文本复制检测报告单(全文标明引文)

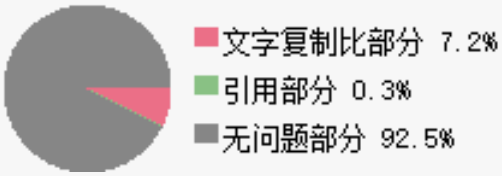
№:ADBD2018R_2018041811104720180418111156120405044731 7 检测时间:2018-04-18 11:11:56

检测文献: 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化
作者: 蒋冬冬
检测范围: 中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
学术论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库
个人比对库
时间范围: 1900-01-01至2018-04-18

检测结果

总文字复制比：7.5% 跨语言检测结果：0%
去除引用文献复制比：7.2% 去除本人已发表文献复制比：7.5%
单篇最大文字复制比：3.1%（python命名实体抽取学习记录（1）- lalalawxt的博客 - CSDN博客）

重复字数：	[2675]	总段落数：	[5]
总字数：	[35748]	疑似段落数：	[4]
单篇最大重复字数：	[1117]	前部重合字数：	[185]
疑似段落最大重合字数：	[1741]	后部重合字数：	[2490]
疑似段落最小重合字数：	[185]		



指 标：☐疑似剽窃观点 ☑疑似剽窃文字表述 ☐疑似自我剽窃 ☐疑似整体剽窃 ☐过度引用

表格：0 公式：0 疑似文字的图片：0 脚注与尾注：0

2.5%（185） 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第1部分（总7355字）
0%（0） 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第2部分（总249字）
6.2%（530） 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第3部分（总8568字）
2.4%（219） 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第4部分（总9076字）
16.6%（1741） 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第5部分（总10500字）

（注释：无问题部分 文字复制比部分 引用部分）

1. 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第1部分 总字数：7355

相似文献列表 文字复制比：2.5%(185) 疑似剽窃观点：(0)

1	35_133520085211015_李卓航 李卓航 -《学术论文联合比对库》- 2015-05-04	1.6%（121） 是否引证：否
2	马宏宇_s12006154_机器学习的研究和基于支持向量机算法的应用 马宏宇 -《学术论文联合比对库》- 2015-04-03	0.9%（66） 是否引证：否

原文内容 红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

分类号：单位代码： 10033

密级：学号： 153520085211015

硕士学位论文

中文论文题目：中国纪录片知识图谱的构建及可视化

英文论文题目： Visualization and Construction of Knowledge Graph Based on Chinese Documentary

申请人姓名：蒋冬冬

指导教师：尚文倩

专业名称：计算机技术

研究方向：数字娱乐与动画技术

所在学院：计算机学院

论文提交日期

中国传媒大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国传媒大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：签字日期：年月日

学位论文版权使用授权书

本学位论文作者完全了解中国传媒大学有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权中国传媒大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：导师签名：

签字日期：年月日签字日期：年月日

致谢

时光匆匆，研究生生活转眼即逝，在入学的第二个年头便要离开，心中感概万千。在这段时间的学习旅程中，作为一名研究生，在学术上，通过老师和同学们潜移默化的影响，了解了现在所学专业的前沿技术和各个分支，拓宽了自己的眼界和思维，并且在了解自身需求和爱好下选择了适合自己的方向，并且深入学习。在深入学习自己所选方向的知识的同时，也不忘时刻关注目前行业内大火的机器学习、区块链等技术。这个行业让我感受到了时刻学习的重要性，同时也让我感受到了探索新知识的乐趣。在生活上，认识了很多热情、活泼、优秀、乐于分享的同学，在实验室，我们是一群互相学习、互相鼓励、一起探讨的同学，学习之余，我们是一群一起分享生活、发现生活、一起疯狂的好朋友。这样的美好让我觉得我的青春期还在，而且一直都在。

在这段值得留恋的研究生生活中，我要对所有帮助我，支持我的老师，同学，亲人们表示诚挚的感谢。

首先要表示感谢的是我的导师，尚老师是一个随和而又严厉的老师，对学生十分关心，无论是学习上还是生活上。每次我遇到问题去请教老师时，老师都是很认真负责的进行解答，虽然我选择的方向不是老师擅长的，但是每次老师也会很热心的给我介绍在这个方向很厉害的老师、师兄、师姐们，鼓励我积极主动地向他们学习请教。而且在我选定自己的方向的时候，老师也非常支持我，也给予了我很大的帮助。在平时生活中，老师也会经常把我们几个学生聚在一起，聊一下学术研究的进展之后也会关注一下我们的生活，跟我们唠唠家常，有烦恼的时候老师总会鼓励我们，找工作的时候老师也会提供各种机会，真的是一位很贴心的老师。能够遇到这位导师是我的幸运。除此之外我还要感谢一下李老师，在李老师的管理下，实验室干净整洁、同学之间亲密无间、实验室氛围活泼又严肃，这给我们创造了良好的学习氛围。另外，在项目上，李老师也分工明确，管理有方，明确责任，而且能够给予中肯的意见，遇到问题时也会很热心的帮助我们。研究生期间还有很多优秀、负责的老师，在此表示诚挚的感谢，谢谢各位老师的栽培。

其次要感谢研究生期间的同学，师哥师姐们，师弟师妹们。实验室的氛围一直都很好，学习的时候大家都很认真，遇到问题的时候大家也都很热心的帮忙解决问题，师兄师姐们也会给我们分享自己的经验教训，这使得我很喜欢在实验室学习，非常感谢他们。另外，我还要感谢班级的同学，同学之间的相互分享、帮助和鼓励才使得我的研究生生活充满乐趣和动力，同学之间的相互包容和磨合使得我们之间的情谊更加珍贵。在此预祝各位同学前程似锦，一切顺心。

最后要感谢我的父母和家人，谢谢你们一如既往的支持和理解，如果没有你们的支持和鼓励，我也不会顺利读到研究生。你们不计回报，默默的付出是我前进的动力。在此献上我最诚挚的感谢和祝福，我也会一如既往的努力，不辜负你们的付出。

摘要

近几年来，随着‘舌尖上的中国’、‘互联网时代’等热播纪录片的出现，让我们重新认识到纪录片带给人们生活的意义。历史类的纪录片能够帮助人们更多的了解历史，从而吸取历史教训。所以为了促进我国纪录片的发展，对我国记录片进行分析，从而进行纪录片相关数据的可视化和知识图谱的构建就有了深远的意义。

随着大数据的火热，数据可视化技术也发展到了一定阶段，各个产业都将可视化技术应用于相应的场景，以此来促进各个产业的发展。数据可视化技术产生的目的就是为了更加直观、清晰的展示大量、复杂的数据。在可视化技术发展的同时，“知

识图谱”这个词被频繁提及，很多公司为了不被时代淘汰也开始积极研究这一领域。知识图谱相关知识属于人工智能领域，人工智能是为了让机器像人一样进行思考、决策，而人能够做出正确的决定是需要一定的知识积累的，所以要想使人工智能更加的“智能”，则需要建立更加完善的知识图谱。

根据以上技术手段，本文针对国内纪录片，应用以上介绍的技术思想，设计实现了中国纪录片数据可视化的展示和知识图谱的构建。数据来源于对三大视频网站（腾讯、爱奇艺、优酷）数据的爬取和纪录片解说词。本文总体流程中所涉及的理论知识与相关方法主要包括可视化技术，知识图谱构建，网络爬虫的编写等。文中对这些技术的理论与应用进行了详细介绍。

关键词：中国纪录片知识图谱数据可视化命名实体识别网络爬虫

the Visualization and Construction of Knowledge Graph of Chinese Documentary

ABSTRACT

In recent years, with the emergence of popular documentaries such as “A Bite of China” and “The Internet Age”, let us re-recognize the significance of documentary films for people’s lives. Historical documentaries can help people learn more about history and draw lessons from history. Therefore, in order to promote the development of our country's documentary, it is of far-reaching significance to analyze the documentaries of our country and to visualize the relevant data of the documentaries and construct the knowledge map.

With the development of big data, data visualization technology has also developed to a certain stage. Various industries have applied visualization technology to corresponding scenarios to promote the development of various industries. The purpose of data visualization technology is to display a large number of complex data more intuitively and clearly. At the same time as the development of visualization technology, the term “knowledge map” was frequently mentioned, and many companies began to actively study this field in order not to be eliminated by the times. Knowledge map is knowledge in the field of artificial intelligence. Artificial intelligence is to make machines think and make decisions like humans, and people can make certain decisions that require certain knowledge accumulation. Therefore, it is necessary to make artificial intelligence more “smart”. , you need to establish a more complete knowledge map.

According to the above technical means, this paper aims at domestic documentaries and applies the above-mentioned technical ideas to design and display the visualization of Chinese documentary data and the construction of knowledge maps. The data comes from crawling data from three major video websites (Tencent, iQiyi, Youku) and commentaries on documentaries. The theoretical knowledge and related methods involved in the overall flow of this article mainly include visualization technology, knowledge map construction, web crawler preparation and so on. The theory and application of these technologies are introduced in detail in this paper.

Keywords: Chinese documentary, Knowledge graph, Data visualization, Named entity recognition, web crawler

目录

1 绪论

1.1 研究背景与意义

继中国纪录片的发展沉寂了多年之后，近几年来，一些深受大众欢迎的纪录片陆续展露头脚，在目前中国影视市场综艺满天飞的境遇中如一股清流出现在大众的视野中。例如：2012年5月在央视首播且目前已经更新到第三季的“舌尖上的中国”、2017年12月在央视首播且在豆瓣评分高达9.5分的“国家宝藏”等等。这些纪录片的成功给我国纪录片的发展带来了希望和曙光。

纪录片在一定意义上可以看作是一件艺术品，“好看”、“真实”是纪录片的两大特征。在一定程度上，纪录片对外可以说是国家对外形象的“明信片”，对内也可以说是提高国民基本素质的“教科书”，所以纪录片的意义还是很深远的。中国纪录片既能够让中国国民更深入的了解中国的历史、文化、地理等，使国民更好的融入当前社会；也能够帮助被中国吸引的外国人更好的了解中国。我国纪录片的发展历史可以追溯到半个世纪之前。从中日合拍的纪录片《丝绸之路》开始，之后制作的《话说长江》等纪录片掀起了中国纪录片史上的高潮，再到后来纪录片盛产的文革时期。

20世纪90年代，电脑还未普及，电视还是主流的播放平台，带有纪录片性质的电视节目深受大众的喜爱，就像现在的真人秀综艺节目一样，风靡一时，但是随着社会的发展，社会风气及人们的生活习惯也都随之发生了改变，纪录片也不再像当初一样蔚然成风，渐渐淡出了大众的视野。虽说我国纪录片在现代社会取得了一定的成就，也产生了一些有影响力的作品，如《大国崛起》《故宫》《舌尖上的中国》等，但是数据显示：截至到2016年12月1日，纪录片以 74.9 亿次的点击量占全网节目视频点击量的 0.94%[2]，从这个数据可以看出，纪录片的占比是相当低，连百分之一都没有占到。为了促进我国纪录片的发展

，使其更好的发挥教育和宣传意义，找出能够使纪录片吸引大众的方式方法成为一件刻不容缓的事情。

一部纪录片是否受观众欢迎的影响因素有很多，例如纪录片的选题一定程度上影响着该记录片的受欢迎程度，比如《舌尖上的中国》，该记录片选择人们最关心的“吃”作为题材，然后选择独特的拍摄方式与故事线条，最后成为脍炙人口的一部作品。近年来，在纪录片的国际舞台上也可以发现，纪录片的选题呈现出多元化的现象，什么样的题材才能够受观众的青睐，怎样的纪录方式更能引起观众的兴趣成为纪录片创作者关注的焦点。

当前社会，随着计算机、互联网的普及，人们所有的操作、行为都会被记录下来，所以每天都有各种各样、数量庞大的信息数据产生。这些数据中蕴藏着大量的机会和可能性，数量越庞大，信息也就越多，机会也就越大。例如：我们可以从用户的行为数据中分析出用户的喜好，从而为其推荐商品。为了更好的利用这些数据，让人们能够更加直观、便捷的从中获取有效信息，从而产生了数据可视化技术。身处大数据的时代，利用与大数据相关的技术方法来助力我国纪录片的发展有重大意义。通过收集互联网上各种纪录片相关的数据，然后通过分析这些数据可以挖掘出有益于我国纪录片发展的信息，例如：哪种主题的纪录片比较受欢迎，哪种形式的纪录片大众比较容易接受等。该论文就致力于这个方向，对互联网上的数据进行收集、分析，最后进行了可视化的展示，使得人们能够清晰、快捷的从这些数据中了解我国纪录片的发展现状。

在大数据技术火热的同时，人工智能也登上了历史的舞台，成为目前非常重要且非常受欢迎的技术。知识图谱是人工智能领域的核心，知识图谱能够使机器更加智能、更加自动化。知识图谱是采用“实体-关系-实体”或“实体-属性-值”这两种知识表示形式对现实世界进行描述的一种方法。通过构建知识图谱，我们能够更加清晰的了解到纪录片讲解的内容，实现定位播放功能，改善纪录片的观看体验，使观众不用把冗长的纪录片全部看完才能找到自己想获得信息。

1.2 研究现状

随着社会的变迁和技术的更新，中国纪录片的发展轨迹也发生了改变，例如：在内容上，现在的纪录片更多的以大众比较在意的题材为主题；在技术上，现在的纪录片综合应用电脑特技、三维动画等新技术手段对故事进行叙述；在传播途径上，也从之前的电视转播转变为通过新媒体平台在网络上进行播放。为了更加清晰的了解到中国记录片的发展方向，我们要更加准确的把握目前纪录片的发展现状，了解当前社会环境、新兴技术[3]。

互联网的发展，使当前这个世界更加的数字化，人类的任何行为、操作都会被当作数据记录下来，所以每天都会产生海量的数据，而这些海量的数据结构复杂、形式多样，人们很难从中发现有用的信息，所以大数据技术随之兴起。目前大数据相关领域的数据挖掘、人工智能技术处于快速发展阶段。各行各业都有将这些技术应用于产业从而取得巨大成效的案例。大数据中的数据可视化技术目前发展相对成熟，相关的理论比较完备，而且可供参考的成功案例也非常多，但是，目前知识图谱的构建技术，虽说国内外都已有相关的成功项目，但是该技术的普及程度还不是非常广。

数据可视化步骤主要包括数据采集、数据分析、可视化展示三个部分。目前基于JS的可视化工具主要有：ECharts.js、D3.js、Hightchart.js等。随着可视化技术的普及和成熟，影视化产业也紧跟脚步，将影视化产业的发展和信息技术的发展融合在一起。目前，三大视频网站：腾讯、爱奇艺、优酷都相应的有数据可视化的平台来分析用户的行为，以此来更好的优化用户体验，促进影视业的发展。另外也有很多关于影视剧数据的可视化分析。

随着人工智能的发展，知识图谱构建技术近年来引起了广泛关注，这个概念最初是由谷歌提出的。谷歌为了优化自己公司开发的搜索引擎，使其更加智能，从而开始研发知识图谱项目。该项目其实不是从头起步的，而是在收购了一家研究语义搜索的公司之后，在该公司的核心技术上进行改进而展开实施的。知识图谱其实并不是一个全新的概念，在提出这个术语之前就有与之相关的研究。知识图谱的前身其实就是语义网，可以说知识图谱是对语义网技术的延伸和升华。早在2006年就有专家提出数据链接的思想，并呼吁业界对相关的技术制定一定的标准。目前国内外对知识图谱的研究都有了一定进展。

我国对于中文知识图谱的研究已经起步, 并取得了许多有价值的研究成果，例如中国科学院计算机语言信息中心董振东领导的知网 (HowNet) 项目、清华大学建成的第一个大规模中英文跨语言知识图谱XLore等。本论文将知识图谱这一新鲜的概念和技术应用到影视领域，是一个突破点和创新点。

1.3 研究内容及目标

本文围绕中国纪录片的发展历程及现状，通过研究与可视化相关的理论和技术及知识图谱构建相关的理论和技术，致力于建立一个集展示三大视频网站中纪录片的相关数据和纪录片知识图谱于一体的系统。

在技术上，本文采用python中Beautiful Soup库来获取源数据，然后用python处理相关数据。服务器端用的是Node的express框架来搭建的轻量级服务器，因为这个技术比较新，所以尝试了一下。前端用的是Facebook研发的React前端框架和用React实现的Ant design UI框架进行实现。数据可视化部分用的是国内百度研发的Echarts进行实现。实体的抽取用的是哈工大社会计算与信息检索研究中心研发一个自然语言处理工具库pyltp。

1.4 论文组织结构

本文一共分为六个章节，每章内容安排如下：

第一章这一章主要介绍了本文选题的背景和意义、本文研究内容的发展现状、本文所要研究的内容和本文的主题结构。

指　标
疑似剽窃文字表述

1. 申请人姓名：蒋冬冬
指导教师：尚文倩
专业名称：计算机技术
研究方向：数字娱乐与动画技术
所在学院：计算机学院
论文提交日期

2. 我要对所有帮助我，支持我的老师，同学，亲人们表示诚挚的感谢。
首先要表示感谢的是我的导师，尚老师

2. 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第2部分

总字数：249

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第二章这一章节从数据可视化的发展、数据可视化的一般流程、数据可视化的方式、实现数据可视化的工具等几个方面介绍了数据可视化技术。

第三章本章首先对知识图谱这一概念做了简要描述，然后介绍了知识图谱领域常见的术语，有助于我们更深的了解知识图谱这一新的技术，然后对知识图谱的分类做了简要说明，最主要的是对构建知识图谱将会用到的技术作了介绍，最后介绍了知识图谱架构。

第四章介绍了本文所实现的系统的功能模块的划分和本系统的开发环境和技术。

另外还逐一列出了本系统所涉及的接口，并进行了说明。

3. 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第3部分

总字数：8568

相似文献列表　文字复制比：6.2%(530)　疑似剽窃观点：(0)

1	520100389_龚杰_测绘工程（信工）	2.4%（208）
	龚杰 -《学术论文联合比对库》- 2015-06-04	是否引证：否
2	财新网数据可视化实践研究	1.9%（160）
	黄婷婷(导师：董小玉) -《西南大学硕士论文》- 2017-05-26	是否引证：否
3	大数据环境下的可视化数据信息安全研究	1.5%（131）
	龙虎;李娜; -《福建电脑》- 2016-12-25	是否引证：否
4	移动互联网数据可视化技术及应用研究	1.1%（92）
	张青;陶彩霞;陈翀; -《电信科学》- 2014-10-20	是否引证：否
5	教务系统中的数据可视化技术研究	1.1%（90）
	李济龙(导师：胡健) -《北方工业大学硕士论文》- 2015-06-30	是否引证：否
6	李济龙-2012312170110-软工研12-教务系统中的数据可视化技术研究	1.1%（90）
	李济龙 -《学术论文联合比对库》- 2015-04-25	是否引证：否
7	2120230412-杨豪斌	1.0%（88）
	杨豪斌 -《学术论文联合比对库》- 2016-03-18	是否引证：否
8	漫谈数据可视化_技术	0.9%（79）
	-《网络（http://www.cnii.com.）》- 2015	是否引证：否
9	漫谈数据可视化_信息化	0.9%（79）
	-《网络（http://www.cnii.com.）》- 2015	是否引证：否
10	中医健康知识图谱的构建研究	0.7%（63）
	郝伟学(导师：于剑) -《北京交通大学硕士论文》- 2017-06-01	是否引证：否
11	海洋数据可视化及应用研究	0.6%（49）
	张才章(导师：方景龙) -《杭州电子科技大学硕士论文》- 2015-03-01	是否引证：否
12	基于分层技术的毕业生管理系统设计与实现	0.5%（43）
	贾海天;陈志峰;应俊; -《电脑编程技巧与维护》- 2011-11-18	是否引证：否
13	基于任务控制的软件项目实训管理系统设计与实现	0.5%（43）
	李宝智; -《现代计算机》- 2013-07-25	是否引证：否
14	矿山地面灾害监测数据可视化表达和实现技术	0.4%（33）
	席茂(导师：张锦) -《太原理工大学硕士论文》- 2015-05-01	是否引证：否
15	席茂-2012510543	0.4%（33）
	-《学术论文联合比对库》- 2015-04-13	是否引证：否

第五章主要从数据获取、数据预处理、数据展示三个流程介绍了本文对三大视频网站纪录片相关数据进行可视化展示的过程及实现效果。

第六章介绍了本系统中知识图谱部分功能的实现方法及结果展示，其中详细介绍了本文中实体抽取的详细步骤。。

第七章对本文进行了总结，并对后续所要做的工作进行了说明。

2 可视化相关理论与技术

数据的来源有很多，例如：数字，图片、声音等都是数据。数据描绘了现实的世界，是对现实世界的简化和抽象表达，可以说是现实世界的一个快照。现在，只要有行为的发生，都有数据的产生，所以每天都有大量的数据产生。在数据爆炸的背景下，想要从多维、多源、多态的复杂数据中方便、快捷、直观的获取有用的信息，只靠单纯的人力是不可能的，这时候就要借助计算机的力量，将这些数据以‘有形’的方式展示出来，让人们能够快速的捕捉到这些数据所蕴含的信息。在这种背景环境下，数据可视化技术应运而生。如果说数据是对现实世界的抽象表达，那么数据可视化是对数据的一种抽象表达。

把握数据及数据所蕴含的信息是数据可视化技术的关键之处，只有这样才能充分发挥数据的力量，揭示有用的信息。目前对数据可视化诠释最好的定义是：***“数据可视化指综合运用计算机图形学、图像、人际交互等技术，将采集或模拟的数据映射为可识别的图形、图像、视频或动画，并允许用户对数据进行交互分析的理论、方法和技术[4]。”***数据可视化的主要目的就是为了展示数据特征，揭示数据蕴含的重要信息。

2.1 数据可视化的发展

任何技术的产生和发展都与人类的生活和需求息息相关，可视化技术也不例外。下面将按照时间顺序从先计算机时代、计算机读表时代、计算机读图时代、大数据时代这几个时代来介绍可视化技术的发展。

（1）先计算机时代

在计算机还未出现的时代，人类就已经开始手工绘出可视化图表。在那个时代，物理基本量（时间、距离和空间）的测量理论和设备的完善、新的图形化形式（等值线、轮廓线）和其他物理信息的概念图（地理、经济、医学）的出现使得系统地构建可视化方法的条件日渐成熟。

在中国，有代表的是陈正祥所绘的中国诗人分布图，如图2-1所示：

图2-1中国诗人分布图

（2）计算机读表时代

随着技术的发展，人类开始进入计算机时代，计算机最强大的就是计算功能，利用计算机来对报表进行统计分析是再合适不过了，所以最先在可视化领域展露头脚的就是表格。在这个时期比较有影响力的有：水晶报表、华表、思达报表等。水晶报表的发展比较曲折，先后经历几次改名和收购，但是水晶报表依旧是是业内非常专业且功能强大的报表系统。

（3）计算机读图时代

随着计算机技术的发展，为了适应新技术，可视化的手段已经不仅仅局限于某些厂商提供的像饼状图、柱状图、折线图这样的图表控件的方式。在Flash占据大量市场份额的时候，出现了Fusioncharts这样一个融合多个平台，跨多个浏览器的Flash图表组件解决方案；在进入H5的时代后，市场上开始涌现像Echarts.js、Highcharts.js、D3.js这类的可视化库，甚至能够实现3D效果；为了实现更加专业的可视化效果，市场上也开始出现一些专业的可视化软件，例如：tableau、FineBI。

（4）大数据时代

随着计算机的普及，人类所有的行为基本都会被记录下来，所以每天都在产生海量的数据，人类正式步入大数据时代。

大数据时代的到来给数据可视化技术带来了巨大的挑战。第一：数据量的骤然增大和数据形式的多样化给可视化增大了难度。以前是对少量的数据进行可视化，相当于在湖中划行，而现在面对的是巨大的数据量，从湖中划行变成了在海中航行。相应的，新的技术也被提出来面对这个挑战，那就是分布式计算和内存计算。第二：复杂且大量的数据将会造成可视化形式的多样性。面对之前简洁且少量的数据，进行展示的时候用比较常规的可视化方案就能解决，但是现在面对的数据复杂且巨大，在进行可视化的时候要找到更加直观明了的可视化方案，甚至要尝试着从多个角度对这些数据进行全面剖析，这就需要更加多样的展示方案，这样才能更加全面的获取信息。

2.2 数据可视化

2.2.1 数据可视化流程

在进行数据可视化时，往往要考虑以下四个问题：（1）你拥有什么数据？（2）关于这些数据，你想要了解什么？（3）这些数据适合哪种可视化方案？（4）从图中你是否获取到了有用信息？对这四个问题的回答即为可视化的一般流程。

这四个问题其实是相互关联，循环往复的一个迭代过程，图2-2可以清晰的阐明这几个问题之间的关系。

图2-2 可视化问题间的关系

关于前两个问题和最后一个问题，是需要本人去斟酌判断的，没有什么方法技术可言，可是对于使用哪种可视化方式，则

需要我们在了解各种可视化方式的特点之后，根据需要进行选择，这就引出了下一个小章节：数据可视化的方式。

2.2.2 数据可视化方式

对于不同类型的数据，可视化的方法也不尽相同。根据数据的类型，将其分为：时空数据可视化、地理空间数据可视化、高维非空间数据可视化、层次和网络数据可视化、跨媒体数据可视化等。

时空数据一般可分为：一维标量数据、二维标量数据、三维标量数据、多变量空间数据、时间序列数据等。在对一维标量数据进行可视化时，因为一维数据只需展示数据的一个维度，所以通常采用二维坐标图或折线图来进行可视化；在对二维标量数据进行可视化时，要从两个维度对数据进行展示，所以通常采用颜色映射、高度映射、等值线提取等方法在二维坐标系的基础上进行可视化；在对三维标量数据进行可视化时，因为其比二维数据更加复杂，要从三个维度进行展示，所以通常采用等值面绘制；在对多变量空间数据进行可视化时，通常采用矢量场或张量场数据可视化；在对时间序列数据进行可视化时，因为时间序列数据与时间有关，所以要观察所涉及时间的类型，然后采取合适的时间可视化方式，例如：对于连续时间的数据可以采用时间线可视化；对于周期性时间，可以采用周期时间可视化等。

对于地理空间数据的可视化来说，地图投影是其基础，地图投影的数学定义：λ,φ→(x,y)。其中，经度λ的取值范围-180,180，正值代表东部，负值代表西部。纬度φ的取值范围-90，90，正值代表北极，负值代表南极。地图投影常用的方法有：墨卡托投影、兰伯特投影、哈默-阿伊托夫投影、摩尔威德投影、余弦投影、亚尔勃斯投影。

对于高维非空间数据来说，因为该类型数据比较特殊，所以在进行可视化时要采取一些特殊的手段，例如：给数据降维、增加交互设计等。给数据降维主要是将反应相同信息的多个变量降为一个，可用的手段有：主成分分析、流行学习、张量化等。增加交互设计主要是给可视化图形增加交互动画，通过交互来更加全面的展示数据，常用的交互手段有：放大、画笔、链接等。

对于层次和网络数据的可视化来说，层次数据常采用节点链接、空间嵌套填充法，网络数据常采用节点链接、相邻矩阵布局。当然也可通过一些交互手段实现可视化。

除此之外，还有对一些跨媒体数据的可视化，例如：文本与文档数据的可视化、社交网络数据的可视化、日志数据的可视化等，这些数据也有相应的可视化手段都值得我们去学习研究。

基于图表的可视化是可视化的核心，所以在介绍完不同类型的数据可以采用的可视化方法之后，下面将详细介绍针对不同类型的数据，可采用哪种统计图表。

（1）单变量数据

单变量数据常用的统计描述方式有均值，众数和中位数，其关注点是数据分布的总体形状、分布比例和密度。下面将介绍常用的单变量数据统计图表方法。

1）柱状图：由一条条高度不同的纵向条纹组成，高度的高低代表数值的大小，常用于较小的数据集。如图2-3，展示了一周中每天的数据量：

图2-3 柱状图示例

2）直方图：很多人会将直方图和柱状图搞混淆，在表现上，不仔细看的话，直方图和柱状图的确很像，唯一的区别就是柱状图的纵向条纹之间有间隔，而直方图没有。所以直方图一般用来描述连续型数据；柱形图一般用来描述间断型数据。直方图如图2-4所示：

图2-4 直方图示例

3）饼图：用圆形及圆内扇形面积表示数值大小的图形，常用于表示部分在总体中的占比。示例如图2-5所示，显示了某站点用户访问来源占比。

2-5 饼状图示例

4）盒须图：盒须图因为长得像盒子从而获得这个名称，同时还有另外一种叫法：箱形图。这种图通常用来了解数据的分散情况。盒须图示例如图2-6所示：

图2-6 盒须图示例

5）其他的一些不常见的，例如：数据轨迹图、抖动图、核密度估计图等。

（2）双变量数据

我们研究双变量数据时最关心的莫过于这两个变量之间是否存在关系以及存在什么样的关系，所以展示双变量数据时要以这个问题为核心进行展示。目前展示双变量数据的方法如下：

1）散点图：散点图主要用来描述双变量数据中两个变量之间的关系，根据散点图的展示，我们可以根据经验对两个变量之间的关系进行判断，从而选择合适的函数对两个变量的关系进行拟合。散点图示例如图2-7所示：

图2-7 散点图示例

2）对数图与半对数图：描述两个变量之间的关系最常用的方式是将一个变量岁另一个变量变化的过程绘制在直角坐标系中，为了更加方便的观察以指数速度变化的变量之间的关系，不再描述原始数据，而是描述其对数值。对数图能有效呈现数据的发幅度变化。两个坐标轴均使用对数值的图成为对数图，只有一个坐标轴使用对数值的图称为半对数图。

（3）多变量数据

因为多变量数据比较复杂，所以在可视化这类数据的时候也要采用相对比较复杂、高级、使用的方法。下面将介绍最常用的多变量统计图表方法。

1）等值线图：等值线图就是将相等的点连成线，然后由这些线形成的图形。等值线图示例如图2-8所示：

图2-8 等值线图示例

2）热力图：热力图使用颜色来表达位置相关的二位数值数据大小。这些数据常以矩阵或方格形式整齐排列，或在地图上按一定的位置关系排列，由每个数据点的颜色反映数值的大小。如图2-9采用热力图显示了全国主要城市空气质量情况。

图2-9 热力图示例

2.3 实现可视化的工具

2.3.1 可视化工具

因为可视化技术的发展，目前市场上也出现了很多帮助我们实现可视化的工具，以便我们更加方便地整理、分析数据。了解这些工具的特性和优缺点有利于我们正确选择合适的工具来进行可视化操作。下面将详细介绍一些比较受欢迎的工具：

（1）Excel

这款工具可以说是人人皆知，已经连续流行二十多年了，可谓是制作表格的代名词了。现在在Excel中也可以制作一些简单的图表，只需选择数据然后选择插入需要的图形就可以。但是Excel局限于它所能处理的数据量上，而且如果遇到比较复杂的数据源和展示方案则需要专门学习一下内置的编程语言。

其实大家所知道的大都是微软公司office里面的excel，这个是线下的软件，所以需要提前安装。Google也上线了一款跟这个功能很类似的产品：Spreadsheets。这个产品在线就可用，这也是这个产品最具有竞争力的地方。

（2）Tableau

相对于Excel来说，Tableau可以对数据做更加深入的分析，而且不用编程。学习成本也比较低，可以在官网上找到相关的学习视频和优秀的作品。但是Tableau是收费的软件，而且费用也不便宜。

（3）Many Eyes

Many Eyes是IBM公司的一款在线可视化处理工具。该工具可以对数字，文本等进行可视化处理。虽然该软件已经在2010年停止更新，不过还能供人们使用。

（4）其他的工具

除了上面介绍的三个比较流行的可视化工具外，市场上还存在着一些针对特定数据的工具，例如：Gephi、ImagePlot、树图、TileMill等等。

2.3.2 编程工具

上章介绍的可视化工具都是拿来即用的，可是这些工具往往对于一些特殊的需求无法实现，如果会编程的话，就可以根据自己的需求定制自己想要的效果，也可以加入交互和动画。当然，编程的代价相对较高，你需要花费精力和时间去学习一门新的语言。不过在有了编程基础之后，学习一门语言并不是一件很难的事情，而且还能丰富自己的知识。

目前比较流行的有：（1）R语言，专门用于统计学计算和绘画的语言，它是开源的。（2）以javascript为基础的可视化库，例如：D3.js、Echarts、HightCharts等。

2.4 本章小结

本章首先介绍了数据可视化技术的发展，然后从数据可视化的流程和数据可视化的方式两个方面介绍了数据可视化的具体过程，最后介绍两类了实现数据可视化的工具。

3 知识图谱相关理论与技术

知识图谱对人工智能发展的重要性类似于知识对人类发展的重要性，都是举足轻重、不可小觑的。虽然知识图谱只是人工智能领域的一部分，而且相较于其他部分的发展来说还属于相对落后的一部分，但是知识图谱对人工智能的发展的意义是不容忽视的。知识图谱能够利用计算机的手段帮助我们梳理现实生活中的知识，帮助我们更加清晰的认识这个世界。

3.1 知识图谱简介

知识图谱这个概念最初是由谷歌提出的。谷歌为了优化自己公司的搜索引擎，使其更加智能，从而开始研发知识图谱项目。该项目其实也不是从头起步的，而是在收购了一家研究语义搜索的公司之后，在该公司的核心技术上进行改进而展开实施的。所以在任何百科类网站中对知识图谱进行搜索时，都会出现Google这家公司。知识图谱技术的出现并不是偶然、突然的结果，而是必然的结果。知识图谱是对语义网、信息抽取等技术的延伸，也是人类追求人工智能的必经之路[1]。

知识图谱的目的在于对现实世界进行模拟，理清真实世界的中的各种复杂的关系，最后形成一张知识网，该知识网由点和边构成，点代表实体或概念、边代表关系或属性。知识图谱例子如图3-1所示：C罗是一个实体，任务和运动员是他的基本概念，金球奖也是一个实体，C罗河金球奖之间的关系就是C罗曾经获得过这个奖项。

图3-1 知识图谱示例

想要更深的理解知识图谱这个概念，可以从知识图谱的学科地位和产生背景两个方面来加深对知识图谱的理解。

从学科地位上来讲，纵观人工智能领域的学科体系，我们可以清晰的发现知识图谱在该领域中的位置。人工智能是希望机器能够像人一样理性的思考和行为的一门学科，知识工程作为人工智能学科体系的一个分支，目的是将知识表示为计算机可以接受的方式，使其作为计算机的“知识储备”，从而使得计算机系统一定程度上可以像人一样，利用人的智谋作出正确的判断和抉择。而在知识工程领域中，知识表达是其核心，知识表达对知识工程的发展起着决定性作用。先要有知识表示，才可能有知识工程的巨大发展。再进一步来讲，在知识表示领域中，有个非常重要的表示方式就是知识图谱。知识图谱将现实世界中各种事物及其关系抽象出来，然后进行展示。

从知识图谱的产生背景上来讲，最开始提出知识图谱这个概念的诉求是为了让机器更好的理解语言，这样才能更精确的返回所要寻找的答案。让机器像人一样理解语言并不是一件简单的事情，语言的理解是建立在认知的背景上的，人和人之间能够相互理解，是因为有着共同的认知。所以为了让机器具备语言的认知能力，也需要让机器跟人类一样，有一定的知识沉淀，而这个知识沉淀可以理解为我们所说的知识图谱。知识图谱的目的就是作为一个强大的知识库，从而成为机器的'大脑'，使机器能够具有与人类一样的认知能力。有了知识图谱之后，机器能够对现实世界的事物进行认知，而不是简单地处理一串串字符。

知识图谱作为一种语义网络，是大数据时代知识表示的重要方式之一，作为一种技术体系，是大数据时代知识工程的代表性进展。

3.2 知识图谱相关术语

知识图谱作为一项在知识管理和知识服务领域新兴的技术，了解其所涉及的相关术语及这些术语与知识图谱的区别和联系有重要意义。下面将简要阐释这些术语。

（1）实体

世界万物皆由具体事物组成，该具体事物即指实体，如"台式电脑"、"笔记本"、"平板电脑"等。由此而知，实体指的是具有可区别性且独立存在的某种事物。实体是知识图谱中最基本的单元，不同的实体间存在不同的关系。如某一个人、某一个城市、某一种植物、某一种商品等等这些都可称为实体。目前，维基百科拥有非常丰富的实体知识，这些实体知识都是通过社区人员来人工维护的。

（2）概念

概念其实是对实体的抽象，例如"计算机"即为"笔记本电脑"的概念。具有相同特性的实体构成的集合可称为概念，如学校、老师、机构、国家等。

（3）本体

本体一词最开始是个哲学名词，在后来的发展过程中，人工智能研究人员将这一概念引入了计算机领域。Tom Gruber把本体定义为"概念和关系的形式化描述[6]"。这个定义跟从哲学上对本体的定义其实非常相似，都是用来表达实体、概念、事件及其属性和相互关系。通俗点讲，本体类似于数据库对象的集合，主要用来定义类和关系，以及类层次和关系层次等。最常用的本体描述语言有OWL和RDF，主要表述方式为三元组。本体通常被用来为知识图谱定义Schema。目前来说，专家构建的WordNet拥有极高准确率的本体知识，但是该词典是针对英文的。

目前，围绕本体出现的问题中讨论最为激烈的是"是否每个领域都需要一个独立的本体论，还是可以有一组共通的理论在所有领域内共享？"这个问题，其实该问题的本质就是本体的合并和对齐。

知识图谱和本体并不是相同的概念，知识图谱是在本体的基础上做了丰富和扩充。本体注重的是概念和概念之间的联系，而知识图谱主要描述的还是实体，只是在描述实体的基础上对其进行了扩展。总而言之，本体描述了知识图谱的数据模式，本体的动态特性赋予了知识图谱动态数据模式支持的能力。

（4）本体库

本体库即由大量本体组成的集合。在理解本体库的同时要区分其与知识库的区别。本体库相当于构建数据库时建立的ER图，是抽象层面的，而知识库则是在本体库构建基础上的具体事物。所以说本体库是用来管理知识图谱的模式层的。只有构建了本体库，知识库中的数据冗余才会缩小。不然会存在大量无用的知识。例如：（自然人、年龄、性别）形成一个本体库，那么大量的（小明，23，男）这样的实例数据则组成了一个知识库。

（5）知识库

知识库将事实和数据联系起来，能够用来进行推理，所以最开始是作为专家系统的一部分，方便人们的生活。专家根据其所在领域的专业知识，系统、全面地设计该领域中的规则集合，这些集合也就构成了我们所说的知识库。知识库侧重于知识的表达，而知识图谱更注重知识之间的关联性。

（6）知识表示

知识表示又可称为知识表现，主要研究系统中知识的组织形式，目的在于将结构化数据组织成便于机器处理、人类理解的形式。知识表现的数据结构，一般都是比较"复杂"的结构，目前知识表示方法有状态空间、与或图、谓词逻辑、产生式规则、语义网络、框架、剧本等。

（7）链接数据

链接数据是基于将互联网上的数据链接起来的思想而由Tim Berners Lee提出的，提出这个概念的目的不仅仅是为了将数据结构化之后放到互联网上供人浏览，更主要的是为了在结构化数据之间形成链接，获得数据之间的联系。另外Tim Berners Lee为建立数据之间的链接制定了四个原则[7]。其实，链接数据这个思想已经非常接近于知识图谱这个概念，在一定意义上来说，也促进了知识图谱的发展。目前最大规模的链接数据的项目是DBpedia项目[8]，基于这个项目，中国也发布了一个基于中文的CN-DBpedia项目。

3.3 知识图谱的分类

随着人类认知和技术的发展，知识图谱从以前耗时耗力且覆盖率低的人工构建一步步发展起来，形成基于结构化数据的知识图谱的构建（也可称为基于百科知识的知识图谱的构建）、基于非结构化数据的知识图谱的构建（也可称为基于自由文本的开放域知识图谱的构建）、基于本体库的知识图谱构建（也可称为自动构建的开放域中文实体知识图谱）。这个分类是根据构造知识图谱所依赖的数据的结构类型和自动化构建程度来分类的。当然，也有别的分类方式，比如：可根据适用领域分为通用

领域的知识图谱和特定领域的知识图谱。

3.3.1 基于百科知识的知识图谱构建

随着技术的发展，万维网的出现，各种各样的信息、知识遍布网络的角角落落，为了使人们更加快捷、方便、准确的获取想要的知识，吉米·威尔士与拉里·桑格两人开启了一个名为维基百科（Wikipedia）的全球性多语言百科全书协作计划，在维基百科这个平台上，知识以词条的形式组织，大家积极协作完成了上百条词条的编写，以后随着社会的发展，这个数量也会呈直线上升。维基百科的发展对结构化知识在网络上的传播带来了便利，同时也促进了链接数据的发展。

链接数据这个概念的提出在一定程度上缓解了当时语义网发展面临的困境。语义网首先是由互联网之父Tim Berners Lee提出的，语义网的的目的是向计算机提供可被机器理解的知识表达，让机器能够像人一样理解语言。

指　标
疑似剽窃文字表述

1. Fusioncharts这样一个融合多个平台，跨多个浏览器的Flash图表组件解决方案；
2. 时空数据可视化、地理空间数据可视化、高维非空间数据可视化、层次和网络数据可视化、跨媒体数据可视化等。
3. 数据一般可分为：一维标量数据、二维标量数据、三维标量数据、多变量空间数据、时间序列数据等。在对一维标量数据
4. 热力图：热力图使用颜色来表达位置相关的二位数值数据大小。这些数据常以矩阵或方格形式整齐排列，或在地图上按一定的位置关系排列，由每个数据点的颜色反映数值的大小。

4. 蒋冬冬_163520085211007_中国纪录片知识图谱构建及可视化_第4部分

总字数：9076

相似文献列表　文字复制比：2.4%(219)　疑似剽窃观点：(0)

1	知识图谱技术综述 -《学术论文联合比对库》- 2016-11-17	1.4%（124） 是否引证：否
2	知识图谱技术综述 徐增林;盛泳潘;贺丽荣;王雅芳; -《电子科技大学学报》- 2016-07-30	1.0%（91） 是否引证：否
3	Nodejs+Express4.x+mongodb简要介绍 - deguotiantang的专栏 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	1.0%（90） 是否引证：否
4	面向慢性病海量数据问答系统智能摘要算法的研究与实现 刘红霞(导师：李捷) -《河南大学硕士论文》- 2016-06-01	0.8%（75） 是否引证：否
5	Express开发框架的安装与配置 - 木鱼大叔的技术博客 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
6	2016年08月存档 - 木鱼大叔的技术博客 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
7	NodeJS - 木鱼大叔的技术博客 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
8	木鱼大叔的技术博客 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
9	Javascript - MaRain - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
10	Express入门 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
11	文章列表 - MaRain - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
12	nodejs - MaRain - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
13	2017年06月存档 - MaRain - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否
14	菜鸟学习nodejs--express(二)路由 - ermuv5 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	0.8%（74） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

但语义网只能说是一个设想，很难进行实施，所以学者们开始将注意力转向数据本身，随后就萌生了“链接数据”这个想法。链接数据的目的不仅仅是使结构化数据遍布于互联网中，还要使这些数据联系起来，形成一张数据网。

随着维基百科、百度百科这样百科类结构化信息网站的出现和链接数据这个概念的提出，大量基于百科知识的知识库建立起来，国外比较有名的例如DBpedia[10]和德国马普研究所的Yago[11]，国内比较有名的例如复旦大学的CN-pedia[12]、上海交通大学的zhishi.me[13]和清华大学的XLore[14]。

3.3.2 基于自由文本的开放域知识图谱的构建

基于百科知识的知识图谱的构建的构建方式包括人工编辑和自动抽取，但是这个自动抽取的方法大多是对维基百科、百度百科中结构化的数据进行抽取，而对非结构化的文本数据无法抽取。

在数据链接技术发展的同时，关于信息抽取的相关技术也在发展。最先提出的思路是为每个目标关系训练相应的抽取器，但是在面对大量的关系类别时为每种关系训练抽取器是不现实的，所以这种方法没有被采用。后来华盛顿大学的Banko等人提出了开放域信息抽取技术（OIE）[15]，这个技术的思想是直接从大规模非结构化文本中抽取实体关系三元组：[头实体，关系词，尾实体]。开放域信息抽取技术OIE通过直接识别关系词组来抽取实体关系。依据这个思想，华盛顿大学陆续推出了TextRunner[16]、Reverb[17]、OLLIE[18]等基于自由文本的开放域三元组抽取系统；以及卡耐基梅隆大学的NELL系统[19]、德国马普研究中心的PATTY[20]等。开放域信息抽取系统主要是对开放域实体关系的抽取。开放域信息抽取技术的发展，推进了基于自由文本的开放域知识图谱的构建工作。

3.3.3 自动构建的开放域中文实体知识图谱

上一小节所介绍的开放域信息抽取（OIE）系统侧重点和技术核心在于对开放域实体关系三元组的抽取，但是却忽略了本体库的构建。本体库是知识图谱的元数据，本体库的构建也可说是Schema的构建是构建知识图谱时不可或缺且至关重要的步骤，也是为三元组赋以语义的关键。

目前关于这方面的研究主要是《大词林》项目，该项目由哈工大的一个研究中心发起的。该项目致力于对实体进行类别划分，并对类别进行抽象和层次化，进而实现对实体上下位关系体系的自动构建，而上下位关系体系正是本体库的核心组成之一。该项目在对实体进行类别划分时，采用的是多信息源自动获取实体类别的方法，所以并不需要领域专家的参与，这种自动化的手段使得数据规模可以不断扩大，突破了人工构建的限制。

3.4 构建知识图谱的关键技术

3.4.1 实体及关系抽取

实体及实体关系的抽取也叫信息的抽取按照实体和关系的抽取顺序可以分为两种：一种是流水线式抽取方法：先抽取实体，然后再对识别出来的实体进行关系分类。另一种是联合学习式的抽取方法：实体抽取和关系抽取同时进行，而且这两个过程还互相促进。两者之间的区别如图3-2所示：

图3-2 流水线抽取和联合式抽取过程

3.4.1.1 流水线式抽取

流水线式信息抽取的思想是将实体抽取和关系抽取这两个步骤分开、独立来进行，先进行命名实体识别，然后对识别出来的实体两两结合，再进行关系分类，最后形成实体关系三元组。流水线式抽取的实体抽取和关系分类这两个子任务是相互独立的。

（1）实体抽取

实体抽取又可以称为：命名实体识别（NER）。对实体进行抽取一般采取两种措施：第一种是通过建立同义词库来进行识别，另外一种是通过语法知识来识别实体。对于第一种方案来说，要特别避免这样的问题：例如“陈冠希望着天空”，应该识别成“陈冠希”、“望着”、“天空”，而不应该解析成“陈冠”，“希望”，这就要解决好词语之间相互覆盖问题。对于第二种方案来说，要想通过语法特征提取实体，那么就要有大量的训练样本。

关于实体抽取的实现，在国外两大主流的是宾夕法尼亚大学实现的基于NLTK的命名实体识别和斯坦福大学实现的基于Standford的命名实体识别。其中基于NLTK的命名实体识别是用python实现的，斯坦福大学是用java进行实现的。在国内有复旦大学的fudanNLP、中科院的NLPIR分词系统、哈工大的LTP。其中哈工大的提供python接口。另外随着深度学习技术的发展，神经网络结构在NER中的应用也取得了不错的效果。

（2）关系抽取

在对关系进行抽取时，提出的方法有很多，大致可以分为：有监督的、半监督的、无监督这三类方法。其实这种分类非常类似于机器学习种算法的分类。

有监督的关系抽取，其实是借鉴机器学习中有监督的分类算法，它将关系抽取问题当做分类问题，通过训练大量的训练集得到相应的分类模型，然后通过该模型对关系进行预测。但是该方法需要大量的人力进行人工标注工作从而获得我们所需的分类模型，所以该方法太耗时。之后就有人提出了远程监督[21]的思想来解决这个问题。

远程监督的思想主要是用一种匹配的方式来抽取关系，主要思想如下：对于一段文本，如果包含现有的三元组<实体1，关系，实体2>中的实体1和实体2，那么该文本表达的关系就是该三元组中的关系。这样就会省去大量的人工，加快我们的工作，但是同时，这种操作会带来误差，从而就制造了大量的噪音，所以后来出现了很多对该技术进行改进的方法，例如：北京大学提出的利用噪音矩阵来拟合噪音[22]，从而达到拟合真实分布的方法；哈工大提出的将高效深度记忆网络应用于远程监督的方法[23]等。

半监督关系抽取主要是采用 BootStrapping[24]方法，这个方法的核心思想就是不断的抽样、更新，这样一次次的迭代使得我们的关系库更加的完善，在对关系进行抽取时会更加的准确。当然，半监督的方式之所以叫“半”监督，是因为这个方法也需要一定的人工手段去设定种子实例，并不是全自动的过程。

无监督的关系抽取基本上是完全自动的，是建立在上下文信息的基础上的，该方法的主要思想是：拥有相似上下文关系的实体对之间的关系也是相同的。在此基础上，将深度学习中的神经网络[25]等方法应用进去将会收到意想不到的效果。

（3）缺点

流水线式的信息抽取方式存在一些显而易见的缺点：第一，实体识别中产生的错误会影响到关系抽取的正确率，这也是所有流水线式作业都存在的问题；第二，实体识别和关系抽取是两个相互关联的，相互加强的任务，流水线式抽取分割了两者存在的联系，所以没有充分利用这种紧密的联系进行抽取。例如如果存在“国家主席”关系，那么我们可以知道前一个实体必然属于“位置”类型，后一个实体属于“人”类型，流水线式的抽取就没有办法利用这样的信息；第三，流水线式抽取产生了大量冗余信息。实体抽取后要两两配对进行关系抽取，但是有些实体对之间是没有关系的，这样就产生了大量冗余信息。

3.4.1.2 联合学习式抽取

联合学习式抽取是将实体抽取和关系抽取同时进行,直接得到有关系的实体三元组的一种操作，这样就巧妙地避免了流水线式抽取所产生的一些弊端。其实这种联合学习，就是将两个相互独立的任务通过某种手段联系在一起，也不是完全的没有先后顺序的。

本文主要介绍基于神经网络算法的联合学习方法。用神经网络算法对实体和关系进行联合抽取时可采用两种方案，一种是通过共享参数的方式，一种是通过使用标注的方法。

（1）参数共享

利用参数实现联合学习其实就是在实体抽取和关系分类两个子任务之间通过参数共享实现两个任务的交互。在论文《Joint Entity and Relation Extraction Based on A Hybrid Neural Network》[26]和论文《End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures》[27]中提出来的思想都是通过参数共享来进行联合学习。两篇论文虽然都是利用参数共享来进行联合学习，但是具体实现还是存在差别的。

在论文Joint Entity and Relation Extraction Based on A Hybrid Neural Network》中的具体实现如图3-3所示：

图3-3 联合抽取

从图中可以看出：对于输入的句子，首先通过词向量层，将词语转化为向量表示，然后先后经过向前的和向后的LSTM（长短期记忆网络）层，之后形成合并。最后，在合并层的基础上，我们用一个LSTM进行解码，从而进行实体识别，在对关系进行抽取时，首先要根据实体识别预测的结果对实体进行配对，然后再用一个CNN（卷积神经网络）对实体之间的文本进行分类，从而实现关系抽取。该模型中，在训练两个子任务的过程中都会向后进行传播，从而可以用来更新共享的参数并以此来实现两个子任务的联系。

在第二篇论文中，作者的具体实现思路是：在进行命名实体识别时，使用的是一个神经网络进行解码，在进行关系分类时，加入了依存信息，根据依存树最短路径，使用一个 BiLSTM 来进行关系分类。

（2）标注策略

在论文《Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme》[28]中提出了一种新联合抽取实体和关系的方法，该方法主要是利用标注策略。该方法将原来涉及到序列标注任务和分类任务的关系抽取完全变成了一个序列标注问题，然后通过一个端对端的神经网络模型直接得到关系实体三元组。

该方法提出的标注策略由三部分组成：

1）词位置信息：B表示该词在实体的开始位置；I表示该词在实体的内部；E表示该词在实体的结尾；S表示该实体只有一个词。

2）实体之间关系的类型信息：这些关系类型是预先定义好的，并且这些关系类型都要经过编码且记录起来的。

3）实体在三元组中的位置信息：1表示头实体；2表示尾实体；O表示不是实体关系三元组内的词。

对句子进行标注之后，可以推断，具有相同关系的实体是一组，可以组成一个三元组，但是当句子中含有多个实体具有相同的关系时，这个时候一般可以断定，在句子中距离较短的两个实体组成一个三元组的可能比较大，所以一般采用就近原则。

3.4.2 动态知识库的构建

随着知识图谱技术的发展，知识图谱正在越来越多的应用中扮演重要的角色，但是现有的知识图谱存在一个很明显的缺陷：即现在知识图谱中数据的实时性很差，很多知识图谱在构建完之后就很少对数据进行更新，即使有些知识图谱对数据进行了更新，更新的周期也相当长，因为每次更新都特别耗费精力、时间，相当于重新构建了一次知识图谱。实时性差导致知识图谱中很多数据没有同步更新，从而存在大量的错误，使得这些知识无法被利用，给知识图谱的应用带来了很大的局限性。

现在有些学者致力于动态补全知识图谱和实时更新知识图谱的研究。对知识图谱进行补全可以大致分为两类：一类是Closed-World知识图谱补全，这类知识补全无法处理从知识图谱外部加入的新实体。另一类叫做Open-World知识图谱补全，这种知识补全可以处理知识图谱外的实体，并将其链接到知识图谱中。

在论文《How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source》[29]中提出了一个实时更新知识图谱数据的方法框架，可以以较高的准确率预测出哪些实体需要被更新，从而可以频繁地对知识图谱进行更新而不用担心耗时耗力。该论文对知识图谱的更新思想及步骤如下：

（1）从互联网上抽取、识别出最近一段时间内热门的实体；

（2）根据步骤（1）抽取出的热门实体，对知识库做出更新；

（3）找出与前两步中更新的实体相关的实体作为候选项，以便后续的更新。

（4）对第（3）步中找出的相关实体按照优先级进行排序，然后根据序号将其更新到知识库中。

在论文《Open-World Knowledge Graph Completion.》[30]中，作者提出了一个ConMask 模型，并利用该模型对知识库进

行动态更新。这个模型的一大特点就是可以将新的实体加入到知识库中。该模型的主要思想是通过计算相似度来引入新的实体。

3.4.3 实体对齐

随着知识图谱技术的发展，越来越多的知识图谱项目被发布，为了形成更加强大的知识图谱，我们通常需要整合一些知识图谱项目，但是每个项目中的实体信息并不是相互一致的。所以为了更加方便的获取更多的知识，综合使用多个来源的知识图谱。要对多个知识图谱进行实体对齐。

近年来，关于实体对齐的研究并不算火热，但是也有一些学者对这方面进行了研究，并提出了一些自动对齐实体的方法，但是这些方法的对齐质量并不是特别高。所以有些学者开始致力于半自动化实体对齐的研究。

在论文《A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases》[31]中，作者提出了一种利用半自动化的手段对来自于不同知识图谱的实体进行对齐的方法，这里的半自动化就是人工和机器学习相结合的手段。在该论文中，作者提出首先通过机器学习的手段粗略地进行实体对齐，然后再对已经对齐和未对齐的实体进行人工判断，在进行人工判断时，作者也提出了严格的步骤以避免人工判断导致的错误。人工判断的步骤主要包括实体集划分、建立偏序、问题选择、容错处理这四个部分。

3.4.4 其他相关技术

关于知识图谱构建还有很多其他延伸的技术，例如知识图谱补全、知识图谱去燥和知识推理等。像知识图谱补全和知识图谱去燥都可以用知识推理进行实现。

3.5 知识图谱的架构

一般谈论到知识图谱的架构不免会提到两个方面：知识图谱自身的逻辑结构以及构建知识图谱所采用的技术体系架构。

3.5.1 知识图谱的逻辑结构

知识图谱自身的逻辑结构又可分为模式层与数据层两个层次。其中数据层一般存储的是（实体1，关系，实体2）或（实体，属性，属性值）这样的三元组结构，这些三元组结构表达的是一系列的事实。因为这样的数据结构比较适合存储在图数据库中，所以可以选择合适的图数据库作为相应的存储介质，当然也可根据实际情况选择合适的存储介质，并不一定要用图数据库。目前比较流行的图数据库有：Neo4j、FlockDB、GraphDB等。知识图谱逻辑结构的模式层是知识图谱构建的核心，而模式层和数据层的关系是非常紧密且相辅相成的，没有数据层的支撑就没法建立模式层。在知识图谱领域，知识图谱的模式层通常指的就是本体库。本体是概念的集合，一般不会改变如“人”、“事”、“物”、“地”、“组织”。本体库是本体的集合，是知识图谱的“骨架”结构。

3.5.2 知识图谱的技术体系架构

在上一小章节中介绍了知识图谱的逻辑结构，在这一章节中将详细介绍知识图谱构建的技术体系结构。知识图谱的技术体系架构指的是知识图谱构建的模式结构，如图3-4所示。

图3-4 知识图谱技术体系架构图

从图中可以看到知识图谱构建的流程有四部分：数据获取、信息获取、知识融合、知识处理。数据获取部分，因为数据来源多种多样，所以获取的数据的结构也不一样，通常可分为：结构化、半结构化、非结构化数据。数据获取完毕之后就要对数据进行处理，也就是从数据中抽取想要的信息，这部分称为信息获取，主要包括实体抽取、关系抽取两部分。对信息进行获取之后，就要对相似的信息进行融合，避免出现大量冗余、重复的信息，这部分称之为知识融合。最后就是对获取且融合过的知识进行处理的过程，包括质量的评估和本体的抽取等。

在构建知识图谱的过程中有两种构建方式，这两种构建方式的最主要区别在于本体库和知识库的构建顺序。第一种是先构建本体库，再在本体库的基础上加入实体构成知识库，这种方案被称为自顶向下的构建方法，例如Freebase项目就是采用这种方式。另外一种是先将尸体加入到知识库中，然后在知识库的基础上构建本体库，这种方案被称为自底向上的构建方法。大多数知识图谱都采用自底向上的方式进行构建，其中最典型就是Google的Knowledge Vault和微软的Satori知识库。

3.6 本章小结

本章从知识图谱的简介开始，首先简洁明了地介绍了知识图谱这个概念，其次，为了加深我们对知识图谱这个技术栈的了解，又分别详细解释了实体、概念、本体、本体库、知识库、知识表示、链接数据这几个知识图谱中常用的术语。然后，详细介绍了基于百科知识的知识图谱、基于自由文本的开放域知识图语、自动构建的开放域中文实体知识图谱这三类知识图谱。另外，本章还介绍了构建知识图谱的所会涉及到的关键技术，如：实体抽取、关系抽取、知识图谱更新、实体对齐。最后从知识图谱的逻辑结构和技术架构介绍了知识图谱的一般架构。

4 系统需求分析及设计

4.1 系统功能模块

本系统主要包含两大功能：一个是可视化部分，一个是知识图谱部分。具体的功能如图4-1所示：

图4-1 系统功能结构图

可视化部分主要是对三大视频网站的纪录片数据进行抓取、分析、展示。展示部分主要从三个角度进行展示：一个是展示三大视频网站播放量排名前十的纪录片，另一个是展示播放量排名前十的纪录片的类型占比，最后主要展示了播放排名前十的记录片的评分情况。

在可视化部分我们可以了解到当前比较热播的纪录片有哪些，也可以清楚的知道各个视频网站中热播纪录片的差别及播放

量。同时也可以了解到受欢迎的纪录片中各个类型的占比，最后也可以看到播放量和评分的关系。

知识图谱功能部分主要是对关键信息进行搜索，然后对搜索结果进行可视化展示。搜索将会分为两种类型：一个是对纪录片的搜索，一种是对关键词的搜索。对纪录片进行搜索，所展示的结果将是对该纪录片所构建的知识图谱的可视化展示；对关键词进行搜索，所展示的结果将是该关键词在哪些纪录片中出现过，且在这些纪录片的什么时刻出现的，然后会有一个跳转链接，这个链接将链接到在这些纪录片中该关键词出现的时刻进行播放。在对知识进行可视化时，常见形式有：概念层次、思维导图、认知地图、语义网络。本系统对知识图谱的展示采用的是语义网络的形式。

另外，为了提供方便，知识图谱部分也将支持人工对信息进行录入。因为对某些纪录片进行切割分析时，不能准确获取某个实体在该纪录片中出现的具体时刻，所以需要进行人工录入。

4.2 系统开发技术及环境

本文所描述的系统是在Mac Pro笔记本的IOS10.12.6的操作系统下进行实现的。其中用到了python、js、html、css等语言，还用到了JetBrains公司研发的WebStorm、PyCharm这两个编辑器。

后台获取数据及一些对数据进行处理的部分使用的是python语言进行编写，其中用到了BeautifulSoup这个用来解析网页并从网页中抓取数据的库。而且最后将获取到的数据存储在JSON文件中。

后台的服务器是用Node编写的轻量级服务器，因为近年来Node技术比较火热，而且是用js进行开发的，所以本文实试验性地用该技术搭建了服务器。其中用到了Express开发框架和Cors模块。Express 是一个基于 Node.js 平台的极简、灵活的 web 应用开发框架，它提供一系列强大的功能，可帮助创建各种 Web 和移动设备应用。Cors模块是一个遵循W3C标准且用来解决跨域资源访问的一个Node模块。

前端部分用的是Facebook研发的React框架，React引入了组件化和虚拟DOM的思想，是前端界的一大重要突破。同时前端还用到了Webpack这个前端打包工具，它可以分析项目结构，自动找到项目中的JS模块进行打包，并且能够将一些浏览器不能识别的扩展语言解析成浏览器可运行的语言。除此之外也使用了Ant Design这个由React封装形成的组件库，这个库是由蚂蚁金服体验技术部经过大量的项目实践和总结，沉淀出来的。在对数据进行展示时，用到了百度的数据可视化团队研发的基于Canvas的纯JS开发的图表库Echarts。该库提供直观，生动，可交互，可个性化定制的数据可视化图表。

4.3 接口API设计

本文前后端进行交互的数据接口及数据格式说明如表4-1所示：

表4-1 接口对照表

所属功能模块接口名称接口描述数据格式

可视化模块 /getTop10 获得播放量排名前十的纪录片名称及播放量 eg:{iqiyi:{xaxis:[“舌尖上的中国”，“万物滋养”]，yaxis:[391.5,78.9]}}xaxis:表示纪录片名称；yaxis:表示纪录片播放量。

/getTypeCount 获得播放量排名前十的纪录片的各个类型的占比 eg:{types:[“军事”，“文化”]，data:[{value:“7”,name:“军事”}，{value:“4”,name:“文化”}]}

/getScore 获得播放量排名前十的纪录片的评分 eg：{iqiyi：{xaxis：[“舌尖上的中国”，“万物滋养”]，yaxis：[7.8,8.9]}}xaxis：表示纪录片名称；yaxis：表示纪录片评分。

知识图谱模块 /getDocKnow 获取所搜索的纪录片的知识图谱的数据（实体及关系） eg:{"纪录片": [{"relation":"包含种类","data":[{"value":"探索类纪录片","data":[{"relation":"所含纪录片","data":[{"value":"奇闻天下"}]}]}]}]}

/getKeyCotent 获得所搜索关键字在纪录片中的具体信息 eg:[{key:“乔布斯”，item：“互联网时代E02”，time：“6300”}]

/additem 提交人工录入的信息 {key:“乔布斯”，item：“互联网时代E02”，time：“6300”}

上表中，“iqiyi”：表示爱奇艺网站上的数据，“yk”：表示优酷网站上的数据，“tecent”：表示腾讯网站上的数据。

指　标
疑似剽窃文字表述

1. 大多数知识图谱都采用自底向上的方式进行构建，其中最典型就是Google的Knowledge Va
2. ress开发框架和Cors模块。Express 是一个基于 Node.js 平台的极简、灵活的 web 应用开发框架，它提供一系列强大的功能，可帮助创建各种 Web 和移动设备应用。

5. 蒋冬冬_1635200852110007_中国纪录片知识图谱构建及可视化_第5部分	总字数：10500
相似文献列表　文字复制比：16.6%(1741)　疑似剽窃观点：(0)	
1 python命名实体抽取学习记录（1） - lalalawxt的博客 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	10.6%（1117） 是否引证：否
2 中文短文本主题分类方法研究 李洪图(导师：刘晓霞) -《西北大学硕士论文》- 2014-06-01	2.0%（213） 是否引证：否
3 社区问答中文问句分类的迁移学习方法研究 杨彬(导师：苏磊) -《昆明理工大学硕士论文》- 2015-03-01	1.9%（199） 是否引证：否
4 不要光顾着吃，品味一下《舌尖2》智慧台词！--相关文章	1.8%（185）

	- 《互联网文档资源（http://www.360doc.co）》- 2014	是否引证：否
5	揭开吃出健康和美丽的秘密：饮食美容专家谈--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2012	是否引证：否
6	舌尖上的故乡 100座城市和它的美食--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2014	是否引证：否
7	视频 中国美食--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2013	是否引证：否
8	环球时报社评：西方一些人对华搞小挑衅，很不自尊--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2012	是否引证：否
9	美食背后 人生百味--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2012	是否引证：否
10	【舌尖上的陕西】完结篇 无尽的智慧--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2014	是否引证：否
11	舌尖上的美食--相关文章	1.8%（185）
	- 《互联网文档资源（http://www.360doc.co）》- 2013	是否引证：否
12	《舌尖上的中国》是这样拍成的--相关文章	1.8%（184）
	- 《互联网文档资源（http://www.360doc.co）》- 2012	是否引证：否
13	舌尖上的中国 A bite of China 720P高清晰版下载（已更新至第七集）_地产智库	1.7%（175）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
14	《舌尖上的中国》热播受捧 展示美食背后的中国_资讯中心	1.7%（175）
	- 《网络（http://www.china.com）》- 2012	是否引证：否
15	中央电视台为吃货们出节目啦 《舌尖上的中国》宣扬中华饮食文化_最好的影视栏目制作	1.7%（175）
	- 《网络（http://blog.sina.com）》- 2014	是否引证：否
16	《舌尖上的中国》亮相央视综合频道(1)_幕后英雄	1.7%（175）
	- 《网络（http://topics.gmw.cn）》- 2017	是否引证：否
17	左所村人	1.7%（174）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
18	敬请观看——【舌尖上的中国】_尊道贵德	1.7%（174）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
19	[转载]7集美食纪录片《舌尖上的中国》 在线观看_王美娥	1.7%（174）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
20	“舌尖上的母校”勾起大学生别样食堂记忆(--相关文章	1.7%（174）
	- 《互联网文档资源（http://www.360doc.co）》- 2012	是否引证：否
21	【蔓萝饮食】 舌尖上的中国---名菜制作大全（60集视频 ）--相关文章	1.6%（171）
	- 《互联网文档资源（http://www.360doc.co）》- 2016	是否引证：否
22	中文文本蕴涵识别技术研究	1.6%（167）
	姚东任(导师：张志昌) - 《西北师范大学硕士论文》- 2016-05-01	是否引证：否
23	向大家推荐一部好片子《舌尖上的中国》_梅姐	1.6%（164）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
24	清风朗月	1.6%（164）
	- 《网络（http://blog.sina.com）》- 2013	是否引证：否
25	《舌尖上的中国》热播 引发网友争议 健美诚品315保健品商城 【www.jmcp315.com】推荐_健美诚品315商城官网	1.5%（162）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
26	餐饮经理人必看 我们的田野-舌尖上的中国_晨丞	1.5%（161）
	- 《网络（http://blog.sina.com）》- 2012	是否引证：否
27	清和日记_闲木草堂	1.5%（161）
	- 《网络（http://blog.sina.com）》- 2013	是否引证：否
28	网友热议《舌尖上的中国》	1.5%（161）
	- 《互联网文档资源（http://www.360doc.co）》- 2016	是否引证：否
29	电视系列片：舌尖上的中国 1——7集全--相关文章	1.5%（158）
	- 《互联网文档资源（http://www.360doc.co）》- 2012	是否引证：否
30	2016年11月存档 - iCoding91 - 博客频道 - CSDN.NET	1.0%（100）

4.4 本章小结

本章第一小节介绍了系统包含的主要功能模块：可视化模块和知识图谱模块。第二小节简要介绍了本系统开发时所用的环境和所涉及到的技术。第三小节详细说明了前后台交互所要用到的接口及接口数据的格式。

5 可视化部分的实现

本文对纪录片数据可视化部分的实现的主要流程是：数据获取、数据预处理、数据展示。技术上后台采用python编写，将要用的数据爬取并存储在文件中，因为最近node作为比较新兴的技术，可以用来搭建轻量级的服务器，所以实现的时候尝试着用node作为服务器写的API接口。

5.1 数据获取

本文获取数据的途径是通过编写网络爬虫，从网页上动态获取数据。数据来源于国内三大视频网站：优酷、爱奇艺、腾讯。纪录片数据的详细信息包括：纪录片排名、纪录片名称、纪录片简介、纪录片类型、纪录片评论数量、纪录片评分、纪录片播放量、纪录片详细页链接。数据结构如表5-1所示：

表5-1 源数据结构表

字段中文名字段英文名示例

纪录片排名 no 1

纪录片名称 name 舌尖上的中国第一季

纪录片简介 summary 在以往的影像素材里，中国美食更多以“烹饪大师”或“美食名家”结构，展现的是“精湛的厨艺”和“繁复的过程”。在本片中，中国美食更多的将以轻松快捷的叙述节奏和精巧细腻的画面，向观众，尤其是海外观众展示中国的日常饮食流变，中国人在饮食中积累的丰富经验，千差万别的饮食习惯和独特的味觉审美，以及上升到生存智慧层面的东方生活价值观。

纪录片类型 type 文化

纪录片评论数量 numOfComments 661

纪录片评分 score 8.1

纪录片播放量 vv 1855.1万

记录片详细页链接 link http://www.iqiyi.com/a_19rrgzvrlx.html

人类只有一双眼睛和一双手，所以收集和处理信息的效率是有限的，而所要获取的信息是大量的，依靠人去搜集过滤这些信息是不现实的，而且也会受主观因素或客观因素的影响而得到错误的信息。但是这些繁杂重复的工作让计算机来进行却是再合适不过了，这样能够极大地提高效率。通过编写爬虫程序，然后在后台运行该程序，能够自动获取给定网页的内容并按照指定的规则进行分析。

由于爬取的数据的数量并不是非常巨大，所以对爬虫运行效率的要求并不是特别高，最终选择用python对网络爬虫进行实现，目前python是编写爬虫的后台脚本类语言中应用最多最广的一种语言。期间，还用到了python的一种库：Beautiful Soup。Beautiful Soup是专门用来从文档中抓取数据的，它能够将文档转化为一种便于我们搜索、查找、修改的导航模式，能够帮助我们节省很多时间，大大提高了我们的工作效率。用于实现爬虫的、非常优秀的python库有很多，例如：scrapy、pyquery、Mechanize，这些框架都有各自的优缺点，本文之所以选择Beautiful Soup，是因为它非常方便且容易上手。另外，在抓取数据时要区分静态页面上的数据和动态获取的数据，有针对性地应用相应的方法来获取数据。

本文数据来源于三大视频网站，由于每个网站的网页结构都不一样，所以爬取数据时，要充分了解每个网页的组织结构，从而根据自己的需求来制定相应爬取规则。在实现时，每个网站的爬取过程及爬取规则定义在相应的函数中，最后形成三个函数：getYkData、getIqiyiData、getTencentData。获取的数据最后以JSON的数据格式存储在对应的JSON文件中。

5.2 数据预处理

数据预处理，从字面意思就可以看出是对数据提前进行的一些操作过程。因为数据一般具有杂乱性、重复性、和不完整性，表现为：数据不完整、不一致、有异常，所以要对数据进行处理，以保证数据的干净、准确和简洁。

在数据挖掘领域，数据预处理是数据处理的过程中必不可少且最重要的一个步骤，输入数据的质量很大程度上决定着输出信息的质量。但是在本文中，数据预处理的概念与数据挖掘中定义的数据预处理的概念既相同却又有不同之处。

在数据挖掘中，数据预处理的目的是为了提高数据挖掘算法的执行效率，而且处理的一般是相当大数量的数据，所以处理过程也相应的会非常复杂，步骤也也比较繁杂。在本文中，数据预处理是为了直观的展示这些数据，至于得到怎样的信息是由用户决定的。不同用户看到同一图表得出的结论可能不是一样的。另外本文中需要预处理的数据相对没有那么多，而且数据相对来说比较简洁规整一些，所以预处理的过程也相对容易一些。

但是，本文数据预处理的过程和数据挖掘中预处理的过程还是有相同之处的。在数据挖掘领域，完善的数据与处理过程一般应该包括：数据集成、数据清洗、数据变换、数据简化。而且每个过程所要用到的方法技术理论现在也比较成熟。现在相关的工具也已经出现，例如Uber一开始的ETL，airbnb的airflow。本文中处理数据的步骤大致也是这些过程，但是并没有使用相关的工具。因为工具的学习需要一些时间成本，而且，工具的功能也是有限的，有一些特别的需求还是需要自己实现，还是‘定制’比较完美。

本文数据预处理过程如下：

（1）数据集成。在本文中，对数据进行集成是在爬取数据的过程中就进行处理实现的。由于对每个网站爬取到的数据的数据结构是不一致的，所以要对这些数据进行处理，然后进行合并。这个过程涉及到数据的选择、数据的冲突问题的解决以及对不一致数据的处理。例如：播放量这个属性，有的数据显示是‘354万次播放‘，有的显示’354万‘，有的显示‘35400’，针对这些问题都要进行处理，使其单位一致；纪录片名称这个属性，有的有别名，但是虽然名字不一样，但却是同一部纪录片，这些都要进行处理。

（2）数据清洗。数据清洗主要是一个去掉噪音数据，弥补缺失数据，转换数据类型的一个过程。在本文中，因为编写爬虫时就已经将一些无关紧要的数据给丢弃掉了，所以爬取到的数据噪音污染程度比较小。对于缺失的数据，因为量比较小，所以选择人工填写遗漏值。最主要的便是转换数据类型这个步骤，因为播放量这个数据单元是非常大的数值，所以对其进行离散化转化。

（3）数据变换。数据变换最常用的手段有：变量派生、变量转换、分箱转换、数据标准化。本文数据比较直观、简洁，所以并没有涉及这一步骤。

对于数据进行处理的操作都是在后台通过Python进行的。

5.3 数据展示

数据可视化的展示主要从播放量排名前十的纪录片的播放量、播放量前十的纪录片的类型占比、播放量前十的纪录片的评分情况这三个角度进行展示。

（1）播放量Top10的纪录片的播放量展示

对播放量前十的纪录片的播放量的展示如图5-1所示：

图5-1 播放量TOP10展示图

这个图是前端实现部分的截图，本系统的前端是用React实现的，同时使用了React框架中比较受欢迎的antdesign这个UI框架，图表展示是用的echarts进行展示的。

从图中可看到，本文将三个视频网站的数据放在不同的tab中进行展示，这样比较方便用户进行比较、发现对用户来说比较有用信息。其中数据处理部分是在服务器端用nodejs对爬取的数据按照播放量进行排序，然后截取播放量排名前十的数据进行展示。判断一个纪录片的好坏的标准有很多，例如：评分、评论数量、播放量等，本文之所以选取播放量作为评判一个纪录片好坏的标准，是因为播放量更能体现一个纪录片的受欢迎程度。

前端请求数据采用的接口是：“/getTop10”，数据格式如图5-2所示：

图5-2 接口‘/getTop10’数据示例

属性“iqiyi”显示的是爱奇艺网站相关的数据，其中“xaxis”显示的是排名前10的纪录片的名称，“yaxis”显示的是排名前10的纪录片的播放量。

对这部分数据进行可视化展示的意义是：因为爬虫爬取的数据是实时的，所以可以及时了解到当前市场上纪录片的现状，比如：哪些纪录片最受欢迎。因为各个视频网站中播放量靠前的纪录片不一样，所以对三大视频网站的数据进行展示是有必要的。

（2）排名前十的纪录片的类型占比展示

排名前十的纪录片的类型占比图如图5-3所示：

图5-3 排名TOP10类型占比示例图

类型占比相关数据，也是在服务器端处理完之后直接传给前台的。类型占比的数据是根据排名前10的纪录片的数据计算的。因为有些记录片不能明确判定属于哪一类型，所以可能同属于两个或多个类型，在处理这种数据时，会将其所属的类型都计算上。

前端所用的数据请求接口是“/getTypeCount”，数据格式如图5-4所示：

图5-4 接口‘/getTypeCount’数据实例

其中属性“types”保存的是所有的类型数据，属性“data”是一个数组，是对象的集合，保存的是每个类型的详细数据。

对这部分数据进行展示的意义是：可以直观的了解到目前播放量靠前的纪录片中哪些类型占比较多，就可以了解到人们更喜欢哪类纪录片。

（3）排名前十的纪录片的评分情况展示

播放量排名前十的纪录片的评分情况如图5-5所示：

图5-5 排名TOP10纪录片评分情况示例图

这部分数据前端采用的请求接口为“/getScore”，返回的数据结构及说明同接口” / getTop10“。

在大部分人的观念中，评分和播放量是有关的，即使可能关联度不是很大，但是起码也是成正比的关系，但是从图中可以

看出，评分和播放量并不是成正比的关系，这个图的意义就是能够给用户提供数据信息，帮助客户更加清晰的分析出什么样的纪录片才能无论是在播放量还是评分上都能取得好的效果。

5.4 本章小结

本章节对数据可视化展示部分的流程、步骤做了详细介绍，包括：数据获取、数据预处理、数据展示。同时也举例说明了前后端数据交互接口的数据格式，最后也对本系统所实现的可视化的成果做了展示。

6 知识图谱部分的实现

因为本文所构建的知识图谱相对来说处于初级阶段，还未考虑到动态知识库、实体对齐等方面的问题，只是简单的对纪录片进行图谱的构建，所以本文所涉及的知识图谱相关的关键技术主要是本体构建、语义标注和实体的提取。

6.1 按照纪录片进行搜索

6.1.1 本体构建

本体构建一般采用半自动化的手段，这样才能确定本体构建的正确性。本文中关于中国纪录片的本体并没有很多概念，所以完全采用人工的手段。

关于中国纪录片的本体构建如图6-1所示：

图6-1 纪录片本体构建

关于这个本体的构造可能不太全面，在以后的过程可以添加更多的关系，在数据接口设计的时候也给这些关系的加入留有余地，所以在进行扩充时只需按照已有的结构添加数据就行，并不需要改变数据结构。

6.1.2 语义标注

对本文来说，语义标注的内容主要是纪录片所属类别的标注，对此可以使用机器学习相关的算法进行标注，也可以人工进行标注。在文中，因为所获取的数据已经带有类别标签，所以该步骤算是自动标注好的。

6.1.3 实体抽取

实体抽取又叫命名实体识别，最开始是由国外提出的，而且国外在这方面也取得了一定成果，例如宾夕法尼亚大学实现的基于NLTK的命名实体识别和斯坦福大学实现的基于Standford的命名实体识别，但是中文和英文的差异决定了处理不同的语言要采取的方法不同，而本文要处理的文本是中文，所以要用本土的语言处理工具。目前国内有很多开源的中文语言处理工具可以实现命名实体识别，比如复旦大学研发的fudanNLP，中科院研发的NLPIR分词系统（又名ICTCLAS2013）和哈工大的LTP。但是在研究中发现，哈工大的LTP在分词、实体识别等方面的效果甚至要优于中科院ICTCLAS，而且LTP还具备了目前在中文信息处理领域较为罕见的语义角色标注（SRL）功能，还提供python接口，所以本文就直接用python调用其封装成的pyltp的模块实现命名实体识别。本文中实体获取主要是对某个纪录片所介绍的实体进行提取。

pyltp命名实体识别过程如下：

1）获取要分析的文本（在本文中就是纪录片的解说词）：打开要获取的文本，然后通过一个循环语句，逐行读入文本并编码，最后append到变量‘text’中。相关代码展示如下：

```
news_files = codecs.open('./text/text1.txt', 'r', encoding='utf8')
 news_list = news_files.readlines()
text = ''
 for i in range(0,news_list.__len__()):
 text = text + news_list[i].encode('utf-8')
```

2）进行分句：对获得的文本进行分句。LTP进行分句的原理是：根据中文标点里的句号、问号、感叹号、分号、省略号进行分句。pyltp提供的分句的接口为：SentenceSplitter.split()函数，在此对其进行了又一层封装，使其结果返回一个列表。相关代码如下：

```
def sentence_splitter(sentence):
 sents = SentenceSplitter.split(sentence)
 sents_list = list(sents)
 return sents_list
sents = sentence_splitter(text)
```

3）进行分词：循环地对每个句子进行分词。在LTP中，将分词任务建模为基于字的序列标注问题。对于输入句子的字序列，模型给句子中的每个字标注一个标识词边界的标记，以此来实现分词。pyltp提供的分句的接口为：Segmentor.segment()。相关代码如下：

```
def segmentor(sentence):
 segmentor = Segmentor()
 segmentor.load('./ltp_data_v3.4.0/cws.model')
 words = segmentor.segment(sentence)
 word_list = list(words)
 segmentor.release()
 return word_list
```

该程序对分句功能进行了封装，首先创建了一个Segmentor实例，然后用其加载模型，本文用的是ltp_data_v3.4.0的cws.model模型。其次调用接口进行分词，并将结果列表化。

4）进行词性标注：循环对句子进行分词的同时对分好的词进行词性标注。在LTP中词性标注与分词模块相同，也是将词性标注任务建模为基于词的序列标注问题。对于输入句子的词序列，模型给句子中的每个词标注一个标识词边界的标记，以此来实现词性标记。pyltp提供的分句的接口为：Postagger. postag()。相关代码如下：

```
def posttagger(words):
 postagger = Postagger()
 postagger.load('./ltp_data_v3.4.0/pos.model')
 posttags = postagger.postag(words)
 postags = list(posttags)
 postagger.release()
 return postags
```

该程序对词性标注功能进行了封装，首先创建一个Postagger实例，然后用其加载词性标注所用的模型，本文用的是ltp_data_v3.4.0的pos.model模型，其次调用接口进行词性标注，并将结果列表化。

5）进行命名实体识别：在对词语的词性进行标注之后进行命名实体的识别。在LTP中命名实体识别与分词模块也相同，将命名实体识别建模为基于词的序列标注问题。对于输入句子的词序列，模型给句子中的每个词标注一个标识命名实体边界和实体类别的标记。在LTP中，我们支持人名、地名、机构名三类命名实体的识别。pyltp提供的分句的接口为：NamedEntityRecognizer. recognize()。相关代码如下：

```
def ner(words, postags):
 recognizer = NamedEntityRecognizer()
 recognizer.load('./ltp_data_v3.4.0/ner.model')
 netags = recognizer.recognize(words, postags)
 recognizer.release()
 nerttags = list(netags)
 return nerttags
```

该程序对命名实体识别功能进行了封装，首先创建了一个NamedEntityRecognizer实例，然后用该实例加载相应的模型，本文中命名实体识别所用的模型是ltp_data_v3.4.0的ner.model模型，其次调用NamedEntityRecognizer.recognize()接口进行实体识别，并将结果列表化。

6）对命名实体进行提取：在成功进行命名实体识别之后，因为结果中保存的是所有的词，所以要把文本中的命名实体识别给提取出来，这就需要把组成命名实体的词给单个提取出来且根据标记连接短语词组。在这里，主要是提取命名实体三大类：组织、人名和地名。我主要利用正则表达式把命名实体的每个词（词的形式为：词语/命名实体标注）给提取出来。

6.1.2 结果展示

因为本文中涉及的关系类型都是确定的，而且所涉及的关系类型较少，所以对于关系的提取这一步骤就进行了省略。在以后随着系统规模的扩大，会考虑添加更多的关系类型。在对结果展示时，因为在展示时最多将展示三层的关系，而搜索的内容所在的层次不一样，所以下面将根据不同的搜索层次来对结果进行展示。

1）当搜索“纪录片”时，结果将展示纪录片的知识图谱，示例如图6-2所示：

图6-2 搜索“纪录片”时结果展示

由图可知，当搜索纪录片时，将会展示围绕纪录片的三层关系，包括纪录片的种类和每个种类下的纪录片名称。

2）当搜索某个种类时，结果将围绕某个种类进行展示，展示的内容包括所含的纪录片名称，和所含纪录片都有哪些集数，如图6-3所示：

图6-3 搜索某一类型纪录片时的结果展示

3）当搜索某个纪录片时，结果将围绕这个纪录片展开，展示纪录片所包含的集数和每一集所涉及到的内容。如图6-4所示：

图6-4 搜索某一纪录片时的结果展示

4）当搜索某个纪录片的具体集数时，知识图谱将展示该集纪录片所包含的实体，展示如图6-5所示：

图6-5 搜索具体纪录片具体集数时的结果示例

从图中可以看到在“互联网时代”第一集的纪录片中都包含了哪些人物。同时在本系统中，对每个实体添加了链接，链接到维基百科。在本例中，只要点击上图中某个人，将会链接到维基百科中介绍该人物的页面。

6.2 按照关键词进行搜索

按照关键词进行搜索，即搜索想要了解的人名、地名、机构名等，就可在系统收录的纪录片中找到与关键字相关的纪录片，并可实现定点播放（即在纪录片中该关键字出现的时间点进行播放）。

示例如下：

1）搜索关键词“谢尔盖-赫鲁晓夫”展示的结果如图6-6所示：

图6-6 搜索关键字结果展示

2）点击“互联网时代e01”，即可跳转到播放页面直接播放该关键词在互联网时代第集中第一次出现的时间点。展示如图6-7所示：

图6-7 跳转播放示例

本系统中出现的类似“s01”这样的描述代表的是第一季；“e01”这样的描述代表的是第一集。随着收录的纪录片数量的增多，查询结果也会更加全面。

6.3 信息录入

因为有些纪录片可通过一些技术手段获得某个关键词在纪录片中出现的具体时间点，而有些却不能通过这些自动化手段获得，所以需要人工的录入。本系统中的信息录入就是针对第二种情况。

需要录入的信息包括：关键词、纪录片名称、时间点。界面如图6-8所示：

图6-8 信息录入页面

6.3 本章小结

本章主要介绍了文中知识图谱部分功能（按照纪录片进行搜索、按照关键词进行搜索、信息录入）的实现，其中详细介绍了本文中对纪录片中实体抽取的步骤。最后对系统中每一部分的实现进行了截图展示。

7 总结与展望

7.1 论文总结

如今随着互联网技术的发展，互联网上流通的数据、知识越来越多，而且这些信息在互联网中的存在形式也是多彩多样。为了使大家能够更加方便、快捷、直观地理解这些信息，因而产生了数据可视化、知识图谱等相关的技术，将信息以更好的形式进行展现。现在这些技术的发展已经相对成熟。本文主要介绍了数据可视化和知识图谱相关的基本理论，另外还将数据可视化及知识图谱相关的技术应用到中国纪录片的市场上，开发了一个中国纪录片知识图谱的构建和可视化系统。本文针对此研究方向所做的具体工作如下：

1）系统环境配置：本系统是在Mac上进行开发的，需要用到node、python、react、webpack等技术，所以要对这些环境进行安装。本系统所用的软件版本：node（v6.10.0），python（2.7.13），react（15.4.2），webpack（1.12.15）。

2）数据可视化部分数据的爬取与处理：本系统数据来源是三大视频网站中纪录片相关的数据，数据是通过python的BeautifulSoup库对网页进行爬取和解析获得的。数据处理部分主要经过了数据集成、数据清理、数据变换这三个过程。

3）数据可视化部分数据的展示：数据的展示部分主要用Echarts对处理过的数据进行了三个角度的展示：播放量排名前十的纪录片的播放量展示、播放量前十的纪录片类型占比的展示、播放量前十的纪录片的评分展示。

4）知识图谱部分数据的获取：知识图谱部分的数据主要来源于纪录片的解说词，这些解说词可以简单、方便地从互联网上下载，然后以文档形式存在电脑上就行。

5）知识图谱部分实体的识别及关系的分类：本文知识图谱部分实体识别及关系分类是用哈工大社会计算与信息检索研究中心经过11 年的持续研发而形成的一个自然语言处理工具库：LTP。它提供包括中文分词、词性标注、命名实体识别、依存句法分析、语义角色标注等丰富、高效、精准的自然语言处理技术。

6）知识图谱部分的可视化：知识图谱的可视化部分也是由Echarts的关系图进行展示的。

7.2 展望与后续工作

在如今的大数据时代，大量的信息存在于互联网的角角落落，相关的大数据的技术也发展起来。但是在本文中，数据可视化部分的数据只涉及三大视频（优酷、腾讯、爱奇艺）中纪录片相关的数据，而且获取的纪录片的相关的信息也不全面，比如纪录片的出品方、首播平台等信息都没有获取到，所以在对这些数据进行展示时不免缺乏片面性和局限性。另外，对纪录片数据进行可视化也需要数据的实时性，这样才能使用户更加及时的捕捉到有用的信息，所以需要不断地运行爬虫程序以获得最新的数据，这时候就要对爬虫的性能有所要求了。除了以上两点之外，本文中对数据的展示的形式主要以统计图表的形式，随着数据量的增加和数据形式的多样化，在数据展示方面也需要添加更加新颖、有效的展示方式。所以在后续的工作中，在数据可视化方面，将会改进以下几点：

1）尽可能多、尽可能全面的获取到纪录片的数据，最好是从比较权威的网站中获取数据，这样才能提高数据的质量，得到更加有价值的信息。

2）为了更加实时的获取数据，要调研一下爬虫技术，对本文中的爬虫进行优化，然后编写一个脚本，定时执行爬虫任务。

3）认真研究纪录片领域待挖掘的信息，同时调研可视化技术，以求寻找到更加合适的展现相关信息的形式。

目前，知识图谱的相关技术在国外发展的比较好一些，虽然国内一直跟随国外的研究脚步也研究出了相应的技术，但是还是有差距的。中文和英文的差别，导致国内和国外的研究不能同步，但是却可以借鉴国外的一些方式方法来优化知识图谱的构建。本文采用了哈尔滨工业大学研发的LTP语言处理技术来对实体和实体关系进行提取，但是提取的效果也不是全然是正确的，所以为了更加准确的展示实体及实体间的关系，需要深刻理解这个技术的原理方法，然后借鉴国外的一些研究方法，使提取的结果更加准确。另外，在知识图谱的构建过程中，因为本文主要是实验性的研究，所以并没有研究出更加通用型的针对纪录片知识图谱构建的技术，所以以后会在这方面加以努力。

本文在系统架构上也有很多需要改善的地方：

1）在服务器方面，本文为了体验用node搭建轻便型服务器，所以采用了node技术，但是在以后的发展，或系统的推广中

，应该采用更加稳定及快速的服务。

2）在前端部分，虽然本文做到了开发的前后端分离，但是前端部分所占用的大小还应该适当减小，需要对前端代码进行优化。

3）本文中大部分的数据都是以文本的形式存在硬盘中，随着数据量的增加及系统的规范化，以后需要将这些数据规范化之后存在合适的数据库中。

参考文献

[1] 刘峤, 李杨, 段宏等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3):582-600.

[2] 何苏六, 韩飞. 2016年中国纪录片产业发展透视[J]. 电视研究, 2017(3):18-21.

[3] 王喆. 探究中国纪录片的未来发展[J]. 声屏世界, 2017(7):43-45.

[4] 陈为, 张嵩, 鲁爱东. 数据可视化的基本原理与方法[M]. 科学出版社, 2013.

[5] Singhal A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012.

[6] Gruber, T. (1995). “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”. International Journal of Human-Computer Studies. 43 (5-6): 907–928.

[7] Tim Berners-Lee(2006-07-27). “Linked Data”. Design Issues. W3C.

[8] Auer S, Bizer C, Kobilarov G, et al.Dbpedia: A nucleus for a web of open data[J]. The semantic web, 2007: 722-735.

[9] Tim Berners Lee. The Semantic Web Roadmap. https://www.w3.org/DesignIssues/Semantic.html, 1998.

[10] Auer S, Bizer C, Kobilarov G, et al.Dbpedia: A nucleus for a web of open data[J]. The semantic web, 2007: 722-735.

[11] Suchanek F M, Kasneci G, Weikum G.Yago: a core of semantic knowledge[C] //Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 697-706.

[12] Xu, B.; Xu, Y.; Liang, J.; Xie, C.;Liang, B.; Cui, W.; and Xiao, Y. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In International Conference on In- dustrial,Engineering and Other Applications of Applied Intelligent Systems, 428–438.Springer.

[13] Niu, X.; Sun, X.; Wang, H.; Rong, S.;Qi, G.; and Yu, Y. 2011. Zhishi. me-weaving chinese linking open data. The Semantic Web–ISWC 2011 205–220.

[14] Wang, Z.; Li, J.; Wang, Z.; Li, S.;Li, M.; Zhang, D.; Shi, Y.; Liu, Y.; Zhang, P.; and Tang, J. 2013. Xlore: Alarge- scale english-chinese bilingual knowledge graph. In Pro-ceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume1035, 121–124. CEUR- WS. org.

[15] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In IJCAI, volume 7, pages 2670– 2676, 2007.

[16] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In IJCAI, volume 7, pages 2670– 2676, 2007.

[17] Anthony Fader, Stephen Soderland, andOren Etzioni. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1535–1545. Association for Computational Linguistics, 2011.

[18] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages523–534. Association for Computational Linguistics, 2012.

[19] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Towardan architecture for never-ending language learning. In AAAI, volume 5, page 3,2010.

[20] Ndapandula Nakashole, Gerhard Weikum,and Fabian Suchanek. Patty: a taxonomy of relational patterns with semantictypes. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages1135–1145. Association for Computational Linguistics, 2012.

[21] Mintz, Mike, Steven, et al. Distant supervision for relation extraction without labeled data[C]// Joint Conference of the, Meeting of the ACL and the, International Joint Conference on Natural Language Processing of the Afnlp: Volume. Association for Computational Linguistics, 2009:1003-1011.

[22] Luo B, Feng Y, Wang Z, et al. Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix[J]. 2017:430-439.

[23] Feng X, Guo J, Qin B, et al. Effective Deep Memory Networks for Distant Supervised Relation Extraction[C]. Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017:4002-4008.

[24]宋卿, 戚成琳, 杨越. 基于Bootstrapping的新闻事件型实体关系抽取方法[J]. 中国传媒大学学报(自然科学版)自然科学版, 2017(4):46-50.

[25]吴骏, 王强, 李振兴等. 一种基于卷积神经网络的企业实体关系抽取的方法:, CN107220237A[P]. 2017.

[26] Zheng S, Hao Y, Lu D, et al. Joint Entity and Relation Extraction Based on A Hybrid Neural Network[J].

Neurocomputing, 2017, 257(000):1-8.
[27] Miwa M, Bansal M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures[J]. 2016.
[28] Zheng S, Wang F, Bao H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[J]. 2017:1227-1236.
[29] Liang J, Zhang S, Xiao Y. How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017:3749-3755.
[30] Shi B, Weninger T. Open-World Knowledge Graph Completion[J]. 2017.
[31] Zhuang Y, Li G, Zhong Z, et al. Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases[C]. ACM on Conference on Information and Knowledge Management. ACM, 2017:1917-1926.

攻读硕士学位期间取得的学术成果

[1] Dongdong Jiang, Wenqian Shang. Design and Implementation of Recommendation System of Micro Video's Topic(ICIS), 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). (Accession number:20174104250857)

指　标
疑似剽窃文字表述

1. 简介 summary 在以往的影像素材里，中国美食更多以“烹饪大师”或“美食名家”结构，展现的是“精湛的厨艺”和“繁复的过程”。在本片中，中国美食更多的将以轻松快捷的叙述节奏和精巧细腻的画面，向观众，尤其是海外观众展示中国的日常饮食流变，中国人在饮食中积累的丰富经验，千差万别的饮食习惯和独特的味觉审美，以及上升到生存智慧层面的东方生活价值观。
   纪录片类型 type
2. 在LTP中，将分词任务建模为基于字的序列标注问题。对于输入句子的字序列，模型给句子中的每个字标注一个标识词边界的标记，
3. 词性标注。在LTP中词性标注与分词模块相同，也是将词性标注任务建模为基于词的序列标注问题。对于输入句子的词序列，模型给句子中的每个词标注一个标识词边界的标记，
4. 分词模块也相同，将命名实体识别建模为基于词的序列标注问题。对于输入句子的词序列，模型给句子中的每个词标注一个标识命名实体边界和实体类别的标记。在LTP中，我们支持人名、地名、机构名三类命名实体的识别。
5. 要把文本中的命名实体识别给提取出来，这就需要把组成命名实体的词给单个提取出来且根据标记连接短语词组。在这里，主要是提取命名实体三大类：组织、人名和地名。
6. 哈工大社会计算与信息检索研究中心经过11 年的持续研发而形成的一个自然语言处理工具库：LTP。它提供包括中文分词、词性标注、命名实体识别、依存句法分析、语义角色标注等丰富、高效、精准的自然语言处理技术。

说明：1.总文字复制比：被检测论文总重合字数在总字数中所占的比例

2.去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例

3.去除本人已发表文献复制比：去除作者本人已发表文献后，计算出来的重合字数在总字数中所占的比例

4.单篇最大文字复制比：被检测文献与所有相似文献比对后，重合字数占总字数的比例最大的那一篇文献的文字复制比

5.指标是由系统根据《学术论文不端行为的界定标准》自动生成的

6.红色文字表示文字复制部分;绿色文字表示引用部分

7.本报告单仅对您所选择比对资源范围内检测结果负责

amlc@cnki.net
http://check.cnki.net/
http://e.weibo.com/u/3194559873/