项目介绍

本项目计划构建热门动漫的相关信息的知识图谱，爬取的信息包括动漫的中文名、角色设计、类型、主要配音、导演、动画监督、首播电视台。最后基于 D3 可视化工具进行可视化，并提供搜索功能。用户可以在搜索框中输入想查询的动漫，网页会给出该动漫的相关信息，并隐藏额外的其他信息。

1 数据

数据爬取

动漫列表从维基、B 站、豆瓣、百度网站上获取，涵盖近年来较为热门的 345 个日漫。然后爬取了百度百科网站上这些动漫的相关的信息。

中文名	犬夜叉	首播电视台	读卖电视台/日本电视台系列
其他名称	いぬやしゃ，Inuyasha	其他电视台	Animax（东南亚）
动画制作	SUNRISE，读卖电视台		中国电视事业股份有限公司（中国台湾）
集　数	167（正篇）+26集（完结篇）		龙祥电影台（中国台湾）
类　型	魔幻，冒险，热血，少年	网络播放	爱奇艺
地　区	日本		优酷网
原　作	高桥留美子		土豆网
导　演	青木康直，池田成		乐视网
剧　本	隅沢克之		腾讯视频
角色设计	高桥留美子，菱沼义仁，沙仓拓实	制　片	谏访道彦，富冈秀行，斋藤朋之，佐藤弘幸
作画监督	池田繁美	出　品	读卖电视台、SUNRISE舞台设定
音　乐	和田薰	发行时间	2000年10月16日-2004年9月13日（正篇）2009年10
主要配音	山口胜平，雪野五月，辻谷耕史，桑岛法子，渡边		月3日-2010年3月29日（完结篇）
	久美子，成田剑，能登麻美子，日高法子，长岛雄		
	一，绪方贤一，森川智之，松野太纪		

数据清洗

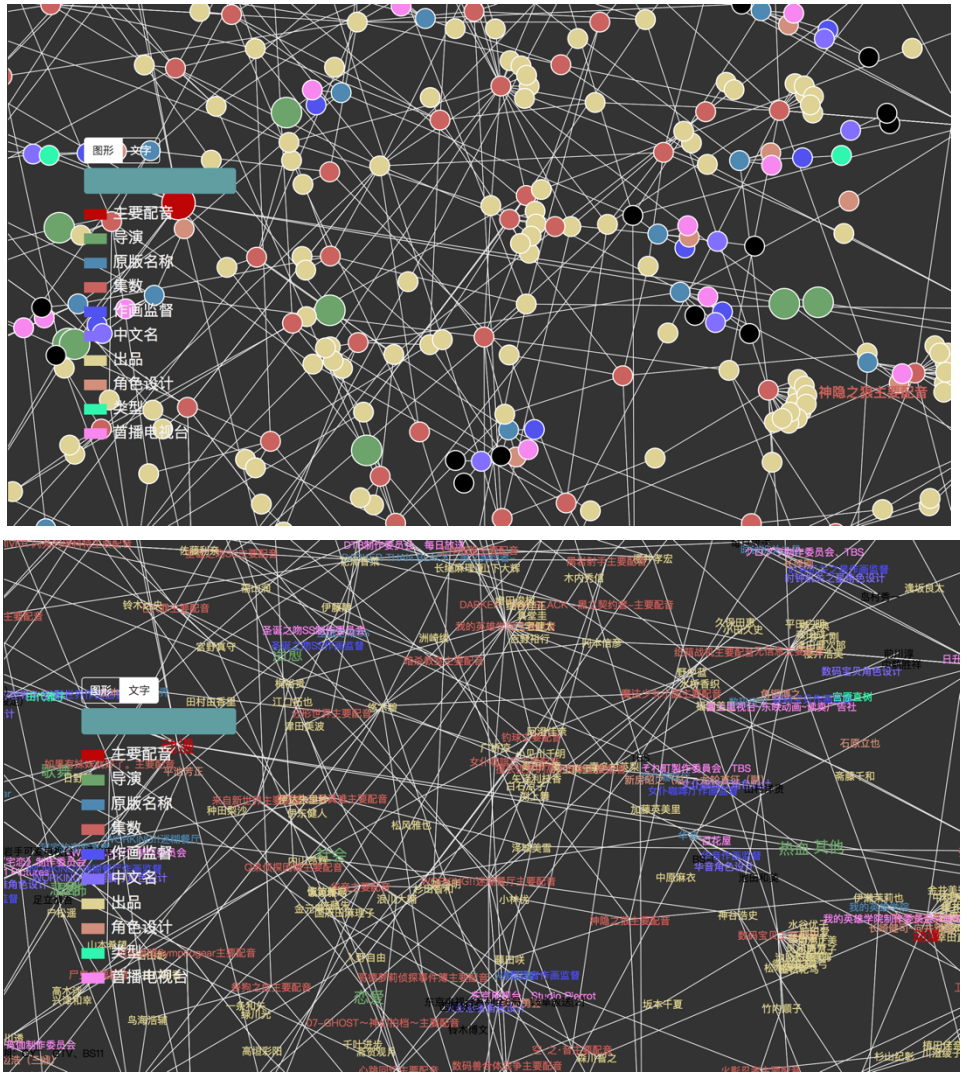
在爬取的过程中，缺失的信息先用无信息替代，清洗环节主要是针对“类型”这一项数据进行清洗，将所有动漫根据原有的类型字符串分为 13 类，包括热血、励志、奇幻、恋爱、歌舞、社会、喜剧、悲剧、悬疑、恐怖、治愈、日常、情感。无信息的先归位“社会”。

最后获取的数据文件为 all.json，每一个动漫包括“主要配音”，“导演”，“原版名称”，“集数”，“作画监督”，“中文名”，“出品”，“角色设计”，“类型”，“首播电视台”信息，总共有 345 个动漫数据条。

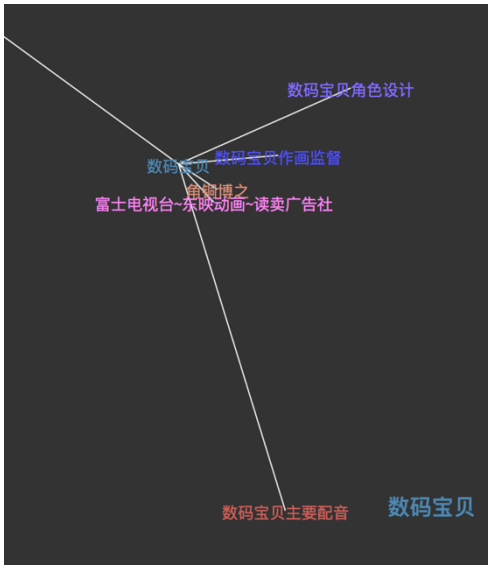
2 可视化

清理后，数据使用 D3 可视化工具进行可视化。首先将数据进行转化，将相关的实体和关系存储为 node 结构，并且使用 link 结构表示实体间存在关系，把 node 两两进行连接。并且可以将 node 节点根据类型绘制彩色实心圆圈，或者使

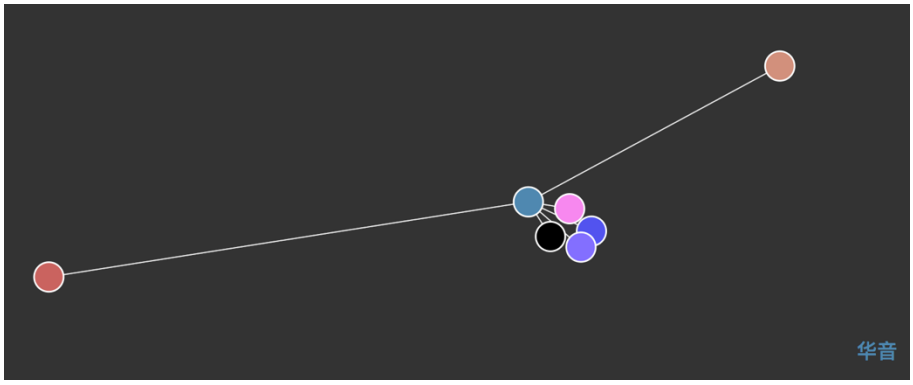
用彩色文字进行显示。实例如下图。



举例说明，图谱中心为“动漫”节点，连接具体类型节点，如“奇幻”、“热血”，然后这四个类型都连接到动漫名称节点“数码宝贝”，再从该节点为中心分散其他节点，导演“角铜博之”，首播电视台“读卖电视台”、“东映动画”、“富士电视台”等。实例如图所示。



并且提供搜索和隐藏额外数据这两个应用。根据数据，可以进行数据查询，显示出相关节点。比如输入“暗杀教室主要配音”，此时网页会隐藏其他无关节点，保留查询动漫的信息。效果如图所示。



另外还可以通过鼠标位置仅显示特定子分支，从而隐藏过多可视化数据。具体应用如图所示。