Analysis and Implementation of Graph Clustering for Digital News Using Star Clustering Algorithm

A B Ahdi¹, K R SW¹, A Herdiani¹

Jl. Telekomunikasi Bandung 40287, West Java, Indonesia

E-mail: aufaba@yahoo.co.id, bagindokemas@telkomuniversity.ac.id, anisaherdiani@telkomuniversity.ac.id

Abstract. Since Web 2.0 notion emerged and is used extensively by many services in the Internet, we see an unprecedented proliferation of digital news. Those digital news is very rich in term of content and link to other news/sources but lack of category information. This make the user could not easily identify or grouping all the news that they read into set of groups. Naturally, digital news are linked data because every digital new has relation/connection with other digital news/resources. The most appropriate model for linked data is graph model. Graph model is suitable for this purpose due its flexibility in describing relation and its easy-to-understand visualization. To handle the grouping issue, we use graph clustering approach. There are many graph clustering algorithm available, such as MST Clustering, Chameleon, Makarov Clustering, and Star Clustering. From all of these options, we choose Star Clustering because this algorithm is more easy-to-understand, more accurate, efficient and guarantee the quality of clusters results. In this research, we investigate the accuracy of the cluster results by comparing it with expert judgement. We got quite high accuracy level, which is 80.98% and for the cluster quality, we got promising result which is 62.87%.

1. Introduction

The abundant information in Internet make difficult for user to find related news about specific topic without entering precise and detailed keywords in the search engine. To overcome this problem, we need a news clustering system which can grouping news based on the content and the characteristic of the news automatically. And for data modelling purpose, graph model is the righteous option because it can represent the news and the relation among the news into nodes and edges.

A grouping or clustering is a process in which objects are classified into groups called clusters [1]. The goal of clustering is to divide the data into specific clusters where each cluster has a certain resemblance or relation [2]. There are many methods of graph clustering such as MST Clustering, Chameleon, Makarov Clustering, and Star Clustering. We consider Clustering Star as the most appropriate, because it is easy to understand, accurate, efficient and guarantee the quality of the resulting cluster [3].

2. Theoretical Foundation

2.1 Graph Theory

Systematically graph G is defined as the set of pairs (V, E) where V is the set of nodes is not empty v1, v2, v3, ..., vn. While E is the set of edges connected between two nodes e1, e2, e3, ..., en So it can
be written $G = (V, E)$. Degree (degree) of a node is the number of side
by side with that node is denoted by $d(\nu)$. Examples of modeling graph can be seen in Figure 1.

![Figure 1. Example of graph](image)

Each node can be connected by an edge and a node can be connected to more than 1 node like example in figure 2 and figure 3.

![Figure 2. Node and edge](image)

2.2 Graph Clustering
Graph Clustering is a method of grouping a graph into several groups / clusters based on particular relevance. Making the cluster is done by cutting or discard useless edges by certain sizes [3]. Based on the literature, the clusters in graph is called community clusters [1]. The application of clustering on a graph unlike the usual clustering and has its own methods such as MST Clustering [3], Chameleon [3], Makarov Clustering [3], and Star Clustering [3].

2.3 Star Clustering
Star clustering algorithm is widely used because it is quite easy to understand, simple, and can produce precise cluster more efficiently. A star cluster in the clustering consisted of 1 centre node that is called star centre and other nodes that are connected to the star centre which is called the satellites [4]. Each satellite node may be connected to more than 1 star centre. Star clustering algorithm use user-defined
threshold for creating clusters where the edge of value below the threshold will be removed from the graph. The example of star clustering graph can be seen in figure 3. Following are the detailed of star clustering algorithm [4]:

- For each node, count the number of degree and sort the graphs according the number of degree (from the most into the least)
- Next, choose a node as the star centre by conducting examination towards each node that has the most number of degree and does not connected with other star centres.
- Repeat the previous process until we obtain all-star centre from the graph database.
- Lastly, create the clusters on the graph. Definition of a cluster in the star clustering algorithm is a collection of a star centre and its satellite nodes. If a node is connected to more than one satellite star centre node, then they will be joined with the star cluster centre which has the most weight towards the satellite node.

2.4 Cosine Similarity

Cosine Similarity is a calculation method to determine the weight / value relationship between two objects based on the cosine angle between these two objects [5]. This method is often used in text mining process. Cosine similarity formula is as follow:

\[
\text{Cosine Similarity} (d_1,d_2) = \frac{\text{dot product}(d_1,d_2)}{||d_1|| \times ||d_2||}
\]

Figure 4. Cosine Similarity Equation

D1 and D2 are the two document that will be determined. If two vectors have similar values which is approaching 1, it can be said there are similarities between the two vectors. To calculate the value of cosine similarity, we must first calculate the value of TF, IDF, and TFxIDF.

3. System Description and Testing Scenario

3.1 System Description

In this research, we will construct graph clustering system by utilizing Star Clustering algorithm, where each node in the graph represents a digital news in Indonesian language. The overview of System Description could be seen in Figure 4: Following are the explanation of Figure 4:

- The format of dataset is JSON. And the digital news is in Indonesian language. The words in dataset is processed so that become form of tokens that can be calculated using the cosine similarity relation.

Figure 5. Overview of System Description
Those token then processed using cosine similarity formula. This process produce weight values inter-document based on the relationship between news.

The next process is modelling the news data into a graph that has not been clustered yet. Each news will become a node and the cosine similarity weight become the weight of the edge.

Cluster the graph database using star clustering algorithm, forming a collection of news based linkages between news.

3.2. Testing Scenario
Following are the testing scenario that we use in our research:

- To know the effect of the number of news and threshold towards the value of intra cluster and inter cluster, we will perform several experiments to test the accuracy of intra cluster. Experiments were performed with four number of news that is 223, 167, 112, 56, and three kinds of threshold value which is 0.05, 0.1, and 0.15.
- To measure the accuracy of the clustering result, we will compare the clusters which is produced by our system with the grouping/clustering that is performed by an expert in Indonesian Language. Expert who perform clustering in this test is a journalists / private media reporter who graduates study program of Communication Studies, University of Indonesia. We calculate the accuracy by using F-measure, precision and recall for each cluster that is produced.

4. Discussion
4.1 Results from first scenario
Based on the experimental results, we got the percentage of 62.87% for the main data test which consist of 202 nodes and the threshold value that we use is 0.1. The experiment results could be seen in table 1. Following are our analysis toward the results:

- In general, the greater the amount of news, the less accurate the intra-cluster value comparison will be and vice versa, but there is a difference in the 0.15 threshold.
- The number of nodes and the number of clusters affect the accuracy of the results. If the number of nodes and clusters increase, then the accuracy value will also increase and if the number of nodes and clusters become smaller, then the accuracy value will also be smaller.
- At 0.15 threshold occurs different result because if the number of news abounded, the number of clusters which are formed are increasing rapidly with fewer nodes.
- At 0:15 threshold, we got good and consistent intra-cluster value which is above 70% for 112 nodes of the news.
- The threshold value has an effect to intra-cluster accuracy where the greater the threshold value, the greater accuracy value will be.
4.2 Results from second scenario

In the second scenario, we compare the cluster results that is produced by the system with the cluster results from the expert. The average accuracy value that we got is 80.98%. See Table 4.2 for more complete result.

| Number of Node | Threshold | Number of Cluster | Node with greater intracluster value | Node with greater intercluster value | Good Cluster Accuracy |
|----------------|-----------|-------------------|--------------------------------------|--------------------------------------|-----------------------|
| 54             | 0.05      | 10                | 43                                   | 12                                   | 79.62%                |
| 102            | 0.05      | 14                | 67                                   | 35                                   | 65.68%                |
| 150            | 0.05      | 11                | 72                                   | 78                                   | 48%                   |
| 202            | 0.05      | 19                | 72                                   | 130                                  | 35.64%                |
| 54             | 0.1       | 16                | 42                                   | 12                                   | 77.77%                |
| 102            | 0.1       | 17                | 71                                   | 31                                   | 69.60%                |
| 150            | 0.1       | 25                | 107                                  | 43                                   | 71.33%                |
| 202            | 0.1       | 33                | 127                                  | 75                                   | 62.87%                |
| 54             | 0.15      | 23                | 36                                   | 18                                   | 66.67%                |
| 102            | 0.15      | 28                | 76                                   | 26                                   | 74.5%                 |
| 150            | 0.15      | 41                | 114                                  | 36                                   | 72%                   |
| 202            | 0.15      | 54                | 153                                  | 49                                   | 75.74%                |
| 252            | 0.15      | 75                | 182                                  | 70                                   | 72.22%                |
| 297            | 0.15      | 84                | 215                                  | 82                                   | 72.39%                |

Table 4.1: Result of first scenario

Based on the generated number of clusters, the expert produced 33 clusters while the system generates 37 clusters, with different contents. Differences clusters results that occur due to some node that is associated based on common words, but they had different main topic, such as for example on the type
of news” kriminal” where there are many types of different clusters, but there are several nodes belonging to the wrong cluster because containing more same word with the star centre on another cluster than the star centre cluster it should be. This error occurs because many news containing more word like “korban” in comparison with its own special topics such as ”mesum“, “jambret”, ”curanmor”, etc. Another factor that affects the value of accuracy is the imperfect pre-processing process. This cause imperfect test data, and in return the clustering result become not 100% accurate.

5. Conclusion

- Digital news can be modelled using graph. This model is easy to understand and we could see the relationship between the news by making edge between nodes to illustrate the interconnectedness of news and rated the weight of each node.
- The star clustering algorithm can be applied in the news grouping case based on the linkages of the news content. This approach defines the node with biggest degree as the star centres and then grouped clusters based by linkages with its star centre.
- The star clustering algorithm can perform grouping of digital news based on linkage of the news content with easy to understand, fast, effective, has a pretty good degree of accuracy (62.87% accuracy for comparison of intra-cluster and 80.98% for the comparison with the clustering by expert)
- Grouping news using graph clustering is easy to understand because it produces good clusters visualization, so that the observer can see the linkage between clusters more easily by the colouring between nodes that have the same cluster.
- Results of star clustering algorithm relies heavily on the star centre that it formed
- The clusters that is formed in this research is not on the basis of topics commonly found on the grouping of digital news in the news media, but by the words that frequently appear on the news.

6. Acknowledgement

The research presented in this paper conducted by the School of Computing, Telkom University. This research was funded by research grant “Hibah Bersaing 2016”. We thank to the Ministry of Research Technology and Higher Education of the Republic of Indonesia (RISTEKDIKTI) for contributing to the success of this research.

7. References

[1] Feldman, Ronen, and James S. 2007. The Text Mining Handbook “Advanced Approaches in Analyzing Unstructed Data”. University Press, Cambridge
[2] Schaeffer, Satu E. 2007. “Graph Clustering”. Laboratory for Theoretical Computer Science, Helsinki University of Technology, Finland
[3] Wijaya, Derry T. 2008. Graph, Clustering and Applications. Depatment of Computer Science National University of Singapore, Singapore.
[4] Aslam, J A, Pelekhov K, and Rus D, 2004. The Star Clustering Algorithm for Information Organization.
[5] Sree, K. S. (2012). Clustering Based on Cosine Similarity Measure.