Region-Based Distance Analysis of Keyphrases: A New Unsupervised Method for Extracting Keyphrases Feature from Articles

Mohammad Badrul Alam Miah
Faculty of Computing, Universiti Malaysia Pahang, Pekan, Malaysia
badrul.ict@gmail.com

Suryanti Awang
Faculty of Computing, Centre For Data Science & Artificial Intelligence (Data Science Centre),
Soft Computing & Intelligent Systems (SPINT)
Universiti Malaysia Pahang, Pekan, Malaysia
suryanti@ump.edu.my

Md. Saiful Azad
Computer Science and Engineering (CSE), Green University of Bangladesh (GUB)
Dhaka, Bangladesh
chairman@cse.green.edu.bd

ABSTRACT

Due to the exponential growth of information's and web sources, Automatic keyphrase extraction is still a challenging issue in the current research area. Keyphrases are very helpful for several tasks in natural language processing (NLP) and information retrieval (IR) systems. Feature extractions for those keyphrases execute a vital role in extracting the top-quality keyphrases and summarising the documents at a superior level. This paper proposes a new region-based distance analysis of keyphrases (RDAK) unsupervised technique for feature extraction of keyphrases from articles. The proposed method comprises six phases: data acquisition and preprocessing, data processing, distance calculation, average distance, curve plotting, and curve fitting. At first, the system inputs the collected different datasets to the preprocessing step by employing some text preprocessing techniques. Afterwards, the preprocessed data is applied to the data processing phase, and then after distance calculation, it is passed to the region-based average calculation process, then curve plotting analysis, and afterwards, the curve fitting technique is utilized. Finally, the proposed system has tested and evaluated the performance through implementing them on benchmark datasets. The proposed system will significantly improve the performance of existing keyphrase extraction techniques.

KEYWORDS: Distance analysis, Region-based distance analysis, Data processing, Feature extraction, Keyphrase extraction technique, Goldkey

ACKNOWLEDGMENT

We want to express our gratitude to the Universiti Malaysia Pahang (UMP), Malaysia, to support this study under grant number RDU192210.

REFERENCES

[1] C. Sun, L. Hu, S. Li, T. Li, H. Li, and L. Chi, “A review of unsupervised keyphrase extraction methods using within-collection resources,” Symmetry, vol. 12, no. 11, p. 1864, 2020.

[2] Z. A. Merrouni, B. Frikh, and B. Ouhbi, “Automatic keyphrase extraction: a survey and trends,” Journal of Intelligent Information Systems, pp. 1–34, 2019.

[3] N. S. M. Nafis and S. Awang, “An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification,” IEEE Access, vol. 9, pp. 52177–52192, 2021.

[4] Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, and L. Panpan, “A graph-based approach of automatic keyphrase extraction,” Procedia Computer Science, vol. 107, pp. 248–255, 2017.

[5] Z. A. Merrouni, B. Frikh, and B. Ouhbi, “Automatic keyphrase extraction: An overview of the state of the art,” in 2016 4th IEEE international colloquium on information science and technology (CiSt), pp. 306–313, IEEE, 2016.