Journal of System Simulation

Volume 28 | Issue 10 Article 38

8-13-2020

Design and Implementation of Historical Figures Knowledge Graph Visualization System

Zhou Yi
1. Department of Information Technology, Beijing Normal University, Beijing 100875, China;;2. Engineering Research Center for Virtual Reality Applications, MOE, Beijing 100875, China;

Mingquan Zhou
1. Department of Information Technology, Beijing Normal University, Beijing 100875, China;;2. Engineering Research Center for Virtual Reality Applications, MOE, Beijing 100875, China;

Xuesong Wang
1. Department of Information Technology, Beijing Normal University, Beijing 100875, China;;2. Engineering Research Center for Virtual Reality Applications, MOE, Beijing 100875, China;

Youliang Huang
1. Department of Information Technology, Beijing Normal University, Beijing 100875, China;;2. Engineering Research Center for Virtual Reality Applications, MOE, Beijing 100875, China;

Design and Implementation of Historical Figures Knowledge Graph Visualization System

Abstract

Abstract: With the advent of big data era, knowledge graph and data visualization technology present the data in a structured, visual way and establish a keyword-oriented knowledge system and render the relationship in a fast and clearly way. In this paper, a historical figures entity relationship visualization system has been established by means of data visualization and knowledge graph. In the system,the complex data are divided into character data and event data by data preprocessing. *In the parsing stage, a label traversing method and a method based on weight of links are applied to the divided data respectively*. With the layered B/S structure design, the system is based on the Nodejs platform in which a historical figures database is founded. The users can obtain the knowledge graphs of relevant historical figures and events according to distinct needs. This system provides a variety of interactive with good scalability and maintainability and makes contributions to comprehension and exploration of the data and relationship quickly by presenting the data in visual forms. To some extent, it owes reference value to research staff as well.

Keywords

knowledge graph, entity relationship, data parse, data visualization

Recommended Citation

Zhou Yi, Zhou Mingquan, Wang Xuesong, Huang Youliang. Design and Implementation of Historical Figures Knowledge Graph Visualization System[J]. Journal of System Simulation, 2016, 28(10): 2560-2566.

第 28 卷第 10 期 系统仿真学报© Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

大数据环境下历史人物知识图谱构建与实现

周亦 [1,2]，周明全 [1,2]，王学松 [1,2]，黄友良 [1,2]
（1.北京师范大学信息科学与技术学院，北京 100875；2.教育部虚拟现实应用工程中心，北京 100875）

摘要：大数据时代下，知识图谱和数据可视化技术能够将数据以结构化、可视化的方式呈现，建立以关键词为中心的知识体系，展示数据间相互关系。在此基础上，设计并实现历史人物实体关系可视化系统。系统基于 Nodejs 平台，采用 B/S 架构，将繁杂数据分为人物数据和事件数据，*分别采用基于标签遍历和基于链接权重的方法进行数据解析，存储至历史人物库*。系统提供多种交互方式并具有良好的扩展性和维护性，以丰富直观地形式将历史人物和事件的信息可视化，帮助人们更好理解、梳理及挖掘历史人物及相关事件关系，对相关研究人员有一定的帮助和参考价值。

关键词：知识图谱；实体关系；数据解析；数据可视化
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2016) 10-2560-07

Design and Implementation of Historical Figures Knowledge Graph Visualization System

Zhou Yi[1,2], *Zhou Mingquan*[1,2], *Wang Xuesong*[1,2], *Huang Youliang*[1,2]
(1. Department of Information Technology, Beijing Normal University, Beijing 100875, China;
2. Engineering Research Center for Virtual Reality Applications, MOE, Beijing 100875, China)

Abstract: With the advent of big data era, knowledge graph and data visualization technology present the data in a structured, visual way and establish a keyword-oriented knowledge system and render the relationship in a fast and clearly way. In this paper, a historical figures entity relationship visualization system has been established by means of data visualization and knowledge graph. In the system,the complex data are divided into character data and event data by data preprocessing. *In the parsing stage, a label traversing method and a method based on weight of links are applied to the divided data respectively*. With the layered B/S structure design, the system is based on the Nodejs platform in which a historical figures database is founded. The users can obtain the knowledge graphs of relevant historical figures and events according to distinct needs. This system provides a variety of interactive with good scalability and maintainability and makes contributions to comprehension and exploration of the data and relationship quickly by presenting the data in visual forms. To some extent, it owes reference value to research staff as well.

Keywords: knowledge graph; entity relationship; data parse; data visualization

引言

随着大数据[1]时代的到来，数据的类型和规模呈指数增长趋势。在海量的数据面前，人们不再满足于数据的量，而更注重数据的质-即数据的规律以及它们之间的关系。借助知识图谱[2-4]和数据可视化[5-6]技术，理解数据成为近年来的研究热点。知识图谱是指通过语义检索，对大量数据进行过滤、分析和管理，实现搜索数据的结构化并且提供详细的主题相关信息，有助于人们建立知识体系，

收稿日期：2016-04-27 修回日期：2016-07-14;
作者简介：周亦(1993-)，女，湖北，硕士，研究方向为虚拟现实与可视化；周明全(1954-)，男，陕西，博导，研究方向为虚拟现实与可视化；王学松(1975-)，男，陕西，博士，研究方向为虚拟现实与可视化。

理解各种实体概念以及它们的关联。数据可视化技术通过直观、可交互的形式展现结构化的数据，梳理数据之间的关系，易于人们接受和使用。将知识图谱与可视化技术结合，与传统文字、图片信息相比，具有准确分析数据、深度解读数据、清晰呈现数据的优势，甚至将文本难以描述的东西，用直观自然的方式协助人们理解。

目前，基于中华历史人物及事件的知识图谱研究较少，为了宣传和弘扬我国历史文化，更好理解、梳理及挖掘历史人物及相关事件关系，本文结合数据可视化及知识图谱技术，设计并实现了历史人物实体关系可视化系统。该系统利用 Web 引擎，建立中华历史人物信息库，构建基于人物和事件的知识图谱，用户可以快速、便捷地搜索和浏览历史人物信息。

1 历史人物可视化系统设计实现

1.1 系统需求

互联网信息源丰富，数据呈现海量性，并且内容充实有价值，数据的可视化展示是利用此类数据的一种重要形式，通常针对数据可视化采用的方式是统计图为主，即柱状图、饼状图等方式，这种统计类型的图表只提供了相关统计信息，但人们无法了解数据中具体细节，提供的有效信息量不足。此外，统计类型的图会使得数据信息过于繁杂，不能清晰梳理数据与数据之间的关联。本文提出的历史人物可视化系统主要针对历史科研工作者和广大学生群体，以及对中华历史感兴趣的人们，受众较为广泛。由此，系统也需保证良好的用户体验，例如用户操作的流畅性，可视化图表直观性等。基于以上背景调研和用户角度的分析，对历史人物可视化系统提出以下几点功能需求：

1. 历史人物展示：展示历史人物的个人信息，以及以该人物为中心的人物关系网络，同时，不同人物的关系需要明确标识。

2. 历史事件展示：以历史事件为中心，将事件涉及的人物、背景等清晰展现出来，以及标识人物在事件的不同角色。

3. 界面交互：用户通过操作可以将同一个图以不同形式呈现，譬如力导向图与和弦图的转换，或者进行拖拽、放大、保存图片等操作。

4. 数据更新：由于历史数据在不断被研究人员发掘完善以及数据源可能失效，会存在信息不正确、不足的问题，系统需要定时更新数据，保证数据的准确性和时效性。

1.2 系统设计

历史人物知识图谱可视化系统以 Nodejs 为开发平台，采用 B/S 架构，系统结构采用分层设计，分别有客户层，服务层，数据处理层和数据存储层。如图 1 所示。

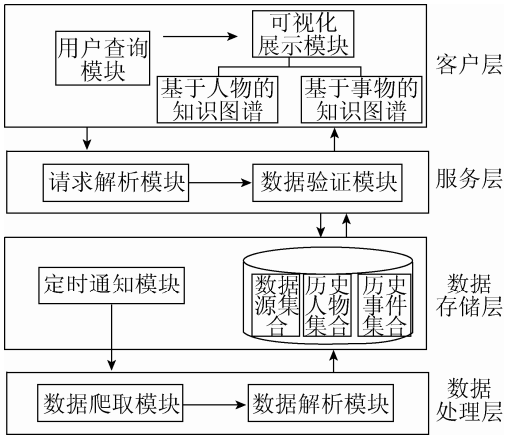

图 1 历史人物实体关系可视化系统框架

(1) 客户层，以用户与系统的交互界面的方式呈现，完成系统的查询和展示功能。其中查询分为两类，包括以人物为关键字的查询和基于事件的查询。依据查询的类别，系统将数据以不同的可视化样式展现，并提供伸缩、移动和保存等交互操作。

(2) 服务层，解析模块负责客户层请求的处理，获取 Post 请求参数，根据不同参数访问数据层中不同集合，将处理后的数据以 JSON 格式回传至数据验证模块。数据验证模块负责对数据进行去冗余、非法字符等容错处理。系统采用 Nodejs 平台内置的 HTTP 模块创建应用服务器，能够方便快

第 28 卷第 10 期 系统仿真学报 Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

捷地构建所需服务。

(3) 数据处理层是整个系统的关键部分，核心模块包括数据爬取模块和数据解析模块，是历史人物库的建立的基础。数据爬取模块访问数据层的数据源集合，爬取数据源中提供的历史人物和事件信息。由于数据源中 URL 数据量大，单线程的爬取存在耗时长的弊端，由此我们采用多线程的方式，以提高速率。爬虫策略[7-9]可分为深度优先、广度优先和最佳优先。深度优先存在着随着深度的增加，网页价值会相应下降的问题，而最佳优先属于局部优先搜索，爬取中可能忽略许多相关网页，所以为了覆盖尽可能多的网页，在完成当前层次的搜索后，再进行下一层次的搜索。由此基于多线程的广度优先爬虫策略更适合本系统。

数据解析[10-11]是建立数据库之前的不可或缺的阶段，针对爬取的网页进行内容解析，将原始数据经过容错、过滤、提取等操作后得到结构清晰统一的数据。原始网页数据不能直接使用，主要由以下几个原因造成：首先相对于结构化的 XML 文件，HTML 网页语言灵活度高，规范化程度低，可能存在没有闭合的非法标签。其次页面中存在大量的无效信息，譬如导航栏、推荐链接、版权说明等与主题无关的信息。为了追求页面的美观性，网页结构和风格呈现复杂多变的特点，样式或者脚本代码等大量嵌入到网页中。这些网页噪声，降低了人们阅读效率，给抽取网页正文造成了一定困难。所以在进行正文提取之前，需要对 HTML 文件进行解析，将其转换为 DOM 树[12]，针对结构化程度较高的人物数据，可以直接读取相应标签内容。但大部分搜索结果中，历史事件信息结构化程度较低，携带关键信息的节点属性并不固定。考虑到此种类型的网页具有正文内容长而且分布密集的特性，由此采用基于权值和链接权重的方法提取正文，随后对正文进行分词处理获取关联人物列表的方法处理。

因此数据解析模块将分别采用基于DOM树标签遍历和基于链接权重的方法将历史人物和事件的数据进一步地封装与抽象化后分别发送到数据层进行存储。模块具体设计及实现请参考 2.1,2.2。

数据存储层，系统建立历史人物库，避免每次用户请求都进行爬取、解析等工作，提高图谱绘制的实时性。人物库存储系统需要的所有数据，主要集合包括数据源集合，历史人物集合，历史事件集合，根据客户层请求类型，将在不同的集合中执行查询任务。同时存储层还包括了与数据处理层交互的通知模块，由于有些 URL 具有失效性并且网络数据处于不断更新状态，为了保证数据的时效性和准确性，通知模块定时发送请求到数据处理层，实现数据库定期更新数据的功能。不同数据源的提供的数据维度之间存在差异，传统的关系型数据库频繁增删表中的属性字段会造成读写瓶颈，所以系统选用 NoSQL 类型[13]的 MongoDB，能够随时存储自定义的数据格式，具有较高的易扩展性和读写性能。

B/S 架构给予用户随时查询和浏览信息的途径，并且采用分层框架，层次之间各司其职，降低了数据展示和服务层的耦合性，能够灵活地改变请求处理规则和数据存储方式，而不影响其他模块，提高了系统的鲁棒性和可维护性。此外 Nodejs 平台采用异步、事件驱动模型，非阻塞式的 IO 处理使得系统在相对低资源耗用下仍具备出众的负载能力，很好地满足大量用户同时访问的高并发需求。

2 数据处理层设计

2.1 数据爬取模块

爬取数据之前，需要选定数据源。为了高效率地获取历史人物相关数据，经过调研，系统选取 Google、百度、Bing 和搜狗网站作为数据来源。这些网站的内容具有开放性，涵盖的信息丰富且可信度高，提供公开的搜索 API。以 Google 为例，通过定制搜索引擎，能够以 RESTful 风格请求进行搜索，返回的结果为 JSON 格式。只需注册生成搜索引擎 ID 和 API KEY，提供的 URL 形式为：https://www.googleapis.com/customsearch/v1?key=INSERT-KEY&cx=INSERT-ID:omuauf_lfve&q= 查询关键字，关键字为历史人物或事件名称。

数据爬取模块的流程如图 2 所示，依据数据库中导出的数据源即 URL 集合初始化 URL 队列，将 URL 队列分段处理后交由不同的线程处理。在处理过程中，爬虫程序遍历访问 URL 对应的网页，采集网页内容的同时将网页中存在的 URL 添加到队尾中，直到抓取了队列中所有的网页或者满足了给定的终止条件。

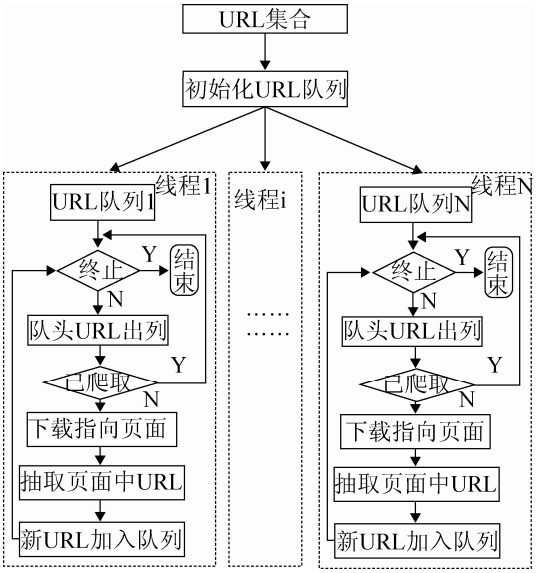

图 2 数据爬取模块流程图

2.2 数据解析模块

系统涉及数据类型主要为人物数据和事件数据，现有的搜索引擎对不同种类的数据结构化程度有所差异，由此针对不同类型的信息，使用不同的提取方式。

2.2.1 基于历史人物的数据解析

DOM 树定义了 HTML 文件在内存中的树形文档结构，清晰地表达了文挡中对象、属性以及它们之间的关系。因此解析的过程就转换成直观的 DOM 树节点的操作。

以人物为关键字的搜索结果具备较高的结构化形式，避免了大量的筛选操作。由此针对历史人物信息，利用同一个网站下的网页在内容布局和样式结构的相似性[14]，如图 3。本文选用 Beautiful Soup 作为解析器，它是一个可以从 HTML 或 XML 文件中提取数据的 Python 库，具有良好的容错性，以及自动处理不规范标记，提供简单易用的导航[15]、搜索以及修改 DOM 树的功能。系统利用以上功能，将文件解析成 DOM 树后按照节点名、CSS 属性、节点之间的关系(兄弟关系，父子关系)等多种查询方式，定位相应节点，极大地提高信息抽取速率。

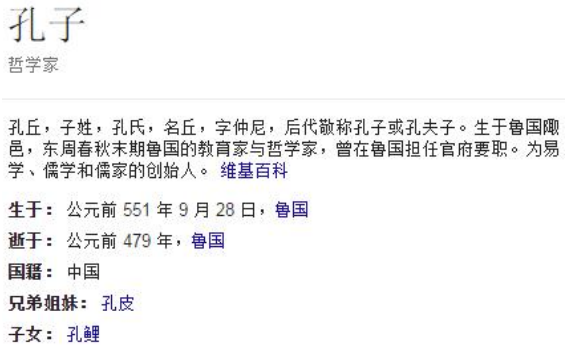

图 3 Google 知识图谱的页面结构

通过查找网页 DOM 树中属性相同的节点，采取基于标签遍历的方法提取相关网页正文。提取历史人物网页正文的主要流程为：

1) 将 HTML 文档转化成 DOM 树结构

2) 针对不同的搜索引擎，提取关键信息的 DOM 节点标签，譬如经过观察通过 Google 搜索的历史人物，页面正文的树结构为图 4 所示。

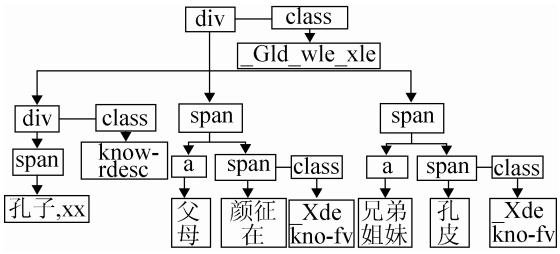

图 4 页面 DOM 树结构

3) 提取正文时只需查找标签为 div class="_G1d_wle_xle"、div class="Xbe kno-fv"、div class="know-rdesc"以及它们的子节点等节点即可。依据解析器的 CSS 查找功能，查询根节点 div class="_Gld_wle_xle"，遍历根节点的子节点，每个子节点可能有多个子节点，通常取子节点的第一个孩子节点的值做为 Key，其余孩子节点的值拼接成

第 28 卷第 10 期 系统仿真学报 Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

Value，键值对进行格式化转换后存储至相应集合。

采取基于标签遍历的方法充分利用页面布局的相似的特点，避免了大量遍历节点的过程，直接导航到 DOM 树的关键节点，简化了提取人物数据流程。

2.2.2 基于历史事件的数据解析

我们认为一个节点中超链接占据的比例越小，该节点中包含的内容为正文的可能性就越大。节点中含有该节点的节点标签和文本内容，记节点标签的字符长度为 *l*，记节点正文内容的字符长度为 *t*，节点权值的定义为：

$$W=(\frac{t+l}{1+(1-a)*l})^2 \tag{1}$$

当节点标签为<a>时，*a*=0，当节点标签不为<a>时，*a*=1。

通常情况下，节点权值最大的节点为正文节点，包含正文的主要部分。计算出权值最大的节点后，以此节点为基准，分别向前向后搜索邻近节点，按照顺序依次将节点所含的文本内容连接起来即可还原网页的正文信息。同时需要引入一个衡量标准确定节点搜索停止条件，规定当遇到低于链接权重[15]的超链接节点则停止搜索。链接权重的定义：从第一个出现的超链接节点开始，选出该节点后的 k–1 个内容非空的节点，那么这 k 个节点的链接权重 *Link* 为：

$$Link=\frac{\sum_{i=1}^{k}(at_i+W_i)}{\sum_{i=1}^{k}t_i} \tag{2}$$

当节点标签为<a>时，*a*=0，当节点标签不为<a>时，*a*=1。当链接权重比较小时，说明这 k 个节点块中超链接很多，那么这些节点应属于链接块，而非正文内容块，否则节点块很可能为正文块，属于网页正文内容的一部分。通过链接权重的方法，可以很好从正文含少许超链接的网页中抽取出正文，具有很强的适应性。

为了建立事件和人物之间的联系，正文提取之后还需进行分词[16-17]操作。针对如何筛选正文中人名的问题，系统利用语言云作为解决方案。该平台以哈工大社会计算与信息检索研究中心研发的“语言技术平台(LTP)”为基础，提供高效精准的中文自然语言处理云服务，包括中文分词、词性标注、命名实体识别、语义角色标注等自然语言处理技术。

根据平台的 API 接口，传入参数，对应的 URL:http://api.ltp-cloud.com/analysis/?api_key=YourApiKey&text=正文内容。如图 5 为一段正文的分词结果，斜线后的内容代表词性，人名的词性标识为 nh，由此提取出“安禄山”、“史思明”以及“唐玄宗”作为事件涉及的人物。

安史/n 之/u 乱/a 是/v 唐朝/nt 于/p 755年/nt 12月/nt 16日/nt 至/p 763年/nt 2月/nt 17日/nt 由/p 安禄山/nh 与/p 史思明/nh 向/p 唐朝/nt 发动/v ，/wp 同/p 唐朝/nt 争夺/v 统治权/n 的/u 内战/n ，/wp 是/v 唐/nt 由/p 盛/a 而/c 衰/a 的/u 转折点/n ，/wp 也/d 造成/v 唐代/nt 藩镇/ns 割据/v 。/wp 由于/c 发起/v 反唐者/n 乃是/v 安禄山/ns 与/p 史思明/nh 二/m 人/n 为主/v ，/wp 故/c 事件/n 被/p 冠/v 以/p 安史/n 之/u 名/n 。/wp 又/d 由于/c 其/r 爆发/v 于/p 唐玄宗/nh 天宝/nz 年间/n ，/wp 也/d 称/v 天宝/nz 之/u 乱/a 。/wp

图 5 正文分词后人物标识结果

提取历史事件网页正文的主要流程为：

1) 解析 HTML 文档成 DOM 树

2) 删除与正文无关的元素。需要删除的元素[18]主要有以下几类：1. 带有标签 script、style、link、meta 等元素，这些元素对应页面的 JS 脚本、CSS 等信息 2. 注释的标签，它的构成为<!-->，注释内容属于 DOM 树的文本节点，但属于无用信息 3. 空节点，主要有两种形式<xxx></xxx>,<xxx/>

3) 选取合适的 k 值和链接权重阈值，计算 DOM 树中节点权值最大的节点，以该节点为中心，分别向前和向后遍历其他节点，记录节点的文本值，直到节点相关的链接权重大于设定阈值则停止。依据搜索引擎，可以自适应的设置 k 值和链接权重。

4) 上一步的正文作为语言云平台的输入，分词之后找出词性标识为 nh 的值。以历史事件作为 Key，人物列表作为 Value，格式转换后存储 Key-Value 键值到历史事件集合。

3 数据存储层设计

历史人物库的集合有人物集合、事件集合和数据源集合三种。人物集合存储系统中爬取的人物信息，包含了姓名、人物介绍、出生日期、朝代、父母以及涉及事件等字段，主要属性字段见表 1。事件集合用于存储历史事件信息，主要的属性有事件名，描述，起始时间，涉及的人物列表等。当需要返回事件数据时，会读取涉及人物字段中 id 号数组，依次查询该 id 号在人物集合中对应的人物数据，从而建立事件和人物之间的关联。表 2 为事件集合的主要属性。

表 1 历史人物集合主要属性

属性名称	属性意义	数据类型	可否为空	是否唯一
_id	id 号	ObjectId	否	是
name	人物名称	String	否	否
invEvent	涉及事件	ObjectId[]	是	否
dynasty	朝代	String	是	否
birthTime	出生日期	Date	是	否
deathTime	去世日期	Date	是	否
descption	人物介绍	String	是	否

表 2 历史事件集合主要属性

属性名称	属性意义	数据类型	可否为空	是否唯一
_id	id 号	ObjectId	否	是
name	事件名称	String	否	是
invPeople	涉及人物	ObjectId[]	否	否
description	事件介绍	String	是	否
beginTime	开始日期	Date	是	否
endTime	结束日期	Date	是	否

4 可视化系统结果展示与分析

经过数据爬取和解析后，人物集合中存储人物记录达到 4 000 多条，历史事件集合记录有 1 000 多条，覆盖从远古到清朝共计 22 个朝代。系统提供的可视化图谱的类型有和弦图、饼状图、力导向图、折线图等。可视化技术以基于 JavaScript 的 ECharts 库为主要工具，它能够兼容当前绝大部分浏览器，提供流畅，可交互，可高度个性化定制的数据可视化图表。

基于人物的系统测试，用户以“孔子”为查询关键字。图 6 是搜索“孔子”后，系统界面显示的知识图谱结果，依据节点的颜色，用户能够清晰地观察到不同人与孔子之间的关系，同时能够进行保存图片、图像拖拽以及转换为和弦图，图 7 所示等交互操作。系统还提供查看孔子弟子籍贯分布的知识图谱的功能选项，依据图 8 可知，孔子的七十二个弟子的籍贯分为 12 种，其中籍贯为鲁国的人数最多，共计 42 人，占据 58.33%，其次是卫国 5 人，占总人数的 6.9%，籍贯未知的有 8 人，占据 11.11%。基于事件的测试以“戊戌变法”为关键词，返回的知识图谱图 9，结果图清晰直观明了地展示出事件相关的人物以及事件背景等信息，并且不同人物角色采用不同颜色标识。

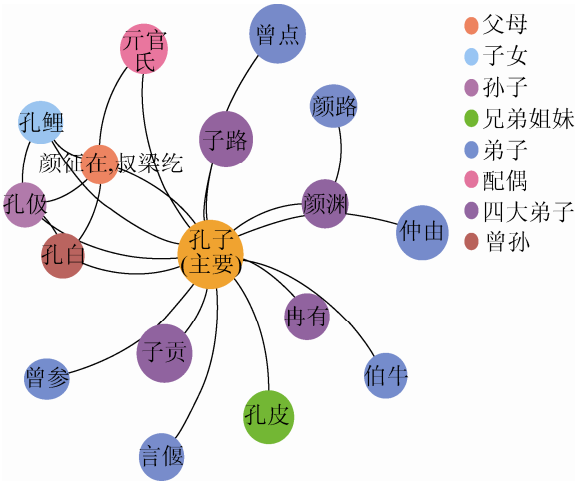

图 6 基于人物的力导向图可视化结果

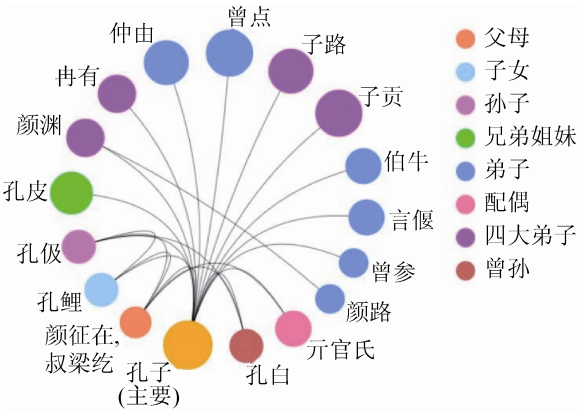

图 7 基于人物的和弦图可视化结果

第 28 卷第 10 期 系统仿真学报 Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

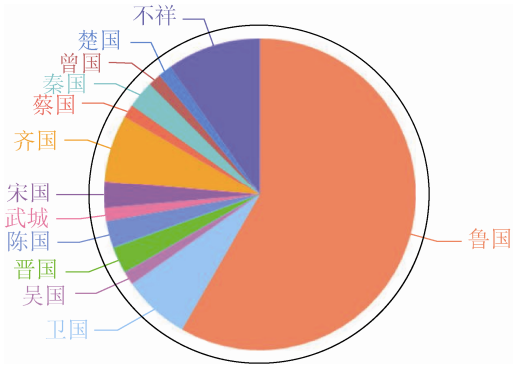

图 8 基于人物分布的饼状图可视化结果

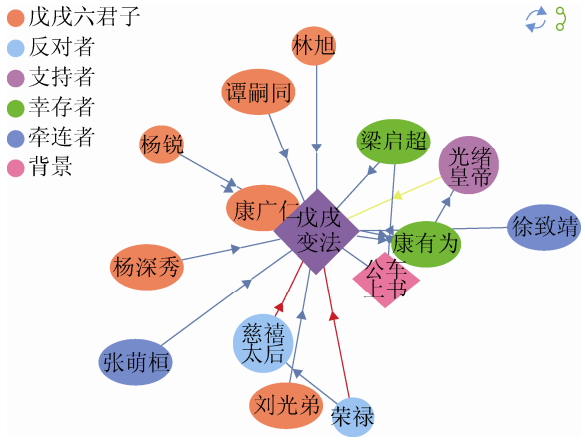

图 9 基于事件的力导向图可视化结果

5 结论

历史人物可视化系统对历史人物和历史事件分别采取不同的方法完成内容提取，高效快速地建立历史人物库。通过对数据的视觉解释，系统提供了交互式历史人物数据展示服务，实现了化繁为简、梳理事件人物关系、揭示人物关联等功能，这种清晰直观的知识体系，易于被人们接受并加深理解。同时采用分层设计，系统具备良好的扩展性和维护性，能够满足大量用户访问的需求。在以后的工作中，系统将从以下几个方面做出改进：1.面对海量数据[19]，需要在史学专家的协助下，对数据的来源、渠道进行有效考证，评价所获数据的质量与意义，以保障数据的真实可靠性；2.为了提高用户的浏览体验，未来系统将提供更丰富美观的可视化形式和灵活自然的交互方式，数据维度不再局限于平面，增加立体化多样化的特性。

参考文献：

[1] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战 [J]. 计算机研究与发展, 2013, 50(1): 146-169.
[2] 刘则渊, 陈悦, 候海燕, 等. 科学知识图谱:方法与应用 [M]. 北京: 人民出版社, 2008: 19-20.
[3] 杨思洛, 韩瑞珍. 知识图谱研究现状及趋势的可视化分析 [J]. 情报资料工作, 2012, 3(4): 22-28.
[4] Herl HE, Jr HFO, Chung GKWK, et al. Reliability and validity of a computer-based knowledge mapping system to measure content understanding [J]. Computers in Human Behavior (S0747-5632),1999, 15(3/4): 315-333.
[5] Keim D A. Information Visualization and Visual Data Mining [J]. IEEE Transactions on Visualization (S1077-2626),2002, 8(1): 1-8.
[6] 张卓, 宣蕾, 郝树勇. 可视化技术研究与比较 [J]. 现代电子技术, 2010, 33(17): 133-138.
[7] 周德懋, 李舟军. 高性能网络爬虫: 研究综述 [J]. 计算机科学, 2009, 36(8): 26-29.
[8] 肖毅, 张林, 聂笑一. 基于 WEB 挖掘的网络爬虫设计与实现 [J]. 计算机系统应用, 2013(9): 60-63.
[9] 范珊珊, 李石君. 基于优先级队列的分布式多主题爬虫 [J]. 计算机工程与设计, 2015, 6(6): 1630-1636.
[10] Gupta S, Kaiser G E, Grimm P, et al. Automating Content Extraction of HTML Documents [J]. World Wide Web-internet & Web Information Systems (S1386-145X), 2005, 8(2): 179-224.
[11] Mane T B, Potdar G P. Template Extraction from Heterogeneous Web Pages [J]. International Journal of Advanced Computer Research (S2249-7277), 2012, 2(6): 2278-0181.
[12] 常育红, 姜哲, 朱小燕. 基于标记树表示方法的页面结构分析 [J]. 计算机工程与应用, 2004, 40(16): 129-132.
[13] Cattell R. Scalable SQL and NoSQL data stores [J]. Acm Sigmod Record (S0163-5808), 2011, 39(4): 12-27.
[14] 王利, 刘宗田, 王燕华, 等. 基于内容相似度的网页正文提取 [J]. 计算机工程, 2010, 36(6): 102-104.
[15] 殷彬, 杨会志. 灵活结构网页的正文提取 [J]. 计算机技术与发展, 2011, 21(9): 111-113.
[16] 张小欢. 中文分词系统的设计和实现 [D]. 成都: 电子科技大学, 2010.
[17] 姜维. 统计中文词法分析及其强化学习机制的研究 [D]. 哈尔滨: 哈尔滨工业大学, 2007.
[18] 胡金栋. 网页正文提取及去重技术研究 [D]. 杭州: 浙江大学, 2011.
[19] 梁正友, 欧杰, 俞闽敏. 基于图文有效信息量的网页正文定位 [J]. 计算机工程, 2011, 37(23): 276-278.

