现代图书情报技术 2007 (2): 49-52 ISSN: 1003-3513 CN: 11-2856/G2

知识组织与知识管理

搜索引擎返回结果自动抽取

藕军[1];任明仑[1,2]

合肥工业大学计算机网络研究所[1]

收稿日期 2006-11-24 修回日期 2006-12-9 网络版发布日期 2007-3-2 接受日期

摘要 提出一种从搜索引擎返回结果页面上自动抽取结果记录及后续页面链接信息并生成Wrapper的方法:对于一个有效的结果页面，通过比较其HTML标签树上节点的相似度从而识别出潜在记录块，利用启发式规则从潜在记录块中将结果记录块和后续页面链接分别识别出来，然后利用其在标签树上的位置信息分别构造Wrapper。实验结论及与已有方法的比较表明，该方法简单可行且高效。

Abstract Present a new method for automatically extracting Search Result Records(SRRs) and Subsequent Result Page Links(SRPLs) from a search engine’s response page. Compare the similarity of nodes on the HTML tags tree of a valid response page to recognize Candidated Records Blocks(CRBs).And recognize SRRs and SRPLs form CRBs based on several heuristic rules.Then building wrapper for them using their location on tags tree. Experiments and comparison with other methods show that the methed is useful and efficient.

关键词 搜索引擎 Web信息抽取 包装器生成 HTML标签树 节点相似度

Key words Search engine;Web information extraction;Wrapper generation;HTML tags tree;Nodes similarity

分类号 TP391.3

DOI:

通讯作者:
藕军 1717go@gmail.com
作者个人主页: 藕军 任明仑