VII-O-15

HỆ THỐNG HUẤN LUYỆN VÀ VẬN HÀNH WEBBOT RÚT TRÍCH THÔNG TIN TỪ WEB

Nguyễn Huy Khánh, Nguyễn Đức Huy, Nguyễn Phạm Phương Nam, Đỗ Hoàng Cường, Trần Minh Triết
Khoa Công nghệ Thông tin, Trường ĐH Khoa học Tự nhiên - ĐHQG Tp.HCM

Tóm tắt

Các ứng dụng web thế hệ thứ hai có đặc điểm là được ghép nối từ nhiều nguồn thông tin và thành phần web khác. Tuy nhiên, những hệ thống web thế hệ thứ nhất hiện đang tồn tại chưa có khả năng sẵn sàng được sử dụng để cung cấp các nguồn thông tin và thành phần web để tạo ra các ứng dụng web thế thệ thứ hai. Trong bài báo này, chúng tôi đề xuất một phương pháp có đặc điểm: huấn luyện các wrapper (WebBot) có khả năng rút trích thông tin từ các website, tham số hóa các giá trị đầu vào trước khi vận hành WebBot, tái vận hành WebBot nhằm rút trích thông tin theo nhu cầu và cung ứng các thông tin rút trích được ra nhiều dạng dịch vụ web khác nhau. Với cách tiếp cận được đề xuất, hệ thống có thể dễ dàng biến các website có sẵn thành những nguồn dữ liệu cung cấp và tổ chức lại thông tin theo yêu cầu với kết quả được kết xuất theo nhiều chuẩn định dạng khác nhau.

Từ khoá: rút trích thông tin web; web wrapper; mashup; tự động hóa web; dịch vụ web

SYSTEM FOR TRAINING AND EXECUTING WEBBOT TO EXTRACT INFORMATION FROM WEBSITES

Nguyen Huy Khanh, Nguyen Đuc Huy, Nguyen Pham Phuong Nam, Đo Hoang Cuong, Tran Minh Triet
Faculty of Information Technology, University of Science - VNU HCMC

Abstract

Content in Web 2.0 application is the combination of information from other sources and components. However, existing Web 1.0 applications have not capable to provide information and components to create Web 2.0 application. In this paper, we propose a methodology having these features: trains wrappers (WebBot) having ability to extract information from websites; parameterizes input before executing WebBot, executes WebBot to extract on demand information and expose this information through web services. With the propose approach, our system can easily turn websites into information sources, recombine information into a new source and export information to web services with common technologies (SOAP, REST).

Key words: web information extraction; web wrapper; mashup; web automation; web service