计算机工程与应用 2009 45 (20): 202-204 ISSN: 1002-8331 CN: 11-2127/TP

工程与应用

Web中的行情数据获取与预测研究

于春燕[1，2]，胡学钢[1]

1.合肥工业大学 计算机与信息学院，合肥 230009
2.滁州学院 计算机科学与技术系，安徽 滁州 239000

收稿日期 2008-4-21 修回日期 2008-7-15 网络版发布日期 2009-7-9 接受日期

摘要 抽取网页中的行情数据进行预测和分析具有重要意义。提出了Web中的行情数据抽取算法，该算法主要基于“行情数据通常在网页中表现为区域最大的数据表格”等实践规律，首先自动识别出最大的数据表格，然后转换为DOM树结构，最后抽取DOM树的结点值。与传统算法不同，算法自动抽取行情区域而无需用户定义抽取数据区域。设计了一个农产品价格预测原型系统，该系统针对某个农产品，自动从特定网站获取价格数据，对月度价格进行预测，实验表明预测性能较好。

关键词 Web内容挖掘 行情数据抽取 行情预测

分类号

Research on market data extraction and forecast on Web

YU Chun-yan [1，2]，HU Xue-gang[1]

1.School of Computer Science and Information Engineering，Hefei University of Technology，Hefei 230009，China
2.Department of Computer Science and Technology，Chuzhou University，Chuzhou，Anhui 239000，China

Abstract

It is significant to extract market data in Web pages for prediction and analysis.An extraction algorithm for Web pages is proposed.Taking into account the common practice that “market data are usually displayed in the largest table on a Web page”，the market data extraction algorithm first detects the largest table on a Web page and then transfers it into a DOM tree，and in the end gets the node values of the tree.This algorithm is different from traditional ones in that it can automatically detect market data and does not need a data extraction region to be specified by the users.A prototype system for agriculture product price prediction is designed and developed.The system extracts market price data from a given website automatically and predicts the price in the future months.Experimental results show the prediction results are satisfying.

Key words

Web content mining market data extraction market data prediction

DOI: 10.3778/j.issn.1002-8331.2009.20.059

通讯作者 yuchy@chzu.edu.cn