主题 Web 挖掘研究

杜光芹[1,2],张化祥[1],赵瑞东[2]
(1. 山东师范大学 信息科学与工程学院,山东 济南 250014;
2. 浪潮集团,山东 济南 250014)

摘 要:网络已经成为人们获取知识的一个重要途径。然而面对巨大的Web资源库,用户若想获得所需要信息已不再是一件简单的事情。通用搜索引擎返回大量的无关信息,不能满足用户的特定信息检索需求。针对这个问题,Web信息检索领域出现了一个新的研究方向——主题驱动的Web资源发现。介绍了通用搜索引擎的基本结构、工作原理及现状。阐述了主题Web挖掘的研究背景、任务及目前研究技术的进展,并对其未来的发展方向进行了探讨。对通用搜索引擎和主题Web挖掘的关系进行了分析。

关键词:搜索引擎;信息检索;Web主题挖掘;聚焦爬虫;本体论
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2008)02-0094-04

State of Topic Web Mining

DU Guang-qin[1,2], ZHANG Hua-xiang[1], ZHAO Rui-dong[2]
(1. School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;
2. Langchao Group, Jinan 250014, China)

Abstract: Internet provides more and more information and it's not easy for people to get what they want. Web often provides a lot of irrelevant information by universal search engine, so a new research direction has been proposed in Web information retrieval community, which is termed as 'topic-driven Web resource discovery'. Introduces the basic architecture and work mechanism and the status of the universal search engine, then introduces the research background, the task and the present research technology of the topic Web mining, and some of the future works are analyzed. Finally analyze the relationship between the universal search engine and the topic Web mining.

Key words: search engine; information retrieval; topic Web mining; focused crawler; ontology

0 引 言

伴随着科技的发展,信息化已经是现代社会发展的一个方向,而网络已成为人们获取信息、传递信息的重要途径,随之而来的是造成了Web的爆炸性增长。据CNNIC 2007年1月23日发布的《第19次中国互联网络发展状况统计报告》[1]显示,截至到2006年底,全国域名数为4 109 020个,半年增长116万,平均每月净增20万个;网页数和网页字节总数分别为44.7亿个和122 306GB,与去年同期相比分别增长86.3%和81.7%。comScore公布的关于2006年世界互联网使用人数及使用状况的报告显示,截至2007年1月,全球15岁以上网民人数达到7.47亿,比去年同期增长10%。

针对以上问题,人们试图找到更为有效的Web信息检索方法,以保证更高的查全率和查准率。由于不同背景、不同目的和不同时期的Web用户有不同的检索需要,因此当要满足一些高级或专业性的信息检索要求时,就需要获得一个面向特定主题(或者特定领域)的全面的Web页面集合,这就是信息检索领域的新方向——主题Web挖掘。

1 通用搜索引擎

通用搜索引擎是一个专用的WWW服务器,它位于Web信息检索系统层次分类的底层,以Web信息为处理对象,向元搜索引擎和信息检索Agent提供很有价值的服务。

1.1 通用搜索引擎的基本结构和工作原理

虽然各个搜索引擎的具体实现不尽相同,但一般

收稿日期:2007-05-15
基金项目:山东省中青年科学家科研奖励基金(博士基金)资助项目(2006BS01020)
作者简介:杜光芹(1980-),女,山东日照人,硕士研究生,研究方向为数据挖掘、信息检索;张化祥,教授,博士,高级工程师,研究方向为机器学习、多代理协同、数据挖掘及知识发现新算法等。

包含 5 个基本部分:Robot、分析器、索引器、检索器和用户接口。其结构图如图 1 所示。

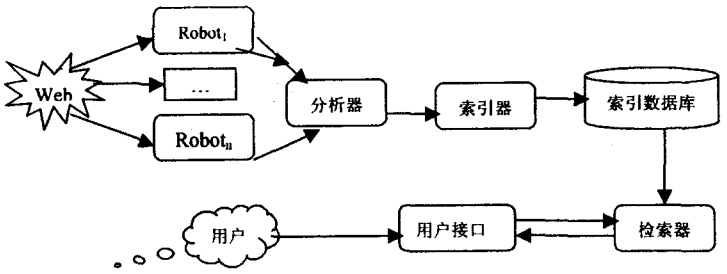


图 1　通用搜索引擎的基本结构图

它的基本原理是:Web 爬虫(即 Robot,也称为 spider,crawler 或 wander)采用广度优先(或深度优先或最佳优先)的爬行(Crawling)策略,周期性对 Web 网页进行遍历并下载,为了提高效率,搜索引擎中可能会有多个 Robot 进程同时遍历不同的 Web 子空间;分析器对 Robot 下载的文档进行分析以提供给索引器使用;索引器根据定义的索引要求建立基于相应检索元的索引库,将文档表示为一种便于检索的方式并存储在索引数据库中;检索器从索引中找出与用户查询请求相关的文档;用户接口为用户提供可视化的查询输入和结果输出界面。

1.2　通用搜索引擎的现状

目前通用搜索引擎已经成为 Web 信息检索(Web Information Retrieval,Web IR)系统中 Web 用户访问 Web 的最广泛使用的工具。CNNIC 2001 年 7 月的调查数据表明,网民最常使用的网络服务中,搜索引擎占 51.3%,仅次于电子邮件。搜索引擎,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和手段,已经成为 Web 用户访问万维网的入口和指南。然而由于 Web 信息资源的指数级增长以及 Web 信息的特点,使通用搜索引擎存在很大的局限性,例如:

* 由于其通用性,返回的结果往往包含大量用户不关心的网页,因而查准率不高;

* 由于 Web 信息的海量性和快速增长,有限的搜索引擎服务器资源造成了查全率不能保证;

* 用户的查询条件一般都是以简单的关键字通过简单的布尔运算构成,因而难以精确地表达所要查询的语义信息;

* Internet 上的信息源是动态变化的,而通用搜索引擎的爬行周期一般较长,因而导致了用户得到的信息往往是几天或几周前的信息。

2　主题 Web 挖掘

主题 Web 挖掘是近几年来刚刚兴起的一个热门研究课题,它对目前 Web 上信息检索比较困难的问题来说是一个富有前景的解决方法,因此得到了国际上许多学者的广泛关注。

2.1　主题 Web 挖掘研究背景

主题 Web 挖掘的一个重要的应用背景就是特定领域的搜索引擎(Domain - specific Search Engine,DSSE),又称专业型搜索引擎或专题搜索引擎。DSSE 就是将系统的检索服务限定在特定的领域内(如科技论文、商业领域等),仅提供给定领域相关信息的搜索服务。目前,比较典型的系统有:NEC 研究院的 CiteSeer (http://citeseer.ist.psu.edu),它提供计算机科技论文的检索;FlipDog(http://www.flipdog.com/)搜索引擎则提供有关 IT 工作职位查询的服务。

DSSE 的首要任务是必须搜集到给定领域的相关网页。通常,DSSE 首先通过人工选定一些专业网站或通过通用搜索引擎获得相关网页的 URL,然后将这些 URL 交给 Web 爬虫作为种子 URL 进行爬行。虽然这种方法能得到质量较高的网页,但比较繁琐,搜集到的网页数量比较有限,而且由于 Web 信息的海量性和快速增长的特点,它很难满足实际应用的需要。因此,需要采用自动化的面向领域/主题的网页搜集系统[2],这也正是主题 Web 挖掘研究的一个重要应用背景。

2.2　主题 Web 挖掘研究任务

主题 Web 挖掘研究任务首先是根据用户或信息检索系统定义的目标主题,以智能的主题爬虫[3~5]在线爬行(Crawl,即下载)Web;然后再对收集到的页面集进行智能的分析和处理,以便用户可以方便地进行检索和使用。

2.2.1　主题爬虫技术

面向主题的主题爬虫是主题 Web 挖掘的核心技术。它的理论和技术基础主要包括机器学习、信息检索、概率统计理论和 Web 新技术。在很多应用领域,例如基于 Web 的行业分析、在线商业竞争分析和自动构建专题数字图书馆等,主题 Web 挖掘系统都富有应用前景,与现有的 Web 搜索引擎可以形成良好的互补。

聚焦爬虫(Focused crawler)[4,6],又称为主题爬虫(Topical Crawler)或主题驱动的爬虫(Topic - driven crawler)。它是一种智能的 Web 爬虫,它的基本工作过程与普通爬虫是相似的:

(1) 将一组种子 URL 作为 URL 队列的初始元素;

(2) 按照某种爬行策略从 URL 队列中取出一个

URL,通过 HTTP/HTTPS 下载模块爬行与该 URL 相对应的 Web 页面;

(3) 然后从已爬行的页面中抽取出新的 URL,并将它们插入到 URL 队列中去;

(4) 重复(2),(3)两步直到爬行到预定义数量的页面。

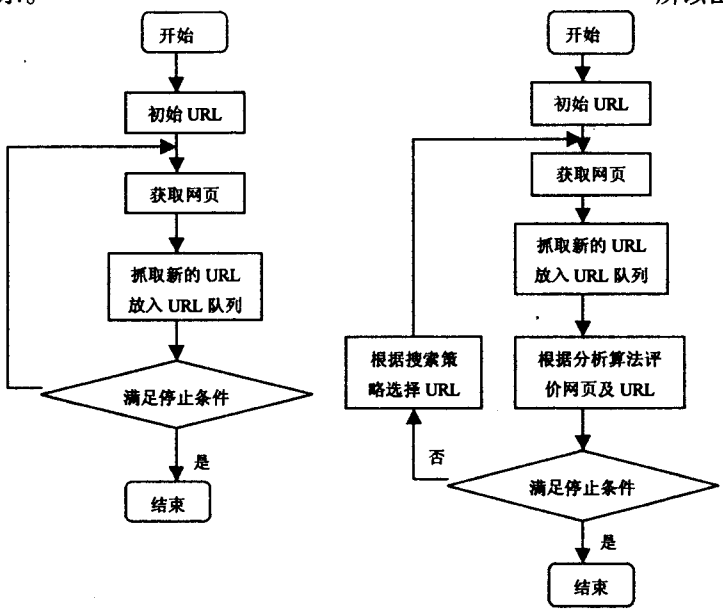
相对于普通爬虫,聚焦爬虫有其自身的研究问题:

①对抓取目标的形式表示和定义;

②对网页或数据的分析与过滤;

③爬行结果性能评价。

普通爬虫和聚焦爬虫工作流程对比如图 2[6]所示。



(a) 传统爬虫工作流程 (b)聚焦爬虫工作流程

图 2 传统爬虫与聚焦爬虫工作流程对比

从图 2 可以看出,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,然后再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。由于 Web 页面是超文本形式,它不仅包含了文字内容信息,还具有其它的一些重要特征,如文档的结构信息、超链接结构信息和元数据,因此如何充分利用这些信息是当前研究的一个重要方向。

目前的网页分析算法可以分为基于网页内容、基于网络拓扑和基于用户访问行为三种类型。基于网页内容的网页分析算法是根据网页的文本内容、文本结构和隐藏 Web[7]等数据特征对网页进行的评价算法[8]。提出了一种能够针对人们关心的热点主题,系统地对网上信息进行收集和分析的模型,从不同的角度和层次得出互联网对该主题报道的强度,对社会科学类研究具有一定的参考价值。

基于网络拓扑的网页分析算法是根据网页之间的链接,通过已知的网页或数据,来对预期有直接或间接链接关系的对象作出评价的算法。文献[9]通过实验数据集测试得出如下结论:链接描述文本对目标网页的主题概括具有高度的准确性,与传统方法相比,使用链接文本在已知网页的定位任务上能够使系统性能提高 96%;把基于链接描述文本的方法与传统方法相结合,能够在检索性能上提高近 16%。

由于用户访问行为模式往往能反映主题资源的相关性,且具有时效性,能及时反映网络连接变更情况,所以出现了基于用户访问行为的网页分析算法。文献[3,5]提出了在用户浏览过程中,通过学习浏览模式来抓取网页的算法,利用了隐含马尔科夫模型适于进行动态模式识别的特性,学习用户的浏览行为,预测不同网页聚类之间的语义联系。

目前也出现了基于本体论知识的领域概念定制的网页分析算法。文献[10,11]使用了较为简单的本体信息—词典,对领域概念信息进行结构化定义,使不同的关键字具有一定的结构和联系。文献[12]提出了利用 Ontology 拥有的领域知识,将 Ontology 融合到信息检索技术中,从而大大提高检索系统对自然语言文本的理解能力,从而提高检索效果。文献[13]提出了建立在本体论基础上的语义信息检索的原型系统。本体论在主题 Web 挖掘中的应用是今后的研究方向。

2.2.2 爬行结果处理技术

主题爬虫的爬行结果是一个预定义大小的页面集合,这个集合一般随应用系统需求的不同而有较大差异,但一般也至少包含数千个页面,有的也可能是百万数量级。由于 Web 信息的多样性和复杂性,主题爬虫爬行下来的页面集也会包含一些主题不相关页面,因而需要对爬行结果进一步处理。例如可以对结果页面集建立索引,构造面向主题的搜索引擎、个性化的搜索引擎等。目前的研究主要包括如下内容:

(1)页面检索和相关度评价:即与搜索引擎中提供的关键字查询相似,提供页面查询功能,使用户可以在爬行页面集上进行二次检索;

(2)超文本分类与过滤:就是通过研究一些算法来对主题爬虫爬行下来的那些不相关页面,进行自动过滤;

(3)超文本聚类与可视化:超文本聚类就是根据页面的主题相似性自动将一个页面集合聚集成若干个不

同类别的小集合的过程。

3 主题Web挖掘和通用搜索引擎的关系

主题Web挖掘是在Web信息日益过载的背景下提出来的,它对于提高Web信息检索理论和技术的发展有着重要的影响。但是主题Web挖掘并不能代替通用搜索引擎,它们之间是必要的互补关系。Menczer在文献[3]中也提到了这一点。

一方面,由于通用搜索引擎具有海量的索引页面,覆盖了Web用户进行信息检索时可能涉及到的大部分主题,因此它能够满足一般用户的检索需要,并且在今后相当长时间内仍将是Web用户进行Web信息检索的主要工具。而对于主题Web挖掘来说,它在利用主题爬虫爬行Web时,通过查询通用搜索引擎可以方便地获得比较好的种子URL和样本页面[9],从而加快主题爬行的效率。

另一方面,通用搜索引擎无法满足一些高级或专业性的信息检索的要求,而这些高级信息检索要求的特点主要体现在:

* 一般是针对某一个主题或领域。

* 以一定的方式表示用户的兴趣,建立用户兴趣模型。

* 要求检索结果比较全面,并且要以一定的精度得到目标主题相关的页面。

以上这些检索要求正是主题Web挖掘的目标,因此可以很好地弥补通用搜索引擎在这方面的不足。

4 小 结

Web的爆炸性增长,使它成为世界上最大的信息库。面对这个海量、异构、半结构化的信息库,Web用户往往感到无所适从。为了解决这个问题,出现了主题Web挖掘的研究。主题信息的获取和表示是影响主题爬虫性能的一个重要因素,用户提供的主题信息越准确则系统的收获率将越高,因此,如何结合用户的反馈信息,从而实现交互式、逐步求精的主题爬行将是一个重要的研究课题;对于是否存在一个理论模型来证明多特征表示方法的内在优越性,这也将是一个值得研究的方向;同时,基于多特征表示的分类器集成方法还需要在更多的分类器上进行全面测试;鉴于Web页面的多主题性以及查询结果页面所具有的主题聚合性,将查询结果聚类技术与超链接分析技术结合起来将是提高页面评价算法性能的一个重要研究方向。

参考文献:

[1] 中国互联网络信息中心(CNNIC).第19次中国互联网络发展状况统计报告[EB/OL].2007.http://www.cnnic.net.cn/html/Dir/2007/01/22/4395.htm.

[2] 宋聚平,王永成,尹中航,等.面向主题的网页搜索系统[J].上海交通大学学报,2003,37(3):401-403.

[3] Liu H Y,Milios E,Janssen J. Probabilistic models for focused web crawling[C]//WIDM'04. Washington,DC,USA:ACM,2004.

[4] Zhang H X, Huang S T. An Incremental Approach to Link Evaluation in Topic-Driven Web Resource Discovery[C]//LNCS.[s.l.]:[s.n.],2005:301-310.

[5] Liu H Y,Milios E,Janssen J. Focused Crawling by Learning HMM from User's Topic-specific Browsing[C]//Proceedings of the web intelligence. IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC,USA:IEEE Computer Society,2004.

[6] 周立柱,林 玲.聚焦爬虫技术研究综述[J].计算机应用,2005,25(9):1965-1969.

[7] Florescud,Levyay,Mendelzonao. Database techniques for the world-wide Web:A survey[J]. SIGMOD Record,1998,27(3):59-74.

[8] 李晓明,朱家稷,闫宏飞.互联网上主题信息的一种收集与处理模型及其应用[J].计算机研究与发展,2003,40(12):1667-1671.

[9] 张 敏,高剑锋,马少平.基于链接描述文本及其上下文的Web信息检索[J].计算机研究与发展,2004,41(1):221-226.

[10] Guo Q,Guo H,Zhang Z Q,et al. Schema Driven Topic Specific Web crawling[C]//Lecture Notes in Computer Science. Berlin/Heidelberg:Springer,2005:594-599.

[11] Graupmann J, Biwer M, Zimmer C, et al. COMPASS:A Concept-based Web Search Engine for HTML,XML,and Deep Web Data[C]//Proceedings of the 30th VLDB Conference. Toronto,Canada:[s.n.],2004.

[12] 陈 康,武港山.基于Ontology的信息检索技术研究[J].中文信息学报,2005,19(2):51-57.

[13] 王 诚,张 璟.基于语义的Web信息检索[J].计算机应用研究,2005(8):111-112.

(上接第93页)

[3] Henning M,Vinoski S.基于C++CORBA高级编程[M].北京:清华大学出版社,2000:25-28.

[4] 朱其亮,邓 斌.CORBA原理及应用[M].北京:北京邮电学院出版社,2001:121-125.

[5] 王育坚,刘 展,马小军.基于CORBA/Web的网管系统的设计与实现[J].计算机应用,2006,26(1):43-45.

[6] 赵 慧,施伯乐.基于CORBA的网络管理的若干关键问题研究[J].计算机科学,2003,30(3):111-113.