Research on Vietnamese Text Classification

ICTU Student 4
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam

***Abstract*—This thesis explores the problem of text classification in Vietnamese, focusing on the introduction of .NET and C#. The author discusses various text representation models such as vector space models and the removal of stop words, and presents experimental results for k-NN and ID3 algorithms. The work aims to contribute to the field of text data mining, particularly in the context of Vietnamese text processing. The author outlines the structure of the thesis, which includes chapters on text representation methods and classification algorithms. The study highlights the importance of handling Vietnamese text complexities and user customization, but also notes limitations, particularly in term-based text processing. The thesis covers the origins and architecture of .NET, the features of C#, and the basics of text classification. Experimental results show that k-NN and ID3 algorithms can be effective in text classification, but k-NN showed better performance in certain scenarios. The author suggests that the use of fuzzy theory and machine learning algorithms could improve the performance of the text classification system.**

***Index Terms*—text classification, vector space model, k-NN, ID3, fuzzy theory, machine learning**

I. Introduction

The thesis begins with an introduction to .NET and C#, covering the origins and architecture of .NET, and the features of C#. It then delves into the basics of text classification and the importance of this task in data mining. The author outlines the structure of the thesis, which includes chapters on text representation methods and classification algorithms. The introduction provides a comprehensive overview of the .NET framework, including its origins, architecture, and key components such as the Common Language Runtime (CLR) and the .NET Framework libraries. The discussion on C# covers its features, benefits, and comparisons with other programming languages. The importance of text classification is emphasized, particularly in the context of handling Vietnamese text complexities and user customization. [1]

II. Methods

The methods section covers text representation models such as vector space models and the removal of stop words. It also discusses classification algorithms like k-NN and ID3, and their application in text classification tasks. The vector space model is described in detail, including Boolean, frequency, and sparse vector models. Stop words are removed to improve the accuracy of the text representation. The k-NN and ID3 algorithms are explained, with k-NN being used for nearest neighbor classification and ID3 for decision tree classification. The implementation of these algorithms is discussed, focusing on their application in text classification tasks. The methods section also outlines the experimental setup, including the selection of datasets and the evaluation metrics used to assess the performance of the algorithms. [1]

Supervisor: Nguyễn Đình Dũng, Thai Nguyen University of Information and Communication Technology (nddung@ictu.edu.vn).

III. Results

Experimental results are presented, showing the performance of k-NN and ID3 algorithms in text classification tasks. The k-NN algorithm was tested on a dataset of Vietnamese text, and the results demonstrated its effectiveness in handling term-based text processing. The ID3 algorithm was also evaluated, and its performance was compared with k-NN. The results indicate that both algorithms can be effective in text classification, but k-NN showed better performance in certain scenarios. The author plans to further develop the topic using fuzzy theory and machine learning algorithms. The experimental setup involved the use of a demo program to test the algorithms, and the results were analyzed to identify strengths and weaknesses of the approaches. [1]

IV. Discussion

The discussion highlights the advantages of the proposed approach, such as its ability to handle Vietnamese text complexities and user customization. However, it also notes limitations, particularly in handling term-based text processing. The author suggests that the use of fuzzy theory and machine learning algorithms could improve the performance of the text classification system. The discussion emphasizes the importance of user customization in the context of Vietnamese text processing, as it allows for more personalized and accurate classification. The limitations of the current approach are acknowledged, particularly in the handling of term-based text processing, which can be challenging due to the complexities of the Vietnamese language. [1]

V. Conclusion

The thesis concludes with a summary of the achieved results and future research directions, emphasizing the need for further development in fuzzy theory and integration with internet systems. The author highlights the importance of user customization in the context of Vietnamese text processing, as it allows for more personalized and accurate classification. The conclusion also notes the limitations of the current approach, particularly in the handling of term-based text processing, which can be challenging due to the complexities of the Vietnamese language. The author suggests that future research

should focus on improving the performance of the text classification system through the use of fuzzy theory and machine learning algorithms. [1]

References

[1] ICTU Student 4, "Research on Vietnamese Text Classification," undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, n.d.

[2] Fabrizio Sebastiani, March 2002 "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1,. (Site: http://faure.iei.pi.cnr.it/~fabrizio/Publications/ACMCS02.pdf)

[3] C. Apte, F. Damerau, and S.M. Weiss, 1994 "Automated Learning of Decision Rules for Text Categorization", ACM Transactions on Information Systems,. (Site: http://www.research.ibm.com/dar/papers/pdf/tois94_with_cover.pdf)

[4] D. Lewis and M. Ringuette, (1994) "A Comparison of Two Learning Algorithms for Text Categorization", In 3rd Annual Symposium on Document Analysis and Information Retrieval, pp.81-93. (Site: http://citeseer.nj.nec.com/lewis94comparison.html)

[5] D.E. Johnson, F.J. Oles, T. Zhang and T.W. Goetz, (2001, Oct.) "A Decision-Tree-Based Symbolic Rules Induction System for Text Categorization", IBM Research Report

[6] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar, 1999 "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification". (Site: http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/wknn.pdf)

[7] D. Lewis, (1991) "Representation and Learning in Information Retrieval", PhD Thesis, Graduate School of the University of Massachusetts. (Site: http://citeseer.nj.nec.com/lewis91learning.html)

[8] T. Joachims, (1998) "Text Categorization with Support Vector Machines: Learning with many Relevant Features", Proceedings 10th European Conference on Machine Learning (ECML), pp.137-142 (Site: http://citeseer.nj.nec.com/joachims98text.html)

[9] Ilya Baraev, November 2000 "Information Retrieval: Standard and AI-based methods",

[10] T.M. Mitchell (1997), "Machine Learning", Mc Graw Hill, pp 52-76