Abstract

Website has evolved since it was first developed in 1990. Since then, the website grows rapidly. Based on the information provided by http://www.worldwidewebsize.com the number of websites is currently at least 4.54 billion pages. With a very large number, the website stores a lot of information that can be used. That problem brings up the concept of data extraction. Web data extraction aims to retrieve the contents of the website so that it can be easy to use for other purposes. The utilization of web data extraction can be used in a product catalog, news, bookstore, travel, etc. There are many systems build by different technique such as manual, supervised, un-supervised, and semi-supervised. This paper discuss supervised learning technique for web data extraction. Several previous surveys have overviewed the wrapper induction system using the concept of supervised techniques to extracted web data up to 2007. The aim of this paper is to present a comprehensive overview of the research in supervised web extraction data by providing the latest research results.

Index Terms

Computer Science

Information Systems
Keywords

Extraction web data, supervised learning, wrapper, machine learning, wrapper induction system