Digital News Graph Clustering using Chinese Whispers Algorithm

M F E Pratama, R S W Kemas and H Anisa
School of Computing, Telkom University, Jl. Telekomunikasi Bandung 40287, West Java, Indonesia
E-mail: fitrah.eka@gmail.com, bagindokemas@telkomuniversity.ac.id, anisaherdiani@telkomuniversity.ac.id

Abstract. As the exponential growth of news creation on the internet, the amount of digital news has reached out billion numbers. Digital news is naturally linked each other but it needs to be grouped so that user can easily classify the news that they read. Graph is the most suitable data model to represent digital news since it can describing relation in easy and flexible manner. Thus, to overcome grouping problems, in this paper we using Chinese Whispers Algorithm as the graph clustering approach. We choose Chinese Whisper Algorithm based on consideration that the algorithm is able to make clusters from a big graph data with a relatively fast process [8], that appropriate with the characteristics of digital news. In this research, we examine the graph quality by comparing intra and inter-cluster weights of every node. This scenario gives us a quite high result that 95% of nodes have intra-cluster weight higher than its inter-cluster weight. We also investigate the graph accuracy by comparing the cluster results with expert judgement. As the result, the average accuracy of digital news graph clustering using Chinese Whisper algorithm is 80%.

1. Introduction
Nowadays digital news is considered as a common thing for people in around the world. The online audience engaged with newspaper content reached a new peak in August 2015, totaling 179.3 million adult unique visitors. The count is a 10% increase in adult unique visitors measured by comScore since August 2014, and is double the growth rate for the Internet overall (5%). The phenomenon happens due to the ease of online news media to spreading the information, its flexible form, timeless accessibility, and borderless accessibility [1].

Digital news data that has been saved in the data storage has reached out ten billion to hundred billions of news. The data is so big that we need a simple representation to make data analytics easier. Graph model then implemented to represent the digital news because the model is more understandable and in line with human’s way of thinking. Besides, graph model can handle a huge scale of data and has semi structured characteristics [2][3] that comply with digital news data. In this graph, every node represents a news, and edges represent news’ connectivity based on its content.

News grouping based on its similarity or connectivity is necessary to help the readers found related articles, but it is hard to be performed if there are no explicit tags on the news or articles. To solve this problem, we use clustering approach. Clustering aims to make groups based on the element similarities [4][5]. Since we use graph to represent the news, we will perform a graph clustering
approach. Graph clustering is a process of nodes grouping, so that the node that have similar characteristics will be grouped in the same unit.

There are quite much graph clustering algorithm such as MST clustering, Chameleon, Makarov Clustering, and Star Clustering [7]. In this paper we use Chinese Whispers Algorithm based on consideration that the algorithm is able to make clusters from a big graph data with a relatively fast process [8], that appropriate with the characteristics of digital news.

2. Theoretical Foundation

2.1. Graph Theory

Theoretically, a graph consists of a group of vertex (V) and edge (E), and symbolized by G = (V, E). According to the edge direction of a graph, there are directed and undirected graph. Direct graph has a vertex used as a start and an end [9] meanwhile undirected graph has no distinction between the two vertex associated with each edge. In other words, undirected graph has bidirectional edges. Example of simple undirected graph can be seen on Fig. 1.

![Fig. 1. Example of simple undirected graph](image)

Besides directed and undirected graph, there is also exist the term of weighted graph. Weighted graph is a graph that has numerical label associated with its edge, called the weight of edge [8]. Fig 2 is an example of weighted graph.

![Fig. 2. Example of weighted graph](image)

2.2. Graph Clustering

Graph clustering is a process of node grouping so that every similar node gets into the same group. Some of researchers were also defined that graph clustering is an effort to grouping nodes in a graph so that every node in the same group have a high value of intra-connectivity and every node in different group have a low value of interconnectivity [6].

Graph clustering has been widely used in biology field especially in gen study, VLSI chip design, social networks, and others [10].

2.3. Chinese Whisper Algorithm

Chinese Whispers (CW) is an easy and effective algorithm in grouping/clustering a weighted, undirected graphs. The idea of CW came from a traditional children’s game, where every child has to spreads words or sentence by whispering it to another child consecutively. Several steps in CW algorithm based on [8] is as follow:

a. Every node in the graph is divided in different cluster. One node in one cluster.
b. Point one node randomly, then take that node into the strongest class in the neighborhood. This is the class whose total edge weights to the current node is maximal.

c. Repeat the process b for every node in the graph.

d. In case of multiple strongest classes, one is chosen randomly.

e. Regions of the same class stabilize during the iteration and grow until they reach the border of a stable region of another class. Note that classes are updated immediately: a node can obtain classes from the neighborhood that were introduced there in the same iteration.

2.4. Cosine Similarity

Cosine similarity is a method that often used in the text mining especially in the information retrieval and also clustering as well for measuring text similarity [11]. Cosine similarity measures the corner size between document vector (point \( (ax, bx) \)) and (point \( (ay, by) \)). Every vector that represents every word in every documents then compared and formed in a triangle. If there are two identical documents, it will form a zero degree \((0^\circ)\) angel so that the similarity’s value is one \((1)\); if two document are not identical at all, the angle that will be formed is 90 degree \((90^\circ)\) so that the similarity’s value is zero \((0)\) [12]. The formula of Cosine Similarity is as follow [13][14]:

\[
\text{Cosine Similarity} \ (\delta_1, \delta_2) = \frac{\text{dot product}(d_1 \cdot d_2)}{|d_1||d_2|} \tag{1}
\]

To prepare data for cosine similarity calculation, we have to do the term weighting by finding out the value of TF (Term Frequency), IDF (inverse-document frequency), and TF • IDF. Fig 3 represent the process for preparing data for Cosine Similarity calculation.

TF is weight of a term that occurs in a document, while IDF is a way of diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

The TF formula is as follow:

\[
\text{TF} (t_k, d_j) = f(t_k, d_j) \tag{2}
\]

With \((t_k, d_j)\) is the number of times that term \(k\) occurs in document \(j\).

The IDF formula is as follow:

\[
\text{IDF}(t_k) = 1 + \log \frac{N}{df(t_k)} \tag{3}
\]

With \(N\) is total number of documents in the corpus and \(df(t)\) is total number of documents in the corpus where the term \(t\) appears.

The TF • IDF formula is as follow:

\[
\text{TF} \cdot \text{IDF} (t_k, d_j) = \text{TF} (t_k, d_j) \cdot \text{IDF}(t_k) \tag{4}
\]

3. System Design

In this chapter we proposed the methodology we used to implement Chinese Whisper Algorithm to cluster the Indonesian digital news. Fig 4 depicting the system design.
Following is the description of the system design:

a. The first step is to prepare data by data preprocessing that consist of tokenization, case folding, stop word removal, and stemming. The raw data is digital news in.json format.

b. Preprocessed data is the input of the term weighting step (TF, IDF, TF • IDF) and cosine similarity calculation.

c. The results value of Cosine Similarity calculation is approximately between 0 and 1. To simplify the number, we normalized the value by multiplying the value by 10 and making it integers. Then, we defined that two documents are related if the value of similarity between 1 - 9.

d. After we know which documents are related one another, we convert the documents into graph. Every node in the graph gives information about the news such as author, title, date of released, content of news, and others. Meanwhile the value on the edges represents the similarity values.

e. All nodes then ready to be clustered using Chinese Whisper Algorithm.

4. Experiment
To measure whether the methodology satisfy our goal, we already conduct some experiments that follow the methodology we have proposed on the previous section.

Datasets. We use 258 Indonesian digital news captured from http://news.okezone.com in.json formats. One example of the news format can be seen in Fig. 5.

Scenarios. (1) In order to measure the graph quality, we examine the edge’s weight between node in the same cluster (intra-cluster value), and edge’s weight between node in different cluster (inter-cluster value). (2) In order to measure the accuracy of clustering result, we compare the clusters with news’ group that manually grouping by expert. The expert is a Bahasa Indonesia teacher, graduated from Indonesian Language and Literature Department.

To get a stable shape of clusters, we perform six (6) iterations of graph clustering based on steps on Chinese Whisper Algorithm.

5. Results and Discussion
5.1. Scenario 1
The result of the first iteration clusters shows that only 82.69% of its nodes have intra-cluster values higher than its inter-cluster values. It happened because the clusters shape was still not stable. The
second iteration shows the promising improvement percentage that is 92.79%. On the second iteration, the clusters are more stable than the previous one.

On the third and fourth iteration, we got even better percentage results that is more than 95%, but there are several nodes that it’s inter-cluster value higher than its intra-cluster because the cluster is still not stable enough. Meanwhile, for the 5th and 6th iteration, the percentage is also more than 95%, but the clusters have been stable because no nodes have higher point of inter-cluster value. The reason why we didn’t achieve 100% is because some nodes has the same value of inter-cluster and intra-cluster. Fig 6 illustrate the results of scenario 1.

5.2. Scenario 2

To perform scenario 2, we asked the expert to grouping all digital news on the dataset. As the result, there are 22 groups of news. Then we compare the result with clusters that we get after the 6th iteration from previous scenario. Table 1 shows the comparison results.

| Experiment | Number of clusters | Number of correct documents clustered | Number of incorrect documents clustered | Accuracy |
|------------|--------------------|--------------------------------------|----------------------------------------|----------|
| 1          | 19                 | 167                                  | 41                                     | 80.3%    |
| 2          | 18                 | 158                                  | 50                                     | 76.0%    |
| 3          | 19                 | 169                                  | 39                                     | 81.3%    |
| 4          | 19                 | 167                                  | 41                                     | 80.3%    |
| 5          | 20                 | 171                                  | 46                                     | 82.2%    |

Inconsistency of accuracy happens because every result of the clustering has different clusters result even though the total number of the clusters was similar. The main factor that differ one clustering result to another is the random phase. For example, experiment 1 and 3 both have 19 number of clusters, but the element of every cluster are different. That is the reason why level of accuracy is also different.

Another factor causing the accuracy not maximum was because of something not worked really well on the preprocessing phase. We investigated that the stemming process was not 100% correct. Thus it influenced the term weighting and overall clustering process. But after all, the average 80% accuracy is quite promising.

6. Conclusion

Transforming digital news into graph can answer the problem of news grouping. Every news document represented by node, meanwhile the similarity between document represented by the weight
of the edges. In this research, we use Cosine Similarity to perform graph clustering and we get quite high result on graph quality (95%) and clustering accuracy (80%).

Acknowledgements
This study is one of research roadmap of the Faculty of Informatics Telkom University, and has been funded by Ministry of Research, Technology and Higher Education of the Republic of Indonesia, through “Hibah Bersaing” grant.

References
[1] Husnul K, “Analisis faktor-faktor website quality dalam menggunakan situs berita digital kompas.com 2013”. Bandung: FEB Universitas Telkom, 2013.
[2] Isjhar K, “Analisis performansi metode graph decomposition index pada graph database”. Bandung, Indonesia: Fakultas Informatika Telkom University, 2014.
[3] Fahri F, Erwin Y H, and Ika N D, NoSQL: Latar Belakang, Konsep, dan Kritik. Semarang, Indonesia: Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, 2012.
[4] Yang Z, Hong C, and Jeffery X Y, Graph Clustering Based on Structural/Attribute Similarities. Hong Kong: Engineering Management University of Hong Kong.
[5] Satu Elisa S, Graph CLustering. Helsinki, Finland: Laboratory for Theoretical Computer Science, Helsinki University of Technology, 2007.
[6] Reena M, Shashwat S, Deepak A, and Mohit K, An Effective Comparison of Graph Clustering Algorithms via Random Graphs. India: Department of Computer Science and Engineering Amity University, 2011.
[7] Derry T W, Graph, Clustering and Applications. Singapore: Deptament of Computer Science National University of Singapore, 2008.
[8] Chris B, Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. Leipzig, Germany: University of Leipzig, NLP Department.
[9] Kenneth H R, Discrete Mathematics and Its Applications, Seventh Edition ed. New York: McGraw-Hill, 2007.
[10] Mousumi D and Shukla K K, Characteristic of Restricted Neighbourhood Search Algorithm and Markov Clustering on Modified Power-Law Distributor. India: Departement of Computer Engineering, 2012.
[11] Bhanu P A. and Venkata G R S., Space and Cosine Similarity measures for Text Document Clustering. Hyderabad, India: Vardhaman College of Engineering, 2013.
[12] Radiant V I, Adelia, Mewati A, and Alexander R, Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks. Bandung, Indonesia: Fakultas Teknologi Informasi Universitas Kristen Maranatha, 2014.
[13] Jana V, (2013, Oktober) Seeking Wisdom. [Digital]. https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/
[14] Jon B, Event-Centric Clustering of News Articles. Uppsala, Sweden: Department of Information Technology, Uppsala University, 2013.