Measuring Character-based Story Similarity by Analyzing Movie Scripts

O-Joun Lee
Dept. of Computer Eng.
Chung-Ang University
Seoul, Korea 156-756
concerto9203@gmail.com

Nayoung Jo
Dept. of Computer Eng.
Chung-Ang University
Seoul, Korea 156-756
joenayoung2@gmail.com

Jason J. Jung*
Dept. of Computer Eng.
Chung-Ang University
Seoul, Korea 156-756
j2jung@gmail.com

Abstract

The goal of this paper is to measure similarity among the stories for categorizing movies. Although genres are well-performing as movies' categories, users have difficulty for predicting substances of the movies through the genres. Therefore, we proposed the story-based taxonomy of the movies and a method for constructing it automatically. In order to reflect characteristics of the stories, we used two kinds of features: (i) proximity among movie characters and (ii) genres of the movies. Based on the features, we constructed the story-based taxonomy by clustering the movies. We anticipate that the proposed taxonomy could make the users imagine and predict substances of movies through comprehending which movies contain similar stories.

1 Introduction

With a rapid growth of media industry, 'crossover' is one of popular strategies in this area. In here, the crossover does not only indicate convergence among media, but also advent of novel genres, which are mixtures of conventional genres [JLYN17]. This paradigm makes the movies have characteristics of multiple genres. It means that the users have difficulty for expecting substances of the movies, if they only rely on the genres.

In order to improve this problem, we suggested a novel taxonomy for exposing similarity among stories of the movies. Also, we proposed a method for automatically constructing the story-based taxonomy. To build the taxonomy, we applied two features that reflect stories of the movies; i.e., (i) proximity among the characters and (ii) genres of the movies.

The story consists of three major components: the character, event, and background. The event is represented by interaction among the characters in a particular background. Therefore, we supposed that the proximity (frequency of the interaction) could reflect lots of stories' characteristics. In our previous studies [DHLJ16, THLJ17, LJ16, JLYN17], we applied character networks (i.e., social networks among the characters) for representing the proximity.

The conventional genres cover various features of the movies; e.g., topics, methods for developing stories, ambiance, and more. In here, a problem is that the genres contain too complex information to identify clear criteria for the classification. Nevertheless, although the genres can not precisely indicate substances of the movies, they can provide us meaningful information.

To construct the story-based taxonomy, we clustered movies based on the character network and the genre distribution. As a preliminary study, we exhibited efficiency and necessity of the proposed method through a small-scaled experiment.

* Corresponding author.

In: A. Jorge, R. Campos, A. Jatowt, S. Nunes (eds.): Proceedings of the Text2StoryIR'18 Workshop, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

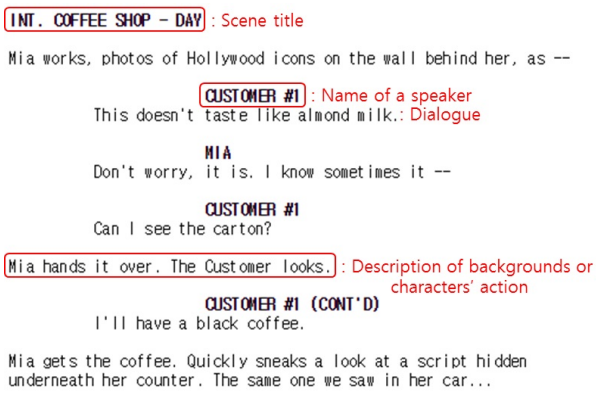

Figure 1: A part of a script of 'La La Land (2016)'.

2 Character Network

Our previous studies [DHLJ16, THLJ17, LJ16, JLYN17, LJ18] used the character network for computationally analyzing the stories. The character network is a social network among characters that appeared in the stories. It was defined as follows;

Definition 1 (Character Network) *Suppose that N is the number of characters that appeared in a movie, $\mathbb{C}_\alpha$. When $N(\mathbb{C}_\alpha)$ indicates a character network of $\mathbb{C}_\alpha$, $N(\mathbb{C}_\alpha)$ can be described as a matrix $\in \mathbb{R}^{N\times N}$. It consists of $N \times N$ components which are the proximity among the characters as:*

$$N(\mathbb{C}_\alpha) = \begin{bmatrix} a_{1,1} & \cdots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & a_{N,N} \end{bmatrix}, \tag{1}$$

where, $a_{i,j}$ is the proximity of c_i for c_j when C_α is an universal set of characters that appeared in $\mathbb{C}_\alpha$ and c_i is an i-th element of C_α.

In this study, we used frequency of the dialogues between the characters for measuring the proximity among them. The dialogues were extracted from the movies' scripts collected from the Internet Movie Script Database (IMSDb) [1].

Since the scripts are structured documents, as displayed in Fig. 1, it is relatively easy to extract dialogues and their speakers. Simply speaking, the movies' script consists of multiple scenes, which start with scene titles. Also, the scene contains descriptions and dialogues. The dialogue includes a speaker of dialogue and its content. In the description, characters' action and backgrounds of scenes are illustrated.

In this study, we mainly focused on boundaries of the scene and the speakers of the dialogues. As formats of the scripts are not completely uniform, we have difficulty for assuring whether we can discover points where the characters appear and disappear, or not. Therefore, we supposed that every characters appeared in the corresponding scene are listeners for all the dialogues spoken in the scene. It can be illustrated as Fig. 2.

Nevertheless, the character networks have a difficulty for comparing with each other, since the number of characters is different from movies. Park et al. [PYKY15] proposed a method for normalizing the character networks by using the Singular Value Decomposition (SVD). In order to compare the character networks, we applied the same method. The normalized character network was denoted as $\mathbb{N}(\mathbb{C}_\alpha)$.

3 Story-based Taxonomy of Movies

The story-based taxonomy consisted of multiple groups of movies that have similar stories. To compare the movies' stories with each other, we used two kinds of features: (i) the proximity among the characters and (ii) the genre distribution. For representing the proximity, we have an efficient model, the character network. However, in case of genres, the movies are not simply included within particular genres, but they partially contain characteristics of multiple genres. Therefore, we represented relationships between the movies and the genres by using a 22-dimensional vector as:

$$\overrightarrow{\mathbb{C}_\alpha^G} = \left\langle \mu_{\mathbb{G}_1}(\mathbb{C}_\alpha), \cdots, \mu_{\mathbb{G}_{22}}(\mathbb{C}_\alpha) \right\rangle, \tag{2}$$

[1] http://www.imsdb.com/

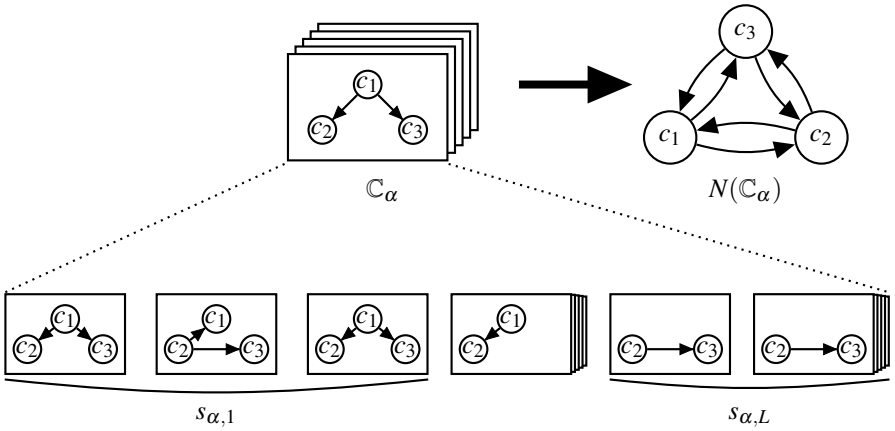

Figure 2: An example of relationships between a movie ($\mathbb{C}_\alpha$), characters (c_1, c_2, c_3), scenes ($s_{\alpha,1}, \cdots, s_{\alpha,L}$), and a character network ($N(\mathbb{C}_\alpha)$).

where $\mu_{\mathbb{G}_g}(\mathbb{C}_\alpha)$ indicates whether $\mathbb{G}_g$ includes $\mathbb{C}_\alpha$. Also, each component was initialized by a boolean value based on annotations collected from IMDB [2].

In order to estimate difference among movies' stories, we applied two distance metrics, which are based on the Jaccard index and the Frobenius norm, respectively. They are formulated as:

$$\begin{aligned}\mathscr{D}_G\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right) &= 1 - \frac{\sum_{\forall \mathbb{G}_g} E(\mu_{\mathbb{G}_g}(\mathbb{C}_\alpha), \mu_{\mathbb{G}_g}(\mathbb{C}_\beta))}{\sum_{\forall \mathbb{G}_g} \max\{\mu_{\mathbb{G}_g}(\mathbb{C}_\alpha), \mu_{\mathbb{G}_g}(\mathbb{C}_\beta)\}}, \\ \mathscr{D}_F\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right) &= \left\|\mathbb{N}(\mathbb{C}_\alpha) - \mathbb{N}(\mathbb{C}_\beta)\right\|_F,\end{aligned} \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $E(\cdot,\cdot)$ is an indicator function that indicates whether two inputs are commonly positive or not.

To combine the two distance metrics, we applied a weighted harmonic mean of them. Thereby, it can be formulated as:

$$\begin{aligned}&\mathscr{D}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right) \\ &= \left[\frac{\theta_F \mathscr{D}_F\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)^{-1} + \theta_G \mathscr{D}_G\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)^{-1}}{\theta_F + \theta_G}\right]^{-1},\end{aligned} \tag{4}$$

where θ_F and θ_G denote weighting parameters for $\mathscr{D}_F$ and $\mathscr{D}_G$, respectively.

For finding optimal θ_F and θ_G, we compared $\mathscr{D}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)$ with users' perception. Since $\mathscr{D}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)$ was not normalized, first, we transformed it into a range of $[0,1]$ by the inverse of $\mathscr{D}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)$. As a result, $\mathscr{S}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right) = \mathscr{D}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)^{-1}$ indicates the similarity between two arbitrary movies, $\mathbb{C}_\alpha$ and $\mathbb{C}_\beta$. Then, a loss function for training was designed as:

$$L_{\mathscr{D}} = \sum_{\forall \mathscr{S}_{u_j}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)} \left\|\mathscr{S}_{u_j}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right) - \mathscr{S}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)\right\|_2, \tag{5}$$

where $\mathscr{S}_{u_j}\left(\mathbb{C}_\alpha, \mathbb{C}_\beta\right)$ indicates a user-estimated similarity between $\mathbb{C}_\alpha$ and $\mathbb{C}_\beta$. Based on the loss function, we optimized θ_F and θ_G with the gradient descent method.

In order to build the story-based taxonomy of the movies, we used the fuzzy c-means clustering algorithm. This algorithm aimed to minimize an objective function:

$$\underset{\mathbb{T}}{\operatorname{argmin}} \sum_{\forall \mathbb{C}_\alpha} \sum_{\forall \mathbb{T}_k} \mu_{\mathbb{T}_k}(\mathbb{C}_\alpha)^m \mathscr{D}\left(\mathbb{C}_\alpha, C_{\mathbb{T}_k}\right), \tag{6}$$

$$\mu_{\mathbb{T}_k}(\mathbb{C}_\alpha) = \left[\sum_{\forall \mathbb{T}_l} \left(\frac{\mathscr{D}\left(\mathbb{C}_\alpha, C_{\mathbb{T}_k}\right)}{\mathscr{D}\left(\mathbb{C}_\alpha, C_{\mathbb{T}_l}\right)}\right)^{\frac{2}{m-1}}\right]^{-1}, \tag{7}$$

[2] http://www.imdb.com/

$\mathbb{C}_\alpha$	$\mathbb{C}_\beta$	$\mathscr{D}_G(\mathbb{C}_\alpha,\mathbb{C}_\beta)^{-1}$	$\mathscr{D}_F(\mathbb{C}_\alpha,\mathbb{C}_\beta)^{-1}$	$\mathscr{S}_U(\mathbb{C}_\alpha,\mathbb{C}_\beta)$
Terminator (1984)	Gravity (2014)	0.25	0.39	2.60
Terminator (1984)	Star Wars: Ep.1 (1999)	0.50	0.70	3.80
Star Wars: Ep.1 (1999)	Gravity (2014)	0.17	0.46	3.40

Table 1: The similarity between 'Terminator (1984)', 'Gravity (2014)', and 'Star Wars: Ep. 1 (1999)', which is estimated by the proposed distance metrics and users.

where $\mathbb{T}$ denotes the total cluster model that corresponds the story-based taxonomy, $\mathbb{T}_k$ refers to a k-th cluster in $\mathbb{T}$, and $C_{\mathbb{T}_k}$ indicates the center of $\mathbb{T}_k$. $C_{\mathbb{T}_k}$ was decided by a weighted average of elements within $\mathbb{T}_k$. A feature vector of $C_{\mathbb{T}_k}$ consisted of two parts as the same with $\mathbb{C}_\alpha$'s, and they can be formulated as:

$$\mathbb{N}(\mathbb{T}_k) = \frac{\sum_{\forall \mathbb{C}_\alpha \in \mathbb{T}_k} \mu_{\mathbb{T}_k}(\mathbb{C}_\alpha)^m \mathbb{N}(\mathbb{C}_\alpha)}{\sum_{\forall \mathbb{C}_\alpha \in \mathbb{T}_k} \mu_{\mathbb{T}_k}(\mathbb{C}_\alpha)^m}, \tag{8}$$

$$C^{\tilde{G}}_{\mathbb{T}_k} = \frac{\sum_{\forall \mathbb{C}_\alpha \in \mathbb{T}_k} \mu_{\mathbb{T}_k}(\mathbb{C}_\alpha)^m \tilde{\mathbb{C}}^G_\alpha}{\sum_{\forall \mathbb{C}_\alpha \in \mathbb{T}_k} \mu_{\mathbb{T}_k}(\mathbb{C}_\alpha)^m}. \tag{9}$$

In order to use the fuzzy c-means clustering, we had to determine the number of clusters. We measured the quality of the total cluster model, as the number of clusters increased one by one. The benefit from increasing the number of clusters was estimated by:

$$\mathscr{B}_{|\mathbb{T}|} = (1-\theta_Q) \times \Delta\mathscr{Q}_{|\mathbb{T}|} + \theta_Q \times \Delta\mathscr{Q}_{|\mathbb{T}|-1}, \tag{10}$$

$$\Delta\mathscr{Q}_{|\mathbb{T}|} = \mathscr{Q}_{|\mathbb{T}|} - \mathscr{Q}_{|\mathbb{T}|-1}, \tag{11}$$

where $|\mathbb{T}|$ indicates the number of clusters in the current cluster model and θ_Q denotes a user-defined parameter that represents the momentum of the cluster model's quality. When the number of clusters increases to $|\mathbb{T}|$, $\mathscr{Q}_{|\mathbb{T}|}$ refers to the quality of the cluster model, $\Delta\mathscr{Q}_{|\mathbb{T}|}$ denotes the amount of changes in the quality, and $\mathscr{B}_{|\mathbb{T}|}$ indicates the gain from the increment of the number of clusters.

If the $\mathscr{B}_{|\mathbb{T}|}$ had a positive value, the proposed method proceeded the next iteration by $|\mathbb{T}| := |\mathbb{T}| + 1$. Otherwise, it determined the optimal number of clusters as $|\mathbb{T}|$.

The quality of the total cluster model, $\mathbb{T}$ was estimated by the Fukuyama-Sugeno index, $FS_m(\mathbb{T})$ [HBV02]. It is formulated as:

$$FS_m(\mathbb{T}) = \sum_{\forall \mathbb{C}_\alpha} \sum_{\forall \mathbb{T}_k} \mu_{\mathbb{T}_k}(\mathbb{C}_\alpha)^m \times \left(\mathscr{D}\left(\mathbb{C}_\alpha, C_{\mathbb{T}_k}\right) - \mathscr{D}\left(C_{\mathbb{T}_k}, C\right)\right), \tag{12}$$

where C indicates the average of all the clusters' centers. A method for calculating the average of the centers is the same with Eq. 8, although it is not weighted, in here. Thereby, the first term of Eq. 12 measures the compactness of each cluster, the second term indicates the adjacency among the clusters, and FS_m is the Fukuyama-Sugeno index for the story-based taxonomy of the movies. If the story-based groups in the taxonomy are well-constructed, FS_m might have a small value.

In addition, m, which is used as exponent of the membership functions, is a user-defined parameter. As m becomes bigger, the membership degree of the movies gets more consideration. In this study, m equals to 2 en bloc.

4 Experimental Result and Discussion

As a preliminary study, we have not constructed an adequate dataset for verifying the proposed method, yet. The experiment focused on efficiency of the proposed distance metrics. Table. 1 exhibits similarity between three movies ('Terminator (1984)', 'Gravity (2014)', and 'Star Wars: Ep. 1 (1999)'), which is estimated by the proposed metrics and users. We collected the user-estimated similarity from 10 students of Chung-Ang University. The users rated the similarity between movies with natural numbers from 1 to 5. A 5th column of Table. 1 indicates average of users' responses.

As displayed in Table. 1, $\mathscr{D}_F^{-1}$ is more correlated with $\mathscr{S}_U$ than $\mathscr{D}_G^{-1}$. Pearson correlation coefficients between them are 0.88 and 0.58, respectively. In particular, between first and third cases, $\mathscr{S}_U$ and $\mathscr{D}_G^{-1}$ have opposite tendency. There is a

possibility that backgrounds of the movies affect users' perception, since 'Gravity (2014)' and 'Star Wars: Ep. 1 (1999) commonly described the astrospace. Nevertheless, it is difficult to describe likeness among movies' stories only with the genres, although the genres cover various characteristics of the movies.

This experiment is too tiny-scaled to verify neither the proposed distance metrics nor the story-based taxonomy. However, the result made sure that the genres are not enough to make the users imagine substances of the movies.

5 Conclusion

In this study, we revealed similarity among movies' stories by clustering them with the character network and the genre distribution. The proposed method enables the users to imagine substances of movies, which they have not seen yet.

Nevertheless, the proposed method has not been verified with an adequate dataset, since this study is a part of ongoing research. Our future work will be focused on composing appropriate datasets and evaluating the proposed method.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1A41015675).

References

[DHLJ16] Tran Quang Dieu, Dosam Hwang, O-Joun Lee, and Jason J. Jung. A novel method for extracting dynamic character network from movie. In *Proceedings of the 7th EAI International Conference on Big Data Technologies and Applications*. EAI, 2016.

[HBV02] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering validity checking methods: Part II. *ACM SIGMOD Record*, 31(3):19–27, September 2002.

[JLYN17] Jai E. Jung, O-Joun Lee, Eun-Soon You, and Myoung-Hee Nam. A computational model of transmedia ecosystem for story-based contents. *Multimedia Tools and Applications*, 76(8):10371–10388, Apr 2017.

[LJ16] O-Joun Lee and Jason J. Jung. Affective character network for understanding plots of narrative contents. In María Trinidad Herrero Ezquerro, Grzegorz J. Nalepa, and José Tomás Palma Mendez, editors, *Proceedings of the Workshop on Affective Computing and Context Awareness in Ambient Intelligence (AfCAI 2016)*, volume 1794 of *CEUR Workshop Proceedings*, Murcia, Spain, Nov 2016. CEUR-WS.org.

[LJ18] O-Joun Lee and Jason J. Jung. Modeling affective character network for story analytics. *Future Generation Computer Systems*, 2018. (TO Appear).

[PYKY15] Seung-Bo Park, Eun-Soon You, Hyun-Sik Kim, and Seong Won Yeo. Rank reduction of a character-net matrix based on svd. In *Proceedings of the 11th International Conference on Multimedia Information Technology and Applications (MITA 2015)*, Tashkent, Uzbekistan, Jun 2015.

[THLJ17] Quang Dieu Tran, Dosam Hwang, O-Joun Lee, and Jai E. Jung. Exploiting character networks for movie summarization. *Multimedia Tools and Applications*, 76(8):10357–10369, Apr 2017.