Exploring Clustering Algorithms and Their Application to Text Clustering

Nguyễn Minh Quang
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student489@ictu.edu.vn

***Abstract*—This thesis explores the application of clustering algorithms to text data, focusing on the evaluation of various algorithms such as K-means, hierarchical clustering, DBSCAN, LSA, and LDA. The author provides an overview of text mining and clustering methods, including preprocessing steps and feature extraction techniques. The work aims to understand the performance of these algorithms in the context of text data and their potential applications in specific tasks like classification, summarization, and content recommendation. The thesis does not provide specific numerical results or measured improvements, but it discusses the effectiveness of the clustering algorithms in forming coherent clusters.**

***Index Terms*—clustering algorithms, text clustering, feature extraction, K-means, LDA**

I. Introduction

The thesis introduces the concept of text mining and clustering, discussing existing research and methods. It covers preprocessing steps, feature extraction techniques, and various clustering algorithms. The author also explores the challenges and applications of text clustering. The introduction provides a comprehensive overview of the background and context of the research, highlighting the importance of text clustering in various domains such as information retrieval and natural language processing [1]. Earlier ICTU student work on related designs is included here for context [7]. Related published studies provide context for the subject and methods [8], [9], [10]. These contextual works were selected for this English summary; they are not asserted to appear in the original thesis bibliography.

II. Methods

The methods section details the preprocessing steps, feature extraction using Bag-of-Words and TF-IDF, and the application of clustering algorithms including K-means, hierarchical clustering, DBSCAN, LSA, and LDA. The author describes the process of clustering text data and evaluates the performance of different algorithms. The preprocessing involves tokenization, stop-word removal, and stemming. Feature extraction techniques such as Bag-of-Words and TF-IDF are used to convert text data into numerical vectors. The clustering algorithms are applied to the extracted features to form clusters. The author does not provide specific numerical results or measured improvements, but discusses the effectiveness of the algorithms in forming coherent clusters [1].

Supervisor: Phạm Thị Liên, Thai Nguyen University of Information and Communication Technology (ptlien@ictu.edu.vn).

III. Results

The results section presents the evaluation of clustering algorithms in the context of text data. The author compares the performance of different algorithms but does not provide specific numerical results. The effectiveness of the clustering algorithms is discussed in terms of their ability to form coherent clusters. The author notes that the performance of the algorithms may vary depending on the specific context and goals. The results indicate the need for further analysis to determine the best approach for specific tasks [1].

IV. Discussion

The discussion section analyzes the structure of the formed clusters to assess their coherence and consistency. The author questions whether the clusters are meaningful and applicable in real-world scenarios, considering their relevance to specific topics or events in the data. The potential applications of the clustering results in tasks such as classification, summarization, and content recommendation are also discussed. The author suggests that the effectiveness of the algorithms may vary depending on the specific context and goals. The discussion highlights the challenges in interpreting the results and the need for further research to improve the accuracy and relevance of the clusters [1].

V. Conclusion

The thesis concludes by evaluating the performance of clustering algorithms and their potential applications. The author suggests that the effectiveness of the algorithms may vary depending on the specific context and goals. The results indicate the need for further analysis to determine the best approach for specific tasks. The conclusion emphasizes the importance of considering the context and goals when applying clustering algorithms to text data. The author also highlights the need for more research to improve the accuracy and relevance of the clusters [1].

References

[1] Nguyễn Minh Quang, ”Exploring Clustering Algorithms and Their Application to Text Clustering,” internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2024.

[2] J. A. Teixeira da Silva, P. Tsigaris, and M. J. S. Erfanmanesh, ”Publishing volumes in major databases related to Covid-19,” vol. 126, pp. 831-842, 2021
[3] C. C. Aggarwal, Y. Zhao, S. Y. J. I. T. o. k. Philip, and d. engineering, ”On the use of side information for mining text data,” vol. 26, no. 6, pp. 1415- 1429, 2012
[4] C. Faloutsos and D. W. Oard, ”A survey of information retrieval and filtering methods,” 1998
[5] R. Socher et al., ”Recursive deep models for semantic compositionality over a sentiment treebank,” in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631-1642
[6] J. Bollen, H. Mao, and A. Pepe, ”Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena,” in Proceedings of the international AAAI conference on web and social media, 2011, vol. 5, no. 1, pp. 450-453
[7] ICTU Student 171, ”A Study on Some Clustering Algorithms in Data Mining,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2010.
[8] R. J. Wadnare et al., ”Development of Text Clustering Method with K-Means for Analysis of Text Data,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2021. doi:10.32628/cseit217237.
[9] Ahmed Adeeb Jalal et al., ”Text documents clustering using data mining techniques,” International Journal of Electrical and Computer Engineering (IJECE), 2021. doi:10.11591/ijece.v11i1.pp664-670.
[10] Peng Yang et al., ”Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering,” IEEE Access, 2020. doi:10.1109/access.2020.2969525.