貴州大學
GUIZHOU UNIVERSITY

Lecture 3 数据采集方法

- 系统日志数据
- 网络数据采集

教学目标

- 认识日志数据的采集、互联网数据的采集
- 掌握互联网数据采集的原理和网络爬虫技术

系统日志数据采集

- 计算机中的任何程序都可以输出日志，这些程序包括操作系统内核、各种应用服务器等

- Web日志包含各种前端Web服务器产生的用户访问日志，以及各种Web应用程序输出的日志。

系统日志数据采集

- 在Web日志记录中，每条日志通常代表着用户的一次访问行为。

211.87.152.44 - [18/Mar/2005:12:21:42 +0800] "GET / HTTP/1.1 " 200 899 "http://www.baidu.com/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Maxthon) "

- 上面这条日志反映了很多有用的信息，例如访问者的IP、访问时间、访问的目标网页、来源的地址以及访问者使用的客户端信息等。

系统日志数据采集目的

- 日志采集的主要目的是为了进行日志分析
  - 从日志记录中获取网站每个页面的页面访问量、访问用户的独立IP数
  - 统计出关键词的检索频次排行榜、用户停留时间最长的页面
  - 构建广告点击量模型、用户行为特征分析

系统日志数据采集工具

- 日志数据中蕴藏了如此大的价值，那么当然需要一些工具帮助我们来分析它们，例如Awstats、Webalizer，都是专门用于对Web服务器日志进行统计分析的开源程序。
- 还有一类产品，虽然不直接分析日志，但提供页面中嵌入js代码的方式统计数据。典型的产品包括Google Analytics、国内的Cnzz、百度统计等。

系统日志数据采集工具

- 业务部门对数据分析的需求总是随着公司业务不断变化的。想要进行稍复杂的个性化分析，依然需要自己动手来采集日志数据。
- 绝大多数的日志分析工具都仅限于单机使用，当数据量的增长超过单机处理的范围，这些分析工具就没办法了。同时，提供在线分析服务的网站对单个站点通常也都有最大流量的限制，故对能够分析的数据样本量也有较为严格的限制。

系统日志数据采集过程

- 日志主机是一个基于Unix或者Windows的服务器系统，它用来集中存储日志消息。

- 日志主机可以集中存储来自多个数据源的日志消息，可以对系统日志信息进行备份，也可以分析日志数据。

系统日志数据采集过程

- 日志消息是如何传输到日志主机的？最常见的方法是通过syslog协议实现的，它是日志消息交换的一种标准。syslog协议实现了覆盖几乎所有客户端和服务器端组件间的通信，并主要采用用户数据报协议（UDP）。
- 为了提高传输的可靠性，syslog协议同样支持传输控制协议（TCP）。日志主机的主要工作就是通过syslog协议采集日志消息，并将其存储在一个本地磁盘上，以进行日志备份、存储和分析。

网络数据采集过程

- 互联网承载了海量的信息，但如何有效地提取并利用这些信息是一个巨大的挑战。

- 搜索引擎是一个辅助人们检索信息的工具，它可作为用户访问互联网的入口。但是，通用性的搜索引擎存在着一定的局限性

网络数据采集过程

- 搜索引擎存在着一定的局限性

  (1) 特定领域、特定背景的用户通常具有特定的检索目的，而通用搜索引擎返回的结果可能包含大量的无用网页信息。

  (2) 随着网络技术的不断发展，互联网中的数据形式越来越丰富。图片、数据库、音频、视频多媒体等不同类型的数据大量出现。

  (3) 目前通用搜索引擎大多仅提供基于关键字的检索，它们难以支持基于语义信息的查询和检索。

网络爬虫的工作原理

- 爬虫根据既定的抓取目标，选择性地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。
- 网络爬虫的技术框架包括控制器、解析器、资源库三大部分。控制器的主要工作是为各个线程分配工作任务，并调度爬虫的线程资源。解析器的主要工作是批量下载网页，并对页面的格式和内容进行处理。资源库的主要工作是存储下载到的网页资源，其通常采用大型的数据库存储模型。

网络爬虫的工作原理

- 网络爬虫往往从一个初始网页的URL开始工作，首先获得初始网页上的URL。在抓取网页的过程中，需要根据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队列中。
- 然后，网络爬虫根据某种搜索策略从队列中选择下一次要抓取的网页URL，并重复上述过程，直到达到系统的某一停止条件，例如搜索时长或搜索页面数量达到某一阈值。

网页搜索策略

- 网络爬虫工作过程中的一个重要组成部分是网页搜索策略。网页的搜索策略按照搜索次序不同，可以分为深度优先、广度优先和最佳优先三种搜索策略。
- 深度优先的搜索策略表述如下：首先跳转进入起始网页的URL链接，分析这个网页中所包含的URL链接，选择其中一个URL链接进入。如此一个链接一个链接地选择并跳转进入，直到访问完路径中的最后一个URL。之后再回到上一层URL链接，处理下一条路径。

网页搜索策略

- 广度优先的搜索策略和深度优先策略不同。在抓取URL的过程中，只有完成当前层级的搜索后，才跳转到下一层级进行搜索。算法的基本思想是：与初始网页URL在有限跳转次数范围内的网页具有主题相关性的概率很大。
- 最佳优先搜索策略是基于降低广度优先搜索策略的算法复杂度而进行优化的。最佳优先搜索策略按照特定的网页分析算法，预测候选URL与主题的相关性，筛选并抓取最相关的某些URL。研究表明，最佳优先搜索策略可以将无关网页的数量降低90%左右。

网页分析算法

- 基于拓扑的网页分析算法是基于网页之间的链接，通过已知的网页，对与其有直接或间接链接关系的对象作出评价的算法。拓扑网页分析算法又分为网页粒度、网站粒度和网页块粒度这三种具体的分析算法。
- 网页粒度算法：PageRank是最常见的网页粒度分析算法，PageRank通过某页面所有的超链接关系来确定一个页面的重要等级。它把从A页面到B页面的链接解释为A页面给B页面投票，并根据投票来源和投票目标的等级来决定新的页面的等级。

网页分析算法

- 网站粒度算法：基于网站粒度的爬虫算法，其算法实现的关键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的抽象，并在一定的模型下计算链接的权重。
- 网页块粒度算法：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。PageRank算法常常给网页分析带来广告等噪声链接的干扰。

网络爬虫框架

- 一些比较著名的网络爬虫体系结构
  - RBSE （Eichmann 1994）是第一个发布的爬虫。它有两个基础程序。第一个是“spider”，抓取网页中的URL，并存储到一个关系数据库中；第二个程序是“mite”，它是一个修改后的www的ASCII浏览器，负责从网络中下载页面。
  - WebCrawler（Pinkerton 1994）是第一个公开可用的建立全文索引的程序，它使用库www来下载页面，使用广度优先来解析获取URL，并对其进行排序。此外，它还包括一个根据选定文本和查询相似程度爬行的实时爬虫。

网络爬虫框架

- 一些比较著名的网络爬虫体系结构
  - World Wide Web Worm (McBryan 1994) 为文件建立包括标题和URL简单索引的爬虫，其索引可以通过grep式的Unix命令来实现。
  - Google Crawler (Brin and Page 1998) 集成了索引处理，支持全文检索和URL抽取。它拥有一个URL服务器，用来提供发送爬虫程序时要抓取的URL列表。在文本解析的时候，URL服务器负责检测某个新的URL是否已经存在。如果不存在的话，就将此URL加入到URL服务器中。

数据采集接口

- 网络应用程序分为前端和后端两个部分。当前的发展趋势是前端设备层出不穷，从桌面电脑发展到笔记本电脑、手机、平板等等。因此，必须有一种统一的机制，方便不同的前端设备与后端进行数据通信。
- 这导致API构架的流行，甚至出现“API First”的设计思想。REST API是目前比较成熟的一套互联网应用程序的API设计理论。微博、微信公众号等常用的商用数据API都支持REST API的方式获取数据信息。

数据采集接口

- REST从资源的角度来观察整个网络，分布在各处的资源由URL定位，而客户端应用通过URL来获取资源。随着不断访问URL来获取资源，客户端应用不断地转变状态。
- REST通常基于使用HTTP、URL、XML、HTML这些广泛流行的协议和标准，故它是一种风格，不是一个标准。
- REST对资源的操作包括获取、创建、修改和删除，这些操作正好对应HTTP协议提供的GET、POST、PUT和DELETE方法