2-3-cn

1 大家好，我是北京理工大学计算机学院数据科学与知识工程研究所的车海莺，本节我们讨论组织外部数据的获取

2

外部数据主要是网络大数据，即互联网上可用的数据，收集网络数据的方式是使用网络爬虫。

在一个URL中，可能有很多链接和上层链接的子链接，那么哪些链接应该先被爬取，爬取的优先级和顺序由Web爬虫的爬取策略决定。

而在本节中，我们将讨论网络大数据、网络爬虫及其策略。

3

网络空间"人、机、物"三元世界交互融合产生的、可在互联网上获取的大数据。"人、机、物"三元世界ternary world在网络空间中彼此之间相互交互和融合所产生并在互联网上可获得的大数据.

网络大数据有不同用户、不同网站产生，数据形式多样，语音，视频，图片，文本等

网络大数据特征包括：

1 多源异构，即互联网数据来自多源，不同的源有不同的数据。

2 及时性是指当事情发生时，它可以立即发布。

3 社会性，网络大数据直接反映社会地位

4 交互性：微信微博、脸书、推特等，网民不仅可以根据需要发布信息，还可以根据个人喜好回复转发信息。

5 突发性：一些新闻传播会导致短时间内产生大量新的网络数据，反映网络大数据和网络群体的突发性

6 高噪音特性容易理解，互联网数据不能100%真实有用，互联网数据质量没有人负责，所以价值密度低，充满脏数据，当你想用互联网数据时，你必须清洗它。

4

如图所示，网络爬虫是自动浏览互联网并获取数据的程序或网络机器人。

5

1Web爬虫爬取过程从一个统一的资源地址列表开始，称为种子URL，并将其作为链接入口进行爬取。当爬虫访问这些种子URL时，它会识别出页面上所有需要的链接，并将它们添加到要爬取的队列中。

2 之后，从待爬取队列中取出网页链接，然后读取URL，做DNS解析，将网页下载到Downloaded web library。

3 将已经下载好的URL放入爬取的URL列表

4 将新的URL提取到待爬取的URL队列中，按照策略放入待爬取的URL队列中

5 所有进程将结束，直到爬取队列为空。

6

种子网址中的扇出网址如何处理，也就是链接的链接，这就涉及到网络爬虫的爬取策略

最常用的抓取策略包括

深度优先

广度优先

部分PageRank策略

OPIC（在线页面重要性计算）

让我们一一解释

7

这是一个扇出URL结构的示例，如果使用深度优先策略，

顺序应该是M1-M2-M5-M8-M6-M3-S7-S4；

如果使用广度优先策略，则顺序应为 M1-M2-M3-S4-M5-M6-S7-M8；

8

PageRank，也称为 Page Ranking 和 Google Ranking，是搜索引擎应用的一种基于网页间相互超链接的技术，由 Google 创始人拉里·佩奇 (Larry Page) 发明。

PageRank 链接分析算法通过统计其他网页指向的链接的数量和重要性来表达每个页面的重要性，从而实现对每个网页的重要性进行排名。

该算法考虑了数量和质量。例如，网页 E 的链接数远多于网页 C 的链接，但网页 C 比网页 E 重要得多。

因为页面 C 是由页面 B 链接的，而页面 B 的重要性很高。

利用 PageRank 的原理，计算出 URL 列表中的重要性值，对抓取的网页进行排序，然后依次遍历每个 URL。

9

OPIC (online Page Importance Computation)

OPIC 策略将每个网页赋予相同的“金币”，

每当下载某个页面 P，则将 P 拥有的“金币”平均分配给网页中所包含的链接页面。

待爬队列中链接依“金币”排序，OPIC 计算速度快于局部 PageRank 策略

10

架构简单，扩展时只需要更新 master， master 节点压力大，易成为瓶颈，Slave 节点数量限制大

架构稍复杂，所有节点进行通信扩展时需要更新所有其他节点，无 master， 不会出现单机热点，slave 节点数量限制小

架构复杂，实现难度大，master 与 master 进行通信，slave 节点数量限制小

11

让我们总结一下外部数据和采集。

我们研究了 Web 爬虫的爬取过程，Web 爬虫的爬取策略，包括深度优先、广度优先、页面排名、OPIC（在线页面重要性计算）。

我们学习了分布式网络爬虫结构、主从、对等和混合结构。

12

这节课的我们讨论了组织外部数据获取，如果您有任何问题，请随时与我联系。