A Combining Approach to Automatic Keyphrases Indexing for Chinese News Documents[1]

Wang Houfeng, Li Sujian, Yu Shiwen and Kang Byeong Kwu

Department of Computer Science and Technology
School of Electronic Engineering and Computer Science
Peking University, Beijing, 100871, China
{wanghf,yusw,lisujian,kbg43}@pku.edu.cn

Abstract. In this paper, we present a combinational approach to automatically supplying keyphrases for a Chinese news documen. In particular, we discuss some factors that have an effect on forming an initial set of keyphrase candidates and filtering unimportant candidates out from the initial set, as well as selecting the best items from the set of the remaining candidates. Experiments show that the approach reaches a satisfactory result.

1 Introduction

In this paper, we present a combinational approach to automatically supplying keyphrases for Chinese news documents.

Our approach is not the same as a pure keyphrase extractor. An extractor only extracts keyphrases from a source document. Supervised machine learning methods [2][4][5] and string-frequency method [1][3] are frequently used in extractors. However, the machine learning methods require a large amount of training documents with known keyphrases. Furthermore, for Chinese texts, in which there is no boundary between words except punctuation, words and phrases need to be recognized before the machine learning methods are availably applied. This is still considered as a difficult question in Chinese Processing. String-frequency method tries to avoid Chinese word segmentation. However, the method is not able to extract a valid phrase that does not occur sufficiently frequently and a resultant string as keyphrase even cannot be ensured as a clear meaning unit. One way of solving the problems is to select keyphrases from a controlled thesaurus. Fortunately, *People Daily* News Agency in China provided us with this thesaurus. We thus can gain some keyphrases by transforming extracted candidates into canonical terms according to this thesaurus, an abbreviation dictionary and a synonymous term dictionary.

Our approach is different from pure assignment tools as well, because our approach will directly extract some keyphrases from an article as long as they are thought as important based on our score strategies.

[1] This work is partially funded by National Natural Science Foundation of Chinese (grant No. 60173005)