A Classifier for Schema Types Generated by Web Data Extraction Systems

Abstract

Generating Web site schema is a core step for value-added services on the web such as comparative shopping and information integration systems. Several approaches have been developed to detect this schema. For a real web site, due to the complexity of the site schema, post process of this schema such as labeling the schema types, comparing among different schema types and generating an extractor to extract instances of a schema type is a challenge. In this paper, a new tree structured called schema-type semantic model is proposed as a classifier for a schema type. Given some instances of a schema type, HTML tags contents, DOM trees structural information and visual information of these instances are exploited for the classifier construction. Using multivariate normal distribution, the classifier can be used to compare between two different schema types; i. e., the classifier can be used for schema mapping which is a core step of information integration. Also, the suggested classifier can be used to detect and extract instances of a schema type; i. e., it can be used as an extractor for web data extraction systems. Furthermore, the classifier can be used to improve the performance of the schema generated by web data extraction systems; i. e., the classifier can be used to get, as much as possible, a perfect schema. The experiments show an encourage
result with the schemas of the test web sites (a data set of 40 web sites).

References


**Index Terms**

Computer Science
Web Services

**Keywords**

Schema Mapping, Schema Type Classifier, Schema Filtration, Web Data Extraction