山东大学学报(理学版) 2007, 42(11) 49-52 DOI: ISSN: 1671-9352 CN: 37-1389/N

论文

基于广义隐马尔可夫模型的网页信息抽取方法

王 静，姚 勇，刘志镜

西安电子科技大学计算机学院， 陕西 西安 710071

摘要：

针对网页所特有的基于版面结构的特点，利用基于视觉的网页分割算法VIPS对网页分块，得到一种新的状态转移序列，取代了传统的状态转移序列。通过二阶Markov链改进广义隐马尔可夫模型（GHMM）的状态转移和输出观测值假设条件，提出了二阶的广义隐马尔可夫模型。最后通过实验说明改进的GHMM对于网页信息抽取有很高的精确率。

关键词： 基于视觉的网页分割 广义隐马尔可夫模型 二阶Markov链 Web信息抽取

Web information extraction based on a generalized hidden Markov model

WANG Jing, YAO Yong,LIU Zhi-jing

School of Computer Science and Engineering, Xidian University, Xi'an 710071, Shaanxi, China

Abstract:

Since web pages are based on the web-specific layout structure feature, instead of using the transitional sequential state transition order, a new state transition order was proposed by using a vision based page segmentation algorithm (VIPS). In addition, the supposed state transition and the emission symbol conditions were improved by using the second-order Markov chain, and then a novel generalized hidden Markov model (GHMM) was proposed based on the improvement. Finally，through an example, it shows that the modified GHMM has a very high precision for web information extraction.

Keywords: vision based page segmentation(VIPS) generalized hidden Markov model(GHMM) second-order Markov chain Web information extraction(IE)

收稿日期 1900-01-01 修回日期 1900-01-01 网络版发布日期 2006-10-24

DOI:

基金项目:

通讯作者: 王 静

作者简介:

本刊中的类似文章

扩展功能
本文信息
Supporting info
PDF(325KB)
[HTML全文](0KB)
参考文献[PDF]
参考文献
服务与反馈
把本文推荐给朋友
加入我的书架
加入引用管理器
引用本文
Email Alert
文章反馈
浏览反馈信息
本文关键词相关文章
基于视觉的网页分割
广义隐马尔可夫模型
二阶Markov链
Web信息抽取
本文作者相关文章
王 静
姚 勇
刘志镜