**农业实体匹配项目**

1. **项目模块分析**

根据需求描述，现将项目的开展分为如下几大模块：1）联合国粮农组织关于农业实体数据的收集；2）数据的处理（包括英文转化为中文，数据集的构建）；3）可视化编辑器的实现；4）用户数据处理功能的实现。

这四个模块中，以最后一个模块即可视化编辑器的实现为主体，这也是本项目的核心，允许用户对个人数据进行可视化编辑，且保证前后台数据的一致性。第一个模块是从联合国粮农组织得到的关于农业实体数据作为本项目源数据存储在后台，作为用户数据相似性匹配的标准，要求该数据要尽可能权威，囊括所有农业中的概念。数据处理部分主要包括两部分，一是将得到的英文数据转换为中文数据，在转换过程中要保证中文表达的准确性，权威性；二是要根据标准分类的数据构建所有实体的同义词库，要求同义词库要尽可能包含所有常用，不常用的同义词。第四部分是在前期工作完成后进行的，用户数据输入后（用户数据以什么格式输入的还有待确认），对每一个实体进行匹配，当该名称与标准名称不一致的时候，系统将依照标准词库将其替换为标准名称；当分类出现错误时候，允许用户对其进行拖拽修改。

1. **各模块具体实现**
2. **农业实体数据的收集处理**

权威性数据的收集处理是本系统的关键，其后的所有功能都是根据这个标准进行进行的。在收集数据之前做了大量的调查，并没有现有的可以供我们直接使用的数据。包括中文的CAT以及英文的AGROVOC都没有提供可以直接使用的数据集。CAT仅仅提供了一些例子集合，并不包含所有农业名词分类信息，并且其提供的Sparql Query Endpoint无法使用。而AGROVOC提供了可以检索的URL，且父子关系都是以URL嵌入的形式给出。为了解决这个问题，我编写了一个爬虫，专门用于对嵌入数据的爬取。由于需要体现出各个节点之间的层级关系，因此需要将其转换成json形式进行存储。由于AGROVOC网站的访问限制，以及数据格式化处理，数据收集这个过程浪费大量的时间。爬取下来的结果如下图1所示，每个json文件均是一棵树，文件名即为树根。

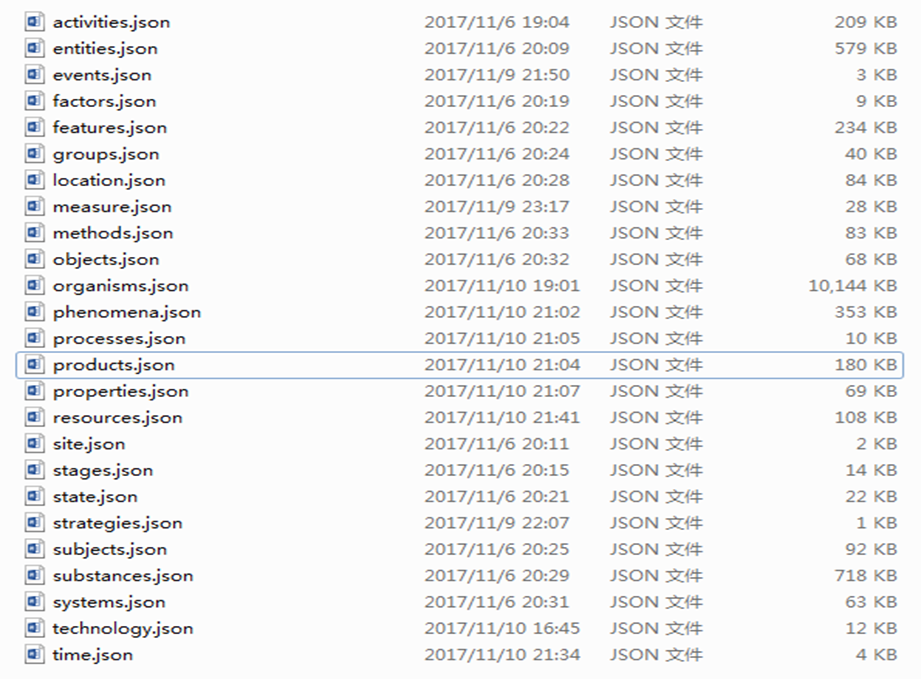


图 1 树的节点信息

1. **数据的处理**

虽然英文版的农业词库分类收集到了，但是我们需要的是中文的，因此要将这个词库分类翻译成中文。在调用有道字典API进行机器翻译时，发现翻译不是很到位，而且存在专有名词翻译不过来的情况，这个问题会在后面的当前遇到的问题这个模块进行详细说明。数据处理的第二部分便是同义词词库的构建了，于是我找到了哈工大整理的同义词词林（如图2），目前看来还是可用的，因为没有源数据，无法进行可信度判断，这个工作后期再进行。

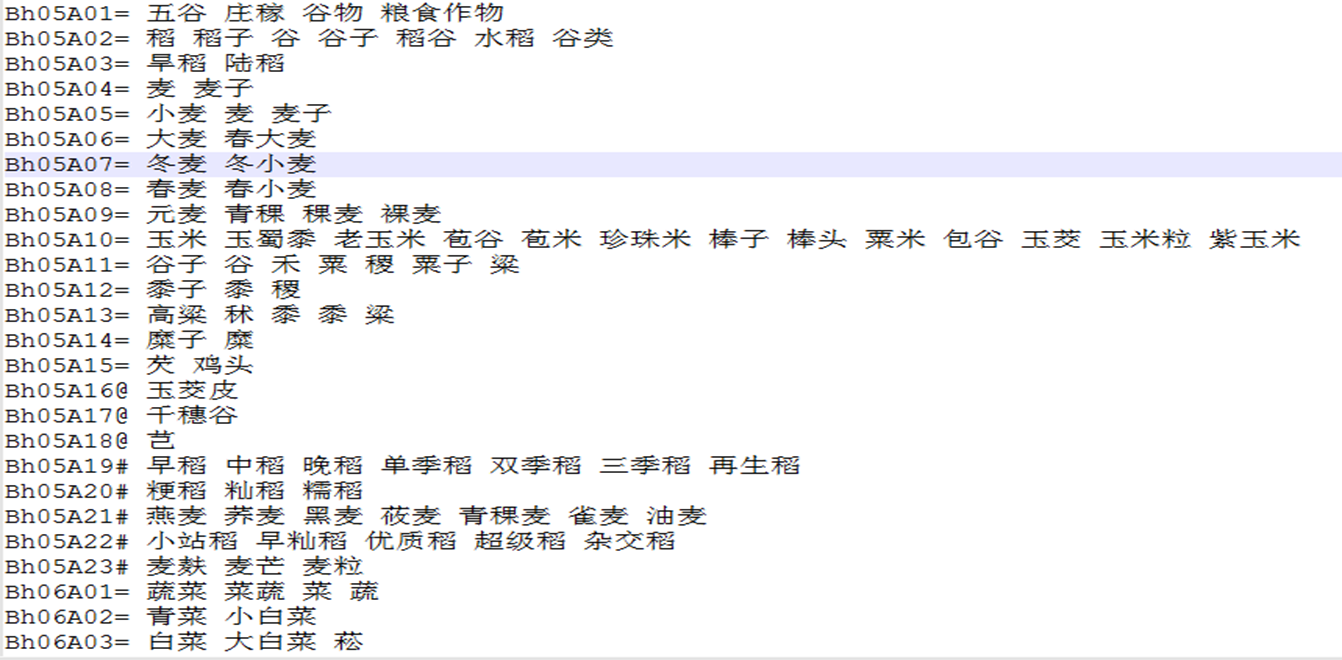


图 2 哈工大同义词词林

1. **可视化编辑器的实现**

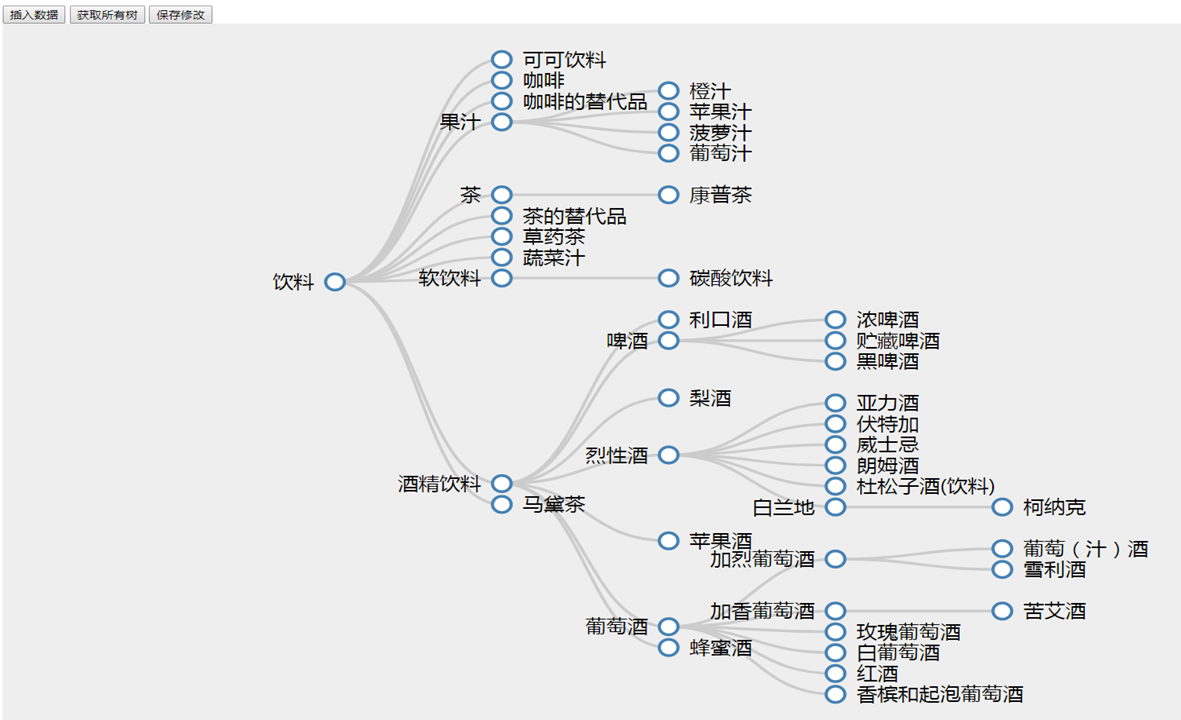


图3 可视化编辑器

考虑到数据的可视化展示存在父子层次结构，因此在数据的存储过程中也进行了层次化处理，采用了json的格式整体存储。使用mongodb对json格式的数据直接进行存储，在读取的时候直接整体读取，修改后的数据根据其id直接覆盖原有的数据，这样就可以保证前后端数据的一致性。可视化部分采用D3.js进行对数据处理展示，包括可以对节点进行拖拽修改，节点样式的修改等。D3.js是一个JavaScript 的函数库，集成了JavaScript用于对数据处理的函数，方便用户的使用需求（如图3所示）。

1. **用户数据处理功能的实现**

这个模块是在前面三个模块实现的基础上进行的，因为前面模块实现过程中遇到了一些问题，因此该部分功能还没有真正实现。以下我就介绍一下我的一些思路，具体实现可能会根据实际需要有些许变动。

该部分是根据我们前面构建起来的数据集对用户数据进行标准化的一个过程，首先假设用户数据是以一棵树的层次结构输入（如：json格式），首先需要对该树的每个节点进行遍历，对遍历的每个节点都执行如下操作：1）在标准的分类中寻找该节点名称，若没有找到则需要在同义词词库中搜索该名称，并把该名称替换为标准名称，若该名称在同义词词库中也不存在，则可以在可视化展示的时候提示用户对该节点换一个名称再进行匹配；2）在对每个节点进行匹配的过程中，需要对其depth以及其父子名称进行判断，如果该节点被分类错误，在可视化展示的过程中要提示用户将其归类到应有的分类中去。如此，每当用户输入新的分类，根据我后台维护的标准对其进行标准化，使得用户得到的最终分类是准备分类的一棵子树。

**三、现阶段工作、问题总结**

在第二部分已经介绍了各个模块具体实现，目前已经完成了前两个模块的工作，即可视化编辑器的设计编写以及农业实体数据的收集处理。目前在进行第三模块的工作，遇到些许问题在下面陈述。

首先便是英文词汇向中文翻译的问题，虽然找到了可以进行调用的有道API接口对其进行翻译，但是翻译过程存在着歧义，一词多义的问题，总结地说翻译过来的中文词汇不能够满足我们的需要，如果再进行人工纠错的话，工作量就大大增加。仔细思考了这个问题，最有效的解决方案就是找到中文的词汇，这样会更加权威可信。

第二个问题就是考虑用户输入格式问题，用户输入数据格式对于系统处理方式影响很大，初步设想用户的输入格式为标准的json格式，这个问题还有待进一步考虑。