**BD 2-1**

**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**

**Deep web data acquisition.**

**2**

**As the picture shows, the content of all internet information could be divided into 3 parts, surface web, deep web and dark web.**

**on top is the surface web, Surface web, its content is basically unstructured HTML information, anyone can access it through the Internet.**

**In the middle is deep web,**

**Deep Web concept is created by Bright Planet in 2000, which is Used to express those websites whose information content is stored in the search database and only responds to direct queries;**

**the content is mostly structured DB information. The information could be academic records, financial records, legal documents, government records and science reports.**

**At Bottom is the dark web, which mostly is the illegal information, related to drugs, weapons and others.**

**In the dark web part, besides the onion is the “TOR”.**

**Tor is the acronym for "the onion router", and users can communicate anonymously on the Internet through Tor.**

**The project was originally sponsored by the U.S. Naval Research Laboratory to hide the whereabouts of government personnel collecting intelligence online;**

**Tor is designed and provided for free by Tor Project, a non-profit organization, and has been adopted by freedom advocates and criminals.**

**Tor sends chat messages, Google searches, purchase orders or emails through multiple computers in a circuitous manner, disguising the activities of Internet users like an onion wraps its core.**

**Information transmission is encrypted at every step, and there is no way to know where the user is , Location and destination of information transfer.**

**Volunteers around the world provide about 5,000 computers as nodes on the transmission path to cover up new pages or chat requests.**

**The Tor Project calls these points relay nodes.**

**3 American Internet experts and librarians Chris Sher-man and Gary Price defined:**

**"Available on the Internet, but those web pages, files or other high-quality, authoritative information that traditional search engines are unable to index due to technical limitations or are unwilling to index after careful consideration"**

**4**

**Features of Deep Web Information**

**1 Highly related to information needs, markets and fields；**

**2 Fastest growing new type of information on the Internet.**

**more specialized and deeper than the traditional surface web.**

**The full value of the deep web content is 1000-2000 times that of the surface web.**

**3 More than half is stored in thematic databases;**

**95% of the information on the deep web is publicly available without payment**

**5**

**Deep web content includes**

1. **Pages that are not referred to by search engines due to lack of directed links .**
2. **Non-web files accessible on the web, such as picture files, Pdf and word documents, etc.**
3. **A dynamic page obtained by querying the back-end online database by filling in the form.**
4. **Content that requires registration or other restrictions to access.**

**6**

**Let ‘s compare deep web and the content searched by search engine,**

**1) from interface, deep web content is Dynamic Web extracted from databases, which usually have complex interfaces, and each query interface supports queries on several attributes.**

**search engine, the content is searched by keyword**

**2) the results from deep web are mainly structured data, but the results from search engine is just web pages.**

**3) in the aspect of how to sort search results, deep web sort search results according to the result of a certain attribute value in Deep Web, search engine sort search results by Similarity between search results and submitted query.**

**7**

**Now let’s learn how to collect the deep web data. Deep web data collection task includes 2 stages, 1) is query interface identification. 2 fill in the form automatically, then execute the query.**

**For a specific website, you can obtain as much deep web data as possible by manually writing or giving crawler scripts with the assistance of wrappers and generators.**

**However, this method not only requires a lot of manpower, but also its scalability is poor because of its Specific website and query interface,.**

**Construct a general deep web crawler in order to crawl the deep web data of many sites at once.**

**1）Query interface recognition: Use a variety of methods including visual layout to parse HTML forms or perform syntax analysis on HTML forms to automatically discover deep web data resources;**

**2) Add text similarity heuristic rules to associate HTML forms with specific fields to realize automatic filling of forms;**

**3 ) By constructing page classifiers and form classifiers to automatically find deep web databases related to tasks**

**4 )** **Try to Automatically fill in the form.**

**4.1Based on domain knowledge: Use heuristic rules to associate the field of the form with the domain, thereby inputting parameters related to the domain concept.**

**4.2 Domain-independent detection: Iteratively obtain query keywords from query results based on sampling, to obtain as many query results as possible with fewer queries**

**8**

**Let’ s summarize the deep web data acquisition topic.**

**In this session we learned**

1. **The concept of the deep web**
2. **Features of Deep Web Information**
3. **What is Deep web content**
4. **We Compared the results of Deep Web Data Collection and Traditional Search Engine Query**
5. **We analyzed Deep web data collection task, and we understand the solution to crawl the deep web content.**

**9**

**In this session we learned the bigdata resources, internal and external data.**

**thank you for your attention, if you have any question, feel free to contact me.**