Automatic Information Extraction from the Web: A HMM-Based Approach

Q. Ho-Van[1], T. K. Dang[2], M. S. Tran-Le[1], and T. T. Vo-Dang[1]

Abstract: With the continued growth of the Internet and a huge amount of available data, extracting meaningful information from the Web has got a wide interest in both research community and business organizations. Although there exist a number of previous research work, to the best of our knowledge, none of them is flexible enough to fulfill users' requirements in a variety of applications domains. In this paper, we discuss and propose a general, extensible and dynamic approach based on the Hidden Markov model (HMM) in order to facilitate the efficient information extraction from HTML pages. Our proposed approach helps experts build a HMM from necessary specifications, train the system search engine, and extract meaningful information from HTML pages with the high precision and at a reasonable cost. More importantly, the proposed approach can be employed to support building knowledge bases for the next generation of the Web applications, i.e. the semantic Web. We developed and evaluated this model on a prototype, called PriceSearch, to extract price information of goods such as Nokia mobiles, computer mice. Experimental results confirm the efficiency of our theoretical analyses and approach.

[1] Faculty of Information Technology
Ho Chi Minh City University of Technology
268 Ly Thuong Kiet St., Dist. 10, Ho Chi Minh City, Vietnam
hcquan@dit.hcmut.edu.vn, dtkhanh@ hcmut.edu.vn