2-4-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，本节我们讨论深网数据获取

2

如图所示，所有互联网信息的内容可以分为三个部分，表层网络、深层网络和暗网。

上面是 Surface web，Surface web，它的内容基本上是非结构化的 HTML 信息，任何人都可以通过互联网访问。

中间是深网，Deep Web 概念是 Bright Planet 于 2000 年创立的，用于表达那些信息内容存储在搜索数据库中并且只响应直接查询的网站；内容主要是结构化的数据库信息。

这些信息可以是学术记录、财务记录、法律文件、政府记录和科学报告。

最底层是暗网，主要是非法信息，与毒品、武器等有关。

在暗网部分，除了贩毒，武器等非法信息之外，还有“TOR”。

Tor 是“洋葱路由器”的首字母缩写，用户可以通过 Tor 在互联网上进行匿名通信。

该项目最初由美国海军研究实验室赞助，旨在隐藏政府人员在网上收集情报的行踪；

Tor 由非营利组织 Tor Project 免费设计和提供，并已被自由倡导者和犯罪分子采用。

Tor 以迂回的方式通过多台计算机发送聊天消息、谷歌搜索、采购订单或电子邮件，将互联网用户的活动伪装成洋葱包裹其核心。

信息传输每一步都是加密的，无法知道用户在哪里，信息传输的位置和目的地。

世界各地的志愿者提供大约 5,000 台计算机作为传输路径上的节点，以掩盖新的页面或聊天请求。Tor 项目将这些点称为中继节点。

3

美国互联网专家和图书馆员 Chris Sher-man 和 Gary Price 定义：“可在互联网上获取，
但传统搜索引擎因技术限制无法索引或慎重考虑后不愿索引的网页、文件或其他高质量、权威信息”

4

深网信息的特点

1） 与信息需求、市场、领域高度相关；

2） 互联网上增长最快的新型信息。比传统的表面网络更专业、更深入。深网内容的全部价值是表面网的 1000-2000 倍。

3） 一半以上存储在专题数据库中；深度网络上 95% 的信息无需付费即可公开获取

5

深网内容包括

1. 由于缺乏被指向链接而没有被搜索引擎引到的页面
2. Web 上可访问的非网页文件，比如图片文件，Pdf 和 word 文档等
3. 通过填写表单形成对后台在线数据库的查询而得到的动态页面
4. 需要注册或其他限制才能访问的内容

6

让我们比较一下深度网络和搜索引擎搜索的内容，

1 从界面上看，深网内容是从数据库中提取的动态网页，通常界面复杂，每个查询界面都支持对多个属性的查询。

搜索引擎，内容按关键字搜索

2 深网的结果主要是结构化数据，而搜索引擎的结果只是网页。

3 在搜索结果如何排序方面，Deep Web 根据 Deep Web 中某个属性值的结果对搜索结果进行排序，搜索引擎按照搜索结果与提交查询的相似度对搜索结果进行排序。

7

现在让我们学习如何收集深度网络数据。

深网数据采集任务包括 2 个阶段，

1)是查询接口识别。

2）自动填写表格，然后执行查询。

对于特定的网站，您可以通过在包装器和生成器的帮助下手动编写或提供爬虫脚本来获取尽可能多的深度网络数据。但是，这种方法不仅需要大量的人力，而且由于其特定的网站和查询界面，其可扩展性较差。构建一个通用的深度网络爬虫，以便一次爬取多个站点的深度网络数据。

1)查询界面识别：利用视觉布局等多种方法解析 HTML 表单或对 HTML 表单进行语法分析，自动发现深网数据资源；

2）添加文本相似度启发式规则，将 HTML 表单与特定字段关联起来，实现表单的自动填充；

3）通过构建页面分类器和表单分类器自动查找与任务相关的深网数据库

4）尝试自动填写表格。

4.1 基于领域知识：使用启发式规则将表单的字段与领域相关联，从而输入与领域概念相关的参数。

4.2 领域无关检测：基于采样从查询结果中迭代获取查询关键字，从而以更少的查询获得尽可能多的查询结果

1）多种方法来解析 HTML 表单或者对 HTML 表单进行语法分析来自动发现深网数据资源

2）将 HTML 表单与特定领域关联已实现表单的自动填写

3）领域无关探测：基于采样迭代式地从查询结果中获取查询关键字，藉此以较少的查询次数获取尽可能多的查询结果

8

让我们总结一下深度网络数据采集主题。

在本节中，我们学习了深层网络的概念，深网信息的特点，什么是深层网络内容，

我们比较了 Deep Web 数据收集和传统搜索引擎查询的结果，

分析了深网数据收集任务，了解了抓取深网内容的解决方案。

9

本节课我们讨论了深网内容，如果您有任何问题，请随时与我联系。