基于访问内容类型统计的 Web Robot 检测算法

郭伟刚[1,2] 鞠时光[2]

(广东佛山科学技术学院信息中心 佛山 528000)[1] (江苏大学计算机学院 镇江 212013)[2]

摘 要 随着搜索引擎的广泛使用,由此而引起的网络机器人(Web Robot)对于 Web 站点的访问所产生的影响必须引起重视。该文分析了网络机器人的访问行为特点,提出了一个基于访问内容类型统计的检测算法。经实验验证,该算法可以有效地检测未知的和不遵守网络机器人排斥标准的 Robot。

关键词 搜索引擎,网络机器人,内容分类,检测,Web 日志

A Web Robot Detection Method Based on Content Classification and Statistics

GUO Wei-Gang[1,2] JU Shi-Guang[2]

(Information and Educational Technology Center, Foshan University, Foshan 528000)[1]

(School of Computing, Jiangsu University, Zhenjiang 212013)[2]

Abstract With the widely use of search engines, the impact Web robots have on the Web sites should not be ignored. After analyzing the navigational patterns of Web robots, a new algorithm based on content classification and statistics is proposed. The experiment shows that the new algorithm can detect the unknown robots and unfriendly robots who do not obey the Standard for Robot Exclusion.

Keywords Search engine, Web robot, Content classification, Detection, Web log

1 引言

网络机器人(Web Robot)是一种能够自动在 Web 上根据某种策略进行远程数据的搜索与获取的程序,也称为网络蜘蛛(Web Spider)或网络爬虫(Web Crawler),各大搜索引擎(如 Google)一般都是通过网络机器人自动对 Web 网站信息进行搜集。但是,网络机器人对于网站的自动访问也带来了许多问题。主要表现在:(1)因为涉及商业秘密,许多电子商务网站不希望未经授权的网络机器人来收集商业信息;(2)电子商务网站需要对网站访问者的访问行为进行分析统计,网络机器人可能引起访问者信息分析的失真;(3)由于涉及信息的时效性和保密性,许多政府网站也不希望自己网站的信息被网络机器人收集和索引;(4)设计不友好的网络机器人进行访问会占用许多带宽,影响正常用户的访问。所以,对于网站管理人员来说,很有必要在众多的访问者中,检测出网络机器人,以采取适当的技术重定向或屏蔽网络机器人的访问。

对于网络机器人的检测,网站管理人员常用的方法是构造一个已知网络机器人的数据库[1],其中数据库记录的字段包括网络机器人的标识 agent 和网络机器人所在的服务器 IP 地址,然后通过检测访问者的 agent 和 IP 地址来进行判别。但是这种方法只对已知的网络机器人有效,对于未知的网络机器人则无能为力。对于网络机器人检测的研究,目前见到的只有文[2]提出的根据网络机器人的访问模式,使用数据挖掘中的分类算法来区别人和网络机器人的访问,但是这种方法比较复杂。本文在分析网络机器人访问行为特征的基础上,给出了一个通过对 Web 日志中访问记录所代表的内容类型进行统计来检测网络机器人的方法。

2 相关概念

网络机器人并不是真正地在 Internet 计算机之间移动,而是驻留在单个主机内的软件。它通过向 Internet 上其它计算机发送 HTTP 请求(主要有 Get ,Post 和 Head 3 种),而获得 Web 文档,并且自动地递归搜索所获得文档中的超级链接(URL)指向的所有其他文档。它的工作方式与人们使用浏览器浏览 Web 站点类似,只是它具有自动性,不需要人的干预。

网络机器人首先选定一组 URL (种子 URL)地址,然后从列表中读取部分 URL 地址,从这些地址下载相应的文档页面,提取 HTML 文档中的 URL 链接,并过滤出有效的 URL 链接,检查 URL 列表,把新的 URL 加入到 URL 列表中,然后再从列表中读取 URL,重复以上过程,直到 URL 列表为空或者满足停止条件时停止。

网络机器人访问 Web 站点时,应遵守网络机器人排斥标准[3] SRE(A Standard for Web Robot Exclusion)。SRE 是由网站的系统管理员在 Web 服务器的根目录下建立一个 robots. txt 文件,在该文件中表明网络机器人不应该访问抓取本站点哪些文件。文件的内容由一个或多个记录组成,每个记录以一个或多个 User-agent 行开始,接下来是一个或多个 Disallow 行。User-agent 的值是一些网络机器人的名称,它们是本条记录描述的访问策略的适应对象。Disallow 的值是一些禁止访问的 URL,在同一个记录的 User-agent 行中列出的网络机器人将不能访问这些 URL。例如,下面的内容就表示禁止所有的网络机器人访问/privatedata 目录下的内容。

User-agent: *

Disallow: /privatedata

根据 SRE 的要求,每一个行为良好的网络机器人在访问

郭伟刚 副教授,主要研究方向:数据挖掘、计算机图形学、网络教学技术。**鞠时光** 教授,博士生导师,主要研究领域为数据库、可视化语言、信息安全。

一个网站时，首先访问的是 robots. txt。所以只要检查来自某个 IP 的记录是否访问了 robots. txt，就可判别是否是网络机器人。这是检测未知网络机器人的最简单有效的方法，并且可以用来检测某些恶意的、不遵守规则的网络机器人。例如，明明是在 robots. txt 禁止访问的目录，它却在后续时间内访问了，就可判定是不守规则的网络机器人。但是，这种方法也存在着误判的可能性，主要表现在：

(1)有些用户希望知道某个网站哪些内容是私秘的，有时也会通过浏览器请求 robots. txt。所以，并非所有访问 robots. txt 的记录都是由网络机器人生成的。

(2)当用户使用 IE 浏览器的"添加到收藏夹"功能将当前访问的地址收藏时，在对话框中有一个选项是"允许脱机使用"，当选中该选项时，IE 浏览器会根据设置，自动从服务器下载当前网页以及该网页所链接的网页，以方便用户的脱机浏览。在下载前，浏览器也会访问 robots. txt[4]，而且在下载网页时，会更改其 agent，通常的方法是在原来 agent 的基础上加上 MSIECrawler。

更为重要的是，由于 SRE 不是一个强制性的协议，许多网络机器人根本就不遵守 SRE，因此也就无法根据是否访问 robots. txt 来进行检测，必须通过其他更加有效的方法来检测。

3 基于访问内容类型统计的检测方法

3.1 人与网络机器人的访问行为的差别

网络机器人访问网站的方式与人使用浏览器进行访问具有很大的不同，这种不同体现在服务器日志记录中的典型表现为：人的访问留下的记录的 URL 域是杂乱的，无规律的；而网络机器人的访问留下的记录的 URL 域是有规律的，通常表现为：在一个会话中，所有的 URL 都具有某一特定的文件类型，要么都是 htm，或者都是 jpg(gif)以及 mp3 等。其主要原因在于：

(1)当人在浏览器中输入一个 URL 请求时，浏览器就向目标服务器发送 HTTP 请求。根据 HTTP 协议[5]，服务器收到请求后，就检查自己的计算机中是否有该 URL 指定的文件，若有，则将该文件送出，否则，就给出出错信息。浏览器收到服务器发出的文件后，就会解析该文件。如果是一个单一文件，如图片等，就直接显示；如果是一个 HTML 文件，则要分析出该 HTML 文件中嵌入的对象(如图片、声音、动画、脚本文件、样式表文件、框架网页等)，然后继续自动向服务器发送 HTTP 请求，直到所有嵌入的对象请求完毕。服务器收到请求后，依次发送所请求的文件，浏览器收到这些嵌入的文件，把它们"组装"起来，就形成了用户看到的一个完整页面。所以，在 Web 服务器的日志中，用户的一个请求，可能会生成多条日志记录。而且，由于网页内容的随机性，这些记录的 URL 域所代表的文件类型没有明显的特征。

(2)网络机器人则不同，通常它从待访问的 URL 列表中取得一个 URL(假设是 HTML 文件)后，就向目标服务器发送 HTTP 请求，当收到服务器发回的文件后，它也分析该 HTML 文件包含的超级链接和嵌入的对象，并根据规则将文档中包含超级链接加入到待访问的 URL 列表中。而对于嵌入的对象(如图片、声音、动画、脚本文件、样式表文件、框架网页等)，不同类型的搜索引擎处理的方法各有不同，有的将这些对象的 URL 也加入到待访问的 URL 列表中，有的则直接放弃，也有的修改该对象在所抓取 HTML 文件中的链接而不直接抓取该对象。但有一点是共同的，即，网络机器人并不会马上向服务器发送对这些对象的请求。所以，网络机器人的访问，一次请求，只在服务器日志中留下一条访问纪录，而且这条访问记录的 URL 域所代表的文件类型通常是网页类型(. htm,. asp,. php 等)。这样，在一次会话中，网络机器人的访问所产生的访问记录的 URL 全部是代表网页类型。当然，对于特定的专题搜索引擎的网络机器人，在一个或多个会话中，由于它已经分析了原来读取的网页，因此它所有访问的文件都是 jpg、gif 和 png(图像搜索引擎)或者 mp3(mp3 搜索引擎)。

3.2 访问内容的分类

依据网站的内容特点，可以将网站提供的内容分成 8 大类(见表 1)，根据这个分类方法，就可以对每一条 Web 日志记录赋予一个类型值。需要注意的，由于许多搜索引擎现在已经开始索引文档类资源，因此有时可以将网页类和文档类作为同一个类型。

表 1 网站内容的分类

类型	类型标记	扩展名(举例)
网页类	webpage	htm, html, shtml, asp, pl, php, /
文档类	document	doc, ppt, xls, pdf, ps, txt
脚本类	script	js, css, vbs
图片类	image	jpg,gif,png,bmp
音乐类	music	mid, mp3, wma, rm
动画类	animation	swf, avi
下载类	download	zip, rar, tgz, exe
其他类	others	不属于以上 7 类的任何文件

3.3 检测算法

我们首先对访问内容进行类型统计。基于这个统计，我们开发了如下网络机器人的检测算法：

Step1：进行数据的预处理。把日志中的每一条访问记录处理成：r = 〈IP, agent, time url, type,〉，其中，IP 为用户的地址，agent 为用户代理，url 为用户请求的页面，time 为请求的时间，type 为访问内容的类型，type 根据表 2 从 url 计算得到。所有的访问记录形成用户访问记录集。

Step2：生成用户会话集 S。首先对日志按照 IP 域、agent 域、time 域分别作为第一、第二、第三关键词进行排序，然后对 IP 和 agent 均相同的访问记录，看作是一个访问者所为。若一个访问者的两次访问的时间之间相差某一个固定的时间长度 T，则当作 2 个不同的会话。T 一般可取 15～30 分钟。

用户会话集表示为：S=〈IP, agent, m, {$type_1$, $type_2$, …, $type_m$}〉。其中，m 是该会话的记录总数，$type_1$, $type_2$, …, $type_m$ 为会话中每一条记录访问内容的类型。

Step3：计算每一个会话中每一种访问内容类型的个数，并找出具有最大个数的内容类型。此时，用户会话集可表为：S=〈IP, agent, m, type, n〉。其中，type 为 8 个访问内容类型标记之一，n 为该内容类型的个数。

Step4：筛选出所有 m=n 的会话，形成网络机器人候选集 C，表示为：C=〈IP, agent, m, type〉。

Step5：合并会话，形成合并会话集 M。在网络机器人候选集中计算同一个访问者的相同访问类型的会话总数和访问记录总数。由于一个搜索引擎通常会使用处于同一 C 类地址的多台机器进行信息的收集，因此可以将来自同一 C 类地址、具有相同 agent 的访问者看作是同一个访问者。M 表示为：M = 〈IP, agent, Snumber, Rnumber, type〉。其中，Snumber 和 Rnumber 分别代表 IP 所在的 C 类地址中具有同一 agent 的访问者的 type 类型的会话总数和记录总数。

Step6：考察 Snumber 和 Rnumber，若超过某一个阈值，就可认为是网络机器人。这个阈值可根据网站的具体情况确定，例如，会话的个数为 2，访问的记录总数阈值可设定为 5 等等。

4 实验结果分析

我们选取我校的 Web 服务器(202.192.168.145)2004 年 1 月 21 日至 2 月 7 日的访问日志作为实验对象。日志中访问记录的总数为 50740 条；来自不同的 IP 和 agent，访问 robots.txt 的记录数为 24 个。使用本文的算法，以 $T=20$ 分钟作为分割的时间间隔，得到会话总数为 7432 个；得到的网络机器人候选集 C 的会话总数为 6724 个，其中，网页类会话 424 个，图片类会话 128 个，音乐类会话 6165 个，动画类会话 7 个。由于音乐类访问的记录形成比较复杂(例如，Windows Media Player 对一个 MP3 文件的访问可能会产生多条记录，并且其 agent 有时也不同)，本文暂不对音乐类会话进行处理，因此实际用于进一步处理的网络机器人候选集 C 包括了网页类、图片类、动画类会话共 559 个。最后得到的合并会话集 M 共有 253 个项目。这 253 个项目平均的会话数为 2.2，最大会话数为 96，最小会话数为 1；平均访问记录数为 5.1，最大访问记录数为 249，最小访问记录数为 1。

表 2 检测得到的未访问 robots.txt 的 IP 以及 agent

IP	agent	访问类型
210.72.21.199	HTML_GET_APP	网页类
216.88.158.142	Mozilla/4.0+compatible+ZyBorg/1.0+(wn.zyborg@looksmart.net;+http://www.WISEnutbot.com)	网页类
66.196.72.103	Mozilla/5.0+(Slurp/cat;+slurp@inktomi.com;+http://www.inktomi.com/slurp.html)	网页类
66.196.90.125	Mozilla/5.0+(Slurp/cat;+slurp@inktomi.com;+http://www.inktomi.com/slurp.html)	网页类
202.96.63.3	User-Agent:+Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0)	网页类
219.133.39.15	-	图片类
205.188.209.37	Mozilla/4.0+(compatible;+MSIE+6.0;+AOL+9.0;+Windows+NT+5.1)	网页类
66.237.60.91	Openfind+data+gatherer,+Openbot/3.0+(robot-response@openfind.com.tw;+http://www.openfind.com.tw/robot.html)	网页类

(1)检测的查全率　在原始日志中，访问了 robots.txt 的不同 IP+agent 共有 24 个。其中，有 20 个出现在最后得到的合并会话集 M 中。其他未出现的 4 个，通过核对原始数据，发现它们仅仅访问了 robots.txt 却没有访问任何其他的内容，故自然不会出现在网络机器人候选集中。故本方法对于遵守规则的网络机器人的查全率达到 100%。

(2)检测的查准率　合并会话集 M 中 253 个项目到底哪些是真正的网络机器人？判别的方法是考察其会话的个数和访问记录的总数。基于下面的假设：1)一个网络机器人访问网站时会产生比较多的用户会话(一般一个访问任务会分解成多次有一定时间间隔的请求)；2)访问比较多的内容，我们从 M 中筛选出会话数 Snumber>=2 或访问记录数 Rnumber>=5(均为 M 的平均数)的项目，共得到 28 个项目。这 28 个项目中有 20 个访问了 robots.txt，其他 8 个都没有他访问 robots.txt。这 8 个项目见表 2。

合并会话集 M 中 253 个项目中其他的项目是否可以认为是网络机器人，要看它以后的访问数，若超过一定的数量，就可以检测出来。这在我们对后续日志的检测中得到了验证。

结束语　本文从搜索引擎网络机器人的访问行为着手，设计开发了一个基于访问内容分类的从 Web 日志检测网络机器人的方法。该方法简单快捷，查准率高。缺点是由于不分析网页的具体构成，因此只有少量访问的时候比较难确定。另外，当网站中有大量的纯文字网页时，可能会将普通的访问者当作是网络机器人。后续的研究，将考虑网页的构成以及网页中超级链接来进行检测。

致谢：张又又老师提供了本文的实验数据，在此表示衷心感谢！

参 考 文 献

1 The Robots Database. http://www.robotstxt.org/wc/active.html. [EB/OL]. 2004.08.01

2 Tan Pang-Ning, Kumar V. Discovery of Robot Sessions based on their Navigational Patterns [J]. Data Mining and Knowledge Discovery, 2002,6(1): 9～35

3 Robots Exclusion. http://www.robotstxt.org/wc/exclusion.html [EB/OL]. 2004.08.01

4 Enhancing Offline Favorites. http://msdn.microsoft.com/ [EB/OL]. 2004.08.01

5 Hypertext Transfer Protocol-HTTP/1.1. http://www.w3.org [EB/OL]. 2004.08.01

(上接第 172 页)

复杂度为 $O(n^2)$。文中所给出的例子将方位关系理论研究与应用有效地结合起来，为进一步的应用研究提供了基础。显然，本文涉及的方位关系表示模型也存在一些不足，例如，模型的完备性、表示方法的认知合理性等都是进一步需要研究的问题。我们认为这些问题的研究对于常识知识表示，对于空间关系的认识以及 GIS 等应用都具有重要意义。

参 考 文 献

1 Pullar D, Egenhofer M. Toward formal definitions of topological relations among spatial objects. In: Proc. of the Third Intl. Symposium on Spatial Data Handling, Sydney, Australia, Aug. 1988

2 Egenhofer M, Herring J. A mathematical framework for the definition of topological relationships. In: Proc. of the Fourth Intl. Symposium on Spatial Data Handling, Zurich, Switzerland, July 1990

3 Papadias D, Sellis T. The semantics of relations in 2-D space using representative points: Spatial indexes. In: Proc. of the European Conf. on Spatial Information Theory, Elba, Italy, 1993

4 Kainz W, Egenhofer M, Greasley I. Modeling spatial relations and operations with partially ordered sets. International Journal of Geographical Information Systems, 1993,7(3): 215～229

5 Frank A. Qualitative Spatial Reasoning about Distance and Directions in Geographic Space. Journal of Visual Languages and Computing, 1992,3:343～373

6 郭平. 定性空间推理技术及应用研究:[重庆大学博士学位论文]. 重庆,2004

7 Guo P, Huang-Fu T, Luo Y. A reasoning method for resolving spatial constraint satisfactory problem. In: the Proc. of the Third Intl. Conf. on Machine Learning and Cybernetics, Shanghai, 2004.2262～2268

8 Vilain M, Kautz H, van Beek P. Constraints Propagation algorithms for temporal reasoning: A revised report. In: Weld D S, de Kleer J, eds. Readings in Qualitative Reasoning about Physical Systems, Morgan Kaufmann, 1990.373～383

9 Kumar V. Algorithms for constraint satisfaction problems: A survey. Artificial Intelligence, 1992,13(1):32～44

10 Hertzberg J, Gusgen H W, et al. Relaxing constraint networks to resolve inconsistencies. In: Proc. GWAI-88, Eringerfeld, Germany, 1988.61～65

11 Liu X, Shekhar S, Chawla S. Consistency Checking for Euclidean Spatial Constraints: A Dimension Graph Approach. In: the Proc. of 12th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI'00), Canada, 2000.333～343