文章编号:1001 - 9081(2005)09 - 1965 - 05

聚焦爬虫技术研究综述

周立柱,林 玲
(清华大学 计算机科学与技术系, 北京 100084)
(dcszlz@tsinghua.edu.cn)

摘 要:因特网的迅速发展对万维网信息的查找与发现提出了巨大的挑战。对于大多用户提出的与主题或领域相关的查询需求,传统的通用搜索引擎往往不能提供令人满意的结果网页。为了克服通用搜索引擎的以上不足,提出了面向主题的聚焦爬虫的研究。至今,聚焦爬虫已成为有关万维网的研究热点之一。文中对这一热点研究进行综述,给出聚焦爬虫(Focused Crawler)的基本概念,概述其工作原理;并根据研究的发展现状,对聚焦爬虫的关键技术(抓取目标描述,网页分析算法和网页搜索策略等)作系统介绍和深入分析。在此基础上,提出聚焦爬虫今后的一些研究方向,包括面向数据分析和挖掘的爬虫技术研究,主题的描述与定义,相关资源的发现,Web 数据清洗,以及搜索空间的扩展等。

关键词:聚焦爬虫;信息检索;链接分析;文本检索;数据抽取;协作抓取;本体描述;元搜索
中图分类号: TP311.13 **文献标识码**:A

Survey on the research of focused crawling technique

ZHOU Li-zhu, LIN Ling
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 10084, *China*)

Abstract: The survey of focused crawling starts with the motivation for this new research and an introduction on basic concepts of focused crawling. The key issues in focused crawling are reviewed, such as webpage analyzing algorithms and the searching strategy on the Web. How to crawl relevant data and information according to different requirements is discussed in detail and three representative architectures of focused crawler systems are analyzed. Some future works for focused crawling research are indicated, including crawling for data analysis and data mining, topic description, finding relevant Web pages, Web data cleaning, and the extension of search space.

Key words: focused crawler; information retrieval; link analysis; text retrieval; data extraction; collaborative crawling; ontology; metasearch

0 引言

随着网络的迅速发展,万维网成为大量信息的载体,如何有效地提取并利用这些信息成为一个巨大的挑战。搜索引擎(Search Engine),例如传统的通用搜索引擎 AltaVista,Yahoo! 和 Google 等,作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。但是,这些通用性搜索引擎也存在着一定的局限性,如:

(1) 不同领域、不同背景的用户往往具有不同的检索目的和需求,通用搜索引擎所返回的结果包含大量用户不关心的网页。

(2) 通用搜索引擎的目标是尽可能大的网络覆盖率,有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深。

(3) 万维网数据形式的丰富和网络技术的不断发展,图片、数据库、音频/视频多媒体等不同数据大量出现,通用搜索引擎往往对这些信息含量密集且具有一定结构的数据无能为力,不能很好地发现和获取。

(4) 通用搜索引擎大多提供基于关键字的检索,难以支持根据语义信息提出的查询。

为了解决上述问题,定向抓取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序,它根据既定的抓取目标,有选择的访问万维网上的网页与相关的链接,获取所需要的信息。与通用爬虫(general-purpose web crawler)不同,聚焦爬虫并不追求大的覆盖,而将目标定为抓取与某一特定主题内容相关的网页,为面向主题的用户查询准备数据资源。

1 聚焦爬虫工作原理及关键技术概述

网络爬虫是一个自动提取网页的程序,它为搜索引擎从万维网上下载网页,是搜索引擎的重要组成[1]。传统爬虫从一个或若干初始网页的 URL 开始,获得初始网页上的 URL,在抓取网页的过程中,不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一定停止条件,如图 1(a)流程图所示。聚焦爬虫的工作流程较为复杂,需要根据一定的网页分析算法过滤与主题无关的链接,保留有用的链接并将其放入等待抓取的 URL 队列。然后,它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL,并重复上述过程,直到达到系统的某一条件时停止,如图 1(b)所示。另外,所有被爬虫抓取的网页将会被系统存贮,进行一定的分析、过滤,并建立索引,以便之后的查询和检索;对于聚焦爬虫来说,这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

相对于通用网络爬虫,聚焦爬虫还需要解决三个主要问

收稿日期:2005 - 07 - 03;**修订日期**:2005 - 07 - 20 **基金项目**:国家自然科学基金资助项目(60173008)
作者简介:周立柱(1947 -),江苏人,教授,博士生导师,主要研究方向:数据库、Web 数据处理、海量信息系统、数字化图书馆;林玲(19[illegible]人,博士研究生,主要研究方向:Web 数据处理.

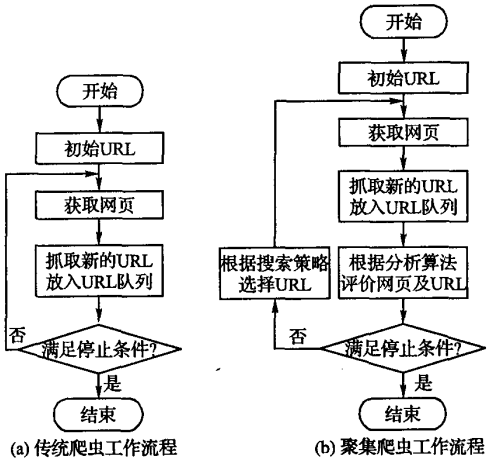
题:

(1) 对抓取目标的描述或定义;

(2) 对网页或数据的分析与过滤;

(3) 对 URL 的搜索策略。

抓取目标的描述和定义是决定网页分析算法与 URL 搜索策略如何制订的基础。而网页分析算法和候选 URL 排序算法是决定搜索引擎所提供的服务形式和爬虫网页抓取行为的关键所在。这两个部分的算法又是紧密相关的。



(a) 传统爬虫工作流程 (b) 聚集爬虫工作流程

图 1 传统爬虫和聚焦爬虫工作流程对比

2 抓取目标描述

现有聚焦爬虫对抓取目标的描述可分为基于目标网页特征、基于目标数据模式和基于领域概念 3 种。

基于目标网页特征的爬虫所抓取、存储并索引的对象一般为网站或网页。根据种子样本获取方式可分为:

(1) 预先给定的初始抓取种子样本;

(2) 预先给定的网页分类目录和与分类目录对应的种子样本,如 Yahoo! 分类结构等;

(3) 通过用户行为确定的抓取目标样例,分为:

a) 用户浏览过程中显示标注的抓取样本;

b) 通过用户日志挖掘得到访问模式及相关样本。

其中,网页特征可以是网页的内容特征,也可以是网页的链接结构特征,等等。

现有的聚焦爬虫对抓取目标的描述或定义可以分为基于目标网页特征,基于目标数据模式和基于领域概念三种。

基于目标网页特征的爬虫所抓取、存储并索引的对象一般为网站或网页。具体的方法根据种子样本的获取方式可以分为:(1)预先给定的初始抓取种子样本;(2)预先给定的网页分类目录和与分类目录对应的种子样本,如 Yahoo! 分类结构等;(3)通过用户行为确定的抓取目标样例。其中,网页特征可以是网页的内容特征,也可以是网页的链接结构特征,等等。

基于目标数据模式的爬虫针对的是网页上的数据,所抓取的数据一般要符合一定的模式,或者可以转化或映射为目标数据模式。

另一种描述方式是建立目标领域的本体或词典,用于从语义角度分析不同特征在某一主题中的重要程度。

3 网页搜索策略

网页的抓取策略可以分为深度优先、广度优先和最佳优先三种。深度优先在很多情况下会导致爬虫的陷入(trapped)问题,目前常见的是广度优先和最佳优先方法。

3.1 广度优先搜索策略

广度优先搜索策略是指在抓取过程中,在完成当前层次的搜索后,才进行下一层次的搜索。该算法的设计和实现相对简单。在目前为覆盖尽可能多的网页,一般使用广度优先搜索方法[26]。也有很多研究将广度优先搜索策略应用于聚焦爬虫中。其基本思想是认为与初始 URL 在一定链接距离内的网页具有主题相关性的概率很大[26]。另外一种方法是将广度优先搜索与网页过滤技术结合使用,先用广度优先策略抓取网页,再将其中无关的网页过滤掉。这些方法的缺点在于,随着抓取网页的增多,大量的无关网页将被下载并过滤,算法的效率将变低。

3.2 最佳优先搜索策略

最佳优先搜索策略按照一定的网页分析算法,预测候选 URL 与目标网页的相似度,或与主题的相关性,并选取评价最好的一个或几个 URL 进行抓取。它只访问经过网页分析算法预测为"有用"的网页。存在的一个问题是,在爬虫抓取路径上的很多相关网页可能被忽略,因为最佳优先策略是一种局部最优搜索算法。因此需要将最佳优先结合具体的应用进行改进,以跳出局部最优点[21,25]。将在第 4 节中结合网页分析算法作具体的讨论。研究表明,这样的闭环调整可以将无关网页数量降低 30% ~90%。

4 网页分析算法

网页分析算法可以归纳为基于网络拓扑、基于网页内容和基于用户访问行为三种类型。

4.1 基于网络拓扑的分析算法

基于网页之间的链接,通过已知的网页或数据,来对与其有直接或间接链接关系的对象(可以是网页或网站等)作出评价的算法。又分为网页粒度、网站粒度和网页块粒度这三种。

4.1.1 网页(Webpage)粒度的分析算法

PageRank[2]和 HITS[3]算法是最常见的链接分析算法,两者都是通过对网页间链接度的递归和规范化计算,得到每个网页的重要度评价。PageRank 算法虽然考虑了用户访问行为的随机性和 Sink 网页的存在,但忽略了绝大多数用户访问时带有目的性,即网页和链接与查询主题的相关性。针对这个问题,HITS 算法提出了两个关键的概念:权威型网页(authority)和中心型网页(hub)。

基于链接的抓取的问题是相关页面主题团之间的隧道现象,即很多在抓取路径上偏离主题的网页也指向目标网页,局部评价策略中断了在当前路径上的抓取行为。文献[21]提出了一种基于反向链接(BackLink)的分层式上下文模型(Context Model),用于描述指向目标网页一定物理跳数半径内的网页拓扑图的中心 Layer0 为目标网页,将网页依据指向目标网页的物理跳数进行层次划分,从外层网页指向内层网页的链接称为反向链接。

4.1.2 网站粒度的分析算法

网站粒度的资源发现和管理策略也比网页粒度的更简单有效。网站粒度的爬虫抓取的关键之处在于站点的划分和站点等级(SiteRank)的计算。SiteRank 的计算方法与 PageRank 类似,但是需要对网站之间的链接作一定程度抽象,并在一定的模型下计算链接的权重[19]。

网站划分情况分为按域名划分和按 IP 地址划分两种。文献[18]讨论了在分布式情况下,通过对同一个域名下不同主机、服务器的 IP 地址进行站点划分,构造站点图,利用类似 PageRank 的方法评价 SiteRank。同时,根据不同文件在各个

站点上的分布情况,构造文档图,结合 SiteRank 分布式计算得到 DocRank。文献[18]证明,利用分布式的 SiteRank 计算,不仅大大降低了单机站点的算法代价,而且克服了单独站点对整个网络覆盖率有限的缺点。附带的一个优点是,常见的 PageRank 造假难以对 SiteRank 进行欺骗。

4.1.3 网页块粒度的分析算法

在一个页面中,往往含有多个指向其他页面的链接,这些链接中只有一部分是指向主题相关网页的,或根据网页的链接锚文本表明其具有较高重要性。但是,在 PageRank 和 HITS 算法中,没有对这些链接作区分,因此常常给网页分析带来广告等噪声链接的干扰。在网页块级别(Block-level)进行链接分析的算法的基本思想是通过 VIPS 网页分割算法将网页分为不同的网页块(page block),然后对这些网页块建立 page-to-block 和 block-to-page 的链接矩阵,分别记为 Z 和 X。于是,在 page-to-page 图上的网页块级别的 PageRank 为 $W_p = X \times Z$;在 block-to-block 图上的 BlockRank 为 $W_b = Z \times X$。已经有人实现了块级别的 PageRank 和 HITS 算法,并通过实验证明,效率和准确率都比传统的对应算法要好。

4.2 基于网页内容的网页分析算法

基于网页内容的分析算法指的是利用网页内容(文本、数据等资源)特征进行的网页评价。网页的内容从原来的以超文本为主,发展到后来动态页面(或称为 Hidden Web)数据为主,后者的数据量约为直接可见页面数据(PIW, Publicly Indexable Web)[11]的 400 ~ 500 倍[9]。另一方面,多媒体数据、Web Service 等各种网络资源形式也日益丰富。因此,基于网页内容的分析算法也从原来的较为单纯的文本检索方法,发展为涵盖网页数据抽取、机器学习、数据挖掘、语义理解等多种方法的综合应用。本节根据网页数据形式的不同,将基于网页内容的分析算法,归纳以下三类:第一种针对以文本和超链接为主的无结构或结构很简单的网页;第二种针对从结构化的数据源(如 RDBMS)动态生成的页面,其数据不能直接批量访问;第三种针对的数据界于第一和第二类数据之间,具有较好的结构,显示遵循一定模式或风格,且可以直接访问。

4.2.1 基于文本的网页分析算法

1) 纯文本分类与聚类算法

很大程度上借用了文本检索的技术。文本分析算法可以快速有效的对网页进行分类和聚类,但是由于忽略了网页间和网页内部的结构信息,很少单独使用。

2) 超文本分类和聚类算法

网页文本还具有大量的 < Title > , < head > , < hn > 等有用的标记信息。这些结构中表示不同的内容的重要程度。可以有效地提高分类精度,降低复杂度。

4.2.2 Hidden Web 的网页分析方法

大约 80% 的数据是动态生成的。这些内容大多"隐藏"存储在后台的可查询数据库中,因此称为"Hidden Web"[10]。目前大多数的通用搜索引擎仅仅覆盖了部分的 PIW[11],却忽略了数据量约为 PIW 400 ~ 500 倍的 Hidden Web(或称为 Deep Web)[9]。针对 Hidden Web 的爬虫与普通的聚焦爬虫相比,需要更多地对网页中表单进行发现、探测查询(probing query)和分析。

对于网页上表单的处理很多时候需要采用用户辅助的半自动方法,如典型的 HIWE 系统[6]。该方法将表单表示为一组(element, domain)二元组,并尝试通过标注、页面布局等信息确定表单的输入数据模式。另一种无需人工辅助的方法则需要更多对网页后台数据库的反复查询,分析结果的数据和属性,在利用熵理论上,文献[7]的思路与文献[6]类似,但采用了无需人工辅助的方法来自动发现领域相关的 Hidden Web 资源。

4.2.3 数据密集型网页的分析方法

数据密集型(data-intensive)网页的数据形式介于 Hidden Web 和文本密集型网页之间。它们具有良好的结构性,又可以直接从页面读取;而且数据的语义在网页上显示标注,因此不需要对这些网页之后的数据库进行探测查询。例如电子商务网站的产品信息页面,具有统一的风格,其中的数据表示具有固定格式,并按照一定目录层次结构来组织,因此也称为分类导向型(taxonomy-directed)网页[13]。的获取工作主要集中在对网页数据的抽取,如页面块或目录发现[13],结构化数据的记录边界确定等等。爬虫将这些数据抽取出来,以一定格式在本地存储、分析,从而指导下一步的抓取工作。基本思路是,将 html 页面转化为 token 序列或标记树(tag tree),如 DOM 树等数据结构,再在这种转化的数据结构上进行模式发现,实现从抽取出结构化的数据。

普遍采用的方抽取法是 wrapper 提取页面信息。Wrapper 可人工维护,或半自动的生成。这种方法通常具有较强的针对性和局限性,动态性常常导致 wrapper 失效,因此需要大量的 wrapper 维护和用互。另一种方法是从具有统一风格和显示规则的若干网页来学习并抽取结构化数据。包括:进行连续数据记录之间的记录边界发现;在页面所转化成的标记串上做模式发现分析;当网页上的数据记录不连续,记录的显示风格也不完全一致时,就需要更鲁棒的算法查找标记树种的重复结点。在很多情况下,这些数据是以 HTML 的表格形式(< table > < /table >)出现的。

4.3 用户协作网页分析算法

链接提供的网页关联度往往带有噪音,网络的异构性和动态性使得对链接结构的建模很难达到令人满意的效果。而用户的访问模式往往可靠反映了资源的主题相关性,且具有时效性,可即时反应网络链接的变更等情况。文献[14~16]提出了通过用户协作、学习浏览模式来抓取网页的方法。协作抓取需要获取用户浏览行为,一般有两种方法:日志挖掘和用户标注。

文献[14]提出了用户浏览模式挖掘法,对与某一特定查询谓词相关的网页作相似性建模。以大量公共域名代理的用户访问日志为参考,经过对大群组用户信息过滤,统计并总结出了三种需要考虑的用户访问信息:对不同网页访问频率;对不同网页特征访问频率;访问同一主题网页的时间局域性。其中,试验表明,协作抓取比基于链接的智能抓取(intelligent crawling)策略有更好的准确性。

文献[15,16]则以用户在浏览过程中,对"有用"网页进行显式标注的网页集合为参考。利用隐含马尔可夫模型(Hidden Markov Model)适于进行动态模式识别模型的特性,学习用户的浏览行为,预测不同网页聚类之间的语义联系。

4.4 基于领域概念定制的网页评价算法

聚焦抓取常以三种方法表示:(1)预给初始种子样本(如种子 URL,目标网页样本等);(2)预定网页分类结构(如 yahoo!)和网页训练集生成的分类器;(3)用户显式标注的或从日志推理得到的"有用"样本。

三种方法都只是对抓取行为的"主题性"或所关心的"领域"给出了模糊的定义。文献[22,23]采用了预定义的本体信息,文献[24]采用了领域核心概念的模式定义,文献[25]采用了领域相关的词典以及预定义的元搜索查询语句来表示领域概念。

在文献[22]中,领域本体由不同的概念、实体及其之间的关系,以及与之对应的词汇入构项(lexical entry)组成。网页中的关键词在通过与领域本体对应的词典作规范化转换之后,进行计数和加权,算出与所选领域的相关度。对规范化后的词进行加权时,根据本体的概念层次,离核心概念越近的权重越高。该方法比广度优先抓取策略和基于关键字的分类分析算法具有更高的准确性和效率。另一个优点在于,即使初始 URL 与领域不直接相关,也具有较好的抗干扰能力,并逐渐趋近于主题相关的抓取路径。

相对于文献[22],文献[23~25]使用了较为简单的本体表示信息:词典。其中,文献[23,24]通过对领域概念信息的结构化定义,提供了类似 SQL 的查询语言,使不同的关键字具有一定的语义结构和联系。文献[23,24]中,领域信息定制使爬虫可有目的地从网页上抓取有结构数据,返回结果具有一定结构的数据信息及附带的若干网页 URL 等。通过词典进行字词扩展,文献[24,25]都使用了元搜索技术。领域信息和元搜索技术结合,使得爬虫抓取过程中,可兼顾准确率和覆盖率,摆脱了局部搜索算法在遇到 Web Community 时所受到的局限性[25]。

4.5 小结

表 1 简要概括了上述介绍的各种方法。这些方法各有其特点和应用场合。它们之间并不是独立的,如果能够有效地结合使用,爬虫的准确率和效率往往比单独使用某种方法更好。

表 1 各种网页分析方法概述

算法类别	具体算法	算法概述
基于网络拓扑	PageRank, HITS, Back-crawling SiteRank Block-level PageRank & HITS	根据网页链接结构计算等级,决定搜索的次序。粒度不同,带来不同准确度和复杂度。优点:分析算法较简单。缺点:主题精确性较低
基于网页内容	Text & Hypertext analysis Hidden Web Data-intensive Web	考虑网页内容特征,根据特征相似度来确定搜索顺序。网页内容依赖于信息抽取技术。算法有效性依赖于网页内容的提取
基于用户访问行为	Collaborative crawling	用户评价评决定搜索顺序。结合网络日志挖掘和 HMM 的统计模型使用。优点:真实反映用户兴趣,及时反映网络更新
基于领域概念定制	Domain Ontology / Lexicon Meta search	通过定义领域并进行搜索,克服了最佳优先局部搜索方法的局限性,提高了抓取的领域相关性

5 聚焦爬虫系统体系结构

下面分析三种具有带表性的体系结构。

5.1 基于分类器的聚焦爬虫

文献[12]是具有代表性的聚焦爬虫的早期研究之一,目前大多数的聚焦抓取都采用了类似的工作流程,其系统结构如图 2 所示。根据一个主题目录和用户指定的初始点(如书签)来描述抓取目标,并在用户浏览过程中,将用户标注的感兴趣网页放入相应的主题目录,修改主题样本。系统的两个主要部分是网页分类器和网页选择器(distiller)。网页分类器负责学习抓取目标的特点,计算网页的关联度,并过滤网页。选择器负责计算网页的重要程度,发现中心型网页,并由此动态决定网页的访问顺序。

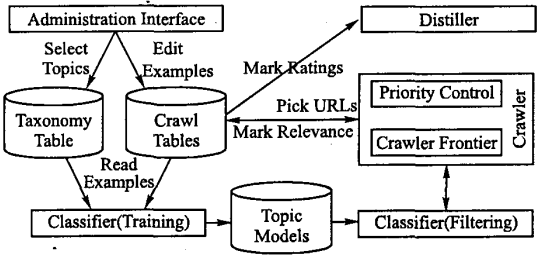


图 2 基于分类器聚焦爬虫体系结构[12]

5.2 基于数据抽取器的聚焦爬虫

如图 3,聚焦爬虫从网络上发现所需要的资源网页,传递给数据抽取器进行数据抽取工作。数据抽取可能是对 Hidden Web 之后的数据库进行探测查询,也可能是通过 wrapper 或网页外观格式信息进行数据抽取。得到的原数据将由数据分析器作分析和标注,才能转化为符合系统数据库模式的结构化数据进行存储和索引。

5.3 基于用户学习的聚焦爬虫

此种聚焦爬虫的体系结构关键在于对用户的浏览模式的学习模块。文献[15]的爬虫体系结构如图 4 所示。系统通过记录用户在浏览到感兴趣的网页时,点击"有用"按钮的行为,记录相相关网页,为用户建立浏览模型。之后,系统对经过标注的网页进行 LSI 聚类,根据聚类结果构建概念图,表示不同类别之间的联系。这样,系统就可以利用不同网页间的链接关系,学习用户的浏览行为。用户浏览模式学习还可以通过对日志的挖掘来获得[14],其具体方法与文献[15]不同,但是系统同样要通过建模和学习两个部分来完成对爬虫的指导。

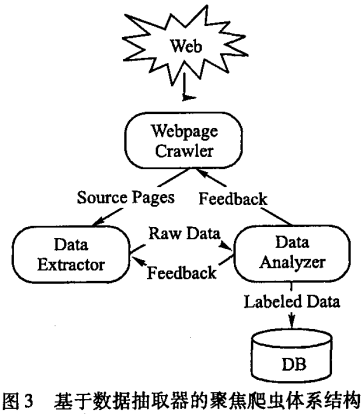


图 3 基于数据抽取器的聚焦爬虫体系结构

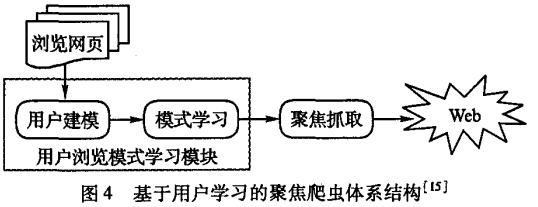


图 4 基于用户学习的聚焦爬虫体系结构[15]

6 结语

本文介绍并分析了聚焦爬虫中的关键技术:抓取目标定

义与描述,网页分析算法和网页搜索策略,并根据网络拓扑、网页数据内容、用户行为等方面将各种网页分析算法作了分类和比较。虽然目前已经存在多种算法和实现系统,聚焦爬虫技术仍有许多值得研究和探讨的课题,归纳起来有以下几个方面:

(1) 面向数据分析和挖掘的爬虫技术研究。更高级的应用应当是从抓取到的网页中学习、挖掘出有价值的结论、关系等知识。数据仓库在 Web 挖掘上往往限于针对结构化的日志数据。突破这一局限,研究为高级需求提供服务的爬虫,使得数据仓库、数据挖掘等方法能够应用于网页数据具有极大的挑战性。

(2) 主题的描述与定义。准确地表示用户关心的主题十分关键。它与抓取目标的描述、抓取目标的分析、用户的查询方式互相联系。这对爬虫的智能处理能力提出了很高的要求。

(3) 相关资源的发现。对于数据密集型网页和 Hidden Web 的抓取,如何在抓取中尽快地发现资源网站或 hub 网页,是提高抓取效率的关键[7]。

(4) Web 数据清洗。数据清洗是对抓取结果的再处理,可以有效地提高抓取数据的质量,这在万维网异构的数据源环境下尤其突出。如网页的去重,网页抽取数据不一致等问题都有待更成熟的方法与技术。

(5) 抓取目标空间的扩展。目前爬虫不能访问占绝大多数的 PC,将来的爬虫将能够更普遍的抓取各种存贮设备上的信息。

参考文献:

[1] PINKERTON B. Finding what people want: Experiences with the web crawler[A]. Proceedings of the Second World-Wide Web conference[C]. Chicago, Illinois, October 1994.

[2] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

[3] KLEINBERG J. Authoritative sources in a hyperlinked environment [A] TARJAN RE, *et al.* ed. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms[C]. New Orleans: ACM Press, 1997. 668-677.

[4] BHARAT K, HENZINGER M. Improved algorithms for topic distillation in a hyperlinked environment[A]. VOORHEES E, *et al.* ed. Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval. Melbourne[C]. ACM Press, 1998. 104-111.

[5] CHAKRABARTI S, DOM B, GIBSON D, *et al.* Automatic resource compilation by analyzing hyperlink structure and associated text[A]. THISTLEWAITE P, *et al.* eds. Proceedings of the 7th ACM-WWW International Conference[C]. Brisbane: ACM Press, 1998. 65-74.

[6] RAGHAVAN S. Hector Garcia-Molina, Crawling the Hidden Web [A]. Proceedings of the 27th International Conference on Very Large Data Bases[C], September 2001.

[7] BERGHOLZ A, CHIDLOVSKI B. Crawling for Domain-Speci. c Hidden Web Resources[A]. Proceedings of the Fourth International Conference on Web Information Systems Engineering[C], December 2003.

[8] IPEIROTIS PG, GRAVANO L, SAHAMI M. Probe, count, and classify: Categorizing hidden-web databases[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data [C]. Santa Barbara, CA, USA, May 2001. 67-78.

[9] The Deep Web: Surfacing Hidden Value. http://www.completeplanet.com/Tutorials/DeepWeb/[EB/OL].

[10] FLORESCU D, LEVY AY, MENDELZON AO. Database techniques for the world-wide web: A survey[J]. SIGMOD Record, 1998, 27(3): 59-74.

[11] LAWRENCE S, GILES CL. Searching the World Wide Web[J]. Science, 1998, 280(5360): 98.

[12] CHAKRABARTI S, VAN DEN BERG M, DOM B. Focused crawling: A new approach to topicspecific web resource discovery[A]. Proceedings of the Eighth International World-Wide Web Conference[C], 1999.

[13] DAVULCU H, KODURI S, NAGARAJAN S. Datarover: a taxonomy based crawler for automated data extraction from data-intensive websites[A]. Proceedings of the 5th ACM international workshop on Web information and data management[C], November 2003.

[14] AGGARWAL CC. Collaborative Crawling: Aggarwal C. Collaborative crawling: mining user experiences for topical resource discovery [A]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining[C], July 2002.

[15] LIU HY, MILIOS E, JANSSEN J. Probabilistic models for focused web crawling[A]. Proceedings of the 6th annual ACM international workshop on Web information and data management[C], November 2004.

[16] LIU HY, MILIOS E, JANSSEN J. Focused Crawling by Learning HMM from User's Topic-specific Browsing, Proceedings of the Web Intelligence[A]. IEEE/WIC/ACM International Conference on (WI'04)[C], September 2004.

[17] ESTER M, KRIEGEL HP, SCHUBERT M. Accurate and Efficient Crawling for Relevant Websites[A]. Proceedings of the 30th VLDB Conference[C]. Toronto, Canada, 2004.

[18] WU J, ABERER K. Using SiteRank for Decentralized Computation of Web Document Ranking[EB/OL]. http://project.alvis.info/alvis_docs/siterank.pdf.

[19] BHARAT K, CHANG BW, HENZINGER M, *et al.* Who links to whom: Mining linkage between web sites[A]. Proceedings of the IEEE International Conference on Data Mining (ICDM '01)[C]. San Jose, USA, November 2001.

[20] ESTER M, KRIEGEL HP, SCHUBERT M. Website Mning: A new way to spot Competitors, Customers and Suppliers in the World Wide WEb[A]. Proceedings ACM SIGKDD[C], 2002.

[21] DILIGENTI M, COETZEE F, LAWRENCE S, *et al*. Focused Crawling Using Context Graphs[A]. Proceedings of the 26th International Conference on Very Large Data Bases[C], September 2000.

[22] EHRIG M, MAEDCHE A. Ontology-focused crawling of Web documents[A]. Proceedings of the 2003 ACM symposium on Applied computing[C], March 2003.

[23] GUO Q, GUO H, ZHANG ZQ, *et al.* Schema Driven Topic Specific Web Crawling[A]. DASFAA[C], 2005.

[24] GRAUPMANN J, BIWER M, ZIMMER C, *et al.* COMPASS: A Concept-based Web Search Engine for HTML, XML, and Deep Web Data[A]. Proceedings of the 30th VLDB Conference[C], 2004.

[25] QIN JL, ZHOU YL, CHAU M. Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method[A]. Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries[C], June 2004.

[26] CHO J, GARCIA-MOLINA H, PAGE L. Efficient crawling through URL ordering[A]. Proceedings of the seventh international conference on World Wide Web 7[C], April 1998.