University of Macau

PAT-Tree and Local statistical information Based Chinese Phrase Extraction

By Yang YiYang

Thesis Supervisor: Associate Professor, Gong ZhiGuo

Master of Science in Software Engineering

ABSTRACT

At present, a mass of data are generated on the Web each day, how to process them and extract some related and useful information becomes urgent. On the other hand, instead of English, more and more internet users prefer to write articles and documents in their native languages, as the result, the data in Asian languages such as Chinese (both simplified and traditional) and Japanese are really an important matter in information related researches. This thesis has its root in Chinese Phrases Extraction, the experiment resource focus on the Chinese documents, and its researches can easily be extended to other similar languages.

Information retrieval is popular topic in recent years. Most of the existing information systems are developed based on English; the drawback is that it is difficult for them to process Chinese Phrases appropriately. Handling the Chinese language structure is an important issue for this research.

Information retrieval systems also suffer from Out-of-Vocabulary (OOV) problem; especially for those dictionary-based applications such as word segmentation, text classification and so on. Because of some reasons, most of current Chinese phrase extraction approaches only focus on the partial of the Possible Phrase Patterns (PPP), and

it will lead to the incompleteness of extraction result problem. In this paper; we propose a PAT-Tree based Automatic Phrase Extraction algorithm to solve these two problems. The proposed approach could process the raw Chinese text without specified requirements on the quality and size of the corpus.

At the earlier stage, for single document or document group, it generates a local PAT-Tree with considering all possible cases of phrase patterns. The Local statistical information and contextual information are then used for phrase global extraction by utilizing the PAT-Tree structure. We compare the proposed approach with three existing phrase extraction approaches; the former could extract much more phrases than the other with satisfied phrase precision (accuracy).