MASTER'S THESIS

Information extraction for on-line job advertisements

Au, Kwok Chung

Date of Award:
2004

Information Extraction for On-line Job Advertisements

AU Kwok Chung

A thesis submitted in partial fulfillment of the requirements

for the degree of

Master of Philosophy

Principal Supervisor: Dr. CHEUNG Kwok Wai

Hong Kong Baptist University

June 2004

Abstract

The Web has widely been used as a low-cost and yet effective way to disseminate various kind of information, where job advertisements (ads) is one of the typical examples. Its rapid growth has quickly resulted in the need for some efficient ways to locate and analyze relevant information from heterogeneous data sources. The commonly used keyword search engines often return more information than needed. In order to support more precise information search using semantics, information extraction systems have been proposed for extracting some specific target items from on-line documents as their semantic tags.

On-line job ads normally contain grammatical, telegraphic as well as ungrammatical text. Also, their writing styles and layout structures are relatively informal. These characteristics make the pattern matching approach become one of the suitable candidates for extracting job ads related information. In the first part of this thesis, we propose a rule-based system for information extraction by matching patterns featured by lexical and typesetting information. The extraction rules were derived automatically via induction. The experimental results show that the proposed system can choose the correct rule representation via the proposed induction process and successfully extract company names from on-line job ads up to an accuracy of 92.13%.

Within a job ad, fields other than “company name” normally do not carry typesetting attributes and extracting them has to be solely based on words. Modeling a large set of word sequences via generalization from a small set of training data has been known to be

non-trivial. In the second part of this thesis, we adopt the *hidden Markov model* (HMM) — a probabilistic model for modeling temporal processes, as a flexible pattern matcher for information extraction. While most of the related works require the HMM topology to be constructed manually, we take the state-merging approach instead and propose an algorithm for learning the HMM topology. The proposed learning process is governed by a localized version of the Kullback Leibler divergence. Extensive experiments have been performed to evaluate the performance of the proposed learning algorithm. A set of job advertisements obtained from a newsgroup was used for the evaluation. The resulting HMM can achieve a recall rate up to 83.33% at a precision of 62.50%.

Table of Contents