A Flexible Keyphrase Extraction Technique for Academic Literature

Gollam Rabby a, Saiful Azad a,b, Paolo Casari c, Kamal Z. Zamlia b, Mohammed Mostafizur Rahman d

a Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, Gambang, Kuantan, Malaysia.

b IBM Centre of Excellence, Gambang, Kuantan, Malaysia.

c IMDEA Network Institute, Madrid, Spain.

d Department of Mathematics, American International University - Bangladesh (AIUB), Dhaka, Bangladesh.

Abstract:
A keyphrase extraction technique endeavors to extract quality keyphrases from a given document, which provide a high-level summary of that document. Except statistical keyphrase extraction approaches, all other approaches are either domain-dependent or require a sufficient amount of training data, which are rare at present. Therefore, in this paper, a new tree-based automatic keyphrase extraction technique is proposed, which is domain-independent and employs nominal statistical knowledge; but no train data are required. The proposed technique extracts a quality keyphrase through forming a tree from a candidate keyphrase; and later, it is expanded or shrunk or remained in the same state depending on other similar candidate keyphrases. At the end, keyphrases are extracted from the resultant trees based on a value, μ (which is the Maturity Index (MI) of a node in the tree), which enables flexibility in this process. A small μ value would yield many and/or lengthy keyphrases (greedy approach); whereas, a large μ value would yield lower and/or abbreviated keyphrases (conservative approach). Thereby, a user can extract his/her desired-level of keyphrases through tuning μ value. The effectiveness of the proposed technique is evaluated on an actual corpus, and compared with Rapid Automatic Keyphrase Extraction (RAKE) technique.

Keywords: Candidate keyphrase, keyphrase, automatic keyphrase extraction technique, tree data structure.