Abstract

In order to extract data from the web pages of Hidden web sources, many semi-automatic and automatic techniques are proposed based on structure and tags of HTML documents. These
techniques include machine learning and schema-matching approaches to solve the problem of data extraction. This paper discusses the research that has been done in the area of data extraction from Hidden Web sources. The goal of this paper is to discuss the advantages and disadvantages of currently existing techniques.

Reference

- Chen Hong-ping; Fang Wei; Yang Zhou; Zhuo Lin; Cui Zhi-Ming. Automatic Data Records Extraction from List Page in Deep Web Sources; 978-0-7695-3699-6/09 © 2009 IEEE pages 370-373.

Index Terms

Computer Science
Information Retrieval

Key words

Surface Web
Hidden Web
Information Extraction