模块二的数据流图：



在模块二中数据的存储需要进行两次，第一次存储是将抽取的数据清洗之后，第二次存储是最终融合后的商品数据。目前考虑两种存储都使用MySQL。第一次存储的表数据结构为

表1.商品基本属性

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 属性 | bookID | bookName | press | author | publishingTime | pagesCount | edition | … |
| 数据类型 | int | String | String | String | Date | int | String | … |

表2.网页基本信息

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 属性 | website | bookID | URLofBook | numOfBrowse | praise | badComment | URLID | price |
| 数据类型 | String | int | String | int | int | int | int | float |

表设计理由：将商品信息和网站信息作为两个不同的实体，信息分别存放。便于存储，但对于查询，需要对数据进行整合后存于第二次的数据库表后进行查询操作。

第二次存储的数据表为

表1.商品实体与商品的对应表

|  |  |  |  |
| --- | --- | --- | --- |
| 属性 | Index | bookName | bookIDs |
| 数据类型 | int | String | List |

表2.商品的详细属性

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 属性 | bookID | bookName | press | author | publishingTime | price | edition | … |
| 数据类型 | int | String | String | String | Date | float | String | … |

表3.商品实体的类别

|  |  |  |
| --- | --- | --- |
| 属性 | bookName | categories |
| 数据类型 | String | List |

这样设计的理由是主要考虑到两种搜索类型，一种是直接使用书名，那么就根据书的名字查找表1商品实体表。找到和该书名对应的商品之后返回商品的详细信息。另一种是使用类别进行搜索，这就需要查找表3商品实体的类别表，通过类别的查找来锁定商品实体，然后再找到对应的商品。

在模块二中主要使用到的技术有相似度计算，instance match，ontology alignment等。