计算机工程 2012, 38(2) 48-50 DOI: 10.3969/j.issn.1000-3428.2012.02.015 ISSN: 1000-3428 CN: 31-1289/TP

软件技术与数据库

基于相似URL的深层网数据区域识别

孔燕燕，施化吉

(江苏大学计算机科学与通信工程学院，江苏 镇江 212013)

摘要： 针对深层网查询结果页面中噪音信息对数据区域识别的干扰问题，提出一种自动识别深层网查询结果数据区域的方法。该方法利用网页的重复结构和相似URL，将页面划分成不同的语义块，依据不同页面块之间URL的相似性识别出数据区域。实验结果表明，该方法能够提高数据区域识别的召回率和准确率。

关键词： 深层网 重复结构 相似URL 语义块 数据区域

Deep Web Data Region Identification Based on Similar URL

KONG Yan-yan, SHI Hua-ji

(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Aiming at the problem that the noise information may interfere with the identification of the data region in Deep Web search result pages. This paper proposes an automatic approach to identify data region in Deep Web search result list pages. It employs continuous repetitive structure and similar URL to divide the sample pages into different semantic blocks, and identifies the block where the data region locates. Experimental results show the approzch can imprave the recall rate and accuracy of the date region identification.

Keywords: Deep Web repetitive structure similar URL semantic block data region

收稿日期 2011-06-08 **修回日期** **网络版发布日期** 2012-01-20

DOI: 10.3969/j.issn.1000-3428.2012.02.015

基金项目:

国家自然科学基金资助项目(61003288)

通讯作者:

作者简介: 孔燕燕(1985－)，女，硕士研究生，主研方向：Web数据挖掘；施化吉，教授

通讯作者E-mail: kyy699@163.com

参考文献：

[1] He Bin, Patel M, Zhang Zhen, et al. Accessing the Deep Web: A Survey[J]. Communications of the ACM, 2007, 50(5): 94-101.

[2] Wang Jiying, Lochovsky F H. Data-rich Section Extraction from HTML Pages[C]//Proceedings of the 3rd International Conference on Web Information Systems Engineering. Singapore: [s. n.], 2002: 313-322.

[3] Zhai Yanhong, Liu Bing. Structured Data Extraction from the Web Based on Partial Tree Alignment [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(12): 1614-1628.

[4] Reis D D C, Golgher P B. Automatic Web News Extraction Using Tree Edit Distance[C]//Proceedings of the 13th International Conference on World Wide Web. [S. 1.]: IEEE Press, 2004: 502-511.
[5] 黄健斌, 姬红兵, 孙鹤立. Web网页中动态数据区域的识别与抽取[J]. 计算机工程, 2007, 33(11): 53-55.
[6] 杨 舟, 卓 林, 赵朋朋, 等. 一种针对商品数据记录的自动抽取方法[J]. 计算机工程, 2010, 36(23): 262-265.
[7] Cai Deng, Yu Shipeng, Wen Jirong, et al. Extracting Content Structure for Web Pages Based on Visual Representation[C]// Proceedings of the 5th Asia Pacific Web Conference. Xi'an, China: [s. n.], 2003: 406-417.
[8] Lin Shianhua, Ho J M. Discovering Informative Content Blocks from Web Documents [C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2002: 588-593.
[9] Liu Bing, Grossman R L, Zhai Yanhong. Mining Data Records in Web Pages[C]//Proceedings of the 9th Int'l Conf. on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2003: 601-606.

本刊中的类似文章

1. 王海龙, 胡景芝, 赵朋朋, 崔志明.基于搜索引擎的Deep Web数据源发现[J]. 计算机工程, 2011,37(5): 77-79,82
2. 郭若飞, 蔡欣宝, 赵朋朋, 崔志明.基于Choquet积分的深层网数据源选择[J]. 计算机工程, 2011,37(4): 40-42
3. 马建华;李赛红;徐兰兰.深层网中基于入口查询的表单填充策略[J]. 计算机工程, 2010,36(7): 66-67,7
4. 杨晓琴, 鞠时光, 曹庆皇, 王秀红.基于包装器的Deep Web自动语义标注[J]. 计算机工程, 2010,36(12): 52-54
5. 华慧, 伏玉琛, 周小科.基于查询接口文本的Deep Web数据源分类[J]. 计算机工程, 2010,36(12): 66-68
6. 曲著伟;李敏强.基于数据区域发现的信息抽取规则生成方法[J]. 计算机工程, 2009,35(22): 59-61
7. 杨巨峰;史广顺;赵玉娟;王庆人.基于规则集的Deep Web信息检索[J]. 计算机工程, 2008,34(13): 51-53
8. 黄健斌;姬红兵;孙鹤立.Web网页中动态数据区域的识别与抽取[J]. 计算机工程, 2007,33(11): 53-55,5
9. 缪建明;;张 全;吴 晨; .自然语言处理中的语句语义表示格式研究[J]. 计算机工程, 2006,32(16): 77-79

文章评论

反馈人		邮箱地址	
反馈标题		验证码	3413
	5		