**BD 2-3**

**1**

**Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering**

**School of Computer Science, in Beijing Institute of Technology, in this session, we will discuss**

**Some idea about External Data Acquisition.**

**2**

**External data is mainly Network big data, which is the data available in the internet,**

**and the way to collect network data is using web crawlers.**

**In one URL, there may be many links and the sub-links of the upper level links, so which links should be crawled first and the crawling priority and sequences decided by the Web crawler crawling strategy.**

**And in this session, we will discuss the network big data, web crawler and its strategy.3**

**3**

**Big data generated by the interaction and integration of the ternary world of "human, machine, and material" in cyberspace and available on the Internet.**

**Network big data characteristics include:**

**1)Multi-source and heterogeneous, which means the internet data comes from Multi-sources, and the different source has different data.**

**2) Timeliness which means when something happens, it can be published immediately.**

**3) Sociality, the network big data directly reflects the social status**

**4) Interactivity: WeChat Weibo, Facebook, twitter, etc. Internet users can not only publish information according to their needs, but also reply and forward information according to their personal preferences.**

**5) Sudden: Some news dissemination will cause a large amount of new network data to be generated in a short time,**

**reflecting the suddenness of network big data and network groups**

**6) High noise characteristic is easy to understand, the internet data cannot be 100% true and useful, no one is responsible for the quality of internet data, so value density is low, full of dirty data, when you want use the internet data, you must clean it.**

**4**

**Like showed in the diagram, A web crawler is a program or web robot that automatically browses the Internet and get data.**

**5**

**1)Web crawler crawling process Starts with a list of uniform resource addresses called seed URL and use it as the link entry for crawling. When the crawler visits these seed URL s, it identifies all the needed links on the page and adds them to the queue to be crawled.**

**2) After that, the webpage links are taken out from the queue to be crawled, then Read URL, do the DNS resolution, and web pages were download into the Downloaded web library.**

**3) Put the already downloaded URL into the crawled URL list**

**4) Extract the new URL into the URL queue to be crawled and put them in the to be crawled URL queue according to strategy**

**5) all the process will end until the queue for crawling is empty.**

**6**

**How to deal with fan-out URLs in seed URLs, which is the links of the link, which involves web crawler crawling strategies**

**The most often used Crawling strategies include**

1. **Depth first**
2. **Breadth first**
3. **Partial PageRank Strategy**
4. **OPIC (online Page Importance Computation)**

**Let’ s explains them one by one**

**7**

**Here is a example of Fan out URLs structure, and if use depth first strategy,**

**the sequence should be M1-M2-M5-M8-M6-M3-S7-S4;**

**And if use Breadth first strategy, the sequence should be M1-M2-M3-S4-M5-M6-S7-M8;**

**8**

**PageRank, also known as Page Ranking and Google Ranking is a technology applied by search engines based on mutual hyperlinks between web pages，**

**Invented by Google founder Larry Page (Larry Page).**

**The PageRank link analysis algorithm expresses each page importance by counting the number and the importance of links pointed to by other webpages, thereby realizing the ranking of the importance of each webpage.**

**This algorithm takes the Quantity and quality into consideration.**

**For example, Links number of webpage E are far more than links to webpage C, but webpage C is much more important than webpage E.**

**because page C is linked by page B, and page B is of high importance.**

**Use the principle of PageRank to calculate the importance value in the URL list and sort the crawled web pages in order, and then traverse each URL in this order.**

**9**

**The OPIC strategy assigns the same "gold coins" to each web page. Whenever a page P is downloaded, the "gold coins" owned by P are equally distributed to the linked pages contained in the web page.**

**The links in the queue to be crawled are sorted by "gold coins"**

**OPIC calculation speed is faster than partial PageRank strategy**

**10**

**Now we know the crawling task is huge, it cant easily be done by one stand-alone crawler ,**

**Through the effective collaboration and cooperation of multiple stand-alone crawler systems to achieve Internet big data capture.**

**which is Distributed Architecture model.**

**There are 3 basic Distributed Architecture model, master- slave, peer to peer and mixed structure.**

**For master- slave, the master needs to be updated when expanding,**

**the master node is under heavy pressure and easily becomes a bottleneck, and the number of slave nodes is limited.**

**For peer to peer, all nodes need to update all other nodes when communication is expanded.**

**No master, no single machine hot spots, and the number of slave nodes is limited**

**For mixed structure, the master communicates with the master, and the number of slave nodes could be small.**

**11**

**Let’s summarize the external data and acquisition.**

**We studied the Web crawler crawling process, Web crawler crawling strategy,**

**including Depth first, Breadth first, Page Rank, OPIC (online Page Importance Computation).**

**And we learned the Distributed web crawler structure, master- slave, peer to peer and mixed structure.**

**In this session we learned and external data acquisition. thank you for your attention, if you have any question, feel free to contact me.**