCE MATRIX (LOWER LEFT TRIANGLE) AND CORRELATION MATRIX (UPPER LEFT TRIANGLE) BETWEEN GENRES

Genres	Short	Drama	Adult	Comedy	Documentary	Thriller	Horror	Action	Romance	Crime	Family	Music	Adventure	Fantasy	Sci-Fi	Mystery	Animation	Sport	Biography	History	Reality-TV	War	News	Musical	Talk-Show	Western	Game-Show
Short		0.840	0.071	0.820	0.836	0.655	0.657	0.679	0.737	0.676	0.829	0.587	0.669	0.836	0.724	0.704	0.722	0.553	0.809	0.779	0.206	0.639	0.725	0.848	0.269	0.671	0.140
Drama	0.160		0.152	0.917	0.786	0.874	0.786	0.832	0.934	0.880	0.841	0.746	0.688	0.884	0.813	0.898	0.707	0.688	0.884	0.774	0.311	0.691	0.546	0.858	0.318	0.723	0.206
Adult	0.929	0.848		0.179	0.072	0.153	0.157	0.112	0.186	0.137	0.031	0.148	0.072	0.103	0.113	0.133	0.057	0.085	0.090	0.057	0.044	0.045	0.018	0.090	0.031	0.053	0.020
Comedy	0.180	0.083	0.821		0.780	0.792	0.756	0.791	0.892	0.794	0.812	0.736	0.708	0.860	0.798	0.816	0.756	0.676	0.831	0.699	0.361	0.605	0.527	0.858	0.447	0.665	0.263
Documentary	0.164	0.214	0.928	0.220		0.597	0.569	0.649	0.676	0.620	0.818	0.589	0.763	0.730	0.684	0.636	0.682	0.646	0.902	0.873	0.351	0.708	0.746	0.796	0.456	0.621	0.248
Thriller	0.345	0.126	0.847	0.208	0.403		0.930	0.916	0.773	0.958	0.636	0.743	0.526	0.803	0.843	0.967	0.606	0.575	0.702	0.624	0.259	0.595	0.367	0.676	0.257	0.727	0.175
Horror	0.343	0.214	0.843	0.244	0.431	0.070		0.849	0.686	0.845	0.584	0.682	0.498	0.778	0.817	0.893	0.593	0.511	0.635	0.570	0.231	0.533	0.361	0.651	0.244	0.687	0.158
Action	0.321	0.168	0.888	0.209	0.351	0.084	0.151		0.723	0.896	0.685	0.783	0.546	0.831	0.897	0.871	0.679	0.630	0.714	0.679	0.281	0.662	0.423	0.692	0.298	0.761	0.198
Romance	0.263	0.066	0.814	0.108	0.324	0.227	0.314	0.277		0.768	0.754	0.702	0.643	0.813	0.710	0.803	0.626	0.641	0.796	0.657	0.288	0.590	0.445	0.813	0.281	0.605	0.180
Crime	0.324	0.120	0.864	0.206	0.380	0.042	0.155	0.104	0.232		0.643	0.712	0.548	0.766	0.791	0.947	0.589	0.585	0.729	0.649	0.257	0.596	0.404	0.690	0.263	0.710	0.170
Family	0.171	0.159	0.969	0.188	0.182	0.364	0.416	0.315	0.246	0.357		0.676	0.661	0.853	0.719	0.686	0.797	0.649	0.840	0.766	0.347	0.637	0.661	0.824	0.363	0.648	0.249
Music	0.413	0.254	0.852	0.264	0.411	0.257	0.318	0.217	0.298	0.288	0.324		0.510	0.773	0.765	0.740	0.660	0.559	0.655	0.593	0.302	0.560	0.369	0.639	0.303	0.619	0.205
Adventure	0.331	0.312	0.928	0.292	0.237	0.474	0.502	0.454	0.357	0.452	0.339	0.490		0.644	0.572	0.564	0.590	0.557	0.769	0.635	0.408	0.503	0.551	0.806	0.457	0.502	0.348
Fantasy	0.165	0.116	0.897	0.140	0.270	0.197	0.222	0.169	0.187	0.234	0.147	0.227	0.356		0.868	0.840	0.852	0.610	0.797	0.727	0.281	0.650	0.529	0.837	0.316	0.720	0.193
Sci-Fi	0.276	0.187	0.887	0.202	0.316	0.157	0.183	0.103	0.290	0.209	0.281	0.235	0.428	0.132		0.836	0.760	0.588	0.717	0.688	0.304	0.654	0.470	0.717	0.325	0.706	0.226
Mystery	0.296	0.102	0.867	0.184	0.364	0.033	0.107	0.129	0.197	0.053	0.314	0.260	0.436	0.160	0.164		0.655	0.585	0.740	0.655	0.276	0.603	0.403	0.723	0.282	0.713	0.189
Animation	0.278	0.293	0.943	0.244	0.318	0.394	0.407	0.321	0.374	0.411	0.203	0.340	0.410	0.148	0.240	0.345		0.530	0.698	0.615	0.317	0.535	0.460	0.725	0.409	0.582	0.239
Sport	0.447	0.312	0.915	0.324	0.354	0.425	0.489	0.370	0.359	0.415	0.351	0.441	0.443	0.390	0.412	0.415	0.470		0.667	0.591	0.524	0.511	0.512	0.614	0.442	0.498	0.427
Biography	0.191	0.116	0.910	0.169	0.098	0.298	0.365	0.286	0.204	0.271	0.160	0.345	0.231	0.203	0.283	0.260	0.302	0.333		0.878	0.325	0.737	0.673	0.836	0.425	0.661	0.220
History	0.221	0.226	0.943	0.301	0.127	0.376	0.430	0.321	0.343	0.351	0.234	0.407	0.365	0.273	0.312	0.345	0.385	0.409	0.122		0.250	0.866	0.748	0.741	0.300	0.660	0.169
Reality-TV	0.794	0.689	0.956	0.639	0.649	0.741	0.769	0.719	0.712	0.743	0.653	0.698	0.592	0.719	0.696	0.724	0.683	0.476	0.675	0.750		0.202	0.409	0.291	0.707	0.230	0.922
War	0.361	0.309	0.955	0.395	0.292	0.405	0.467	0.338	0.410	0.404	0.363	0.440	0.497	0.350	0.346	0.397	0.465	0.489	0.263	0.134	0.798		0.561	0.617	0.227	0.574	0.133
News	0.275	0.454	0.982	0.473	0.254	0.633	0.639	0.577	0.555	0.596	0.339	0.631	0.449	0.471	0.530	0.597	0.540	0.488	0.327	0.252	0.591	0.439		0.596	0.402	0.469	0.380
Musical	0.152	0.142	0.910	0.142	0.204	0.324	0.349	0.308	0.187	0.310	0.176	0.361	0.194	0.163	0.283	0.277	0.275	0.386	0.164	0.259	0.709	0.383	0.404		0.335	0.636	0.215
Talk-Show	0.731	0.682	0.970	0.553	0.544	0.743	0.756	0.702	0.719	0.737	0.637	0.697	0.543	0.684	0.675	0.718	0.591	0.558	0.575	0.700	0.293	0.773	0.598	0.665		0.237	0.663
Western	0.329	0.277	0.947	0.335	0.379	0.273	0.313	0.239	0.395	0.290	0.352	0.381	0.498	0.280	0.294	0.287	0.418	0.502	0.339	0.340	0.770	0.426	0.531	0.364	0.763		0.163
Game-Show	0.860	0.794	0.980	0.737	0.752	0.825	0.842	0.802	0.820	0.830	0.751	0.795	0.652	0.807	0.774	0.811	0.761	0.573	0.780	0.831	0.078	0.867	0.620	0.785	0.337	0.837	

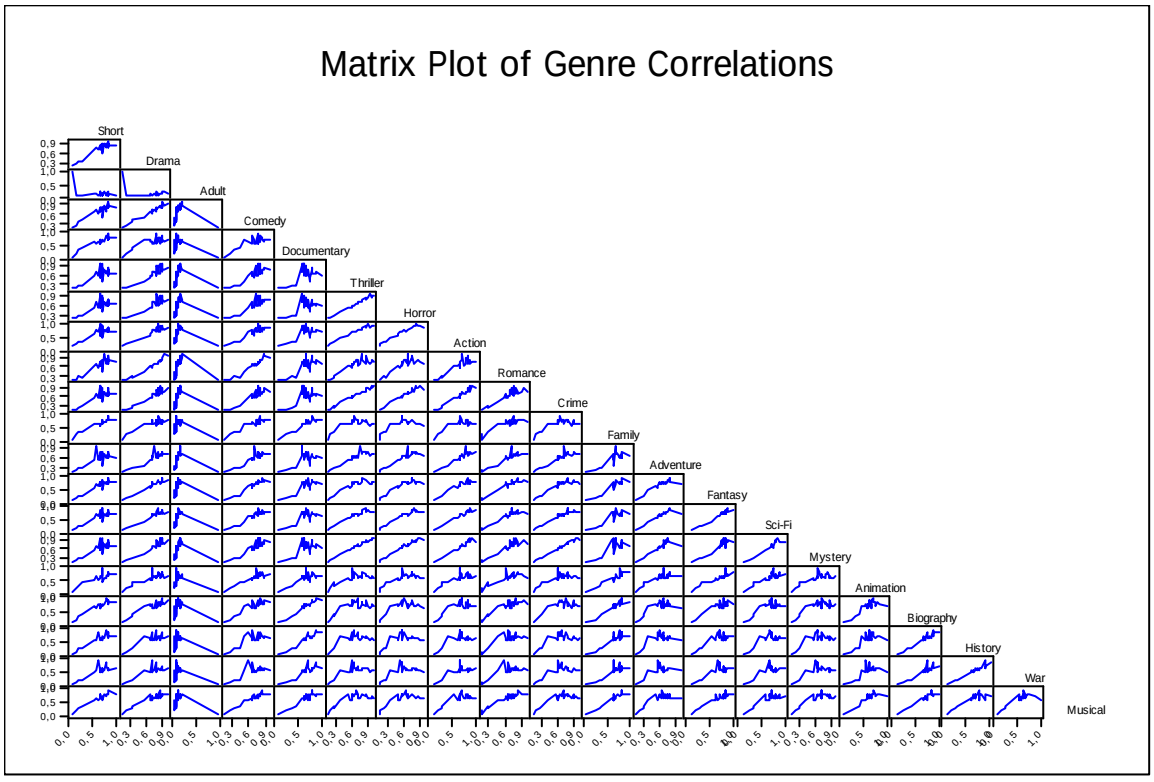

Fig. 3. Matrix plot of genre correlations

Generally, as a measure of internal consistency, a statistic called Cronbach's alpha is used. The values 0.7 or 0.75 are often used as cutoff values for Cronbach's alpha and thus for the reliability of the test [11]. For the above keyword distributions, the overall Cronbach's alpha value is computed as 0.8391, which is good considering the the cutoff value of 0.75.

4.2 Hierarchical Clustering of Genres

Considering that there are 27 movie genres, each represented with an array of length 19561, the data that need to be clustered would be too few. In this case, in order to discover the genre relationships, hierarchical clustering would be an appropriate method. Since, hierarchical clustering organizes data into the hierarchical structure based on the proximity of data with each other. Agglomerative clustering, a widely used method for hierarchical clustering, starts with N singleton clusters, each containing a single data, and performs a series of merge operations at each step until one cluster is left. The result is usually depicted by a dendrogram, which visualizes the potential clustering structures. By cutting the dendrogram at different levels, different clustering structures can be obtained.

When combining a pair of clusters at each level, we have used complete linkage algorithm as distance algorithm between two clusters. Complete linkage algorithm is considered effective for small clusters. It ensures that all items are within a maximum distance of each other, that is, it uses the largest distance between items of the clusters to define inter-cluster distance.

TABLE 3
AMALGAMATION STEPS FOR HIERARCHICAL CLUSTERING OF GENRES

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	26	98.3683	0.032633	6	16	6	2
2	25	97.3355	0.053289	6	10	6	3
3	24	96.7134	0.065732	2	9	2	2
4	23	96.1016	0.077968	21	27	21	2
5	22	95.1015	0.097971	5	19	5	2
6	21	94.8401	0.103199	8	15	8	2
7	20	94.5857	0.108287	2	4	2	3
8	19	93.6382	0.127235	5	20	5	3
9	18	92.6514	0.146972	11	14	11	2
10	17	92.3969	0.152063	1	24	1	2
11	16	92.2667	0.154667	6	7	6	4
12	15	91.1781	0.176439	1	11	1	4
13	14	89.5475	0.209051	6	8	6	6
14	13	86.8286	0.263428	1	2	1	7
15	12	85.3866	0.292267	5	22	5	4
16	11	84.3267	0.313467	6	26	6	7
17	10	83.1609	0.336781	21	25	21	3
18	9	83.0035	0.339929	12	17	12	2
19	8	82.1727	0.356546	1	13	1	8
20	7	79.1008	0.417984	6	12	6	9
21	6	78.0444	0.439113	5	23	5	5
22	5	77.6394	0.447211	1	18	1	9
23	4	74.88	0.5024	1	6	1	18
24	3	68.069	0.638619	1	5	1	23
25	2	56.6527	0.866946	1	21	1	26
26	1	50.9082	0.981835	1	3	1	27

Performing hierarchical clustering with complete linkage on IMDB data set produces amalgamation steps given in Table 3. The cluster Ids used in Table 3 is as follows: Short(1), Drama(2), Adult(3), Comedy(4), Documentary(5), Thriller(6), Horror(7), Action(8), Romance(9), Crime(10), Family(11), Adventure(12), Music (13), Fantasy(14), Sci-Fi(15), Mystery(16), Animation(17), Sport(18), Biography(19), History(20), Reality-TV(21), War(22), News(23), Musical(24), Talk-Show(25), Western(26) and Game-Show(27).

At step 1, Thriller(6) and Mystery(16) form a cluster. Notice from Table 2 that, these two genres are the closest pair, with a distance of 0.033 (and the largest correlation value of 0.967) among all genre pairs. At step 2, we observe that Crime(10) is merged with Thriller(6)-Mystery(16) pair to form another cluster. Also, as seen from Table 2, Thriller(6) and Crime(10) is the second closest pair, with a distance of 0.042, among all genre pairs. The result cluster contains Thriller, Mystery and Crime.

When we follow steps in Table 3, we observe the following results: Drama(2) and Romance(9) is merged at step 3 and later at step 7 they are merged with Comedy(4) forming a cluster composed of Drama, Romance and Comedy genres. At step 4, Reality-TV(21) and Game-Show(27) is merged into a cluster, which is then merged with Talk-Show(25) at step 17, forming a cluster composed of Reality-TV, Game-Show and Talk-Show genres . Documentary(5) and Biography(19) are merged at step 5, which is later merged with History(20) at step 8 forming another cluster containing Documentary, Biography and History genres. At step 6 Action(8) is merged with Sci-Fi(15), at step 9 Family(11) is merged with Fantasy(14), at step 10 Short(1) is merged with Musical(24) and step 18 Adventure(12) is merged with Animation(17). These groupings show us which genre pairs or triples are mostly blended together in a movie. Also notice that Adult(3) is merged at the final stage to form the root cluster. This shows that Adult genre cannot be correlated with other movie genres.

The corresponding dendrogram for Table 3 is also shown in Fig. 4. The cluster formations explained above can be visually followed from Fig. 4.

Applying a cutoff value between 0.45 – 0.50 to the dendrogram in Fig. 4 (shown as dashed line) results in 5 genre clusters. These clusters are given in Table 4.

TABLE 4
GENRE CLUSTERS OBTAINED FROM HIERARCHICAL CLUSTERING

Cluster 1	Short, Drama, Comedy, Romance, Family, Music, Fantasy, Sport, Musical
Cluster 2	Thriller, Horror, Action, Crime, Adventure, Sci-Fi, Mystery, Animation, Western
Cluster 3	Documentary, Biography, History, War, News
Cluster 4	Reality-TV, Talk-Show, Game-Show
Cluster 5	Adult

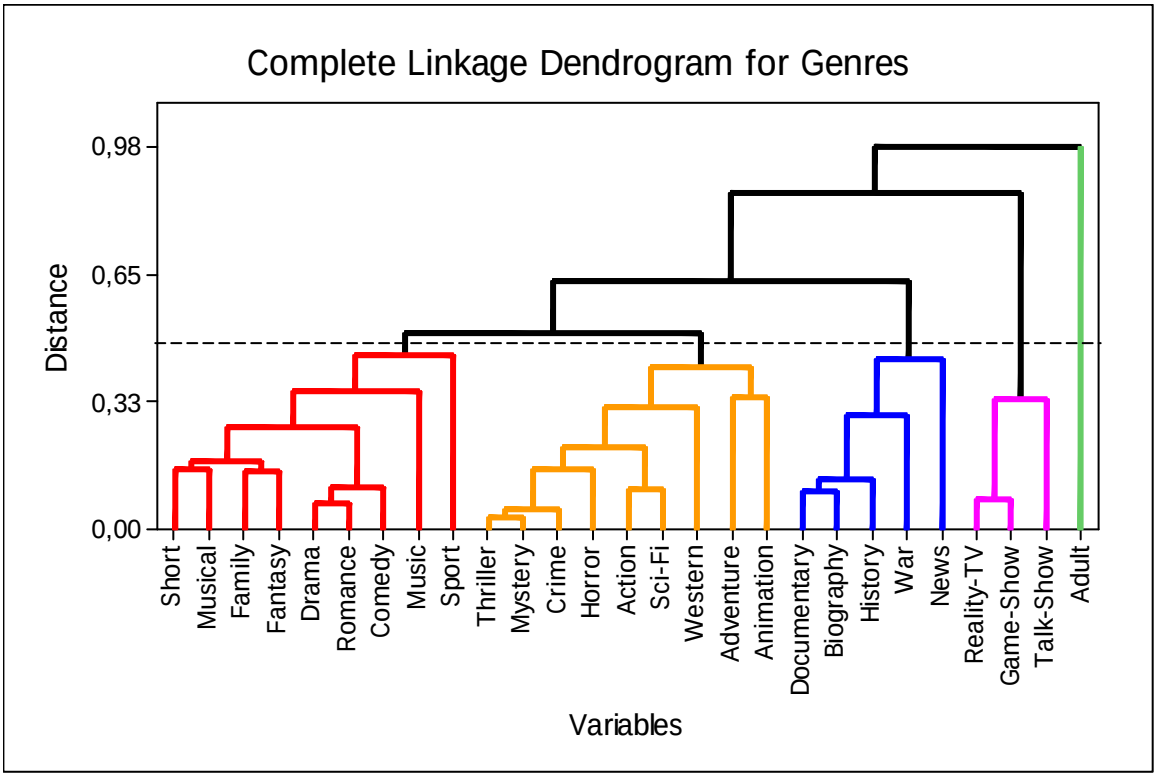

Fig. 4. Complete linkage dendrogram for genres

4.3 Principal Component Factor Analysis for IMDB Data

As an alternative to hierarchical clustering method, we also applied principal component factor analysis (PCFA) to the IMDB data. Principal component factor analysis is the technique to reduce a large number of variables to smaller random quantities called factors.

The main purpose of applying PCFA here is to compare the relationship between hierarchical clustering results and the results of PCFA. Factor loadings are computed using the covariance matrix obtained from IMDB data. Factor loading pattern of five factors are given in Table 5. PCFA have identified five factors with 84.6% explained variance among IMDB genres.

TABLE 5
GENRE FACTOR LOADINGS BY PCFA

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Short	-0.075	**0.174**	0.074	-0.085	0.017
Drama	0.046	**0.079**	-0.027	-0.031	-0.052
Adult	-0.048	-0.032	0.041	-0.007	**-0.958**
Comedy	-0.002	**0.219**	-0.146	0.004	-0.069
Documentary	-0.107	0.070	**0.207**	-0.012	-0.027
Thriller	**0.226**	-0.169	-0.041	0.005	-0.003
Horror	**0.210**	-0.135	-0.060	0.001	0.000
Action	**0.197**	-0.169	0.011	0.011	0.055
Romance	0.009	**0.203**	-0.136	-0.034	-0.117
Crime	**0.197**	-0.159	-0.005	0.000	-0.008
Family	-0.079	**0.233**	-0.016	-0.025	0.080
Adventure	**0.113**	0.023	-0.110	0.016	0.000
Music	-0.131	**0.293**	-0.067	0.037	-0.033
Fantasy	0.036	**0.177**	-0.120	-0.038	0.065
Sci-Fi	**0.135**	-0.050	-0.031	0.007	0.064
Mystery	**0.184**	-0.085	-0.067	-0.001	0.013
Animation	-0.038	**0.328**	-0.215	-0.003	0.133
Sport	0.015	-0.033	0.046	**0.126**	-0.038
Biography	-0.060	0.077	**0.137**	-0.026	-0.036
History	-0.031	-0.179	**0.408**	-0.052	-0.020
Reality-TV	0.010	-0.109	-0.025	**0.374**	-0.006
War	0.041	-0.309	**0.436**	-0.047	-0.006
News	-0.127	-0.122	**0.412**	0.050	-0.024
Musical	-0.086	**0.302**	-0.083	-0.048	-0.006
Talk-Show	-0.053	0.064	-0.087	**0.294**	0.020
Western	0.131	**-0.172**	0.124	-0.016	0.097
Game-Show	0.010	-0.136	-0.011	**0.384**	0.011
Variance	9.0822	5.3935	4.1812	3.1283	1.0515
Cumulative variance (%)	33.6	53.6	69.1	80.7	84.6

In Table 5, for each row, the maximum absolute value is found and the value is shown bold and thick border. For each factor (column) we have made a clustering of genres. After this clustering we obtain the clusters in Table 6. Comparing clusters obtained by hierarchical clustering method and clusters obtained by principal component factor analysis, only 3 out of 27 genres, shown in bold in Table 6, are placed into different clusters. Using PCFA, Animation and Western are placed into cluster 1 instead of cluster 2 in hierarchical clustering and Sport is placed into the cluster 4 instead of cluster 1 in hierarchical clustering. Hence, classification of 24 out of 27 genres (88.9%) matches with each other. It is interesting that factor 5 could be identified as Adult factor and this genre was the most distinct cluster to others in hierarchical clustering.

TABLE 6
GENRE GROUPS OBTAINED FROM PRINCIPAL COMPONENT FACTOR ANALYSIS

Cluster 1 (Factor 2)	Short, Drama, Comedy, Romance, Family, Music, Fantasy, **Animation**, Musical, **Western**
Cluster 2 (Factor 1)	Thriller, Horror, Action, Crime, Adventure, Sci-Fi, Mystery
Cluster 3 (Factor 3)	Documentary, Biography, History, War, News
Cluster 4 (Factor 4)	**Sport**, Reality-TV, Talk-Show, Game-Show
Cluster 5 (Factor 5)	Adult

5 CONCLUSION

Movie directors combine elements from different genres into a single movie plot. Hence, a movie may belong to several genres. In this study, we have used movie data from The Internet Movie Database. We have chosen movies with English language titles between years 2006 and 2010, a five year period, which makes a total of 48483 titles, with 27 genres and 19561 keywords. We have established genre keyword sets from movie keywords and used them to analyze the proximity of genres with each other. We have classified genres into five clusters and discovered the closest genre pairs and triples. We have compared the results obtained hierarchical clustering method and principal component factor analysis. Results of the two analyses are close to each other: classification of 24 out of 27 genres (88.9%) match with each other.

REFERENCES

[1] Z. Rasheed and M. Shah, "Movie genre classification by exploiting audio-visual features of previews", *Proc. the 16th International Conference on Pattern Recognition* vol.2, no., pp. 1086- 1089 vol.2, 2002.

[2] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Transactions on Circuits and Systems for Video Technology,* vol.15, no.1, pp. 52- 64, Jan. 2005

[3] H. Zhou, T. Hermans, A. V. Karandikar, J. M. Rehg, "Movie Genre Classification via Scene Categorization", *Proc. 10th international conference on Multimedia,* pp. 747-750, 2010.

[4] A. Austin, E. Moore, U. Gupta, and P. Chordia, "Characterization of movie genre based on music score," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp.421-424, 2010

[5] B. W. Herr, K. Weimao, E. Hardy, and Borner, "Movies and Actors: Mapping the Internet Movie Database," the *11th International Conference* on *Information Visualization,* pp.465-469, 2007

[6] A. Ahmed, V. Batagelj, X Fu, S. -H. Hong, D. Merrick, and A. Mrvar, "Visualisation and analysis of the internet movie database," the *6th International Asia-Pacific Symposium on Visualization,* pp.17-24, 2007

[7] M. Saraee, S. White, and J. Eccleston, "A Data Mining Approach to Analysis and Prediction of Movie Ratings", the *5th International Conference On Data Mining,* pp. 343-352, 2004

[8] W. Zhang, and S. Skiena, "Improving Movie Gross Prediction through News Analysis," *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies,* pp.301-304, 2009

[9] S.-T. Park, and D. M. Pennock, "Applying collaborative filtering techniques to movie search for better ranking and browsing", *Proc. the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp.550-559, 2007.

[10] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and Predicting the Helpfulness of Online Reviews," the *8th IEEE International Conference on Data Mining*, pp.443-452, 2008.

[11] A. Christmann, and S. Van Aelst, "Robust estimation of Cronbach's alpha," *Journal of Multivariate Analysis*, vol. 97, pp. 1660-1674, 2006.

Hasan Bulut is a member of the IEEE and the IEEE Computer Society. He is an Asst. Prof. of Computer Engineering Dept. at Ege University, Izmir, Turkey. He received his B.S. degree in Electronics and Telecommunications Engineering in 1996 from Istanbul Technical University, Istanbul, Turkey, M.Sc. in Computer Science in 2000 from Syracuse University, Syracuse, NY, USA, and Ph.D. in Computer Science in 2007 from Indiana University, Bloomington, IN, USA.

Serdar Korukoglu is a full-time professor of Computer Engineering Dept. at Ege University, Izmir, Turkey. He received his B.S. degree in Industrial Engineering, M.Sc. in Applied Statistics and Ph.D. in Computer Engineering from Ege University, Izmir, Turkey. He was in Reading University of England as a visiting research fellow in 1985.