

University of Macau

Abstract

AUTOMATIC HIDDEN-WEB DATABASE
CLASSIFICATION

by Zhang Jing Bai

Thesis Supervisor: Associate Professor, Gong Zhiguo
Software Engineering Master

In this thesis, we address a novel method for automatic classification of Hidden-Web Databases. In our approach, the classification tree for Hidden Web Databases is selected by tailoring some existing classification tree for Web documents. Then the feature for each class is extracted from randomly selected Web documents in the corresponding category of the classification hierarchy. For each Web database, query terms are selected from the class features based on their weights. A hidden-web database is then detected by analyzing the results of the class-specific query. To raise the performance further, we also use Web pages which have links pointing to the hidden-web database (HW-DB) as another important source to represent the database. We combine link-based evaluation and query-based detection as our final classification solution. Our experiments show that the combined method can produce much better performance for classification of Hidden-Web Databases.