A Minimum Spanning Tree Representation of Anime Similarities

Canggih Puspo Wibowo
Sagasitas Research Center
Yogyakarta, Indonesia
(canggih.p.w@ieee.org)

Abstract—In this work, a new way to represent Japanese animation (anime) is presented. We applied a minimum spanning tree to show the relation between anime. The distance between anime is calculated through three similarity measurements, namely crew, score histogram, and topic similarities. Finally the centralities are also computed to reveal the most significance anime. The result shows that the minimum spanning tree can be used to determine the similarity anime. Furthermore, by using centralities calculation, we found some anime that are significance to others.

Index Terms—anime, similarity measurement, minimum spanning tree

I. INTRODUCTION

Minimum spanning tree is an undirected graph that has no cycles, connects to every vertex, and has the minimal total weighting for its edges. It is known as a graph which has low complexity and easy to implement [1]. Mostly, minimum spanning tree is used to represent wires, roads, and water pipes so that the total cost is minimum. However, recently it has been used in various areas such as geographical information [2], radio networks [3], EEG [4], chip architecture [5] and stock exchange [6]. A similar concept also applied in a movie recommendation system in the form of a dendogram [7].

On the other side, Japanese animation, which is known as anime, has become internationally widespread nowadays. Not only in the eastern countries, American audience also enjoying anime through Hayao Miyazaki’s Studio Ghibli, which is well-known in western. According to Oricon’s data [1] for the past five years (2011-2015), the Blu-ray Disc and DVD selling of anime were quite stable; it sold more than 600 thousand discs for each year. With the huge popularity of anime, a recommendation system is needed to find the similar anime based on particular indicators.

The aim of this work is to represent anime similarities using a minimum spanning tree to be used as a recommendation system. Moreover, the significance of anime will be revealed by extracting the centralities of the minimum spanning tree.

II. SIMILARITY MEASUREMENTS

In order to construct the minimum spanning tree, a distance measurement between anime has to be calculated. In this case, similarity measurement between anime will be used. Many works have been done in similarity measurements of movies in general. Researchers used user reviews [8], movie mood [9], and movie score [10] to determine the similarity. There is also an NLP approach proposed using topic and summary similarity [11]. In this work, besides using the score and topic, a new measurement is proposed, namely crew similarity. It is considered that crew similarity is an important characteristic in anime recommendation. Thus, the similarity measurements are described as follows:

1) Crew Similarity

There are two kinds of crew working in the anime industry, viz., production staff and voice actor/actress. Both of them are considered as important factors determining the anime success. Here a similarity measurement by using those factors is proposed, namely crew similarity. Let $S_n$ be a set of crew involved in anime $n^{th}$. The crew similarity between two anime is defined as the number of crew (both staff and voice actor/actress) who work for both anime, as calculated as follows:

$$d_{ij} = \log |S_i \cap S_j|,$$

($i = 0, 1, 2, \ldots, k - 2$),

$$j = i + 1, i + 2, \ldots, k - 1$$

where $|S|$ and $k$ means the number of members in set $S$ and the total number of anime, respectively. Here, log transformation is applied since there are some data which are too far away from the others.

2) Score Histogram Similarity

Score histogram is determined by using user votes. There are some categories that user can select reflecting anime score. For instance, in Anime News Network, anime’s votes are classified into 11 categories: Masterpiece, Excellent, Very good, Good, Decent, So-so, Not really good, Weak, Bad, Awful, and Worst ever. Based on the number of votes for each category, the total score of an anime is calculated. Thus, here the votes for each category are assumed to be a histogram of scores. The similarity between score histogram would represent the user preference for the particular anime. Let $X_i$ be the score histogram of anime $i^{th}$ which is defined as

$$X_i = \{x_{i1}^1, x_{i2}^2, x_{i3}^3, \ldots, x_{iN}^N\},$$

where

$$x_{iN}^N = \frac{C_n}{\sum_{n=1}^{N} C_n},$$

1http://www.oricon.co.jp; data were gathered in http://www.someanthing.com
and $C_n$ is the number of votes for category $n$, while $N$ is the number of score categories. Then, the score histogram similarity $(s_{ij})$ is calculated using chi-squared distance as follows:

$$s_{ij} = \frac{1}{\sqrt{\sum_{x_i,j} \frac{(x_i - x_j)^2}{x_i + x_j}}}.$$  

(4)

3) Topic similarity

Each anime is commonly labeled with some genres to make it easier to be classified. In Anime News Network, besides genre, anime are also classified into themes. Both genre and theme are considered as important parameters for classifying anime. Therefore, topic similarity is used here, employing both genre and theme of anime, to show the similarity with respect to the content. The topic similarity between two anime is defined as the number of topics (genres and themes) that present in both anime. Let $G_n$ be a set of genre and theme terms of anime $n^{th}$. Then topic similarity, $h_{ij}$, is calculated as

$$h_{ij} = |G_i \cap G_j|$$  

(5)

Since the three measurements have different scales, a normalization with respect to size is performed. The calculation is given by

$$\hat{z}_{ij} = \frac{z_{ij} - z_{\text{min}}}{z_{\text{max}} - z_{\text{min}}}, (z = d, s, h)$$  

(6)

where,

$$z_{\text{min}} = \min_{0 \leq i,j \leq K-1} z_{ij}, \quad z_{\text{max}} = \max_{0 \leq i,j \leq K-1} z_{ij}, \quad (z = d, s, h)$$  

(7)

From Eq. (1), (4), and (5) we know that crew and topic similarity results in higher values when the both anime are considered similar, however, for the score histogram similarity, the result is otherwise. Hence, the crew and topic similarities are recalculated so that all similarities are aligned. Let $d$, $\hat{s}$, and $\hat{h}$ be the normalized version of $d$, $s$, and $h$, respectively. The calculation is given by

$$\hat{q}_{ij} = 1 - \hat{q}_{ij}, (q = d, h)$$  

(8)

Afterwards, all three measurements are combined into a similarity vector, defined as $s_{ij} = [d_{ij}, \hat{s}_{ij}, h_{ij}]$, where $S'$ means the transposition of $S$. Thus total distance $\delta_{ij}$ is calculated as

$$\delta_{ij} = \|s_{ij}\|$$  

(9)

where $\|\|\|$ is the Euclidean distance. Afterwards, in the next section, the total distance between anime, $\sigma_{ij}$, will be used as the edge length of a graph.

III. MINIMUM SPANNING TREE REPRESENTATION

In order to construct the minimum spanning tree, Kruskal’s algorithm is employed (Alg. 1). The algorithm is implemented using disjoint-set data structure. Let $V = \{v_k, 0 \leq k \leq k-1\}$, be a set of vertices, $E = \{(v_i, v_j), 0 \leq i, j \leq k - 1\}$ be a set of edges connecting a pair of vertices, and $w = \{\delta_{ij}, 0 \leq i, j \leq k - 1\}$ be a set of total distances obtained from the previous section.

| Algorithm 1 Kruskal’s Algorithm |
|----------------------------------|
| 1: procedure MAKESET($v$)        |
| 2: Create new set containing $v$ |
| 3: end procedure                |
| 4: function FINDSET($v$)        |
| 5: return a set containing $v$   |
| 6: end function                 |
| 7: procedure UNION($u,v$)       |
| 8: Unites the set that contain $u$ and $v$ into a new set |
| 9: end procedure                |
| 10: function KRUSKAL($V, E, w$) |
| 11: $A \leftarrow \{\}$         |
| 12: for each vertex $v$ in $V$ do |
| 13: MakeSet($v$)                |
| 14: $A \leftarrow A \cup \{(u,v)\}$ |
| 15: $A \leftarrow A \cup \{v\}$ |
| 16: end for                      |
| 17: Arrange $E$ in increasing costs, ordered by $w$ |
| 18: for each $(u,v)$ taken from the sorted list do |
| 19: if FindSet($u$) \neq FindSet($v$) then |
| 20: $A \leftarrow A \cup \{(u,v)\}$ |
| 21: $A \leftarrow A \cup \{u\}$ |
| 22: end if                       |
| 23: end for                      |
| 24: return $A$                  |

In graph theory, measuring central vertices has been an active area of research. Many researchers proposed centrality indicators. Some of them are based on walk structure, namely degree and eigenvector centralities. Apart from it, there is also some which is based on geodesic distance, such as betweenness and closeness centralities. In this work, central vertices are going to be identified based on the aforementioned indicators. The centrality measurements are described below:

1) Degree Centrality

Degree centrality of vertex $v$ is the proportion of other vertices that are adjacent to $v$. It is defined as

$$C_D(v) = \frac{1}{k-1} \sum_{u \in V} a(u,v)$$  

(10)

where,

$$a(u,v) = \begin{cases} 1, & \text{if } u \text{ and } v \text{ are connected by a line} \\ 0, & \text{otherwise} \end{cases}$$  

(11)

Anime having high degree means that it have many similar anime around it.

2) Eigenvector Centrality

Conceptually, eigenvector centrality is similar to degree centrality. The centrality also measures the number of walks of a vertex. However, instead of having the length of one, eigenvector measures the number of walks of length infinity. Thus, in eigenvector centrality, a vertex has a high centrality if it is connected to another vertex.
that also have a high centrality. The eigenvector centrality is defined as the summed connection of a vertex to others, weighted by their centralities. Let \( R = r_{uv} \) be a matrix of relationship, i.e., \( r_{uv} = \alpha(u,v) \). Eigenvector centrality of vertex \( v \), denoted as \( e_v \), is calculated as follows
\[
\lambda e_v = \sum_u r_{uv} e_u
\]
(12)
where \( \lambda \) is a constant required so that the equation have a non zero solution. This problem can be rewritten as an eigenvector equation:
\[
\lambda e = Re
\]
(13)
where \( e \) is the eigenvector of \( R \) and \( \lambda \) is the corresponding eigenvalue.

3) Betweenness Centrality
In the concept of betweenness centrality, a vertex is called in central position when it is located on the shortest path between two vertices. Based on it, betweenness centrality of a vertex is defined as the number of times a vertex acts as a bridge between the shortest path of two other vertices. The betweenness centrality of vertex \( v \) is defined as
\[
C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}
\]
(14)
where \( \sigma_{st} \) is the number of shortest paths between \( s \) and \( t \), \( \sigma_{st}(v) \) is the fraction of those shortest paths that pass through \( v \). According to Freeman [14], the betweenness can be normalized as
\[
\hat{C}_B(v) = \frac{2C_B(v)}{k^2 - 3k + 2}
\]
(15)

4) Closeness Centrality
Closeness measures how close a vertex to all other vertices in a graph. It is defined as the inverse of the total distance from a vertex to other vertices. Since this measurement depends on the number of vertices in the graph, the relative closeness is calculated with a normalization. The relative closeness is given by
\[
C_C(v) = \frac{k - 1}{\sum_{u=1}^{k-1} d(v,u)}
\]
(16)
where \( d(v,u) \) is the shortest-path distance between \( v \) and \( u \). High value of closeness means that a vertex is relatively close to the other vertices.

In order to calculate the total centrality, Euclidean distance are applied so that \( \varphi(v) = \sqrt{C_D(v)^2 + C_B(v)^2 + C_C(v)^2} \), where \( \varphi(v) \) is the total centrality of vertex \( v \).

IV. RESULTS
In this work, 4029 anime data were collected randomly from Anime News Network[2]. For each anime pair, three similarity measurements are calculated, then used as features to build the minimum spanning tree representation. To visualize the

\[\text{http://www.animenewsnetwork.com, accessed on May 10, 2016}\]
one until the largest closeness. This kind of distribution is expected for closeness. Vertices that are located on the outer part of the graph are having the small value of closeness, nevertheless large values are obtained as the vertices located near the center of the graph. Thus, many of the vertices are located between the outer and the center of the graph.

Table I shows the corresponding anime having the largest centrality value for each measurement. It can be seen that anime One Piece and Naruto are ranked the first and second respectively in all centrality measurements as well as the total one. It shows the significance of those anime among others.

V. CONCLUDING REMARKS

A new way of representing anime similarity has been proposed by applying the minimum spanning tree. Here, we used similarity measurement, called crew similarity, as an addition to the commonly used similarity measurements, namely scores and topics. The results show that using a minimum spanning tree, a similar anime can be obtained easily by looking at the graph. Moreover, we found that anime such as One Piece, Naruto, and Bleach are considered as the most significant anime based on the centrality calculations.

REFERENCES

[1] I. Herman, G. Melancon, and M. S. Marshall, “Graph visualization and navigation in information visualization: A survey,” IEEE Transactions on Visualization and Computer Graphics, vol. 6, pp. 24–43, Jan 2000.
[2] L. Moncla, M. Gaio, J. Nogueras-Iso, and S. Mustire, “Reconstruction of itineraries from annotated text with an informed spanning tree algorithm,” International Journal of Geographical Information Science, vol. 30, no. 6, pp. 1137–1160, 2016.
[3] M. K. Murmu, “A distributed approach to construct minimum spanning tree in cognitive radio networks,” Procedia Computer Science, vol. 70, pp. 166 – 173, 2015. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems.
[4] A. Crobe, M. Demuru, L. Didaci, G. L. Marcialis, and M. Fraschini, “Minimum spanning tree and k-core decomposition as measure of subject-specific eeg traits,” Biomedical Physics and Engineering Express, vol. 2, no. 1, p. 017001, 2016.
[5] V. MAICAN, “Minimum spanning tree algorithm on mapreduce one-chip architecture,” Romanian Journal of Information Science and Technology, vol. 18, no. 2, pp. 126–143, 2015.
[6] S. L. Gan and M. A. Djauthari, “New york stock exchange performance: evidence from the forest of multidimensional minimum spanning trees,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2015, no. 12, p. P12005, 2015.
[7] M. Vlachos and D. Svonava, “Recommendation and visualization of similar movies using minimum spanning dendrograms,” Information Visualization, vol. 12, no. 1, pp. 85–101, 2013.
[8] N. Jakoh, S. H. Weber, M.-C. Miller, and I. Gurevych, “Beyond the stars: Exploiting free-text user reviews to improve the accuracy of movie recommendations,” in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp. 57–64, 2009.
[9] Y. Shi, M. Larson, and A. Hanjalic, “Mining mood-similar movie similarity with matrix factorization for context-aware recommendation,” in Proceedings of the workshop on context-aware movie recommendation, pp. 34–40, ACM, 2010.
[10] G. Takacs, I. Pilaszy, B. Nemeth, and D. Tikk, “On the gravity recommendation system,” in Proceedings of KDD Cup Workshop at SIGKDD07, 13th ACM Int. Conf. on Knowledge Discovery and Data Mining, pp. 22–30, 2007.
[11] M. Fleischman and E. Hovy, “Recommendations without user preferences: a natural language processing approach,” in Proceedings of the 8th international conference on Intelligent user interfaces, pp. 242–244, ACM, 2003.
[12] J. Nieminen, “On the centrality in a graph,” Scandinavian journal of psychology, vol. 15, no. 1, pp. 332–336, 1974.
[13] P. Bonacich, “Power and centrality: A family of measures,” American Journal of Sociology, pp. 1170–1182, 1987.
[14] C. L. Freeman, “A set of measures of centrality based on betweenness,” Sociometry, pp. 35–41, 1977.
[15] C. L. Freeman, “Centrality in social networks conceptual clarification,” Social Networks, vol. 1, no. 3, pp. 215–239, 1978.
[16] S. C. North, Drawing graphs with NEATO, http://www.graphviz.org/pdf/neatoguide.pdf, 2004
[17] T. Kamada and S. Kawai, “An algorithm for drawing general undirected graph,” Information Processing Letters, vol. 31, pp. 7–15, 1989.
### TABLE I
Highest values of each centrality

| Rank | Degree   | Eigenvector | Betweenness | Closeness | Total   |
|------|----------|-------------|-------------|-----------|---------|
| 1    | One Piece| One Piece   | One Piece   | One Piece | One Piece |
| 2    | Naruto   | Naruto      | Naruto      | Naruto    | Naruto  |
| 3    | Aquarion Evol | Detective Conan | Bleach | Bleach | Bleach |

Fig. 3. The minimum spanning tree result.