NGHIÊN CỨU THU THẬP VÀ XÂY DỰNG CƠ SỞ DỮ LIỆU CHỮ VIẾT TẮT TIẾNG VIỆT

COLLECTING AND BUILDING AN VIETNAMESE ABBREVIATION DATABASE

Tác giả: *Huỳnh Công Pháp**, *Nguyễn Văn Huệ*

Tóm tắt bằng tiếng Việt:

Chữ viết tắt trong tiếng Việt ngày càng tăng lên đáng kể về số lượng, đa dạng về ký hiệu, nhiều chữ viết tắt có nhiều nghĩa khác nhau. Điều này đã dẫn đến một thực trạng là làm cho người đọc văn bản nhiều lúc hiểu nhầm nội dung hoặc khó có thể đoán ra được nghĩa của từ viết tắt. Tuy nhiên, hiện nay chúng ta vẫn chưa tìm thấy một hệ thống tra cứu chữ viết tắt tiếng Việt. Để xây dựng được hệ thống tra cứu chữ viết tắt cũng như công cụ hỗ trợ gõ tắt, bước đầu tiên là cần phải xây dựng được cơ sở dữ liệu chữ viết tắt tiếng Việt. Trong bài báo này chúng tôi tập trung nghiên cứu thu thập tự động và xây dựng một cơ sở dữ liệu chữ viết tắt tiếng Việt. Trên cơ sở đó, chúng tôi sẽ tiến đến xây dựng một hệ thống quản lý và tra cứu chữ viết tắt tiếng Việt trực tuyến nhằm đáp ứng nhu cầu của đông đảo người sử dụng.

Từ khóa: chữ viết tắt; từ điển chữ viết tắt; trích rút văn bản; xử lý tiếng Việt; cơ sở dữ liệu chữ viết tắt; hệ thống tra cứu chữ viết tắt

Tóm tắt bằng tiếng Anh:

Vietnamese abbreviations increase very fast, diversify in forms and some of them have multiple meanings. This poses a problem for readers to recognize abbreviations or to understand the relevant meaning in some situation. However, we currently can't still find out a system of vietnamese abbreviation consultation. To have such a system, the first step we should build a vietnamese abbreviation database. In this paper, we focus on the research of acquiring vietnamese abbreviations from documents and the internet to build an abbreviation database. From this database, we aim to propose an online system of abbreviation management and consultation as well as a "hooked" software (like Vietkey) supporting autotext when typing.

Key words: abbreviation; acronym; abbreviation dictionary; text extraction; vietnamese language processing; vietnamese abbreviation consultationn system