

Research on Vietnamese text classification problem

nan

Abstract: A research and implementation project on building a solution for the text classification problem in Vietnamese. The first part introduces the .NET framework and C# language, followed by clarifying basic concepts of text classification and information representation methods such as the vector space model (Boolean, TF-IDF frequency). The author analyzes popular classification algorithms such as decision trees (ID3), K-NN (K-nearest neighbors), and proposes a specific solution direction for Vietnamese text, including model selection, term segmentation approach, and stop word removal techniques. The project successfully implemented a test program in C# to evaluate and compare the performance of the algorithms.

Keywords: Text classification, natural language processing, Vietnamese, K-NN, decision tree, TF-IDF

References

- [1] Damerau, and S.M. Weiss, 1994 “Automated Learning of Decision Rules for Text Categorization”, ACM Transactions on Information Systems.
- [2] 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.
- [3] D. Lewis, (1991) “Representation and Learning in Information Retrieval”, PhD Thesis, Graduate School of the University of Massachusetts.
- [4] 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.
- [5] Ilya Baraev, November 2000 “Information Retrieval: Standard and AI-based methods”.
- [6] T.M. Mitchell (1997), “Machine Learning”, Mc Graw Hill, pp 52-76.