计算机工程与应用 2008 44 (21): 166-170 ISSN: 1002-8331 CN: 11-2127/TP

机器学习

基于规则归纳的信息抽取系统实现

石 倩[1], 陈 荣[1,2], 鲁明羽[1]

1.大连海事大学 信息科学技术学院，辽宁 大连 116026
2.吉林大学 符号计算与知识工程教育部重点实验室，长春 130012

收稿日期 2008-4-30 **修回日期** 2008-5-26 **网络版发布日期** 2008-7-17 **接受日期**

摘要 面对Web信息的迅猛增长，信息抽取技术非常适合于从大量的文档中抽取需要的事实数据。通过文档对象模型（DOM）解析以及检索、抽取、映射等规则的定义，设计并实现了一种具有规则归纳能力的信息抽取系统，用于Web信息的自动检索。在用于抽取规则归纳的框架下，还重点对用于生成抽取模式的WHISK学习算法进行了实验对比分析，结果表明系统对于单槽和多槽数据都具有不错的归纳学习能力。

关键词 信息抽取 抽取规则 DOM 学习算法

分类号

Implementation of rule induction-based information extraction system

SHI Qian[1],CHEN Rong[1,2],LU Ming-yu[1]

1.School of Informational Science and Technology，Dalian Maritime University，Dalian，Liaoning 116026，China
2.Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education，Jilin University，Changchun 130012，China

Abstract
With the rapid increase of Web information，Information Extraction （IE） techniques are good for automatically extracting data of interest from a mass of Web documents.In this paper，the design and the implementation of a rule induction based IE system is presented for automating Web information retrieval by DOM parsing and rules for retrieval，extraction and mapping.In this framework for rule induction，the authors particularly focus on the experiments with the WHISK algorithm for generating patterns.Experimental results show that the system performs well on both single-slot and multi-slot extraction tasks.

Key words information extraction extraction rule DOM learning algorithm

DOI: 10.3778/j.issn.1002-8331.2008.21.046

通讯作者 石 倩