Vol. 03, No. 02, pp. 112 –115 (2022)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Semantic-Based K-Means Clustering for IMDB Top 100 Movies

Niyaz M. Salih

Information System Engineering Department, Erbil Polytechnic University, 44001 Erbil, Iraq, niyaz.salih@epu.edu.iq

Abstract

Textual documents are growing rapidly through the internet in today's modern technology era. Electronic structured databases archive offline and online documents, e-mails, webpages, blog and social network posts. Without appropriate ranking and demand clustering when there is classification without any specifics, it is quite difficult to retain and access these documents. K-means is one of the methods that is frequently used for clustering. In terms of determining the proximity of meaning or semantics between data, the distance-based K-means method still has flaws. To get around this issue, semantic similarity can be estimated by measuring the level of similarity between objects in a cluster. This research provides a method for clustering documents based on semantic similarity. The approach is carried out by defining document synopses from the IMDB and Wikipedia databases using the NLTK dictionary, and we provide a semantic-based K-means clustering approach that assesses not only the similarity of the data represented as a vector space model with TFIDF, but also the semantic similarity of the data Precision, recall, and F-measure, we demonstrate how well the semantic-based K-means clustering technique works using experimental findings from the IMDB and Wikipedia top 100 movies datasets.

Keywords: *K-means Algorithm, Document Clustering, Semantic Similarity, TF-IDF*

Received: October 08th, 2022 / Accepted: December 28th, 2022 / Online: December 30th, 2022

I. INTRODUCTION

These days, using a search engine to get the answer to any issue is quite helpful and quick. By recognizing and resolving the issues at hand or getting information from a global existing knowledge, the Internet offers the quickest method of learning. When searching, we may find more unrelated information than related information, which can be annoying. As a result, all search engines use document clustering to display the results of searches in an efficient and organized way [1].

Without any previous information, clustering splits data into sections (clusters) that are relevant or helpful. Text clustering is a significant and fundamental piece of work in the field of processing information, which also serves as the foundational technology for biology, data mining, statistics, machine learning, and other crucial concerns in a variety of fields [2]. The complexity of the research problems and the amount of information and data available globally are both growing, placing more and more requirements on the present clustering methods. The conventional methods can be easily dragged into locally optimum solutions and are sensitive to different initial conditions. Additionally, there is increasing concern over the text clustering technique based on semantic similarity [3].

The semantic learning algorithm depends on texts with stronger meaning associations, paired with context to investigate deeper semantic information, which may achieve more accurate comparable values of two texts, and therefore provide more accurate clustering results obtained [4]. Currently, the majority of clustering algorithms used today are grounded on statistical models. For instance, the Distance measure or cosine distance vector is used in the clustering approach based on VSM to determine the association between texts. The main concept is to use information such as term frequency statistics to obtain component term weights and structured vectors. This method minimizes the clustering accuracy by ignoring the semantic association between terms and terms and between texts and texts [5].

Different researchers have proposed a variety of similarity metrics for the process of document clustering. The easiest is the Euclidean metric, which calculates the degree of similarity between texts using a distance function. The cosine similarity measurement, which utilizes documents defined in vector space with TFIDF and determines the angle between the vectors directionally of these documents, is another commonly used similarity metric. The semantics of the phrases (words) in the document are not taken into consideration by these two measurements. Researchers proposed using WordNet-based semantic similarity to determine semantic similarity; first, they extracted semantic words from texts and built a semantic class hierarchy of the phrases to determine semantic similarity based on similar conceptual hierarchy [6], [7].

In this research, we describe a semantic-based K-means clustering technique for grouping a huge number of movies that not only assesses the similarity between the movies expressed

doi:10.38094/jastt302138

by a vector space model with TFIDF but which also evaluates the semantic similarity between the movies using the NLTK dictionary. This research's remainder is arranged as follows: Section 2 reviews related work on movie clustering algorithms. Our proposed method is shown in Section 3. Section 4 presents the experimental outcomes. Finally, Section 5 summarizes the research by identifying future work directions.

II. Related Works

Strehl et al.[8] were the first to investigate the influence of similarity measurements on the process of clustering. They employed YAHOO datasets that had previously been classified by human specialists into several categories. They used these measurements to execute many different clustering methods in order to compare various similarity measures. They employed four commonly employed similarity metrics: Euclidean, Cosine, Pearson correlation, and Modified Jaccard. They used the following clustering techniques: Revised K-mean, self-organizing characteristics map, hyper-graph segmentation, and weighted graph grouping. They used analytical tests to confirm the importance of their experiment outcomes. The findings of the studies demonstrated that Extended Jaccard and Cosine similarity are quite similar to human-performed categorization results. This study's findings are equivalent to those of Strehl et al. In this study, a newly suggested similarity measure based on the topic map's description of the texts is examined in the same way as many of the best-performing similarity metrics of clustering documents. This research also builds on their previous work on the influence of similarity measurements on the clustering of extended datasets.

Wang and Koopman used a variety of clustering approaches based on the link within documents. The semantic link, on the other hand, was used to group the publications. The strategy presented in this research is to create the semantics of papers based on the participation of objects in those papers during the initial phase, as well as three vectors of descriptions for each document. The average vector is measured for all entities, and most elements but without quotes and only quotation elements. Once the document vectors are formed, the next step is to identify vector-based groups of papers applying (k-mean) clustering and group recognition using the Louvain (clustering technique Network based) [9].

Dhuria et al.[10] introduced a TVC technique that combines natural language processing (NLP) with ontology-based grouping to create semantically meaningful topics recorded in cluster words. However, since external knowledge, such as ontology, is too broad to evaluate similarity, these algorithms demand a high processing cost and a lengthy time for clustering.

The semantic-based concept processing was proposed by Shady at el. [11]. This method employs the semantic function labeler for each phrase, followed by a noise-filtering technique that eliminates frequent and less meaningful terms. The extracted idea words are used to show documents in a concise format. Their semantic similarity is computed using matched concepts and their hierarchies in the two documents.

Another technique worth mentioning is Ding's work [12]. He proposed the aspherical k-means method for k-means. This approach was used for document clustering, where the cluster centers are generated as concept vectors and merged to LSI index vectors. The study's principal result is that the multiplicity represented by the defining vectors is comparable to the feature space of the LSI.

Awajan used the MapReduce approach to conduct a K-means bisecting. The goal is to propose a system for resolving the clustering of focused data documents. To improve clustering, the bisecting K-means clustering technique is used with WordNet to get the semantic relationship between words. Elastic MapReduce is being used to evaluate the implementation of the bisecting k-means method. The inclusion of semantic connections in WordNet reduces the number of features, and the grouping of massive data became obvious as the number of dimensions decreased. WordNet lexical classifications are also used to enhance the measurement of internal findings [13].

We suggest a semantic-based K-means clustering technique to solve these issues. For grouping a huge volume of movies, our method analyzes semantic similarity between terms in movies using NLTK clustering and movie similarity using the vector space model.

III. Semantic-Based K-Means Clustering

As shown in Fig.1, picking the suitable clustering technique and text similarity measurement is significant, as it has a significant impact on the final clustering outcome. Text similarity calculations must go through word segmentation, stop word removal, feature selection, and preprocessing stages before creating a text representation model. Following that, we computed similarity using the semantic similarity calculation approach, constructed a semantic model, and constructed a similarity matrix in preparation for further clustering.

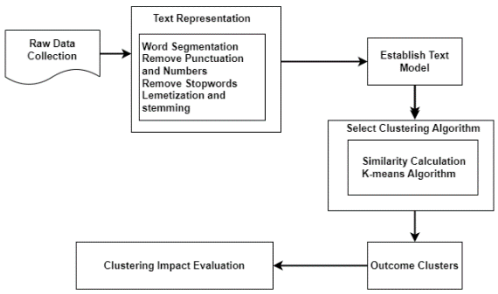

Fig. 1. Document clustering processes.

The clustering technique examines movie data from the pre-processing method in order to calculate the similarity of every movie and cluster the given movies. We present a semantic-based K-means clustering algorithm for clustering movies, which is a revised form of the K-means algorithm. The K-means technique is a powerful clustering algorithm that systematically classifies data into k clusters depending on a distance measure such as cosine similarity using a vector space modeling with TFIDF of inputs and the appropriate number of k clusters. Nevertheless, the K-means method does not take semantics into account while determining similarity, instead relying on cosine similarity. The technique considers the received movie data as vector space points and determines how tightly they are

associated. Furthermore, because movies are limited to 2000 words, they are difficult to evaluate for clustering. When clustering a great number of movies, this produces a low quality. To solve this issue, we use foreign technology to more precisely evaluate movie semantics for similarity determination.

IV. Experiments And Results Analysis

We empirically evaluated our proposed approaches for semantic-based K-means classification using the top 10 movies from the Wikipedia and IMDb datasets. The top ten movies from the Wikipedia and IMDb dataset were collected between December 2012 and March 2017 and cover 20 genres, including drama, romance, history, sport, war, music, and so on. My records are not dependent on the genuine brilliance and/or popularity of the someone, institution, or object being listed, not on my personal favorites. What you can see here is the result of an investigative process that starts with collecting synopses for the top 100 films of all time and ends with identifying the implicit subjects inside every document. We examine the efficacy of the semantic-based K-means clustering technique on these datasets. To cluster a big number of movies, we assigned k to 20 and clustered the 20 categories from the top 10 movies from Wikipedia and IMDb datasets.

Table I. presents the complete findings of semantic K-means clustering on the top 10 movies from the Wikipedia and IMDb datasets, including precision, recall, and F-measure statistics for every semantic K-means group. Apart from the eighteenth cluster, our proposed clustering technique achieves usually acceptable results. We'll illustrate why and how this cluster had such a poor result later. The fourth and eleventh clusters have perfect precision. The sixth, fourteenth, and twentieth clusters all had 100% recall. In conclusion, the top 10 movies from the Wikipedia and IMDb datasets were distributed accurately, with 85.2% precision, 90.5% recall, and 87.77% F-measure.

When compared to the original K-means, the recall and precision of our suggested clustering method improved from 7.52% to 14.47% (10.73% on average). In regards to recall, the suggested approach outperformed the competition by an average of 13.87%. The sixteenth cluster looked excellent in terms of recall and precision. Overall, the new clustering approach outperformed the original technique in terms of recall, precision, and F-measure. This suggests that the suggested semantic-based K-means algorithm can conduct clustering more successfully for a huge number of movies.

TABLE I. Semantic-based K-means clustering Results (K=20)

no. of Clusters	Precision (%)	Recall (%)	F-Measure (%)
1	88	96.05	91.85
2	95	76.9	85.00
3	83	93.1	87.76
4	100	99	99.50
5	99	100	99.50
6	85	97	90.60
7	89	89	89.00
8	62	96.6	75.53
9	91	97.2	94.00
10	99	96.8	97.89
11	100	95.7	97.80
12	74	89	80.81
13	85	92.1	88.41
14	96	100	97.96
15	80	84.9	82.38
16	86	96.6	90.99
17	76	78	76.99
18	38	38.3	38.15
19	94	93.7	93.85
20	84	100	91.30

V. Conclusion

In this research, we introduced a semantic-based K-means clustering technique that uses the NLTK dictionary to efficiently cluster a huge number of documents such as synopses. The semantic-based K-means clustering technique takes into consideration not only the similarity of terms in synopses described by a vector space model with TFIDF but as well as the semantic similarity of terms in a document via semantic development using the NLTK dictionary. Whereas a synopsis has limited information, semantic expanding to add related words to a collection is critical for understanding the synopsis more appropriately. Our suggested synopsis clustering improves the existing method in studies on the top ten movies from the Wikipedia and IMDb datasets. The selection of a clustering technique is frequently followed by the selection of a similarity computation method. As a result, the appropriate computation approach is critical in text clustering. In this research, semantic word clustering outperforms the traditional technique; the semantic strengthen approach increases processing efficiency. In the future, we may investigate a more powerful semantic-based k-means clustering technique that is independent of domains and subjects.

Acknowledgment

I would like to express my heartfelt gratitude to Dr. Shahab Wahab for his invaluable and helpful comments during the conception and development of this paper work. His willingness to give so generously of his time has been greatly appreciated.

References

[1] S. Mohammed, K. Jacksi, and S. Zeebaree, “A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, pp. 552–562, Apr. 2021, doi: 10.11591/ijeecs.v22.i1.pp552-562.

[2] S. AFREEN and D. B. SRINIVASU, “SEMANTIC BASED DOCUMENT CLUSTERING USING LEXICAL CHAINS,” 2017.

[3] N. M. Salih and K. Jacksi, "State of the art document clustering algorithms based on semantic similarity," *Jurnal Informatika*, vol. 14, pp. 58–75, May 2020, doi: 10.26555/jifo.v14i2.a17513.

[4] R. Ibrahim *et al.*, *Clustering Document based on Semantic Similarity Using Graph Base Spectral Algorithm*. 2022, p. 259. doi: 10.1109/IICETA54559.2022.9888613.

[5] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, "The K-Means Clustering Algorithm With Semantic Similarity To Estimate The Cost of Hospitalization," *Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 4, Art. no. 4, Oct. 2019, doi: 10.22146/ijccs.45093.

[6] E. M. B. Nagoudi, J. Ferrero, D. Schwab, and H. Cherroun, "Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences," in *Arabic Language Processing: From Theory to Practice*, vol. 782, Cham: Springer International Publishing, 2018, pp. 19–33. doi: 10.1007/978-3-319-73500-9_2.

[7] N. M. Salih and K. Jacksi, *Semantic Document Clustering using K-means algorithm and Ward's Method*. 2021. doi: 10.1109/ICOASE51841.2020.9436588.

[8] M. Rafi and M. S. Shaikh, "An improved semantic similarity measure for document clustering based on topic maps." arXiv, Mar. 17, 2013. doi: 10.48550/arXiv.1303.4087.

[9] S. Wang and R. Koopman, "Clustering articles based on semantic similarity," *Scientometrics*, vol. 111, pp. 1017–1031, 2017, doi: 10.1007/s11192-017-2298-x.

[10] S. Dhuria, H. Taneja, and K. Taneja, "NLP and ontology based clustering — An integrated approach for optimal information extraction from social web," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 1765–1770.

[11] S. Shehata, "A WordNet-Based Semantic Model for Enhancing Text Clustering," in *2009 IEEE International Conference on Data Mining Workshops*, Dec. 2009, pp. 477–482. doi: 10.1109/ICDMW.2009.86.

[12] C. Ding, "A Probabilistic Model for Latent Semantic Indexing," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 65–74, Apr. 2005, doi: 10.1002/asi.20148.

[13] A. Awajan, "Semantic Similarity Based Approach for Reducing Arabic Texts Dimensionality," *International Journal of Speech Technology*, Jun. 2015, doi: 10.1007/s10772-015-9284-6.