华东师范大学学报 自然科学版
JOURNAL OF EAST CHINA NORMAL UNIVERSITY NATURAL SCIENCES

华东师范大学学报(自然科学版) » 2010, Vol. 2010 » Issue (5): 96-102 DOI:

自动抽取web数据的树对齐算法

景寒星, 陈少红, 俞 琨

华东师范大学 计算中心, 上海 200062

Automatic web data extraction based on tree alignment

JING Han-xing, CHEN Shao-hong, YU Kun

Computer Center, East China Normal University, Shanghai 200062, China

- 摘要
- 参考文献
- 相关文章

全文: PDF (0 KB) HTML (0 KB) 输出: BibTeX | EndNote (RIS) 背景资料

摘要 针对从模板生成的网页中自动抽取web数据的问题, 提出了一种新的树对齐算法. 该算法能够确定输入网页的最大匹配结构. 经过一系列的对齐操作之后, 多棵树被合并成为一棵记录着合并前多个网页上的统计信息的合并树, 树对齐算法可以发现合并树中的重复模式, 在最可能内容块上构建包装器, 并按照重复模式从网页上抽取数据. 实验结果表明, 该算法的抽取结果具有较高的准确性和良好的稳定性.

关键词: 数据抽取 包装器 树对齐 数据抽取 包装器 树对齐

Abstract: This paper proposed a new tree alignment algorithm for determining the optimal matching structure of the input web pages, in order to extract web data automatically. Based on the alignment, the trees were merged into one union tree whose nodes record statistical information obtained from multiple web pages. The algorithm detects repeating patterns on the union tree, and a wrapper built on the most probable content block and the repeating patterns extracts data from web pages. Experimental results showed that the proposed algorithm achieves high extraction accuracy and has steady performance.

Key words: wrapper tree alignment data extraction wrapper tree alignment

收稿日期: 2010-03-01;

通讯作者: 陈少红

服务
- 把本文推荐给朋友
- 加入我的书架
- 加入引用管理器
- E-mail Alert
- RSS

作者相关文章
- 景寒星
- 陈少红
- 俞 琨

引用本文:

景寒星,陈少红,俞 琨. 自动抽取web数据的树对齐算法[J]. 华东师范大学学报(自然科学版), 2010, 2010(5): 96-102.

JING Hanxing,CHEN Shaohong,YU Kun. Automatic web data extraction based on tree alignment[J]. Journal of East China Normal University(Natural Sc, 2010, 2010(5): 96-102.

没有本文参考文献

没有找到本文相关文献