论“中国地方志数字化关键技术研究与演示平台设计”项目的创新点

肖 禹

国家科技支撑计划项目“中国地方志数字化关键技术研究与演示平台设计”（2015BAK07B00）始于 2015 年 7 月，旨在以方志学、图书馆学理论为指导，在已有的地方志数字化成果上，以中文信息处理、数据库、GIS、数据挖掘、人工智能等先进技术为手段，探索下一代地方志数字资源系统范式，研究地方志数字化与资源服务的新技术、新方法与新模式，实现现代技术与传统文化的紧密结合。

本项目以 1912 年以前编纂或出版的地方志为研究对象，研究成果也适用或部分适用于于 1912 年至 1949 年间编纂或出版的地方志。本项目在以下几方面有所创新：

1.面向数据的地方志研究

传统方志学是研究方志现象运动规律的科学，研究方志的产生和发展、类别和功能，以及编纂理论。其研究的主要内容包括：方志的产生和发展、方志的性质和分类、方志的特征和功能、志书编纂理论、旧志整理和方志利用、方志批评和志书评论、方志和其他学科的关系等。

传统方志学研究往往很少关注方志数字化和方志数字资源应用研究，相关成果也非常有限。本项目以方志数字资源和数字化过程为研究对象，一方面研究方志数字化的基础性问题，着力解决地方志大规模数字化过程中的重要窒碍；另一方面基于已有的方志数字资源和软件工具，拓展数字资源的应用模式，探索新技术手段如何更有效、便捷地服务于传统方志学研究。

1.1 地方志数字化研究

1.1.1 地方志数字化用字分析

地方志数字化以汉字字符集为基础，我国先后发布了 GB2312-80（基本集，包含 6763 个简体字）、GB12345-90（第一辅助集，与基本集对应的繁体字）、GB7589-87（第二辅助集，包含 7237 个简体字）、GB/T13131-91（第三辅助集，与第二辅助集对应的繁体字）、GB7590-87（第四辅助集，7039 个简体字）、GB/T13132-91（第五辅助集，与第四辅助集对应的繁体字）、GB13000.1-93（包含 20902 个字）、GBK（包含 21003 个字）、GB18030-2000（包含 27533 个字）、GB18030-2005（包含 70244 个字）等字符集标准。

我国后续发布的字符集标准，兼容 Unicode 的 CJK 部分，而 CJK 收录简体汉字、繁体汉字、方块壮字、日本国字、韩国独有汉字、越南喃字等，并非完全的汉字字符集。目前，Unicode 的最新版本为 10.0.0，CJK 包括基本集、扩展 A 集、扩展 B 集、扩展 C 集、扩展 D 集、扩展 E 集和扩展 F 集，收字 87849 个。每个 CJK 文字有一个或多个来源，如 G、H、M、T、J、K、KP、V、MY、U 等，其中 G 表示中国大陆和新加坡，H 表示香港（特别行政区），M 表示澳门（特别行政区），T 表示中国台湾，J 表示日本， K 表示韩国，KP 表示朝鲜，V 表示越南，MY 表示马来西亚，U 表示 Unicode。CJK 文字来源统计如表 1 所示。

表 1 IRG 来源统计表

来源	基本集	扩 A 集	扩 B 集	扩 C 集	扩 D 集	扩 E 集	扩 F 集	合计
G	20913	6192	30525	1120	76	2814	1304	61640
H	15353	572	1702	1	0	0	0	17628
M	0	0	1	16	0	48	22	65
T	18370	5906	30178	1750	24	1257	0	57485
J	12563	738	303	367	107	415	1645	14493
K	15391	1835	166	404	0	0	1793	17796
KP	15011	3189	5766	8	0	0	0	23974
V	4757	308	4231	785	0	1028	0	11109
U	13	13	52	81	19	227	2885	405

依据汉字的传播历史，将来源 G、H、M、T 视为汉字，在 87849 个 CJK 文字中，汉字有 74340 个，占 84.62%。上述数据基于 CJK 来源数据统计获得，未经过严格的字源考证与学术论证，只能作为参考。

相对于字符集有了集外字的概念，集外字是字符集所不包含的文字，若不采用其他的技术和方法，集外字无法输入、处理和显示。集外字的数量与字符集的收字数量直接相关，若数字化对象的用字总量和文字处理规则固定，字符集收录的文字越多，集外字的数量越少。以国家图书馆地方志文本化项目为例，该项目使用 Unicode 字符集中的 CJK 基本集、扩 A 集和扩 B 集，用字量和字频统计如表 2 所示。相对于基本集文字 99%以上的使用频率，虽然扩 A 集文字的使用频率仅为 0.13%和 0.17%，扩 B 集文字的使用频率仅为 0.36%和 0.47%，但是项目的文字错误率要求为 3%%，若不使用扩 A 集和扩 B 集，不对这些文字做处理，那

么项目的文本化率（文本数据的字数与原书实际字数的比值）不超过 99.5%，即原书中的 0.5%文字未文本化，那么 3%%的错误率也就没有多少实际意义。若不使用扩 A 集和扩 B 集，对这些文字做造字处理，那么造字总量将超过 2 万个。

表 2 地方志文本化项目一期二期用字统计表

	地方志文本化项目第一期	地方志文本化项目第二期
总叶数	506485 叶	400349 叶
总字数	204808490 次	161222529
基本集汉字使用个数	16801 个	17123 个
基本集汉字使用次数	203781248 次	160135692 次
基本集汉字使用频率	99.50%	99.32%
扩 A 集汉字使用个数	2959 个	3265 个
扩 A 集汉字使用次数	274847 次	279401 次
扩 A 集汉字使用频率	0.13%	0.17%
扩 B 集汉字使用个数	9117 个	9353 个
扩 B 集汉字使用次数	732675 次	759774 次
扩 B 集汉字使用频率	0.36%	0.47%
集外字使用个数	4866 字	3009 字
集外字使用次数	19720 次	47662 次
集外字使用频率	0.01%	0.03%

通过上面的分析不难看出，地方志中有大量的异体字，即使采用 Unicode 的最新版本，仍然有大量的集外字，必须引入有效的集外字描述方法。同时，Unicode 并非规范汉字集，为了检索、数据分析、数据挖掘等后续应用，还要考虑文字规范问题。

1.1.2 构建地方志文本数据模型

文本化是将各种载体的文献转化成文本数据的过程，相对于索引数据、书目数据和图像数据，文本数据既能揭示文献包含的绝大部分信息，又能支持检索和显示。文本数据的复杂性远高于书目、索引、图像等类型的数据，是数字化研究与实践的重点和难点。同时，文本数据是检索、显示、信息抽取、数据分析、数据挖掘等一系列后续应用的基础，文本数据的

内容、格式、数据质量、描述方式等都会对后续应用有直接的影响。

地方志文本是信息的载体，这些信息用文字、符号、图形、图像等形式表示。文字是地方志内容的重要载体，地方志中的文字以汉字为主，还有满文、蒙古文、藏文等少数民族文字，日文、韩文、拉丁文、英文等外国文字。在当前技术条件下，文字描述以字符集为基础，字符集内的文字用编码直接表示，而集外字（字符集内未包含的文字）必须用其它方法处理。同时，文字可以用字体、字号、位置、颜色、变形、旋转等属性来描述。

除了文字之外，地方志中还有大量的符号。中国古代各个历史时期的符号和符号的用法在不断地发展变化中，符号形式多样，用法也各不相同，与文字共同构成复杂的表意系统。地方志中的符号大致可以划分为标点符号、校对符号、版式符号、专类符号等几类。在地方志文本化过程中，符号的处理方式与文字类似，字符集内的符号可以直接使用，字符集中未包含的符号要采用其他方式处理。符号也具有大小、颜色、变形、旋转等属性。

地方志中的常见图形包括线段、圆弧、圆形、矩形等，通常与文字、符号一起使用。线段用起点、终点表示；圆弧用起点、圆心角、圆心、半径表示；圆形用圆心、半径表示；矩形用对角顶点表示。图形可以用颜色、线宽、线形等属性来描述。

图像是地方志中包含的一类重要内容。中国书籍之有图，历史悠久，古人著书有左图右书，左图右史之制，“文不足，以图补之；图不足，以文叙之”，图文并茂，相辅相成，是中国书籍的优良传统。地方志插图按成图方式可分为写本、刻本、石印本、珂罗版印本，还有少量的套印本、铜版画和照片。在地方志文本化过程中，图像大致划分为三类，包括版框内插图、书叶内插图和其他插图。图像具有尺寸、分辨率、颜色模式等属性。

此外，地方志文本化对象还包括大小字、墨围、墨盖子、表格、图形组合、牌记、印章、版式等。

依据地方志文本化的范围可先将地方志文本数据划分为全文文本和部分文本两类，再依据文本化的特征将全文文本划分为纯文本、位置文本、版式文本、语义文本等几类，如图 1 所示。

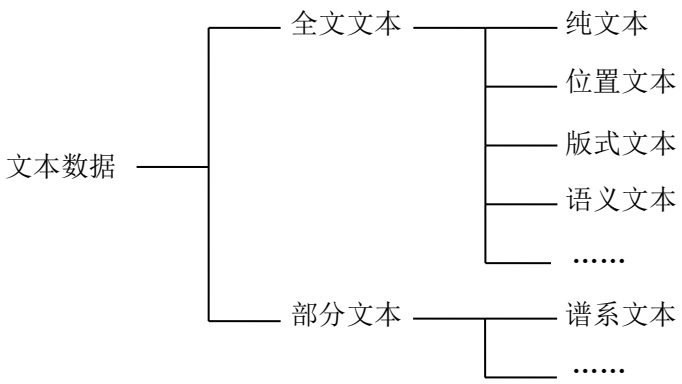


图 1 地方志文本数据分类示意图

全文文本是文本化范围为整部地方志的文本数据;部分文本是文本化范围为地方志中特定内容的文本数据，下面将重点讨论纯文本、位置文本和版式文本。

纯文本是只包含文字和非修饰性、非格式控制（回车符和换行符除外）符号的文本数据。纯文本是最早出现的古籍文本数据形式，其结构简单，可用任何文本编辑器创建、查看或修改。随着计算机和软件技术的发展，新的古籍文本数据形式很快出现，并逐步取代纯文本。但是互联网的出现对数据的通用性、易用性、跨平台等方面提出了新的要求，纯文本数据完全符合上述要求，很快成为网络古籍文本资源的主要形式。而随着电子阅读器、智能手机、平板电脑等移动设备的普及，纯文本又成了移动设备首选的古籍文本格式。“汉籍电子文献资料库”、“CBETA 电子佛典集成”、“中国基本古籍库”、“古籍电子定本工程”等均为纯文本数据。

位置文本是只包含文字和非修饰性、非格式控制（回车符和换行符除外）符号和位置信息的文本。位置信息是文字或符号在古籍图像上的坐标信息，通常描述为矩形区域的两个对角顶点。与纯文本格式相比，位置文本格式加入了位置信息，文本数据与图像数据建立了一一对应关系。位置文本主要应用于双层 PDF 和双层 DjVu。

版式文本是具有版式结构化描述的文本。版式文本数据以版式描述为基础，完整描述了古籍的内容信息、版式信息、结构信息等，是专业古籍文本数据库采用的主要数据格式。目前，相对于其他类型的文本数据，版式文本的应用实例最多、最成熟，尤其是大规模应用型项目。书同文“文渊阁四库全书电子版”、“爱如生大型古代数据库”、国家图书馆“数字方志”等项目各具特色，《中文文献全文版式还原与全文输入 XML 规范》是唯一公布的文本数据标准规范。

纯文本、位置文本和版式文本都属于全文文本，但描述方式各不相同。纯文本最为简单，位置文本是在纯文本的基础上加入位置信息，版式文本是具有版式结构化描述的文本，三者

的区别如表 3 所示。

表 3 文本数据对照表

描述对象 \ 文本类型		纯文本	位置文本	版式文本
内容对象	文字	基于字符集描述文字字形，集外字用字符串描述	基于字符集描述文字字形，集外字用字符串描述，同时描述文字的位置	基于字符集描述文字字形，集外字用字符串、图像等描述，同时描述字体、字号、位置、颜色、变形、旋转等属性
	符号	基于字符集描述标点符号和专类符号，集外符号替换为“●”（U+25CF），不描述校对符号和版式符号	基于字符集描述标点符号和专类符号，集外符号替换为“●”（U+25CF），同时描述符号的位置，不描述校对符号和版式符号	基于字符集描述符号，集外符号用字符串、图像等描述，同时描述大小、颜色、变形、旋转等属性
	图形	可描述为“[图]”或不描述	只描述图形中的文字或符号，同时描述文字或符号的位置	用起点、终点描述线段，用圆心、半径描述圆形，用半径、起点、终点描述圆弧，用左上顶点、右下顶点描述矩形、用顶点描述边形等，同时描述线形、线宽、颜色、填充颜色等属性
	图像	可描述为“[图]”或不描述	只描述图像中的文字或符号，同时描述文字或符号的位置	作为一个整体描述，描述位置、来源、层等属性
	大小字	用符号或标签简单区分大小字	用标签和属性简单区分大小字	用标签和属性描述大小字，同时描述字体、字号、

				位置、颜色、变形、旋转等属性
	墨围	用符号或标签简单描述	用标签和属性简单描述	墨围中的文字按大小字描述，墨围按图形描述
	墨盖子	用符号或标签简单描述	用标签和属性简单描述	墨盖子中的文字按大小字描述，墨盖子按图形描述
	表格	可描述为“[表]”或不描述	只描述表格中的文字或符号，同时描述文字或符号的位置	表格中的文字按大小字描述，表格按图形描述
	图形组合	可描述为“[图]”或不描述	只描述图形中的文字或符号，同时描述文字或符号的位置	图形组合中的文字按大小字描述，其他按图形描述
	特殊图像（牌记、印章等）	可描述为“[图]”或不描述	只描述图像中的文字或符号，同时描述文字或符号的位置	作为一个整体描述，描述位置、来源、层等属性
	版式	不描述	不描述	版框、版心、天头、地脚、界栏、鱼尾、象鼻、书耳等，其中的文字按大小字描述，其他按图形描述
结构对象		每个文件对应书、卷或多叶，头文件包含书目信息	以叶为描述单位，每个文件对应一个或多个古籍图像，头文件包含书目信息和卷目信息	以叶为描述单位，每个叶文件对应一个古籍图像，文件头对应一种书，包含书目信息和卷目信息

1.2 基于数据的地方志定量研究

在传统方志学领域中，无论是理论方志学、方志学史，还是方志编纂学，都不可回避方志的内容及其发展变化。方志学论著在讨论上述问题时，往往采用定性描述或部分列举的方

式，如地方志的内容不仅包括各地区的疆域、气候、山川、物产等地理资料，也涵盖户口、人物、赋税、艺文等人文历史各方面的记载，是地方的百科全书，一地之全史。

本项目上述问题进行定量研究，以国家图书馆馆藏地方志卷目数据（包含地方志 6868 种，卷目数据 461924 条）为基础，通过统计卷目数据的使用频率，分析地方志的内容特性，结果如表 4 所示。

表 4 地方志卷目数据使用频率统计表

No	卷目	数量	No	卷目	数量
1	人物	8442	41	形勝	1179
2	藝文	7813	42	食貨	1152
3	選舉	4519	43	記	1135
4	目录	3956	44	壇廟	1074
5	職官	3686	45	橋梁	1050
6	古蹟	3628	46	武備	997
7	列女	3599	47	兵防	990
8	學校	3535	48	官師	985
9	山川	3364	49	祀典	970
10	建置	3221	50	秩官	954
11	風俗	3137	51	祠祀	951
12	疆域	3061	52	舉人	941
13	序	2950	53	文苑	906
14	沿革	2787	54	津梁	906
15	城池	2683	55	圖	843
16	田賦	2331	56	倉儲	802
17	名宦	2084	57	傳	788
18	寺觀	2070	58	祠廟	762
19	星野	1887	59	雜志	756
20	輿地	1845	60	宦績	755
21	公署	1842	61	忠義	751
22	戶口	1841	62	賦	718

23	物産	1820	63	氣候	694
24	目錄	1814	64	孝義	691
25	凡例	1789	65	兵制	683
26	列傳	1637	66	武職	681
27	地理	1435	67	方技	668
28	祥異	1399	68	封贈	664
29	詩	1398	69	陵墓	658
30	水利	1395	70	武科	658
31	仙釋	1346	71	薦辟	647
32	賦役	1300	72	宦蹟	631
33	流寓	1255	73	儒林	628
34	金石	1251	74	雜記	618
35	孝友	1248	75	學宮	617
36	進士	1228	76	鹽法	609
37	物產	1225	77	形勢	601
38	書院	1196	78	忠節	590
39	坊表	1195	79	災祥	587
40	隱逸	1191	80	義行	584

2.大规模地方志数字化

据统计，现存的旧志（1949 年以前编撰或出版，含专志）超过 1.5 万种、3 万个版本，约 1 千万筒子页。目前，已有三分之二完成图像采集，而文本化不超过三分之一。作为重要的学术资源，地方志数字化的数量与质量都不能满足学界的要求。要解决上述问题，实现地方志大规模数字化，就必须建立完善的标准规范体系，突破文字识别等关键技术，形成高效的地方志数字化工艺流程。

2.1 构建地方志数字化加工规范体系

本项目研制 8 个数据加工规范，包括元数据加工规范 2 个，地方志专门元数据规范和地方志卷目数据标引规范；对象数据加工规范 2 个，地方志图像数据规范和地方志文本数据规范；知识数据加工规范 2 个，地方志语料数据规范和古今地名数据规范；数据加工所需数据

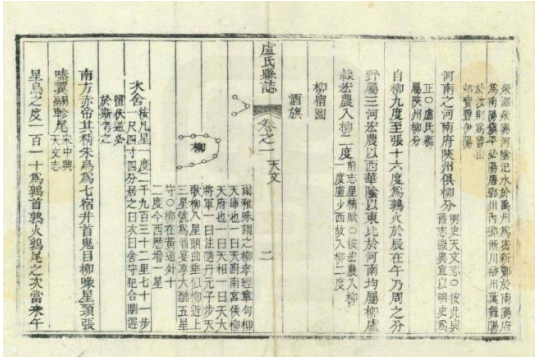
规范 2 个，包括汉字集外字描述规范和文字认同描述规范。

地方志专门元数据规范用于整体描述地方志图像数据和文本数据，内容包括书目信息、数字化参数、版权信息、用户使用提示等；地方志卷目数据标引规范用于描述地方志图像数据的卷次结构，内容包括卷目信息、层次结构、分类、页码等；地方志图像数据规范用于描述地方志图像及规范图像采集方法，内容包括图像类型、画幅内必备要素、图像参数、数字化设备参数、图像压缩、色彩管理等；汉字集外字描述规范用于描述地方志文本数据中的集外字，内容包括集外字定位、集外字图像（或图形）、IDS 描述等；文字认同描述规范用于描述地方志文本数据中的文字认同情况，内容包括认同字定位、认同方式、认同类型等；地方志文本数据规范描述地方志文本数据的基本结构和全文文本描述方式，内容包括文本数据格式、文字描述、版式描述、图形图像描述等；地方志语料数据规范描述地方志语料数据的基本结构和标引方式，内容包括文本、文字描述、版式描述、图形图像描述等；中国古今地名数据描述规范是基于已有研究成果，规范古地名的名称，并与今地名进行映射，规定了中国古今地名数据的内容、构成要素及各要素的描述规则，适用于中国古今地名数据库的建立及格式交换。8 个地方志数字化加工规范和应用指南作为本项目研究成果结集出版。

在 8 个地方志数字化加工规范中，项目组经过反复讨论，最终选取其中 4 个申报文化行业标准规范，包括中国古今地名数据描述规范、文字认同描述规范、汉字集外字描述规范和地方志文本数据规范。在行业标准草案申报过程中，文化部全国图书馆标准化技术委员会的专家认为“地方志文本数据规范”的适用范围较小，应将适用范围扩展到汉文古籍。项目组经过反复研究，认为地方志具有汉文古籍的普遍特征，对“地方志文本数据规范”进行必要的修改，最终申报“汉文古籍文本数据规范”。

2.2 版刻文字识别技术

文本化是地方志数字化的核心，而 OCR（Optical Character Recognition，光学字符识别）是地方志文本化的核心技术。不同于现代文献，旧志年代久远、用字复杂、字形多变、版式各异，使得普通 OCR 难以满足旧志文本化的需求。在旧志中，刻本超过 80%，因此，本项目研究的重点之一即版刻文字识别。



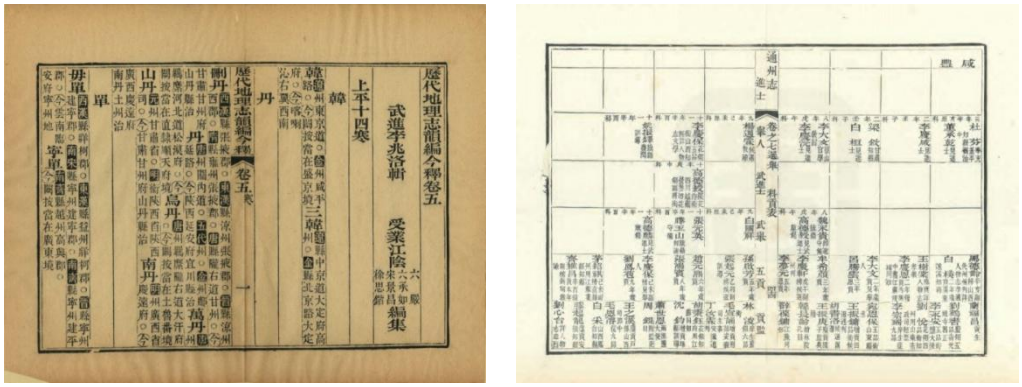


图 2 地方志样例图

为了解决旧志版刻文字识别问题，项目组采用了基于深度学习的单字符识别的技术路线，使用 Caffe 深度学习框架，该框架是开源的，核心语言是 C++，支持命令行、Python 和 Matlab 接口，速度快，利用了 MKL、OpenBLAS、cuBLAS 等计算库，支持 GPU 加速，非常适合做二维图像数据的特征提取。

要实现 UNICODE8（80376 个字符）的单字识别核心，常用字为明、清、民国版刻以及现代印刷体。针对一万多的常用字，可以直接从扫描图像中提取出字符图像作为样本，提取出的字符图像个数至少有 600 个，基本满足训练的要求。对于非常用字，提取出的图像个数较少，因此通过提取不同字号，不同字体的字模图像来满足训练的要求。因此，每个字符抽取的样本个数的在 3000 和 200 之间，按照 5:1 的比例拆分成训练集和验证集。针对八万多的字符集，抽取出的训练集样本总数为 2000 万左右，验证集样本总数为 430 万左右。

本项目采用 CNN（Convolutional Neural Network，卷积神经网络）模型，如图 3 所示。data 层是网络输入层，每张图片都被缩放成 64×64 大小的图像，以单通道的灰度图像输入。本项目采用了四个卷积层，conv1、conv2、conv3、conv4 都是卷积层，conv1 卷积层用 5×5 的大小的滤波器，卷积步幅为 2，两头补齐 2 个像素点，本层共有 96 个卷积滤波器，本层的输出则是 96 个 32*32 大小的图片。所有卷积层都跟着一个 BatchNorm 层、Scale 层。conv1、conv2、conv3 卷积层后都跟着一个 relu 层和 pooling 层，conv4 层后只能了 relu 层。Pooling 层都采用了 MAX 的方式，滤波器大小为 2*2，步长为 2。Conv2 卷积层用 5×5 的大小的滤波器，卷积步幅为 2，两头补齐 2 个像素点，本层共有 256 个卷积滤波器，本层的输出则是 256 个 16*16 大小的图片。Conv3 卷积层用 3×3 的大小的滤波器，卷积步幅为 1，两头补齐 1 个像素点，本层共有 384 个卷积滤波器，本层的输出则是 384 个 8*8 大小的图片。Conv4 卷积层用 4×4 的大小的滤波器，卷积步幅为 1，本层共有 1024 个卷积滤波器，本层的输出则是 1024 个 1*1 大小的图片。fc6、fc7、fc8 为全连接层，fc6 和 fc7 层后都跟着一个 relu

层和 Dropout 层。BatchNorm 层是整个训练过程中比较关键的一个网络层，BatchNorm 简称 BN，是对数据做一个归一化的操作，归一化到均值为 0、方差为 1，然后再送入下一个网络。归一化的操作主要是起到了调整网络输入分布的作用，使得训练过程中可以很快的收敛。如图 4 所示。

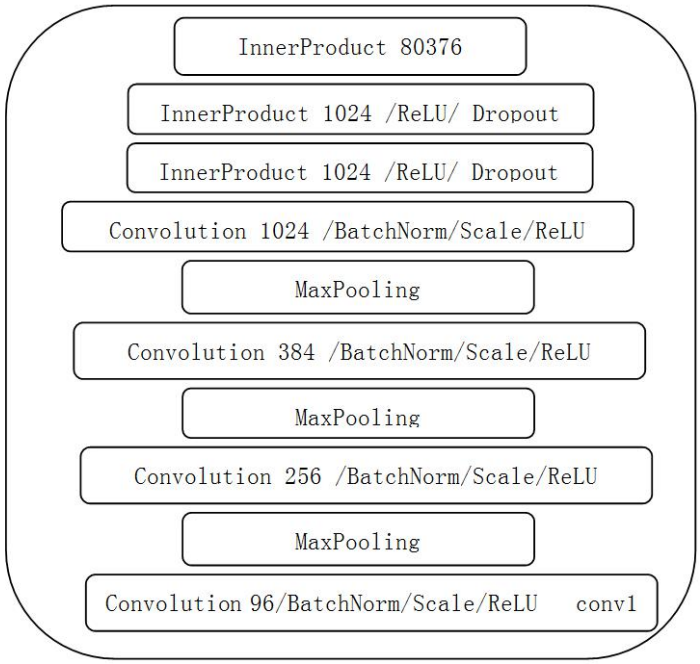


图 3 CNN 网络模型结构图

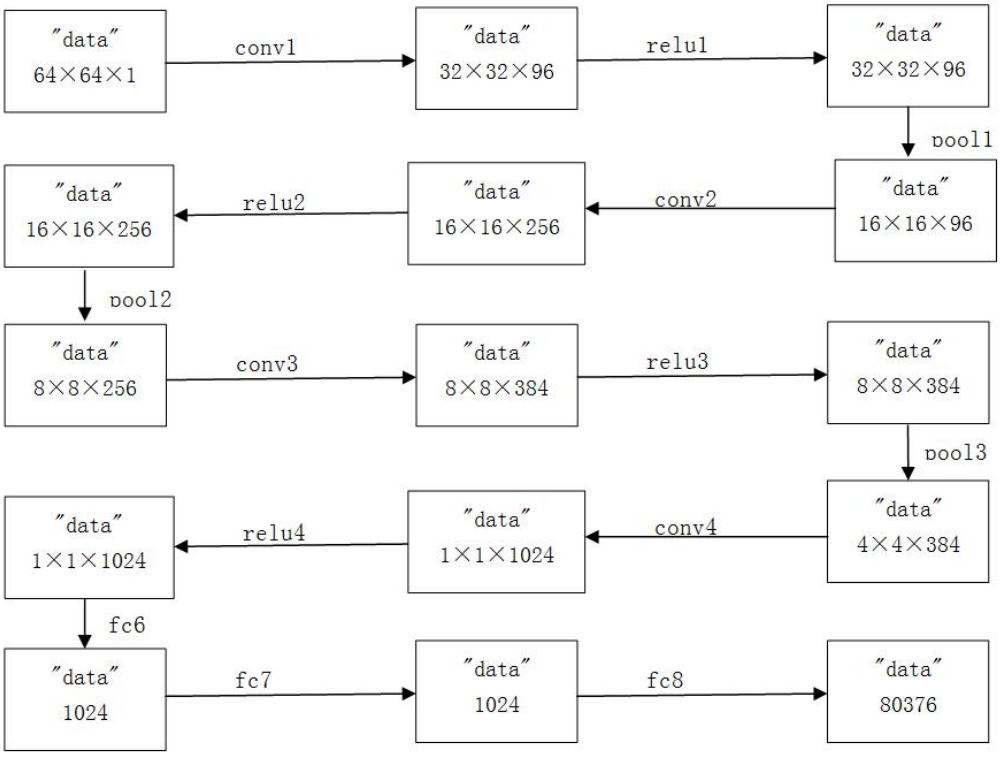


图 4 深度 OCR 数据流图

本项目采用 SGD（随机梯度下降）的方法来实现网络参数的优化、更新。它是利用扶梯度和上一次权重的更新值的线性组合来更新权重。训练过程中设置 Batch_Size 为 256，迭代 340000 次后，训练趋于平稳。版刻文字识别核心性能如下：

（1）支持 Unicode 8.0， 共 80376 个字符。

（2）识别率统计如表 5 所示。

表 5 识别率统计结果表

类型	总字数	正确识别字数	识别率
版刻汉字样本	2732791	2621370	95.9%
生成汉字样本	68897	67520	98.0%

2.3 优化地方志文本化工艺流程

要实现地方志大规模数字化，就必须有完善、合理、高效的工艺流程。项目组将地方志文本化过程划分为图像处理、版式切分、文字识别、文字校对、集外字处理、数据标引等工序。同时引入流水线的思想，将每个工序软件工具化，并将图像扫描工具、版式工具、自动识别工具、字符切分工具、纵校工具、横校工具、数据合并工具、成品导出工具等集成到文本化平台中，如图 5 所示。

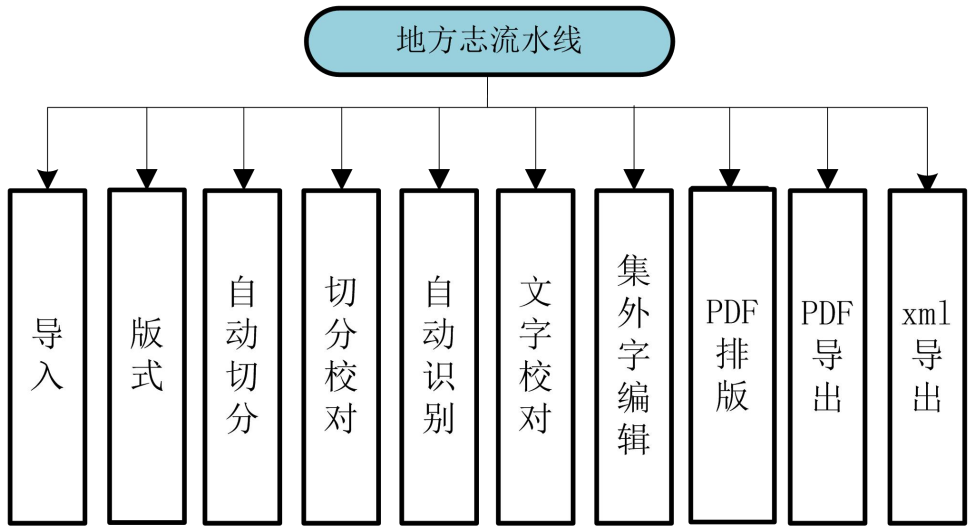


图 5 地方志加工流水线示意图

地方志文本化平台采用 C/S 软件架构，总体上为三层结构，如图 6 所示。持久化层，包括数据库服务器和存储服务器集群，用于存储管理数据和文件数据，其中数据库服务器存储工艺、项目、任务等管理信息，存储服务器用于存储加工的文件数据等；业务逻辑层为应用

服务器集群，每个应用服务器运行业务逻辑组件，从持久化层获取信息和数据，用户表示层提供管理接口和调用接口；用户展示层，包括加工平台系统和工具集，该平台将管理系统和工具集独立开来，方便模块化和版本维护等。为了使系统具有高并发、高性能和高扩展性，本方案在具体的系统部署上采用了集群相关的技术。

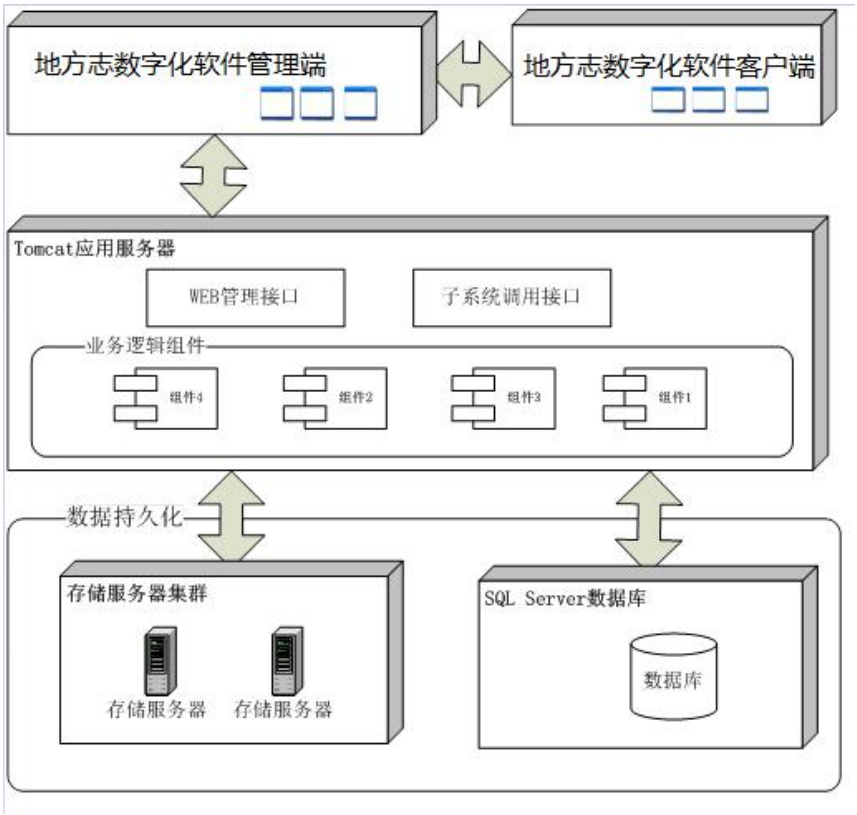


图 6 方志数字化软件架构图

地方志文本化平台具有以下特性：先进性，采用先进成熟的、业界主流的技术和产品为基础，在一段时间内保持技术的先进，并具有良好的发展潜力，以适应未来信息化发展和技术升级的需要；实用性，设计和实现需要考虑到本机构的应用场景定位、特点、服务对象、目前和未来的实际需求，建立一个可用性强、有人文社科特色的信息系统；安全性，包括设备、平台和数据安全，要确保各种设备运行稳定和容错，当发生故障时整个系统或主要应用系统能保持运行和服务，要建立完善的安全保障体系确保平台的正常运行、故障恢复和数据资源的容灾；可扩展性，系统必须是可扩展的，当业务发生变化或发展时无论是硬件设施，还是平台系统都能在原有投入的基础上进行扩展；开放性与互连性，网络系统要具备与多种协议计算机通信网络互连互通的特性，确保网络系统基础设施功能的充分发挥，应用平台系统和数据库系统要遵循开放性原则，以便实现系统间的数据共享和整合、以便进行应用功能的二次开发，系统应采用基于构件的开发方法，以提高模块的可复用性和可替换性，支持系统的演化式开发；经济性，应以较高的性能价格比构建整个系统，使资金的投入产出比达到

最大值，能以较低的成本、较少的人员投入来维持系统运行，提供高效能和高效益，尽可能保留并延长系统的投资，减少资本与技术投入方面的浪费；易用性，整个系统应该是使用方便、维护简单，网络系统等硬件系统具有高性能的管理平台，以便进行系统检测、监控和系统维护，应用平台系统具有良好的人性化界面和很高的自动化水平；标准化，强调各个系统都必须满足或遵循相关的国际标准、国家标准和行业标准；高度集成，强调各个系统不是孤立的或完全独立的，它们之间存在着各种关联性，通过集成和整合技术，形成一个多层次的、关联的整体。

3.地方志知识抽取与数据挖掘

3.1 地方志知识抽取

目前在知识抽取领域，常用的方法有两种：基于统计机器学习的方法和基于规则的方法。在本项目中，我们采用基于规则的方法实现地方志领域知识抽取。首先构建知识本体，本体内部包括概念和模板，并且每类知识对应几个模板。本体构建过程可以高效融入多个专家的知识，领域专家及语言学家直接参与知识本体的构建，并可以在后续过程中持续补充完善。然后对知识本体进行解析，在此基础上对地方志文本进行知识抽取，得到文本所包含的元数据及其类别。最后对初步抽取的结果进行补全处理和指代消解，并用于完善本体库。

项目构建的语义分析器首先解析本体，生成各种知识属性的规则表示。通过对本体中的概念和模板的解析，对每个概念，解析成正则表达式；对每个模板，解析概念间的关系，生成对应的规则表示。然后对地方志文本进行数据挖掘，得到时间、地点、人物、事件等地方志元数据。

知识本体中的规则由多个子规则用“+”符号连接而成，子规则由概念和符号组成，其中可能出现“{}”、“ˆ” 、“#”三个符号：

其中“{}”需要与数字进行匹配使用，可以出现在概念后面，用来对该概念可能出现的次数进行描述，如：{2}表示该概念会出现 2 次，{1,4}表示该概念会出现 1-4 次。“{}”符号也可以有{1000}，{-1}两种形式单独出现，而不需要与概念一起出现，分别表示“匹配任意长度的字符串”和“匹配句首”。

“ˆ”表示该符号修饰的概念不会出现，如果出现，则不符合抽取要求。

“#”表示该匹配不显示在抽取结果中，如：<template name=″日″>#ˆ日期限制+数词{1,3}+日+#{1000}</template> 中“ˆ日期限制”和“#{1000}”所匹配的文字都不出现在最后的结果中，最终匹配的结果为“一日”，“十三日”之类的信息。

例如，对人物类别中各个属性的抽取方法如表 6 所示。

表 6 人物属性抽取示例表

类别	抽取属性	抽取方法	规则数	示例
人物	人物名称	预处理		
	字	使用规则	2	<template name="002">#^关系{3}+#^字前限制+#字+^字限制{1,2}+#{1000}</template> I匹配：丁敬字士安河間人由編修至大六年謫任
	号	使用规则	3	<template name="002">#号前缀+#号+#曰+^标点{1,3}+#{1000}</template> 匹配：李白字太白号曰太白居士
	所属书目	预处理		
	类目	预处理		
	人物描述	预处理		
	籍贯	使用规则	9	<template name="002">#字+#^字限制{1,2}+^籍贯限制{1,4}+#人+#{1000}</template> 匹配：丁敬字士安河間人由編修至大六年謫任
	相关地名	使用规则	14	<template name=″已有″>古地名+#职官</template> 匹配：……冀州县令……

根据概念和规则的定义，可以将描述同一概念的规则集合转化为有顺序的多条正则表达式，用正则表达式去匹配待抽取文字，在抽取过程中使用动态规划的方式寻找最合适的匹配方式，若匹配成功，则表示抽取到需要的信息，将该信息保存下来作为数据抽取的初步结果。如图 7 所示。

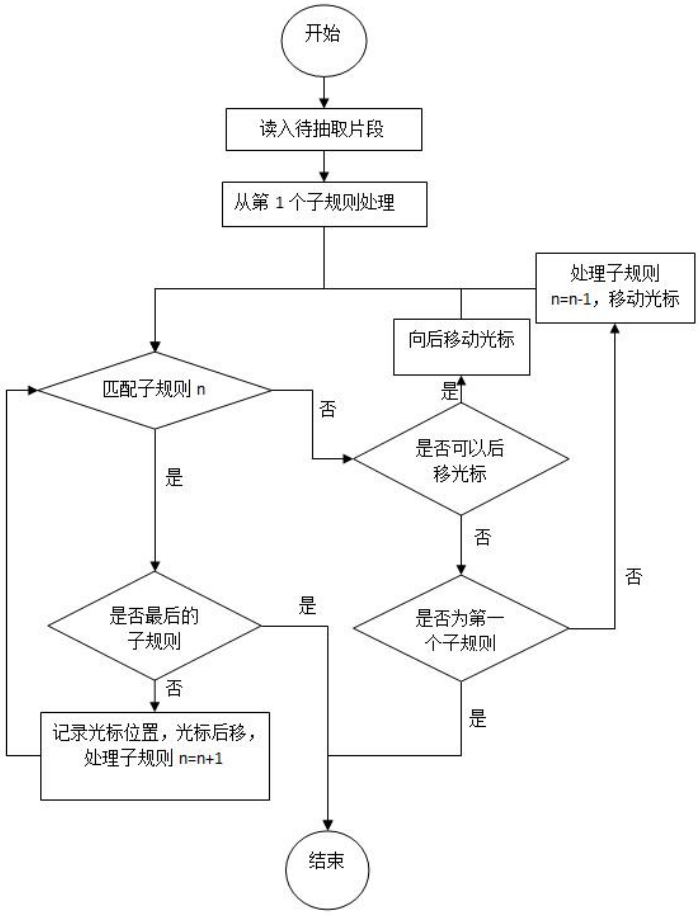


图 7 抽取器流程图

具体流程为：

（1）读取待抽取的片段。

（2）把光标置为第一个字的位置，从第一个子规则进行匹配。

（3）如果匹配成功，则匹配下一个子规则，直到所有规则匹配完成结束程序。

（4）若匹配不成功：

a.根据规则判断如果允许光标后移尝试匹配，则移动光标；

b.如果不允许光标后移，则回退到上一个子规则，后移光标重新匹配；

c.若不能后退，则匹配失败，结束程序。

抽取器完成工作后可以得到初步的处理结果，初步结果要想达到更好的使用效果还需要经过处理，主要是时间的补全处理和地点的指代消解。

事件知识的时间这一属性通过上下文进行补全，以一二行为例：第二行抽取出来的时间属性“三年七月”，其根据上下文推断出应补全为“咸宁三年七月”。此处时间的补全利用了地方志中事件记载的特点——按照时间进行记录，如果年号或年份与上一行相同则省略。如

图 8 和图 9 所示。

条目	描述		地名	时间
咸寧二年八月有星孛於太微至翼				咸寧二年八月
三年七月荆州大水			荆州	三年七月
四年荆州大水螟			荆州	四年
九月太白當見不見				九月
太康元年三月零陵泉陵白鹿見			零陵泉陵	太康元年三月
二年六月江夏大水殺人				二年六月
三年秋七月零陵令蔣微獲白鹿以			零陵	三年秋七月
四年冬荆州大水			荆州	四年冬
八年三月震災西閣楚王所止坊			西閣	八年三月
九年長沙地震			長沙	九年

图 8 抽取结果示例图

条目	描述		地名	时间
咸寧二年八月有星孛於太微至翼				咸寧二年八月
三年七月荆州大水			荆州	咸寧三年七月十
四年荆州大水螟			荆州	咸寧四年
九月太白當見不見				咸寧四年九月
太康元年三月零陵泉陵白鹿見			零陵泉陵	太康元年三月
二年六月江夏大水殺人				太康二年六月
三年秋七月零陵令蔣微獲白鹿以			零陵	太康三年秋七月
四年冬荆州大水			荆州	太康四年冬
八年三月震災西閣楚王所止坊			西閣	太康八年三月
九年長沙地震			長沙	太康九年

图 9 抽取结果后处理示例图

地名知识的与相关地点的方向关系、与相关地点的距离通过书名或者其他已知特征进行指代的消解，以图 10 第一行为例，抽取出的两个属性值为“縣東”，“縣東十里”，而这本地方志为灵寿县志，所以补全后为“灵寿縣東”，“灵寿縣東十里”。如图 11 所示。

条目	描述	条目名称		别名	方位	距离
衛河在縣東十里南入滹沱河禹貢曰恒衛既從漢書地里志曰靈壽中山桓公居此		衛河			縣東	縣東十里
慈河一名滋水在縣北五十里慈峪鎮北舊志發源山西靈縣自枚回山流入本縣境		慈河		滋水	縣北	縣北五十里
汊河在縣西北七十里發源阜平縣之白蛇嶺至义頭鎮與慈水合		汊河			縣西北	縣西北七十里
大明川在縣西北九十里府志云横山嶺西圍箔口東俗呼爲錦繡大明川按此即慈		大明川			縣西北	縣西北九十里

图 10 抽取结果示例图 2

条目	描述	条目名称		别名	方位	距离
衛河在縣東十里南入滹沱河禹貢曰恒衛既從漢書地里志曰靈壽中山桓公居此		衛河			灵寿縣東	灵寿縣東十里
慈河一名滋水在縣北五十里慈峪鎮北舊志發源山西靈縣自枚回山流入本縣境		慈河		滋水	灵寿縣北	灵寿縣北五十里
汊河在縣西北七十里發源阜平縣之白蛇嶺至义頭鎮與慈水合		汊河			灵寿縣西北	灵寿縣西北七十里
大明川在縣西北九十里府志云横山嶺西圍箔口東俗呼爲錦繡大明川按此即慈		大明川			灵寿縣西北	灵寿縣西北九十里

图 11 抽取结果后处理示例图 2

抽取结果可以用来完善知识本体，如：根据数据“梁德珪良鄉人仕至叅知政事奏對無不稱旨世祖嘗怪州郡囚數過多德珪對曰當國者急於徵索蔓延抆繫以致此耳世祖感悟因大赦”与规则可以提取出该人物籍贯为“良鄉”，那么“良鄉”作为一个提取出来的地名又可以完善本体库中已知地名。随着已知地名越来越多，那之后抽取的地名信息也会越来越准确。

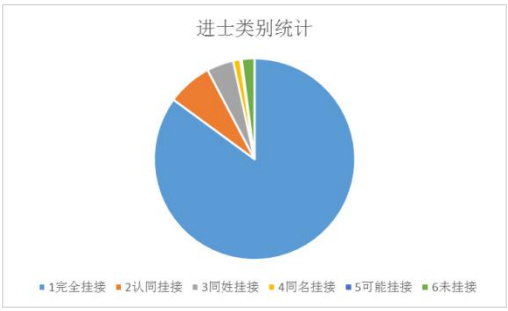
3.2 地方志数据挖掘

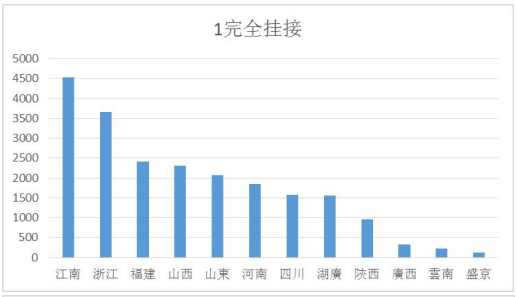
基于抽取系统获得的地方志知识库，项目组选取物产和人物中的进士部分进行了数据挖掘。数据挖掘的步骤包括：数据整理、统计分析、发现规律、发掘研究线索。数据整理是选

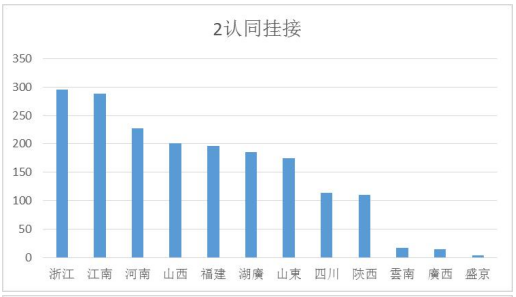
定要分析的主题，获得数据抽取的结果数据，对数据进行清洗，以及数据规范化；统计分析是针对选择的主题和数据选定要关注的角度，统计关心的数据，对统计的数据进行分析；发现规律是从统计分析工作中获得统计的结果数据，发现数据之间的联系，总结数据展现出来的规律；发掘研究线索是分析规律可能的原因，从而获得研究线索。

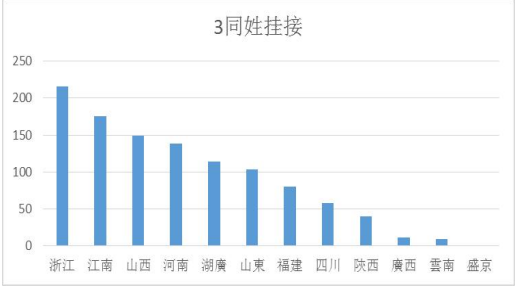
以人物数据中的进士部分为例，要明确姓名、籍贯、科榜、朝代等信息，进行必要的文字规范和主题规范，去除重复数据，最初人物库中包含进士数据 27834 条，数据整理后获得进士数据 25382 条。

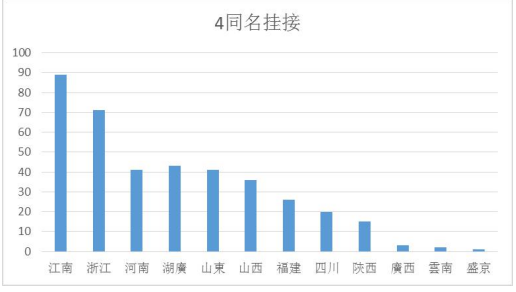
项目组以《明清进士题名碑录索引》为标准数据源，以姓名、年号、科榜、地域信息等为重要属性，用方志人物库中的进士数据进行匹配，这一过程称为挂接。根据属性一致情况分为如下几个类别：完全挂接，姓名完全一致，地域一致、科榜一致；认同挂接，姓名经过同音字认同后一致，地域一致、科榜一致；同姓挂接，姓一致，名不完全一致，但根据地域和科榜分析应该属于同一人的情况；同名挂接，姓不一致，名一致，根据地域和科榜信息分析属于同一人；可能挂接，姓名中有一致的字，根据地域和科榜信息分析可能属于同一人；未挂接，以上挂接条件均不能找到匹配人员的情况。结果如图 8 所示。











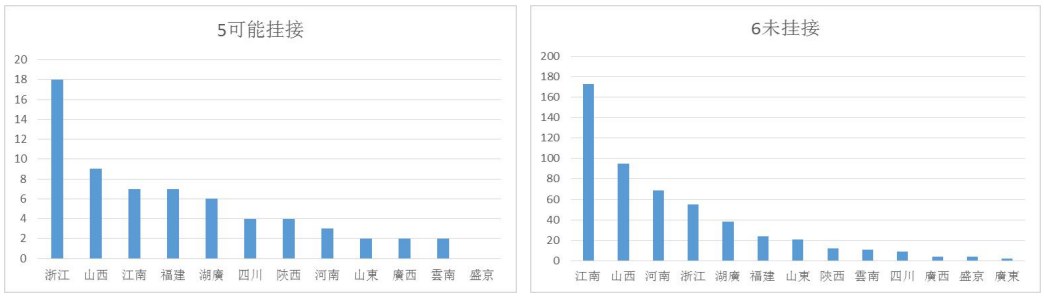


图 8 进士数据分析结果图

根据挂接类别可以看出，部分同一进士的姓名并不完全一致，经过分析有以下几种情况：同音同义字，如：王章燦-王章粲；特定字代替，如：楊尹賢-楊胤賢；姓发生变化，如：孟麟-臧麟，孟凤-臧凤。

经过分类得到了几种姓名不一致的情况，经过具体分析归纳得到以下几个可能的原因：异体字、避讳字和明朝复姓制度等。

项目组用相同的方法挖掘了物产数据，其中明朝物产 1313 种，清朝物产 1769 种，物产在地域维度和时间维度上的分布情况如图 9 所示

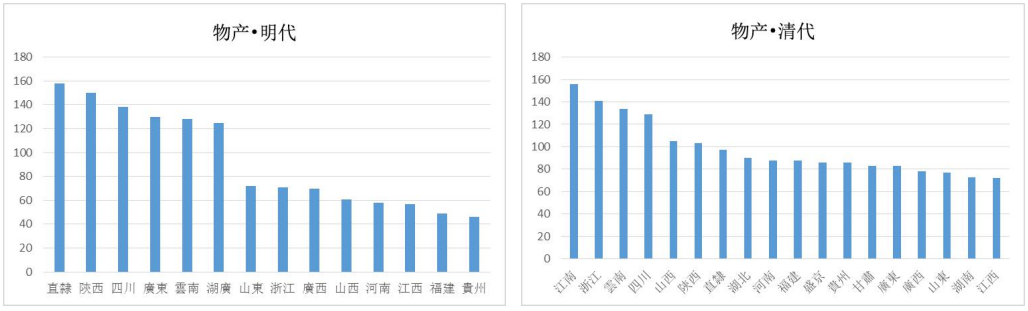


图 9 物产数据分析结果图

项目组进一步挖掘了物产中的纺织品种类与进士数量的关系，结果如图 10 和图 11 所示。

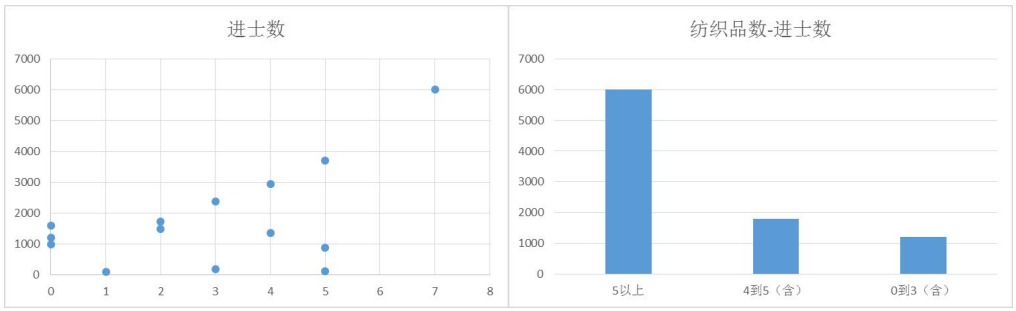


图 10 明代纺织品种类与进士数量分析结果图

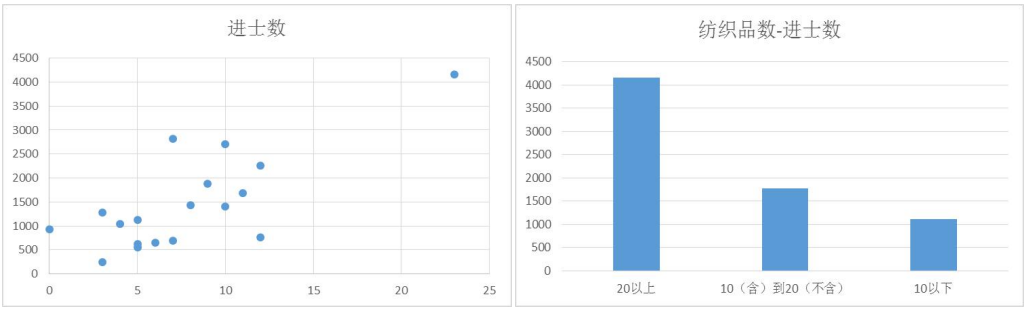


图 11 清代纺织品种类与进士数量分析结果图

通过上述结果不难看出，纺织品种类和进士数量成正相关，经分析原因可能有如下几种：纺织品丰富的地区可能更加开放，重视教育，从而影响了进士的数量；纺织品丰富的地区可能经济更发达，从而带动了教育；纺织业消耗的劳动力较少，从农业释放出来的劳动力可以专注于科举考试，中进士的几率更大。

4. 地方志可视化

地方志可视化是利用计算机图形学和图像处理技术，将地方志数据转换成图形或图像在终端屏幕上显示出来，并通过人机界面与用户进行交互。地方志是地方百科全书，内容丰富，形式规整，能够全面反映地方文化，既可以作为学术资源服务于人文社科领域的专业人员，又可以作为地方文化资源服务于社会公众。

4.1 面向专业用户的地方志可视化

对专业用户而言，地方志作为学术资源，要求内容准确详实，可以方便地收集、整理和使用同一主题的资源，而可视化系统只是作为学术研究的工具之一。

本体图谱是以一个本体概念或实例为中心，多个关联概念和实例为端点的网络。它用连接线展示概念和实例之间的关联性；用不同的可视方式区分概念和实例以及各种不同的关联性；概念和实例可用鼠标点击，点击后将该概念或实例作为检索关键词重新检索。项目组基于 Echarts 实现地方志知识库可视化，可提供直观、生动、可交互、可高度个性化定制的关系图谱，将数据之间的隐形关系多方向、多维度、可视化地展现出来。以人物关系图谱为例，展示效果如图 12 所示。

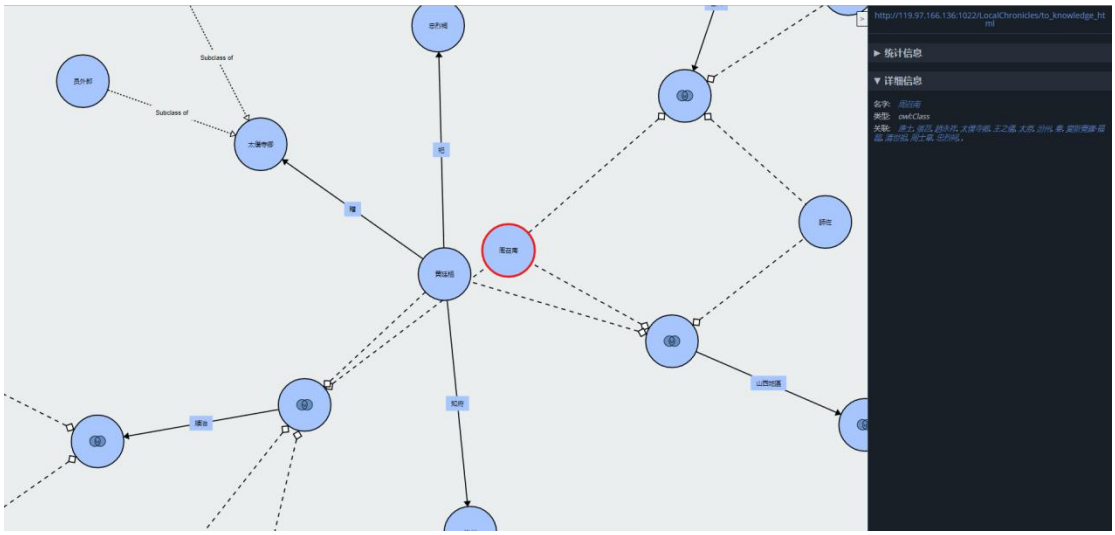


图 12 地方志可视化效果图

地方志中的大量数据都具有空间属性，而描述空间属性的信息绝大部分都属于抽象信息。地方志可视化系统作为表达抽象信息的有力工具，不仅给信息以直观、形象的表达，而且能够揭示信息间的关联。基于 WebGIS 的地方志可视化框架如图 13 所示。

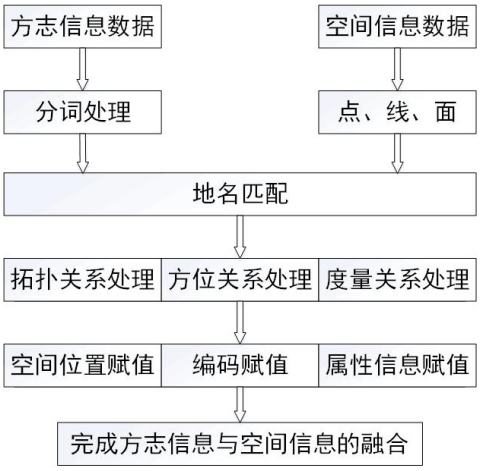


图 13 基于 WebGIS 的地方志可视化框架图

项目组基于 WebGIS 实现了地方志知识库中的人物、地点、事件、物产等数据的可视化。以地方志人物数据为例，展示效果如图 14 所示。

图 14 地方志可视化效果图 2

地方志知识库中的大量地点数据包含了地理位置信息的文本描述，如“宏济寺，在介休县西南二十二里高岩上”，项目组通过建立数据模型，结合 GIS 分析，可根据文本描述信息得出大致的位置，对古代不同时期的位置与现有地名进行自动或半自动匹配，经人工审核确认后即可实现方志数据与空间数据的匹配，提高数据定位的效率及准确性。定位效果如图 15 所示。

图 15 地方志可视化效果图 3

4.2 面向普通大众的地方志可视化

对普通大众而言，地方志蕴藏了丰富的历史文化信息，但是旧志用文言文撰写，内容晦涩繁复，让人望而却步。为了让“古籍里文字活起来”，项目组根据专家提供脚本（功能需求和内容设计），以地方志资料为基础，同时加入相关的文本、图片、音频、视频与动画等多种类型的地方文化资源，通过电子书、游戏、动画、集成多媒体、三维虚拟漫游等形式，实现地方志可视化，效果如图 16 所示。

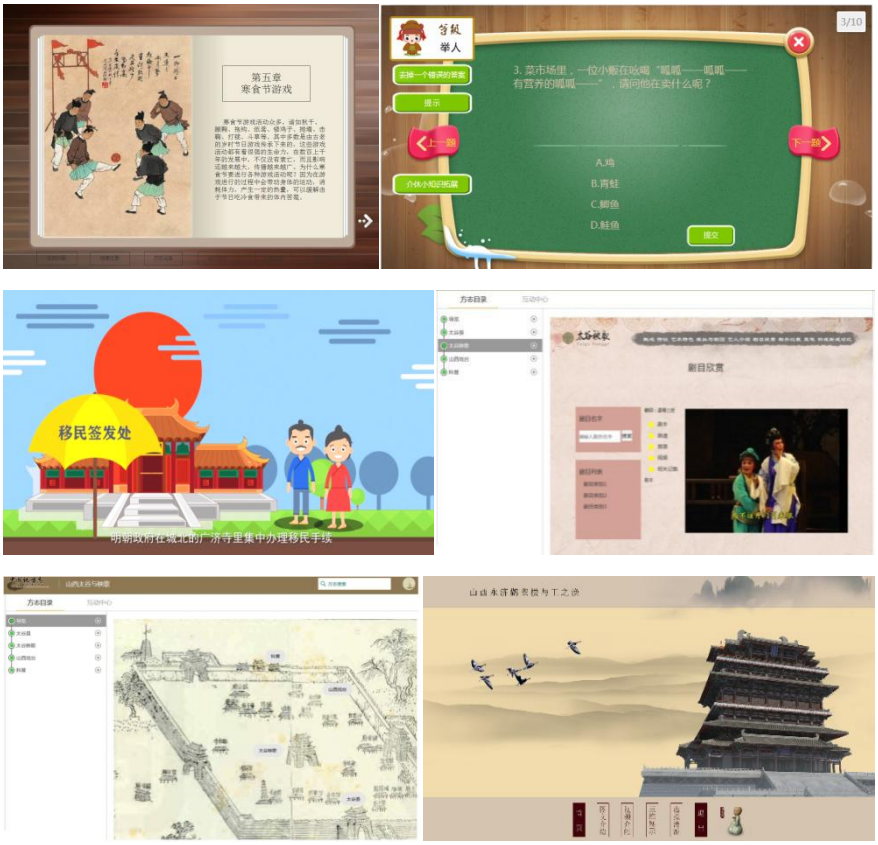


图 16 地方志可视化效果图 4

5. 地方志数据平台

本项目的演示数据包括清康熙时期纂修方志目录数据 1525 条；地方志图像数据 34 种 57595 筒子页；文本数据 34 种 57595 筒子页；古今地名规范数据 10023 条；语料数据 27101366 字，另外完成了山西介休与介子推多媒体电子书、山西洪洞与大槐树动画、山西太谷与秧歌音视频、山西介休与方言互动游戏和山西永济鹳雀楼与王之涣三维虚拟漫游场景。

传统的地方志数据发布系统数据与功能相互绑定，存在结构简单、功能单一、不易扩展、用户体验差等问题。而本项目数据类型多样，数据功能需求众多，且要满足海量数据处理的需求，必须重新设计平台架构。

地方志数据演示平台的总体架构如图 17 所示，分为应用层、服务层、数据层。数据层主要是对地方志资源进行存储，由于数据存储格式多样，按关系型和非关系型数据格式分开存储；由于地方志资源的数量庞大，采用 MongoDB 存储非关系型数据；地方志资源中的图像和视频采用云存储方式，以减少服务器的压力；由于平台的开放性，在用户使用过程中，会实时产生大量日志，日志文件的存储方式采用分布式文件系统进行存储；用户信息和地方志索引的数据量较少，但需要较高的访问速度，存储在 MySQL 中便于存取，可提高检索的效率。

服务层主要是对用户、数据和平台功能进行管理，同时向用户提供一系列数据应用和处理工具，向其他平台提供系统接口，向用户和其他平台提供数据接口。由于地方志数据演示平台采用数据与功能分离的设计理念，用户可任意调用工具集中提供的检索、浏览、数据分析、可视化等工具，也可以设定自己需要的数据集合，通过数据接口还可以导入外部数据。

应用层主要解决系统跨平台问题，可支持多浏览器，包括 Windows Internet Exporer8.0 及以上版本、Google Chrome10.0 及以上版本、Firefox4.0 及以上版本；支持多操作系统，包括 Windows7 及以上版本、iOS 6.0 及以上版本、Android3.2 及以上版本；在中低性能以上的终端环境下（Core 2 Duo CPU 2.33GHZ 及以上，Intel G33/G31 集成显示卡及以上，2G 内存及以上），能够可靠运行。

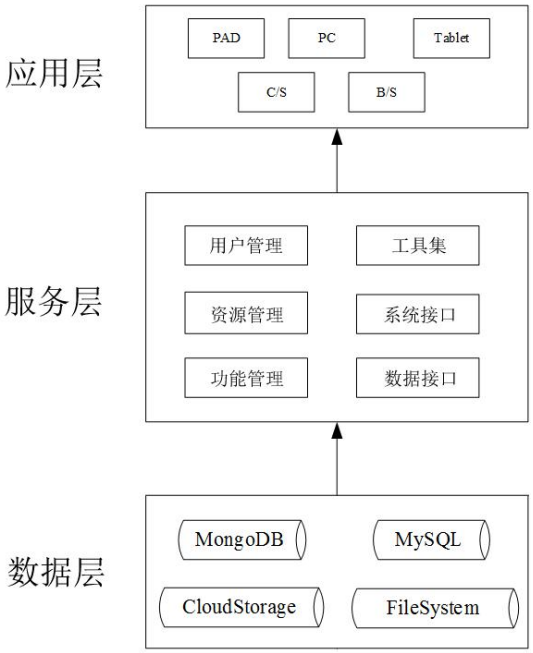


图 17 地方志数据演示平台架构图

6. 余论

旧志（1949 年以前编纂或出版的地方志）具备古籍一般的文献特性，基于旧志的数字化研究成果可直接应用于古籍。

第一代古籍数字化系统以目录、索引为核心，用户通过数据检索找到所需要的文献，通过阅读文献原本（通常为纸质文献）满足需求。第二代古籍数字化系统以文本、图像为核心，用户通过全文检索找到所需要的内容，直接调用文本或图像数据满足需求。目前，与第三代古籍数字化系统有关的研究还很少。

本项目从地方志文献特性出发，研究地方志数字化过程中的基础性问题，在此基础上建立标准规范体系，同时研究地方志数字化关键技术和工艺流程，解决地方志数据的低成本、

高质量、批量化生产问题；研究地方志数据抽取和数据挖掘技术，丰富地方志数据类型和数据处理方法，同时研究地方志可视化技术，增加地方志数据的应用方式。

随着人工智能技术日益成熟，大数据、云计算等技术普遍应用，我们希望能够解决地方志数据量产和数据抽取、数据挖掘、数据分析等问题，逐步构建开放式的数据平台，让用户平等地享有海量数据和多种数据工具，无障碍调用所需的资源，减少资料收集、资料整理、资料分析等工作的负担，最终形成开放式的社会科学领域研究平台，助力学术研究。