使用双层分类器在垂直搜索中自动识别交互式查询接口

王琳, 王行甫, 杜云开

(中国科学技术大学计算机科学与技术学院, 安徽合肥 230027)

Email : xiaquhet@mail.ustc.edu.cn

摘 要: 一框式检索功能普遍提供于各类互联网信息门户的首页, 由于需要交互式操作以及返回页面仅包含与用户所提交关键字相关的查询结果等原因, 较少受到传统搜索引擎的关注。但是在垂直搜索中, 若能够有效利用远程服务器自带的站内检索功能, 将在显著降低本地计算资源和带宽消耗的同时, 提高查全和查准率。本文提出并实现了一种用于在主题相关的页面采集过程中自动定位交互式查询接口的双层分类器。针对 8 个不同领域主题的规模化实验显示, 该分类器能够准确过滤非相关域名和非可查表单, 实现搜索接口的有效识别。

关键词: 垂直搜索; 查询接口识别; 表单特征分类; html 解析; 支持向量分类; 决策树

中图分类号: TP391.3 **文献标识码:** A **文章编号:**

Automatic Identifying Query Interface in Vertical Crawling by Hierarchical Classification

WANG Lin , WANG Xing-fu , Du Yu-kai

(Department of Computer Science and Technology, University of Science and Technology of China, HeFei, 230027, China)

Abstract : One-Frame search interface are widely provided on the home page of most well-known Web information gateways. Due to their requirement of users' interactive queries via page form submission, this functionality are not paid enough attention to by traditional hyperlink-based search engines. In vertical search however, utilizing the in-site retrieval power of the remote server and extracting records from the result page returned can not only lower the depletion of local bandwidth and computing resources, but also improve precision and recall significantly. This paper proposes a hierarchical classifier to locate domain-specific search interfaces automatically. Experiments conducted on eight different topics demonstrate that the classifier can get rid of non-relevant domains and non-searchable forms both accurately and efficiently.

Key words : deep web data source, domain-specific crawler, searchable forms, html analysis, SVM classifier, decision tree algorithm

1 引言

一框式检索功能不仅构成商业搜索引擎的基本界面, 同时也普遍提供于各类互联网信息门户的首页（图1）。由于需要交互式操作以及返回页面仅包含与用户所提交关键字相关的查询结果等原因, 较少受到传统搜索引擎的关注。但是在垂直搜索中, 若能够有效利用远程服务器自带的站内检索功能, 将在显著降低本地计算资源和带宽消耗的同时, 提高查全率和查准率。在主题相关的页面采集过程准确自动地定位交互式检索接口, 是聚焦爬虫的起步环节需要克服的困难之一, 其中主要包括:

- 首先, 相对于其他 html 页面元素而言, 查询表单分布相对稀疏且相互之间缺少链接指引。例如, 一个权重优先的聚焦爬虫在爬取 100 000 个与电影相关的页面后, 仅获得 94 个电影主题的查询接口 [1]。
- 另外, 在已获取的 html 表单中存在相当比例并不是可查询的, 如用于用户登陆、电子邮件列表订阅、购物订单提交、后台结构化数据库查询等用途的表单 [2]。

图 1 无处不在的一框式检索

Fig.1 One-Frame search interface are widely provided

收稿日期: ; 基金项目: 国家自然科学基金 (61472382,61272472,61232018) , 国家科技重大专项 (2012ZX10004-301-609) 资助 作者简介: **王琳**, 女, 1988 年生, 博士, 研究方向为垂直搜索、 Web 信息抽取; **王行甫**, 男, 1964 年生, 副教授, 研究方向为无线传感网与安全; **杜云开**, 男, 1992 年生, 博士, 研究方向死锁检测.

- 最后但并非次要的是，即使是可查询表单，由于来自于不同的域名，特征多样异构，不能获得统一模式，导致自动识别方法不能高效推广。

本文提出一种新的框架，用于克服上述技术挑战：首先，使用一种改进的权重优先爬取策略，使搜索空间尽可能约束在最具有希望的范围内；其次，使用双层分类器，分别利用页面文本特征和表单特征对页面采集的方向进行指引。

本文的行文结构如下：第2节对相关领域的既有工作进行简要回顾；第3节分别给出页面文本和表单特征分类器构建方法；双层框架的工作流程在第4节介绍；实验评估在第5节中讨论；第6节 总结本文并给出下一步工作的开展方向。

2 相关工作

随着人们对个性化信息服务需要的日益增长, 基于主题爬虫的专业搜索引擎的发展将成为搜索引擎发展的主要趋势之一。主题网络爬虫爬行策略的研究, 对专业搜索引擎的应用和发展具有重要意义。国内关于主题爬行技术的研究也已取得不少成果 [3]，典型代表主要包括基于向量空间模型的方法 [4, 5]、基于超链分析的方法 [6]、 基于分类器预测的方法 [7]以及将不同策略相互结合的混合方法 [8]等。整体上看，国内工作目前仍然较多集中于网页爬取策略的讨论，对于主题表单等交互式内容尚未见到有涉及。

国外研究者对于深层网的探索更趋多样化，一些重要的相关工作包括： 文献 [9]介绍了一种页面分区方法，用于区分传统搜索引擎的查询接口和深层网络实体的入口，并且进一步构建了查询，通过获取并分析返回的查询结果来对自动辨识的结果做进一步的确认。 Cope 等 [10]使用一种自动的特征生成技术来描述候选表单，在此基础上进一步使用 C4.5 决策树进行分类。在他们选取的两个实验平台 —— ANU 测试集和随机 Web 集合上，该技术分别取得了高于 85% 和 87% 的准确率。 Bergholz 等 [11]研究了一种从 PIW(Publicly Indexable Web， 公开可引页面) 出发，寻找隐藏页面入口的爬行策略。该爬虫是主题相关的，并且使用预分类好的文档和关键字进行初始化训练。 文献 [2]描述了一种能够自适应学习的聚焦爬虫 —— ACHE 框架。该框架能够高效胜任相同域名下的交互接口定位任务，但是不能适应广域网下目标稀疏分布于不同域名的情形。而且， ACHE 构建复杂，计算延迟较高。

3 双层分类器

为弥补已有工作的不足，本文使用两个相互独立的分类器，以分层工作的方式为垂直搜索的页面采集过程指引最优方向。它们分别是页面文本分类器和表单特征分类器。图2给出该框架的上层结构。

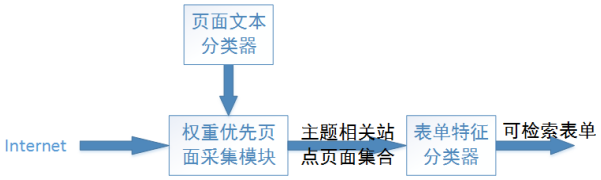

图 2 双层分类框架

Fig.2 The high-level architecture

- 首先，给定一个 URL，使用页面文本分类器判断其对应的域名首页是否主题相关，当且仅当首页相关时，才对该站点进行深入爬取。前人工作 [2, 12]显示，可以使用 libsvm 算法训练该分类器；
- 然后，对于页面文本分类器筛选出的主题相关页面，使用表单特征分类器提取其中的可查询接口。根据 Luciano 等 [13]和 Cope 等. [10]的结论，该情形下决策树是最优选择。

本文采用这种分层框架的原因是为了获得模块化的效果：当一个复杂问题分解为多个简化的子问题时，可以为每一个子组分的特征集合选择一个最合适的学习算法，使得系统的鲁棒性和准确性得到整体提高。

3.1 表单特征分类

html 表单通常包含结构和文本两个部分。例如图3中著名的 Apache Lucene 主页上的站内检索入口，不仅包含如“sort”、“Search”这样的文本内容，也包含如“select” 元素、“submit”按钮这样的控件结构。

图 3 一个交互式检索表单示例

Fig.3 An illustration of form entry

```
<form method="get" action="/search" name="f"
  class="searchbox">
  <input type="text" name="query" value=""
    size="35">
  sort: <select name="mode">
    <option value="none"> time-biased relevance
      </option>
    <option value="pure"> relevancy </option>
    <option value="newestOnTop"> newest </option>
    <option value="oldestOnTop"> oldest </option>
  </select> & nbsp;
    <input type="submit" value="Search">
</form>
```

基于对正负样本中结构或文本特征的分布统计，本文选取了 12 个区分度最高的特征 N1 ～ N12 （见表2）来训练表单分类器。

3.2 页面文本特征分类

本文通过剥除 html 标签及其中内嵌的 JavaScript 代码的方式获取给定页面的纯文本内容，其他出现频率较高的非文本信息，如 php/asp 标签(<?php ?> <%php ?> <% %>) 中嵌入的代码以及可能套用的风格版式（如 style sheets）等，也同时剥除。在此基础上，可以对已抽取文本施予文本分类所通用的预处理步骤：

1. 用空格代替所有非字母和数字的字符；
2. 统一使用小写字母；
3. 删除无查询意义的虚体词（使用 org.apache.lucene.analysis.StopAnalyzer）；
4. 用词干替换其派生出的其他词形变化（使用 org.apache.lucene.analysis.PorterStemmer）；
5. 使用 TFIDF [14]将每个文本样本转换为其对应的特征向量。

4 改进的权重优先爬虫

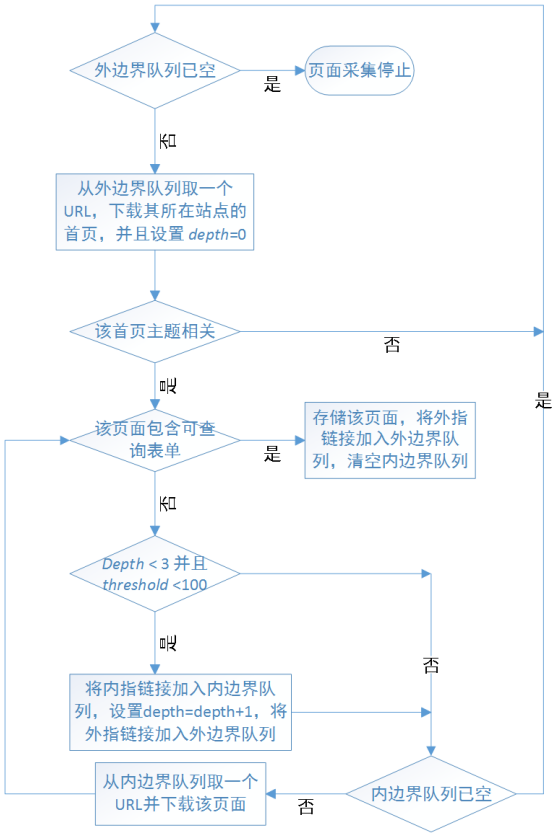

图 4　页面采集过程控制流程图

Fig.4　Modified best-first crawler

在主题网络爬虫研究领域, 权重优先（Best First）算法具有一定的竞争力, 所以很多研究者将其作为性能的比较基准。但是权重优先策略的设计初衷是尽可能多的爬取与主题相关的页面，并没有考虑到 web 表单相对于 web 页面的分布所具有的特殊性，因此标准的 Best First 算法并不适合直接应用于本文框架中。基于以下观察，本文在该方法基础上进行了适应性的调整和优化。

权重优先的基本思想是给定一个待爬行 URL 队列, 从中挑选最好的 URL 优先爬行。爬行主题采用关键词集合来描述, 待爬行 URL 的优先级是根据特征关键字在训练文档中的权重向量 q 和在已爬行网页中的权重向量 p 的余弦相关度来估计的:

$$sim(q,p) = \cos\theta = \frac{\sum_{k=1}^{M} q_k \cdot p_k}{\sqrt{\sum_{k=1}^{M} q_k^2 \sum_{k=1}^{M} q_k^2}} \tag{1}$$

其中，p_k 为特征向量的第 k 维关键字在抓取的网页文档中的权重，q_k 为第 k 维关键字在训练文档集中的权重。

观察 1 *交互式检索接口普遍置位于每个站点的浅层，同时文献 [15]也报告绝大多数（约 94%）该类表单出现的深度小于 3 层。*

因此，本文为每个站点设定两个阈值 —— 爬取深度 $depth \geq 3$ 和访问页面总数 $threshold \geq 100$，用于避免爬虫在某些站点发生局部的陷入。同时，以相同一级域名下所有页面组成的站点作为爬行单位，仅对每一站点的首页进行主题相关度的判定。

观察 2 *关键词匹配在描述用户兴趣方面具有局限性。*

因此，本文使用 SVM 文本特征分类模型来实现用户兴趣的学习和主题相关度的预测网页。页面文本分类器和表单特征分类器一起，构成本文提出的双层分类框架。文本分类模型能够从更深的层次描述主题信息，并有利于提高主题搜索的准确率。

控制流程的细节如图4所示。

5 实验评估

5.1 表单分类器的训练

本文从各类互联网门户分别手工选取了 248 个可查询的表单作为正样本和 196 个不可查询表单做为负样本，样本来源和特征统计情况分别列于表1和表2。

表 1　表单训练样本来源分布

Table.1　Distribution of training sample for form classifier

正样本 合计 248				负样本 合计 196	
网页搜索	8	官方机构	50	用户登录	52
新闻门户	38	学术文库	22	后台数据库	46
百科门户	43			电子商务	44
论坛社区	53			电子邮件	54

比对表2中不同特征在正负样本中的分布，可以得到一些经验性推论，如：正样本在其输入元素的 name 属性或者 value 属性中普遍包含类似“搜索”字样的文本，同时包含数量较多的 hidden 类型标签；负样本则含有较多的复选框和下拉列表控件。

本文使用两个不同的工具构建决策树：R rpart 算法 [16] 和 Matlab fitctree 函数 [17]. 由于两种算法采用了完全相同的分裂策略和训练参数（详见表3），因此图5与图6给出的构建

结果具有基本完全一致的结构，唯一的区别是 rpart 算法较 fitctree 而言拟合程度略高。这是由于两种算法对于最优剪枝程度的逼近方式不同造成的，fitctree 对过拟合风险做出了相对保守的估计。并且交叉验证子集划分的随机性也可能是一个原因。

表 2 正负样本表单特征分布

Table.2 Feature distributions of searchable and non-searchable forms

Feature		平均含量	正样本	负样本	正/负比率
文本特征	N2	**“搜索”** [1]	1.16	0.02	**50.5:1**
	N5	“验证码”	0.02	0.34	1:17
	N7	“email”	0.00	0.22	0
	N8	“密码/账号”	0.00	0.20	0
	N6	“登录/注册”	0.00	0.28	0
结构特征	N12	**hidden 标签**	4.45	1.63	**2.72:1**
	N13	submission method-get	0.87	1.40	1:1.61
	N11	text 控件	1.01	3.00	1:3
	N10	image 控件	0.36	1.21	1:3.36
	N9	textarea 控件	0.02	0.07	1:3.5
	N3	**checkbox**	0.03	1.09	**1:36.3**
	N1	**select option**	0.17	10.64	**1:10**
	N4	radio 标签	0.00	0.48	0

[1] 包括“检索”、“Go”、“Search”等同义近义词，下同

表 3 决策树算法参数设置

Table.3 Parametres set for classification tree

		最小分支	最小叶	最大深度	分裂判据	交叉验证
R	参数名	minsplit	minbucket	maxdepth	Parms.split	xval
	默认值	20	minsplit/3	30	Gini	10
Mat-lab	参数名	minparent	minleaf	MaxNum-Splits	split-criterion	Test.-nsamples
	默认值	10	1	样本数-1	’gdi’[1]	10
实验设置		2	默认值			

[1] 基尼分歧指数（Gini’s diversity index）

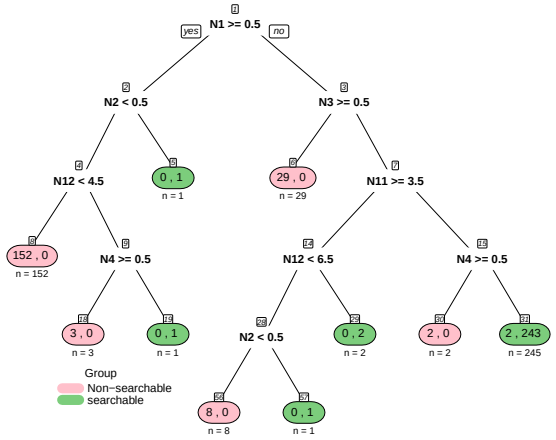

图 5 R rpart 构造决策树

Fig.5 The decision tree generated by R

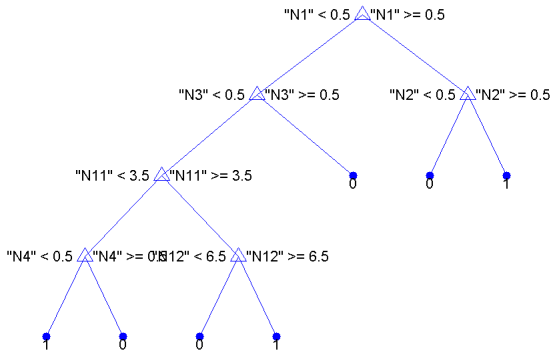

图 6 Matlab fitctree 构造决策树

Fig.6 The decision tree generated by Matlab

两种算法的具体剪枝过程相互对照如图7所示。图7显示该差别对于样本整体误差的影响是十分微弱的。

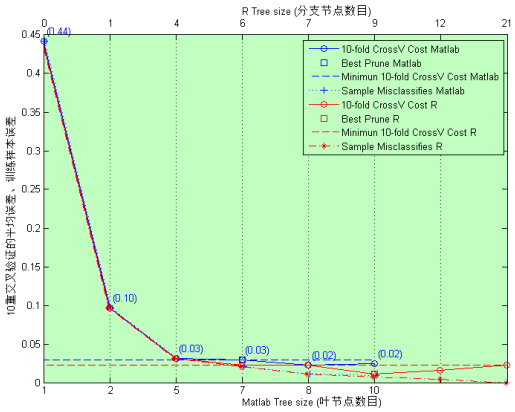

图 7 Matlab fitctree 与 R rpart 的剪枝过程

Fig.7 The decision tree generated by R

5.2 页面文本分类器的训练

本文借助于在线开放目录工程 DMOZ（`http://dmoz.org/`）收录的 web 内容训练页面文本分类器。正样本需要包含领域主题相关的内容，因此，本文选定 8 个不同的领域主题（见表4），使用 Python 脚本将该 8 个领域主题分别填入 DMOZ 主页面上的查询接口并自动抽取返回结果中的 url 作为正样本。对于负样本，则不经过人工筛选，直接从其 RDF 存库（`http://rdf.dmoz.org/rdf/`）中随机选取。DMOZ 库存中的样本信息是通过 RDF 格式描述的：

```
<ExternalPage about = "http://www.airwise.com/
  airports/us/SLC/index.html">
  <d:Title> Salt Lake City Airport - airwise.com
    </d:Title>
  <d:Description> Information about the airport
    including airlines, ground transportation,
    parking, weather and airport news.
```

```
    </d:Description>
  <topic>Top/Regional/North_America/United_States/
    Utah/Localities/S/Salt_Lake_City/
    Transportation/Airports</topic>
</ExternalPage>
```

本文使用 ‘d:Description’ 元素的内容以及 ExternalPage 元素的 ‘about’ 属性所指向的页面来确定一个负样本。DMOZ 的一级目录分为 16 类，为了使训练样本的组成结构更具有代表性，本文根据一级目录的容量确定样本抽取的数量。除去其中不可下载的部分，每个领域主题下正负样本的最终确定数目见表4。

表 4 8 个领域主题下训练样本数量和 SVM 分类准确率
Table.4 Training data and precision of page classifier for each topic

主题	正样本	负样本	SVM	总体准确率
Airfare	116	316	0.961 9	0.90
Auto	251	356	0.946 3	0.88
Book	156	332	0.913 5	0.91
Rental	91	228	0.973 7	0.95
Hotel	170	272	0.994 1	0.94
Job	170	317	0.979 1	0.81
Movie	160	312	0.900 4	0.86
Music	22	87	0.87	0.80
合计/平均	1136	2220	0.94	0.88

以上页面集合经过第3.2节中描述的处理步骤后，可以完成各自 SVM 分类器的学习。其中，从正样本中提取到的出现频率最高的 5 个特征列于表5。SVM 使用高斯核，并且仍然通过10-重交叉验证确定最优的核参数。

表 5 正样本的页面文本特征抽取
Table.5 Five most frequent textual features extracted

Topic	Textual features (Feature: Frequency)				
Airfare	pm: 41	am: 40	airline: 27	air: 12	airway: 10
Auto	docum: 108	car: 105	leas: 84	search: 63	make: 56
Book	search: 130	title: 110	book: 95	author: 75	new: 72
Rental	pm: 402	option: 202	am: 168	airport: 144	car: 143
Hotel	hotel: 234	pm: 228	island: 151	new: 135	room: 84
Job	job: 207	new: 125	locat: 84	service: 82	island: 81
Movie	press: 211	book: 123	s: 109	video: 107	entertain: 107
Music	record: 16	music: 16	sub: 15	search: 7	new: 8

5.3 双层分类框架的整体性能

本文针对 8 个不同的领域主题，分别实施了相同规模的垂直爬取实验。初始集合设置为从 DMOZ 中抽取的 50 个 URL 种子。每当爬取到主题相关且包含至少一个交互式查询接口的页面时，将其进一步提交表单分类器进行过滤并提取其中包含的 URL 链接加入到当前队列中。除去 Music 主题相关的查询接口分布相对稀疏，本文实现的权重优先爬虫仅定位了 50 个查询接口以外，其他 7 个实验中均能够在有效时间内完成 100 个发现的既定目标。其中，例举 Books 主题下的 5 个表单发现如下：

```
http://www.amazon.com/books/
http://www.thebookpeople.co.uk/
http://books.half.ebay.com/
http://onlinebooks.library.upenn.edu/search.html/
http://www.newyorker.com/books/
```

最后，对一共 750 个表单的有效性进行人工逐一确认，得到系统整体的分辨率如表4最右列。

6 结论

本文提出并实现了一种在垂直搜索中自动发现交互式检索表单的双层分类器框架。8 个初等规模的页面收集实验显示，该方法的准确性和效率都能够达到较理想水平，8 个不同领域主题的平均准确率是 0.88。做为起步，本文的工作为智能交互的后续环节，如查询提交和结果抽取的自动化实现奠定了基础。但是仍有诸多环节存在不足，如对于领域主题的描述仍然过于宽泛，真实应用场景下用户的兴趣表现会更加复杂，需要更加精确和细化的分类；实验所使用的爬行环境 DMOZ 与真实的网络环境仍然存在较大的差别，后者的多样性和动态性对页面采集的适应性和实时性将提出更高的要求。这些问题都需要未来进一步的工作来加以完善。

References:

[1] CHAKRABARTI S, Van den BERG M, DOM B. Focused crawling: a new approach to topic-specific Web resource discovery[J]. Computer Networks, 1999, 31(11) : 1623 – 1640.

[2] BARBOSA L, FREIRE J. An adaptive crawler for locating hidden-web entry points[C]. Proceedings of the 16th international conference on World Wide Web. 2013 : 441 – 450.

[3] LIU J-H, LU Y-L. Survey on topic-focused web crawler-[J]. Appl. Res. Comput, 2007, 24(2629) : 25.

[4] LI W, LIU J, HE H, et al. Research and Realization of Intelligent Focused Web Crawler[J]. Application research of Computers, 2006, 2 : 163 – 166.

[5] ZOU H, SUN L. A Customized Focusing Crawler[J]. Electronic Science and Technology, 2009, 1 : 016.

[6] FANG Y, YANG Q, WU G, et al. Customized focused crawler for peer-to-peer Web search[J]. Journal of Huazhong University of Science and Technology (Nature Science Edition), 2007 : S2.

[7] FU X, FENG B, MA Z, et al. Focused crawling method with online-incremental adaptive learning[J]. JOURNAL-XIAN JIAOTONG UNIVERSITY., 2014, 38(6) : 599 – 602.

[8] LIU D, LIN P, GAO H. Research on Crawling Strategy of Subject Searching Spider by Content-Based and Hyperlink-Based Analysis[J]. Computer & Digital Engineering, 2009, 1 : 007.

[9] DU X, ZHENG Y, YAN Z. Automate discovery of deep web interfaces[C]. 2010 2nd International Conference on Information Science and Engineering (ICISE). 2010 : 3572 – 3575.

[10] COPE J, CRASWELL N, HAWKING D. Automated discovery of search interfaces on the web[C]. Proceedings of the 14th Australasian database conference-Volume 17. 2003 : 181 – 189.

[11] BERGHOLZ A, CHILDLOVSKII B. Crawling for domain-specific hidden web resources[C]. Proceedings of the Fourth International Conference on Web Information Systems Engineering 2003, WISE 2003. 2003 : 125 – 133.

[12] BARBOSA L, FREIRE J. Combining classifiers to identify online databases[C]. Proceedings of the 16th international conference on World Wide Web. 2012 : 431 – 440.

[13] BARBOSA L, FREIRE J. Searching for Hidden-Web Databases.[C]. WebDB. 2005 : 1 – 6.

[14] TORGO L, GAMA J. Regression by classification[G]. Advances in artificial intelligence. Berlin Heidelberg : Springer, 1996 : 51 – 60.

[15] RAGHAVAN S, GARCIA-MOLINA H. Crawling the Hidden Web[C]. VLDB '01 : Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2011 : 129 – 138.

[16] The R Project for Statistical Computing[EB/OL]. `http://www.r-project.org`. [Access Date: 2014-09-02].

[17] Classification Trees and Regression Trees[EB/OL]. `http://cn.mathworks.com/help/stats/classification-trees-and-regression-trees.html`. [Access Date: 2014-10-16].

附中文参考文献:

[3] 刘金红, 陆余良. 主题网络爬虫研究综述[J]. 计算机应用研究, 2007, 24(10): 26–29.

[4] 李卫, 刘建毅, 何华灿, 等. 基于主题的智能 Web 信息采集系统的研究与实现[J]. 计算机应用研究, 2006, 23(2):163–166.

[5] 邹海亮, 孙莉. 可定制的聚焦网络爬虫[J]. 电子科技, 2009, 22(1):47–50.

[6] 方启明, 杨广文, 武永卫, 等. 面向 P2P 搜索的可定制聚焦网络爬虫[J]. 华中科技大学学报(自然科学版), 2007, 35:148–152.

[7] 傅向华, 冯博琴, 马兆丰, 等.可在线增量自学习的聚焦爬行方法[J].西安交通大学学报, 2014, 38(6):599–602.

[8] 刘朋, 林泓, 高德威. 基于内容和链接分析的主题爬虫策略[J]. 计算机与数字工程, 2009, 37(1):22–24.