Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 2, 2023, 221-228.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

MoGaL: Novel Movie Graph Construction by Applying LDA on Subtitle

Mohammad Nazari, Hossein Rahmani*, Dadfar Momeni and Motahareh Nasiri

School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

Article Info

Article History:
Received 10 December 2022
Revised 10 January 2023
Accepted 18 March 2023

DOI:10.22044/jadm.2023.12481.2396

Keywords:
Subtitle analysis, Movies graph, Graph analysis، Graph entropy, Graph homophily.

**Corresponding author: h_rahmani@iust.ac.ir (H.Rahmani)*

Abstract

Graph representation of data can better define the relationships among the data components, and thus provide a better and richer analysis. So far, movies have been represented in graphs many times using different features for clustering, genre prediction, and even for use in recommender systems. In constructing movie graphs, little attention has been paid to their textual features such as subtitles, while they contain the entire content of the movie, and there is a lot of hidden information in them. Thus in this paper, we propose a method called MoGaL to construct movie graph using LDA on subtitles. In this method, each node is a movie, and each edge represents the novel relationship discovered by MoGaL among two associated movies. First, we extract the important topics of the movies using LDA on their subtitles. Then we visualize the relationship between the movies in a graph using the cosine similarity. Finally, we evaluate the proposed method with respect to the measured genre homophily and genre entropy. MoGaL succeeds to outperform the baseline method significantly in these measures. Accordingly, our empirical results indicate that movie subtitles could be considered a rich source of informative information for various movie analysis tasks.

1. Introduction

Movies are a big part of the entertainment industry, and their number is increasing daily. With the spread of online streaming websites and people's acceptance, this industry has become more important, so this data has attracted the attention of the data scientists and researchers in the recent years [1-5].

One of the most useful and widely used approaches in the field of data mining and relationship discovery is the representation of the problem in the form of a graph. Graph representation can be very useful in visualizing and understandably presenting data [6, 7]. Movie graphs allow for studying relationships, discovering similarities between movies, and using graph analysis algorithms. In various papers, movie graphs have been used to recommend movies to the users [8-10], genre predictions [11], find cooperation patterns between movie actors, and suggest actors to directors [4, 12]. It has been shown that using more data types enriches the graph, and leads to better results [13-15]. In order to construct graph, they usually used the relationships among features that describe the movie such as writer, director, and actors or user ratings to the movies [4, 12, 9].

To the best of our knowledge, there is no previous work that used the textual data of movies in the construction of movie graph, and this important feature of movies has been neglected. The textual features of the movies indicate the content of the movie. Subtitles, plot synopsis, and summary can be mentioned as the textual features of the movies. These unstructured features contain a lot of information that can be used to represent the content of a movie [1, 10].

According to the recent advances in the field of natural language processing, various analyzes have been performed to find the content similarity of movies using textual features, which include all kinds of traditional methods, machine learning, and deep learning [1-3, 10].

Therefore, we want to obtain the content relationships between movies using subtitles and represent it in a graph. We enriched movie graphs and we succeeded to get better results by adding content relations of movies to them. In this paper, we propose a method called MoGaL (Movie Graph Construction by Applying LDA) on subtitle that uses subtitles similarity to construct a movies graph. This method first extracts the important topics of movies using LDA [16] on their subtitles. Then constructs a movie graph using the topic similarities. Finally, considering that the genre is a stylistic or thematic categorization based on the main story of the movie, we evaluate the graph made with three methods based on the genre of the movies.

The structure of this paper is as what follows. Section 2 overviews the available methods for constructing the movie graphs and measuring the similarity of movies using textual features. The proposed method to construct a movies graph is described in detail in Section 3. The Empirical results are presented in Section 4. Section 5 includes a discussion of the results, and proposes promising directions for future research works.

2. Background

Graph representation of data makes it easier to understand and analyze them; this is also true for movies [6, 7]. Movies have frequently been modeled as graph in papers for various purposes [12, 17, 18]. One of the most popular methods for representing movies in graphs is to consider each movie as a node connected by links to features that describe the movie such as writer, director, actors, studio, year of production, and awards [7]. Toine [19] used this graph to build a contextual recommendation model. Toine was able to increase the accuracy of his movie recommender system by using the contextual information of the movies that he put in the graph. Also by analyzing this graphs, Tang *et al*. [4] found the relationships between the types of movies and the cooperation networks of the directors and the actors, and Spitz *et al*. [18] predicted the probability of success and popularity of a movie. Another common way to construct movie graph is making a bipartite graph of movies and users, where each edge represents the user's score for each movie. This graph is widely used in recommender systems [8, 9]. In this graph, by using different machine learning methods, it is possible to predict the score of users, and suggest movies to the users. Lee *et al*. [14] added new edges to the bipartite graph that represent users' emotions toward movies to have a better recommender system. Also, in addition to the bipartite graph of users' scores, Darban *et al*. [20] used side-information of the users to recommend movies.

Zhou *et al*. [11] also used the graph of movies to predict the genre. In their graph, each movie is a node, and each edge represents having the same director. The tags assigned to each movie on the IMDb website are also considered attributes of each movie in the graph.

As we have reviewed, the movie graph has many applications, and each of them has used different features of the movies to construct the graph. However, we could not find any paper that uses textual data in graph construction. Therefore, the content relationship between movies on these networks has been neglected. The textual features of the movies indicate the content of the movie. Subtitles, plot synopsis, and summary can be mentioned as the textual features of the movies. These unstructured features contain a lot of information that can be very useful if extracted correctly [13]. Due to the advances in natural language processing, subtitles have been used in many papers for various applications. For example, Bougiatiotis *et al*. [2] represented each movie directly using its subtitles, and used it to calculate the similarity between the movies [21]. Luhmann *et al*. [1] introduced the SubRosa method, which used subtitles to solve the cold start [15] problem in the recommender systems. SubRosa is a content-based recommender system based on subtitles of movies. Also Hasan *et al*. [3] used movie subtitles to cluster the movies.

We have seen those subtitles, as the longest textual feature of movies, have already been used in the representation and analysis of movies but they have never been used to construct movies graph. Therefore, in this work, we want to construct a graph of movies using their subtitles.

3. Proposed Method

In this section, we propose a method called MoGaL for constructing the graph of the movies based on LDA on their subtitles, and then we evaluate the constructed graph based on three measures. In this method, after collecting the data and pre-processing it, we construct the graph of the movies using the proposed method MoGaL. In the following, we describe MoGaL in more detail. After constructing the graph, we evaluate it using 3 proposed measures based on genre homophily and genre entropy. Figure 1 shows the structure of the proposed method, each of its steps will be explained in detail in the next sections.

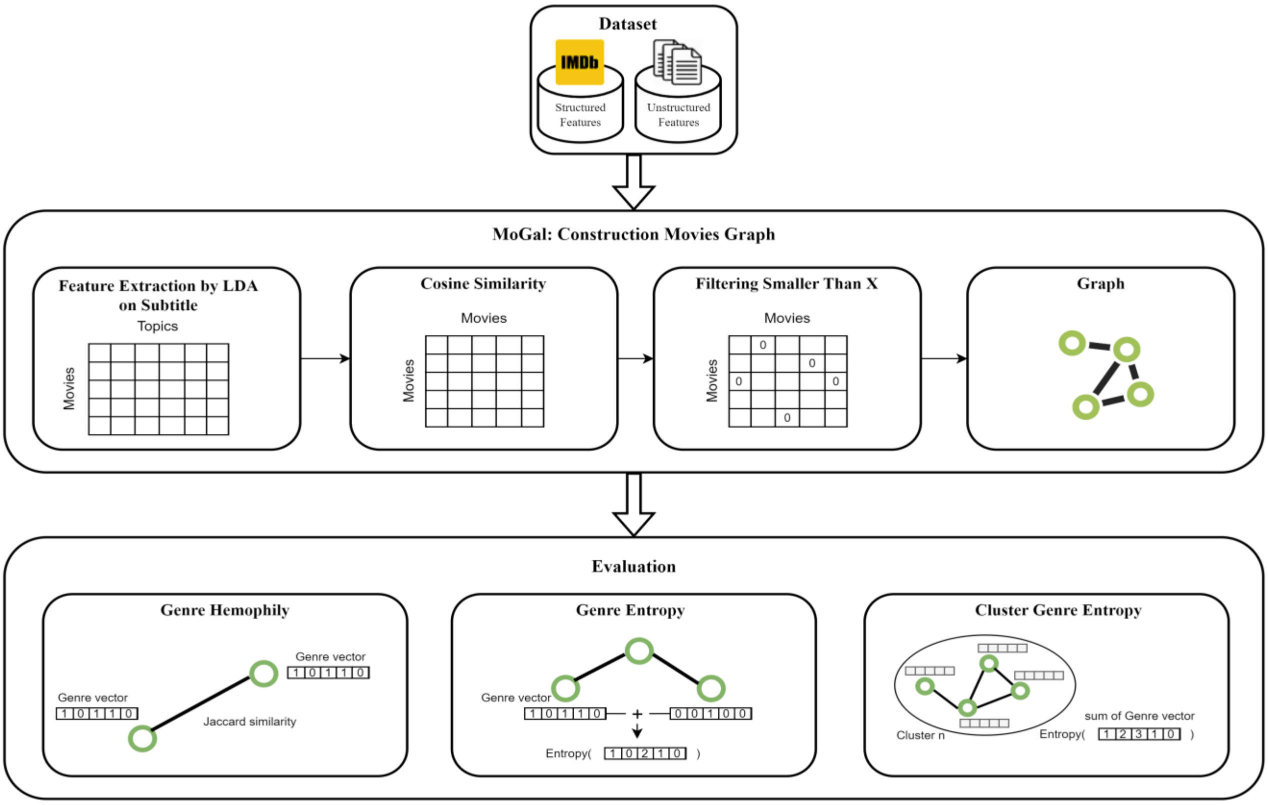

Figure 1. Steps of constructing movies graph by MoGaL method and its evaluation.

3.1. Dataset

Our dataset includes the structured and unstructured features. Structured data refers to movie features such as title, release year, and genres, which we used the IMDb online dataset [22] to collect. Unstructured data refers to the English subtitles of the movies, which we collected them using website crawling. We represent our dataset with M, which contains structured and unstructured features for 4050 movies. Table 1 describes our dataset.

3.2. Pre-processing

Subtitles include daily conversation dialogues and special characters that we clean and ready to analyze in several steps.

- Removing HTML tags that include information about the subtitle-provider website, movie specifications, and details about how subtitles are displayed including font type and color.
- Removing additional descriptions that describe some of the important sounds of the film for the hearing-impaired such as hums, whispers, cheers, and passing cars.

Table 1. Description of each database column.

Variable	Description	Definition
M	Movies set	$M = \{ m_1, m_2, \dots, m_{4050}\}, \quad \lvert M \rvert = 4050$
$GENRES$	Genres set	$GENRES = \{g_1, g_2, g_3, \dots, g_{23}\} = \{$ Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, FilmNoir, History, Horror, Music, Musical, Mystery, News, Romance, Sci – Fi, Sport, Thriller, War, Western$\}$, $\lvert GENRES \rvert = 23$
Gen_{m_i}	Genres set for movie m_i	Each movie m_i in our dataset can have 1 to 3 genres, which we denote by GEN_{m_i}. $Gen_{m_i} = \{g^1_{m_i}, \dots, g^n_{m_i}\}, \quad 1 \le n \le 3, \quad g \in GENRES$
$GenVec_{m_i}$	Genre vector for movie m_i	$GenVec_{m_i}$ denote one-hot representation of m_i movie genre set. $\overrightarrow{GenVec_{m_i}} = \{b_1, b_2, b_3, \dots, b_n\}, \quad n = \lvert GenVec_{m_i} \rvert = \lvert GENRES \rvert = 23$ $b \in \{0,1\} \quad \begin{cases} b_x = 1 \ if \ g_x \in Gen_{m_i} \\ b_x = 0 \ if \ g_x \notin Gen_{m_i} \end{cases}$

- Removing frequently used words such as plural signs and prepositions, which do not have an influential connotation and can be ignored (stop words).
- Removing specific names that are repeated in a large number in a subtitle, and do not have an influential conceptual load using the NER tool and the spacy library [23].
- Tokenizing a subtitle into a list of words, and lemmatizing each word.

3.3. Movie Graph Construction: MoGaL

In this section, we construct the graph of movies using the MoGaL. First, we use the LDA algorithm to extract important topics from the subtitles. In order to determine the number of topics of LDA, we calculate the coherence and perplexity for the different number of topics and denote it by t. Then we represent each movie m_i with a vector v_{m_i} of length t by applying the LDA algorithm on their subtitles. Next, to calculate the similarity between each pair of movies m_i and m_j, we use the cosine similarity between vectors v_{m_i} and v_{m_j} according to (1). In this way, we calculate the cosine similarity between the movie vector two by two. Then we use these values to build the similarity matrix of the movies.

$$Cosine\,Similarity\left(m_i, m_j\right) = S_C\left(v_{m_i}, v_{m_j}\right) = \frac{v_{m_i} \cdot v_{m_j}}{v_{m_i} v_{m_j}} \tag{1}$$

Using the similarity matrix generated from the previous step, we create the similarity graph of the movies $G(V, E)$. Graph G is a complete graph that each vertex $v_i \in V$ represents a movie m_i, and each edge $e_{i,j} \in E$ has a weight $w_{i,j}$, which shows the cosine similarity between two vertices connected to that edge, which is defined as (2).

$$E = \forall v_i, v_j \in V \rightarrow \exists e_{i,j},\ w_{i,j} = S_C\left(v_{m_i}, v_{m_j}\right) \tag{2}$$

Considering all the edges and vertices in the graph, analyzing and processing the graph is more laborious and time-consuming, and therefore, the graph should be pruned. We prune the graph based on the weight of its edges. For this purpose, we choose a threshold θ for minimum edge weight (minimum similarity), and we call the pruned graph $G_M(V_M, E_M)$. Graph G_M is the final output of the MoGaL method.

3.4. Graph Evaluation

In this section, we evaluate the graph constructed by MoGaL. We use three different measures based on movie genre to measure the quality of the constructed graph. Genre is a stylistic or thematic categorization based on the main story of the movie, suggested by humans for each movie. Genres are widely used to categorize and recommend movies to users. Thus we can measure the effectiveness of the graph constructed by MoGaL in tasks such as recommending movies, predicting genres, and classification movies using genre.

3.4.1 Homophily

To measure the homophily in this graph, we use the Jaccard similarity measure between the set of genres of the connected movies. Thus we calculate the homophily (H) between two connected movies m_i and m_j using (3).

$$H\left(m_i, m_j\right) = Jaccard\left(Gen_{m_i}, Gen_{m_j}\right) = \frac{\left|GEN_{m_i} \cap GEN_{m_j}\right|}{\left|GEN_{m_i} \cup GEN_{m_j}\right|} \tag{3}$$

To measure homophily graph G_M, we use the average homophily of all connected movies m_i and m_j in graph G_M $(e_{m_i,m_j} \in G_M)$, according to (4).

$$H\left(G_M\right) = \frac{\sum_{e_{i,j} \in E_M} H\left(m_i, m_j\right)}{E_M} \tag{4}$$

3.4.2 Entropy

The entropy measure is another method for evaluating the G_M. In this regard, we define the entropy of each movie m_i in G_M to be equal to genre entropy of its neighbors in the graph. First, we define the set of neighbors of the movie m_i according to (5), and call it N_{m_i}.

$$N_{m_i} = \left\{m_j : e_{i,j} \in E_M\right\} \tag{5}$$

Then we add up the $GenVec$ of the m_i's neighbors ($GenVec_{N_{m_i}}$) according to (6).

$$\overrightarrow{GenVec_{N_{m_i}}} = \sum_{v_j \in N_{m_i}} \overrightarrow{GenVec_{v_j}} \tag{6}$$

Finally, the entropy for each movie m_i is calculated by (7), where P_i is defined by (8). In (8), $GenVec^j_{N_{m_i}}$ represents the *j-th* element of the vector $GenVec_{N_{m_i}}$.

$$Entropy\left(m_i\right) = \sum_{i=1}^{|GENRES|=23} -P_j . \log\left(P_j\right) \tag{7}$$

$$P_j = \frac{GenVec^j_{N_{m_i}}}{\sum_{h=0}^{|GENRES|=23} GenVec^h_{N_{m_i}}} \tag{8}$$

Next, after calculating the entropy for each of the movies of the graph G_M, we calculate the entropy of the G_M using the average entropy of all movies according to (9).

$$Entropy(G_M) = \frac{\sum_{m_i \in V_M} Entropy(m_i)}{V_M} \qquad (9)$$

3.4.3. Clustering

We use clustering and entropy as another method to evaluate the graph G_M. For this purpose, we embed each node in the graph with the Node2Vec method [24] for the input of the clustering algorithm. This method works similar to the Word2Vec [25] method, which is common in text analysis. Then we partition the graph into k clusters c_1 to c_k using the K-means algorithm. We used WSCC (the sum of squared distance between each point and the centroid in a cluster) to find the optimal number of clusters.

Then for each cluster c_i, we define the entropy of its members with respect to their genre vectors as follows. First, we calculate genre vector cluster c_i ($GenVec_{c_i}$) by using the sum of the $GenVec$ of all the movies in that cluster according to (10). Then we calculate the entropy for each cluster as the same (7) and (8). After calculating the entropy of each cluster, we calculate the clusters average entropy of graph G_M.

$$\overrightarrow{GenVec_{C_i}} = \sum_{v_k \in C_i} \overrightarrow{GenVec_{v_k}} \qquad (10)$$

4. Empirical Results

In this section, we implement MoGaL, and examine its results. The results of each evaluation measures are mentioned in the following section.

We use the correlation criterion to determine the number of LDA topics. As shown in Table 2, the correlation is maximized at the value of 150. Thus the value of $t = 150$ is considered for the number of LDA topics. Then after feature extraction using LDA, we construct the graph.

Table 2. Coherence values for different number of topics in LDA.

Number of topics	Coherence
50	0.4554
100	0.5195
150	0.5446
200	0.5267

After constructing the complete graph of movies, to prune it, we keep only highly weighted edges that have weight more than θ ($weight\left(e_{i,j}\right) > \theta$). Figure 2 and Figure 3 show the number of nodes and edges, respectively, with respect to threshold θ in pruned graph $G_M(V_M, E_M)$. As Figure 2 and Figure 3 indicate, $\theta = 0.9$ is the best optimal value for this hyperparameter. Accordingly, the number of edges of the pruned graph $G_M(V_M, E_M)$ is reduced to 34512, and the number of its vertices with edges is reduced to 2933. The diameter of G_M graph is 25, and the average of its shortest paths is 6.72. Also according to the importance of genres in our evaluation measures, The genres distribution diagram of G_M's movies is given in Figure 4.

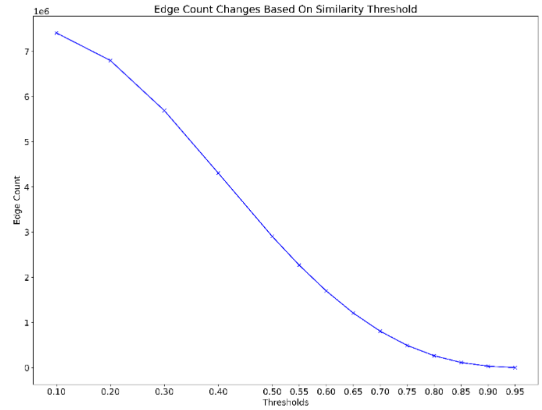

Figure 2. Changes in the graph's edge count based on a similarity threshold of 10% to 95%.

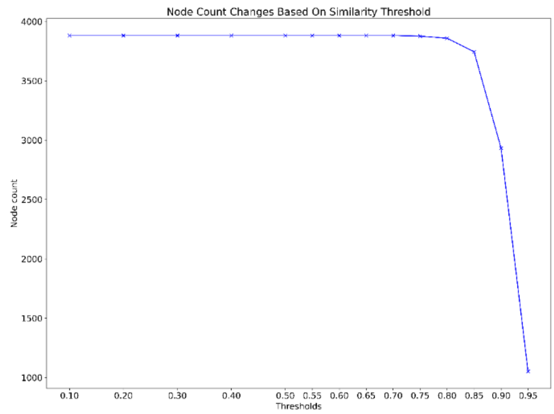

Figure 3. Changes in the graph's vertices count based on a similarity threshold of 10% to 95%. As it is clear in the figure, the chart has a turning point for the similarity threshold of 90%.

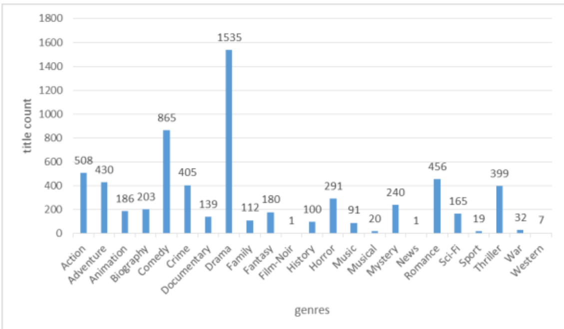

Figure 4. Genre distribution of movies.

To evaluate the constructed graph, we use the movie genres as the key feature to assess the graph using the three measures, discussed in the Section 3.4. Also, we constructed two other graphs of movies to compare their results with the graph constructed by MoGaL. The first one is graph G_E, which is constructed using common entities of movies. Zhou et al. [11] used this graph to classification movies by genre. In this graph, each movie is a node and each edge indicates the existence of the same director, writer or star[1] between the two movies. Using this idea, we construct the graph G_E for 2933 movie which exist in the graph G_M. Finally, the G_E included 31843 edges, that 22198 of which are also exist in G_M.The second graph is the random graph G_{NS}, which we use as a baseline for comparison.

We create a random graph $G_{NS}(V_{NS}, E_{NS})$ with the negative sampling method. For this purpose, we select all the nodes of the graph G_M, So, $V_{NS} = V_M$. Then we select the set of edges of the graph G_{NS} in the following ways: first, not be a member of the set of edges of the G_M ($e_{i,j} \notin E_M$). Second, they are randomly selected, And third, the number of edges of the random graph is equal to the number of edges of the G_M, Therefore $|E_{NS}| = |E_M|$.

Now, we measure the quality of the MoGaL for constructing movies graph by comparing the three evaluation measures discussed in the three constructed graphs. As it is shown in Table 3, the homophily in the graph constructed by MoGaL is more than the others, which indicates that the movies connected by edges have more genre similarities. The difference between the graph constructed by MoGaL and the random graph is significantly large. The values of the standard deviation in this table indicates the dispersion of the Jaccard similarity, which is good if the value is low, but on the condition that the average value is high. For example, in the random graph, a low value of the standard deviation indicates that the Jaccard similarity values are concentrated around the mean, which is a low value.

Table 3. Genre homophily in the graphs.

	G_{NS}	G_{En}	G_M
Mean Jaccard similarity (μ)	0.1775	0.3761	**0.3957**
Jaccard similarity STD (σ)	0.2207	0.2755	0.2973

In Table 4, we present the results of the genre entropy analysis of the neighbors of a node in three constructed graphs.

Table 4. Genre entropy in the graphs.

	G_{NS}	G_{En}	G_M
Mean entropy (μ)	3.4509	2.2585	**2.2329**
Entropy STD (σ)	0.1693	0.5620	0.6034

[1] A movie star is an actor or actress who is famous for their starring, or leading, roles in movies. In our dataset, the maximum number of stars for each movie is 3.

A high entropy means a uniform distribution of neighbors' genre. Thus in this case, low entropy is good. A low entropy indicates that the neighbors of a node have the same genres. As you can see in Table 4, the lowest amount of genre entropy belongs to graph G_M. As a result, it can be stated that the G_M graph has much more homogeneity in terms of genre.

As regards the last assessment method, after performing the clustering, we calculate the clusters' average entropy for G_M. We use within cluster sum of squares (WSCC) to find the optimal number of clusters. Finally, we set the number of clusters to $k = 40$ according to Figure 5. Table 5 shows the average and standard deviation of entropy in clusters for the three constructed graphs.

Table 5. Clusters genre entropy in graphs.

	G_{NS}	G_{En}	G_M
Mean cluster entropy (μ)	3.593	**2.668**	2.952
Clusters entropy STD (σ)	0.0901	0.2498	0.2506

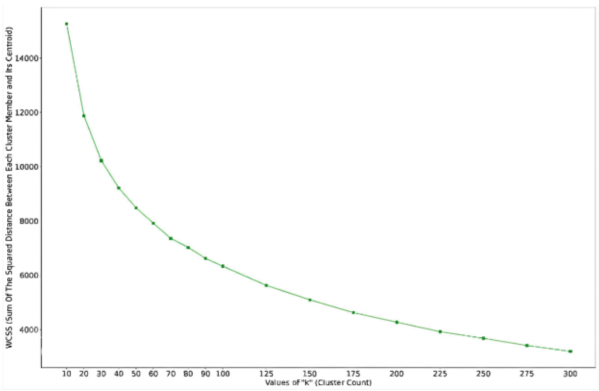

Figure 5. WSCC changes (the sum of the squared distance of each sample from the center of the cluster) according to the number of clusters (K) in the K-means algorithm.

Since low entropy indicates less genre diversity in clusters, it can be better from this point of view. In Table 5, the lowest clusters genre entropy is for graph G_{En}. This shows that the clusters in graph G_{En} have more similar genres than the clusters in the other two graphs. However, our method has a significant difference with the baseline in this criterion. Also we found by manually checking the results that the reason for the high value of the standard deviation in these three criteria for the MoGaL method is the existence of movies that have a high degree. These movies have common topics that can be close to different genres.

The homogeneity of the G_M with respect to movie genres indicates that movie subtitles are one of the most informative features for discovering similarities among movies. Also our evaluations show that subtitles are a rich and descriptive feature for topic extraction and similarity analysis of movies.

Genres of movies are usually chosen according to the suggestions of writers and viewers. Therefore, this relationship among the topics extracted by LDA from subtitles and genres indicates that our topics are close to the real world. On the other hand, due to the power of LDA in understanding hidden topics, this constructed graph has information beyond genres, from thematic connections between movies, which are very valuable.

5. Conclusions and Future Work

In this paper, we proposed a method called MoGaL for constructing the graph of the movies based on LDA on their subtitles and evaluating movies' similarity graphs based on their subtitles. In this regard, we first collected the data and subtitles of movies. Then we extracted important topics from the subtitle text using topic extraction methods LDA. In the following, using the cosine similarity measure, we measured the similarity among topic vectors of movies, and based on this, we constructed a movies graph. Next, considering that the genre of a movie represents its dominant topic, we introduced three criteria for evaluating the graph made using homophily and entropy based on the genre of movies and evaluated the graph.

The results obtained show that there is a correlation between the topics extracted with LDA and the genres of the movies. Comparing the results of the constructed graph by MoGaL with two baseline graphs and the graph constructed with the entities of the movies shows that the values of entropy and homophily are very different from the baseline and are very close to the value of the graph constructed with the entities. This shows that the subtitles of movies indicate the genre and content of a movie. Therefore, it can be used as an important feature to analyze movies.

For future work, considering that MoGaL considers the relevance of hidden topics in subtitles, it can be used in the field of genre prediction. This graph can also be used to recommend movies because people usually like movies with similar themes to their favorite movies.

References

[1] J. Luhmann, M. Burghardt, and J. Tiepmar, "SubRosa: Determining Movie Similarities based on Subtitles," *INFORMATIK 2020*, 2021.

[2] K. Bougiatiotis and T. Giannakopoulos, "Content representation and similarity of movies based on topic extraction from subtitles," in *Proceedings of the 9th Hellenic Conference on Artificial Intelligence*, 2016, pp. 1-7.

[3] M. M. Hasan, S. T. Dip, T. M. Kamruzzaman, S. Akter, and I. Salehin, "Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach," in *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, 2021, pp. 219-224.

[4] Y. Tang, J. Yu, C. Li, and J. Fan, "Visual analysis of multimodal movie network data based on the double-layered view," *International Journal of Distributed Sensor Networks*, 2015.

[5] H. Koosha, Z. Ghorbani and R. Nikfetrat, "A Clustering-Classification Recommender System based on Firefly Algorithm," *Journal of AI and Data Mining*, vol. 10, pp. 103-116, 2022.

[6] J. B. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, and E. Koh, "Attention models in graphs: A survey," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2019, pp. 1-15.

[7] B. Rao and A. Mitra, "Graph Mining and Its Applications in Studying Community-Based Graph under the Preview of Social Network" in *Product Innovation through Knowledge Management and Social Media Strategies*, 2016, pp. 94-146.

[8] D. Sulieman, M. Malek, H. Kadima, and D. Laurent, "Toward social-semantic recommender systems" *International Journal of Information Systems and Social Change (IJISSC)*, vol. 7, pp. 1-30, 2016.

[9] M Zhang and Y. Chen, "Inductive matrix completion based on graph neural networks," in *International Conference on Learning Representations*, 2020.

[10] S. Eden, A. Livne, O. Sar Shalom, B. Shapira, and D. Jannach, "Investigating the Value of Subtitles for Improved Movie Recommendations," in *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 2022, pp. 99-109.

[11] C. Zhou, H. Chen, J. Zhang, Q. Li, D. Hu, and V. S. Sheng, "Multi-label graph node classification with label attentive neighborhood convolution" *Expert Systems with Applications*, vol. 180, 2021.

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

[13] S. Eden, A. Livne, O. Sar Shalom, B. Shapira, and D. Jannach, "Investigating the Value of Subtitles for Improved Movie Recommendations," in *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 2022, pp. 99-109.

[14] C. Lee, D. Han, K. Han, and M. Yi, "Improving graph-based movie recommender system using cinematic experience" *Applied Sciences*, vol. 12, pp. 1493, 2022.

[15] M. Goyani and N. Chaurasiya, "A review of movie recommendation system: Limitations, Survey and Challenges" *ELCVIA: electronic letters on computer vision and image analysis*, vol. 19, pp. 18-37, 2020.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation" *Journal of machine Learning research*, pp. 993-1022, Jan 2003.

[17] S. B. Park, K. J. Oh, and G. S. Jo, "Social network analysis in a movie using character-net" *Multimedia Tools and Applications*, vol. 59, pp. 601-627, Jul 2012.

[18] A. Spitz and E. Á. Horvát, "Measuring long-term impact based on network centrality: Unraveling cinematic citations," *PloS one*, Oct 2014.

[19] T. Bogers, "Movie recommendation using random walks over the contextual graph," 2010.

[20] Z. Z. Darban and M. H. Valipour, "GHRS: Graph-based hybrid recommendation system with application to movie recommendation," *Expert Systems with Applications*, vol. 200, Aug 2022.

[21] K. Bougiatiotis and T. Giannakopoulos, "Enhanced movie content similarity based on textual, auditory and visual information," *Expert Systems with Applications*, 2018.

[22] "IDb Datasets," IMDb, 16 05 2021. [Online]. Available: https://www.imdb.com/interfaces/. [Accessed: May. 16, 2021].

[23] M.Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol 7, pp. 411-420, Jul 2017.

[24] A. Grover and J. Leskovec, "node2vec: Scalable feature larning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855-864.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

MoGal: ساخت گراف فیلم‌ها با استفاده از LDA بر روی زیر نویس

محمد نظری، حسین رحمانی*، دادفر مؤمنی و مطهره نصیری

دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت، تهران، ایران.

ارسال ۲۰۲۲/۱۲/۱۰؛ بازنگری ۲۰۲۳/۰۱/۱۰؛ پذیرش ۲۰۲۳/۰۳/۱۸

چکیده:

نمایش گرافی داده‌ها می‌تواند روابط بین اجزای داده را بهتر نشان دهد و در نتیجه تحلیل بهتر و غنی تری ارائه دهد. تاکنون فیلم‌ها بارهـا بـا اسـتفاده از ویژگی‌های مختلف برای خوشه‌بندی، پیش‌بینی ژانر و حتی در سیستم‌های توصیه‌گر به صورت گرافی مدل شده‌اند. تا کنون در ساخت گراف‌هـای فـیلم ها به ویژگی‌های متنی آن‌ها مانند زیرنویس توجه چندانی نشده است، در حالی که زیرنویس‌ها کل محتوای فـیلم را در بـر می‌گیرنـد و اطلاعـات پنهـان زیادی در آن‌ها وجود دارد. بنابراین در این مقاله، ما روشی به نام MoGaL برای ساخت گراف فـیلم‌هـا بـا اسـتفاده از LDA روی زیرنویس‌هـا پیشـنهاد می‌کنیم. در این روش، هر گره در گراف ساخته شده نشان‌دهنده‌ی یک فیلم است و هر یال نشان‌دهنده رابطه جدیـد کشـف شـده توسـط MoGaL در بین دو فیلم مرتبط است. ابتدا موضوعات مهم فیلم‌ها را با استفاده از LDA از زیرنویس آن‌ها استخراج می‌کنیم. سپس با استفاده از شباهت کسینوسـی، رابطه بین فیلم‌ها را در یک گراف مجسم می‌کنیم. در نهایت، روش پیشنهادی را با توجه به هموفیلی و آنتروپی ژانر ارزیـابی مـی‌کنیم. MoGaL موفـق شد در این اقدامات به طور قابل توجهی از روش پایه پیشی بگیرد. بر این اساس، نتایج تجربی ما نشان‌می‌دهد که زیرنویس فیلم‌ها می‌توانند منبعی غنـی از اطلاعات آموزنده برای کارهای مختلف تحلیل فیلم در نظر گرفته شوند.

کلمات کلیدی: تحلیل زیرنویس، گراف فیلم‌ها، تحلیل گراف، آنتروپی گراف، هموفیلی گراف، استخراج موضوع.