PHƯƠNG PHÁP PHÂN CỤM TỪ TIẾNG VIỆT DỰA TRÊN PHƯƠNG PHÁP DENDROGRAM VÀ WIKIPEDIA

VIETNAMESE WORDS CLUSTERING METHOD BASED ON DENDROGRAM AND WIKIPEDIA

Tác giả: *Nguyễn Thị Lệ Quyên, Pham Minh Tuấn**

Tóm tắt bằng tiếng Việt:

Ngày nay, cùng với phát triển thông tin một cách nhanh chóng, việc phân loại văn bản tự động đang là một vấn đề cấp thiết. Nhiều phương pháp học máy như cây quyết định, mạng nơron nhân tạo hay máy vector hỗ trợ được áp dụng cho tiếng Anh và mang lại hiệu quả cao. Tuy nhiên các phương pháp này lại gặp khó khăn khi áp dụng cho phân loại tiếng Việt vì tiếng Việt có rất nhiều từ đồng nghĩa nhưng cách biểu diễn khác nhau. Báo cáo này đề xuất phương pháp phân cụm các từ tiếng Việt dựa vào tần số xuất hiện cùng nhau trên một trang Wikipedia tiếng Việt nhằm rút gọn vector thuộc tính của văn bản. Báo cáo này đồng thời đề xuất sử dụng phương pháp phân tích nhóm (Cluster Analysis) sử dụng đồ thị dendrogram trong việc phân cụm các từ Tiếng Việt. Kết quả thực nghiệm cho thấy phương pháp đề xuất đã phân cụm đúng các từ đồng nghĩa và các từ có chung một chủ đề.

Từ khóa: Văn bản tiếng Việt; Phân cụm từ; Phân tích nhóm; dendrogram; wikipedia

Tóm tắt bằng tiếng Anh:

Nowadays, within the development of quick information technology, the automatic document classification is an urgent issue. Many machine learning methods such as decision trees, artificial neural networks and support vector machines are applied to classify English documents and bring high efficiency. However, these methods are difficult to apply to classify Vietnamese documents because Vietnamese has many synonyms but performing different ways. This paper proposed a Vietnamese word clustering methods based on frequency appearing together on a Vietnamese Wikipedia page to shortened the length of feature vector of the document. This paper also proposed methods using cluster analysis based on graph clustering dendrogram. The experimental results show that the proposed method has the correct clustering of the synonyms and the words with a common theme.

Key words: Vietnamese documents; words clustering; cluster analysis; dendrogram; wikipedia