Application of Genetic Algorithm in Text Classification of Vietnamese Language

ICTU Student 197
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student197@ictu.edu.vn

***Abstract*—This thesis explores the application of genetic algorithms in text classification for the Vietnamese language. It covers the basics of C# and SQL Server, and delves into genetic algorithms, including their theoretical foundations, operators, and implementation in a text classification program. The author successfully implemented a genetic algorithm for Vietnamese text classification, but acknowledges limitations due to limited experience and time constraints. The study also discusses alternative methods such as CRFs and MIRA, and the challenges of text segmentation in Vietnamese. The author seeks feedback to improve the report.**

***Index Terms*—genetic algorithm, text classification, C#, SQL Server, Vietnamese, optimization**

I. Introduction

The thesis aims to find optimal solutions for complex problems in computer science, focusing on genetic algorithms. These algorithms are used to reduce search space and improve efficiency in large search spaces, but traditional search methods are limited in their effectiveness. The author explores the application of genetic algorithms in text classification for the Vietnamese language, a task that requires handling the unique challenges of the Vietnamese language, such as text segmentation and the lack of a standardized alphabet. The study covers the basics of C# and SQL Server, which are used for the implementation of the genetic algorithm [1]. Related earlier ICTU undergraduate theses provide context for this English summary [8]. Published articles on the methods and subject of this project provide context for this English summary [9]. Published books on the methods and subject of this project provide context for this English summary [10].

II. Methods

The thesis details the implementation of genetic algorithms, including crossover and mutation operators, and their application to text classification. The author uses C# and SQL Server for the implementation. The genetic algorithm is structured with a binary encoding, roulette wheel selection, and K-point crossover. The author also discusses alternative methods such as CRFs and MIRA, which are used for part-of-speech tagging and named entity recognition. The text classification program is designed to handle Vietnamese text, which poses unique challenges due to its complex structure and lack of a standardized alphabet [1].

Supervisor: Trần Hải Thanh, Thai Nguyen University of Information and Communication Technology (ththanh@ictu.edu.vn).

III. Results

The author successfully implemented a genetic algorithm for Vietnamese text classification, but did not report any numerical evaluations or performance metrics. The program was able to classify text into predefined categories, but the accuracy and efficiency of the classification were not quantitatively assessed. The implementation faced challenges due to the complexity of the Vietnamese language, particularly in text segmentation and the lack of a standardized alphabet. The author acknowledges the need for further optimization and improvement [1].

IV. Discussion

Despite the author's efforts, the project faced limitations due to insufficient experience and time, leading to some unoptimized aspects. The author seeks feedback to improve the report. The study highlights the importance of handling the unique challenges of the Vietnamese language, such as text segmentation and the lack of a standardized alphabet. The author suggests that future work could focus on improving the accuracy and efficiency of the text classification program through more rigorous testing and optimization [1]. Earlier ICTU undergraduate work by La Ngọc Tùng addressed genetic algorithm applied to optimization [7].

V. Conclusion

The author successfully implemented a genetic algorithm for Vietnamese text classification, but acknowledges the need for further optimization and improvement. The study covers the basics of C# and SQL Server, and delves into genetic algorithms, including their theoretical foundations, operators, and implementation in a text classification program. The author seeks feedback to improve the report and suggests that future work could focus on improving the accuracy and efficiency of the text classification program through more rigorous testing and optimization [1].

References

[1] ICTU Student 197, "Application of Genetic Algorithm in Text Classification of Vietnamese Language," internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2010.
[2] TS. Nguyễn Đình Thúc (chủ biên), Lập trình tiến hóa, NXB Giáo dục, 2001
[3] Phạm hữu khang (chủ biên), C# 2005 - Lập trình hướng đối tượng, NXB Lao Động Xã Hội, 2005

[4] Phạm Hữu Khang (chủ biên), C# 2005 – Lập trình Windows Forms, NXB Lao Động Xã Hội, 2005
[5] TS. Đinh Mạnh Tường, Cấu Trúc Dữ Liệu Và Giải Thuật, NXB Khoa Học Và Kỹ Thuật, 2001
[6] Melanie Mitchell, An introduction to genetic algorithms, 1999
[7] La Ngọc Tùng, "Application of Genetic Algorithm in Transportation Problem," internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2008.
[8] ICTU Student 4, "Research on Vietnamese Text Classification," internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2007.
[9] Fuchun Peng, Xiangji Huang, "Machine learning for Asian language text classification," Journal of Documentation, vol. 63, no. 3, pp. 378-397, 2007. DOI: 10.1108/00220410710743306.
[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, 2nd ed. MIT Press, 2001. ISBN 9780262032933.