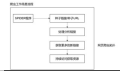
Spider

基本概述

网络爬虫(WEB SPIDER)
基本概述：爬虫又叫网络蜘蛛，工作与万维网中，是一个自动化的网络信息下载工具，用户可以指定下载位置与下载规则，爬虫帮助用户将有用的信息筛选爬取下来
爬虫的用途：下载大量的基础数据为其他分析提供原料，数据分析，数据挖掘，我都可用化等领域都可以通过爬虫下载的数据进行深入工作。
万维网介绍：Web所指www即头所有网页和相关的应用程序，都属于万维网应用
搜索引擎：搜索万维网应用，也是用户获取信息的主要渠道，爬虫模拟浏览器行为自动获取万维网数据。
网络数据分析：文本资源(txt、html、xml、shtml)，图像资源(jpg、bmp、gif、png)，音频资源(mp3)、视频资源(rmvb、mp4、flv)。
万维网中网页的关键性极强，网页中会出现大量的新网址链接(便于用户访问新位置新资源。

简易流程介绍

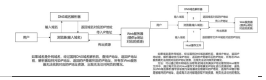
爬虫通过强关联的特性在万维网持续的工作。



机器学习：简单来说通过大量现有的数据去训练一个好的模型，找大量的训练数据，输入一定的行相同与执行逻辑，进行训练数据，提取特征，才能点击进行训练，决策。通过网络爬虫爬取海量数据，进行筛选，找到合适的数据作为数据集。
在通过爬虫增强形可以进行训练的数据。
前期的学习知识可离不开大量的深度学习基础知识，包括卷积，BN，激活函数等，通过爬虫爬取一系列相关知识可以有效的系统性的进行学习。

关键词

URL：统一资源定位符(数据资源源地址，爬虫根据地址访问数据)
Web服务器：B/S，Web服务器用于与浏览器交互，提供资源数据给大多数万维网资源部署于Web服务器中，根据浏览器的资源请求(HttpWeb进行交互请求)
DNS域名解析器(通过域名查找公网IP每一个域名都绑定了一个或多个公网IP))
Web服务统一端口:80
Host文件(本机缓存文件)，浏览器先访问Host缓存文件，看是否有输入域名对应的IP地址
(1) 如果存在，直接取出对应IP地址进行资源获取。
(2) 如果不存在，将域名输入到DNS域名解析器获取对应IP地址，并缓存到Host文件中，下次再遇到域名访问，直接从Host缓存文件中获取即可。
通过Host缓存文件可以大大减少浏览器输入同一域名不断访问DNS域名解析器获取相应IP地址所耗时间。
可以通过将外网域名(自取等)或域名和真实IP地址缓存到Host缓存文件中，通过每次先访问Host缓存文件获取真实IP地址进行资源获取，避免从DNS域名解析器中获取被污染或篡改后被破坏的IP地址，造成Host缓存都是被破坏的IP地址，造成每次访问都是获取不到资源，网页无法访问的情况。
Http协议：超文本传输协议(万维网主流协议)，http基于tcp实现。
http/https：端口号访问：http:80 https:443
例如http://www.baidu.com
https://www.google.com 访问不同的协议使用不同的端口


DNS域名解析器
输入域名
域名数据(入域名)
用户
域名数据(入域名)
查找本地文件(Host文件)
有ip数据
Web服务器
根据域名数据(域名数据)
找到对应的IP地址
用户
域名数据(入域名)
查找本地文件(Host文件)
有ip数据
Web服务器
根据域名数据(域名数据)
找到对应的IP地址
Web服务器
根据域名数据(域名数据)
找到对应的IP地址
Web服务器
根据域名数据(域名数据)
找到对应的IP地址