Hierarchical Keyword Extraction Algorithm using Text Social Network

[redacted]

Abstract. The Keywords is most basic and efficient element for document classification and information retrieval. But from a the reader of the documents, as the keywords connote the most abridged information of the document, it is most important feature for understanding the document. At this starting point of grasp of the document, we suggest the keyword extraction algorithm based on graph and reveal the result with the way of easy to understand visually. Using the information of location in document for each word, we estimate the influence of the word and quantify the relationship of domination by computing the degree of overlapping between each influence. We analyze the element of graph theory in the maximum spanning forests graph on the basis of this relationship information, we can visualize the relationship between each pair of word and extract the core keywords. As a result of this extraction process, we confirmed that this method of performance is improved than extraction based on only the frequency.

Keywords: Keyword Extraction, Document Analysis, Document Visualization, Graph Theory

1 Introduction

As more information need to be acquired from the electronic documents, more efficient ways of document classification, retrieval and summarization are necessary. In this series of information processing, the keywords of which present the integrated information of document are the most important material. But the extraction of the keywords from documents by human such as the writer of document or others needs a lot of effort, so that the automatical extraction method is actively studied. Most research on the keyword extraction focus on processing a lot of information efficiently, the way of expression for result of keyword extraction ended up in enumeration of the keywords ordered by importance. On the other hand, in this paper, we not only suggest the efficient way for processing a lot of document and comparing the importance among words, but also visualize the result of extraction and present the relationship between each keyword in the light of understanding the document.