方志类古籍地名识别及系统构建

朱锁玲　包　平

摘　要　以地方志资料汇编《方志物产》(广东分卷)为语料,设计并构建了古籍地名识别系统。采用规则与统计相结合的命名实体识别方法,实现了物产地名的自动识别。分析了命名实体识别技术在中国方志类古籍整理中的应用前景,为方志类古籍进行数字化整理、挖掘物产分布、物产引进和传播等相关研究提供了新的途径。图3。表3。参考文献16。

关键词　地名识别　方志　命名实体　古籍数字化　古籍整理

分类号　G255.1;TP311.13

ABSTRACT　Based on the research about *Products in Local Chronicles of Guangdong*, this paper designs and implements a recognition system about location names in ancient books. Applying the Rules-based and Statistics-based Method, the authors have achieved the automatic recognition of products' location names. The authors illustrate the bright prospect of the application of the named entity recognition technique in digital collation of ancient books such as local chronicles and explore a new way for the collation of ancient books and the content mining of products' distribution, importation and dispersal. 3 figs. 3tabs. 16 refs.

KEY WORDS　Location name recognition.　Local Chronicle.　Named entity.　Digitization of ancient books.　Collation of ancient books.

CLASS NUMBER　G255.1; TP311.13

作为信息抽取中最有实用价值的一项关键技术,命名实体识别最初是在MUC-6(Message Understanding Conference)中作为一个子任务提出的[1]。国外有关英文命名实体识别的研究开始较早,并达到了较高的水平,MUC会议测试的准确率和召回率可达到97%左右[2]。目前中文命名实体识别的研究仍处于探索阶段,国内有关中文命名实体识别的研究主要集中于人名和地名[3-4],其应用涉及生物医学、电子产品、音乐等领域[5-7],研究针对的语料也主要是现代文献。古籍命名实体识别研究较少,仅有古典文献[8]和中医古籍[9]。

中国方志类古籍起源早、持续久、类型全、数量多。据《中国地方志联合目录》的统计,仅保存至今的宋至民国时期的方志就有8264种,11万余卷,占中国古籍的1/10左右。可见,中国方志无疑是地方文献中的大宗,它既具有丰富坚实的史料基础,更具备取之不尽、足资参证的史料价值[10]。本文以地方志资料汇编《方志物产》(广东分卷)为语料,探讨命名实体识别技术在方志类古籍内容挖掘中的应用前景。通过借用规则与统计相结合的命名实体识别方法,从中识别出物产的地名,构建物产地名识别系统,为物产分布、物产引进和传播等相关研究提供信息平台。

1　物产地名识别方法的选择

物产的地名是专有名词,属于命名实体的范畴。目前,命名实体识别的方法主要有规则方法、统计方法以及规则和统计相结合的方法[11]。规则方法,主要通过分析命名实体的内部和外部特征,人工构造规则模板实现命名实体的识别。统计方法,主要是针对命名实体语料库来训练某个字作为命名实体组成部分的概率值,并用它们来计算某个候选字段作为命名实体的概率,其中概率值大于一定阈值的字段为识别出的命名实体。规则与统计相结合的方法,是通过概率计算减少规则方法的复杂性与

盲目性,而且可以降低统计方法对语料库规模的要求。

方志类古籍中涉及的地名很多,但没有明确规范的地名定义。有些地名涵盖的地域范围宽泛,如"南夷"、"西域",有些地名则比较具体;地名的长度没有严格限制,短的如"广"、"粤",长的如"南海龙之都会新安龙穴洲";古籍中涉及的外国地名,大都是旧称,但对外国地名的翻译缺乏统一规范,如"颇稜国"与"颇陵国"、"交阯"与"交趾";时常多个地名一起出现,但有地名出现的地方,其文字表述的含义又不尽相同,既有说明某一物产原产地的,如"瓮菜本生东夷古伦国",又有说明物产现有分布地的,如"龙猪出南雄龙王岩在城东百里";既有说明物产引进传播情况的,如"番薯种自外洋吕宋移来由闽而广",也有说明该地区没有某一物产的,如"日月蚝今惠来等处有之揭无此物"。这些复杂的地名表述情况,加之古籍的书写又不分句读,大都没有标点符号,这就加大了地名识别的难度。已有的地名识别相关研究大都在进行地名识别之前先做分词处理,这就势必造成一些问题。如忽略了地名用字的特殊性,把地名用字等同于一般字做同样的简单分词处理,造成分词结果的错漏;当地名中含有常用词或地名与地名前后字组成常用词时,常用的分词方法还会降低地名识别的正确率[12]。

分析方志类古籍发现,古籍中地名的结尾常有地名特征词出现,如"国"、"府"、"州"、"县"、"郡"等,地名还常与一些介词、动词、方位词之类的指示词一起出现,如"丹竹出仁化"、"蕉布产潮州肇庆"、"安石榴种自涂林安石国得来"等,这些特征词和指示词即为命名实体识别方法中所谓的"规则"。我们可以提取这些规则,采用规则与统计相结合的命名实体识别方法来进行方志类古籍地名的智能化识别研究。

2　物产地名识别系统的设计

2.1　文档处理及物产粗分词

方志类古籍地名识别的基础工作是建立古籍的数字文档及数据库。本文所研究的语料中,物产的行文叙述格式多种多样,缺乏统一规范。借鉴前人的研究成果[13],对文本内容格式做规范处理,格式如下:

手抄本名称
属省序号
志书名称
成书年代
起始页码
序言
物产属名 1
　　物产名 1 说明文字(可有可无)
　　物产名 2 说明文字(可有可无)
　　……
物产属名 2
　　物产名 1 说明文字(可有可无)
　　物产名 2 说明文字(可有可无)
……
……
综论

按上述格式对文档进行规范处理后,每一种物产都分行列出,物产名和该物产的解释说明文字之间有一空格。这一过程实现了物产的粗分词,通过计算机切分和人工析取粗略地分出了物产名词。与此同时,也为数据库的设计提供了依据,为文档的批量入库做好了准备。

2.2　物产地名识别

借用规则为主、统计为辅的命名实体识别方法,选择方志类古籍为语料,具体识别步骤如下:

2.2.1　构建地名标引词典

方志中的地名大都是古代地名,参阅《古今地名对照表》、《古代地名通俗称谓大全》以及明、清和民国时期广东省行政区划等相关资料,收集、整理、统计古代地名,构建地名标引词典。

2.2.2 构建地名识别规则库

(1)选取清朝康熙 23 年至民国 32 年间的《大埔县志》、《埔阳县志》、《惠来县志》、《饶平县志》、《揭阳县志》等富含多种地名表述方式的志书作为训练语料,抽取并统计地名的上下文

信息,生成地名识别规则库(见表1)。

表1 物产地名识别规则库

左开右闭型规则	左闭右开型规则	两端封闭型规则
-出者佳	取之-	惟-有之
-产者佳	得之-	产-者佳
-多产	出-	出-者佳
-所产	种出-	种自-来
-所出者	本出-	自-得此种
-献	产-	自-而来
-贡	产自-	从-来
-进	生-	自-来
-入献	本生-	贩-间
……	……	……

表1中,"-"代表要识别的地名,根据地名在规则词中的位置(前、后、中间),把规则分为三种类型:左开右闭型、左闭右开型、两端封闭型。对于左开右闭和左闭右开这两种类型的规则,除规则外另截取5个汉字。对于两端封闭型,若中间词串长度不超过5个字符,则全部截取。

(2)选取其余部分广东方志作为测试语料,用规则库中的规则信息匹配测试语料中的物产解释,通过对匹配结果的统计分析,计算规则的频度,以此来判断规则的可信程度。

不同的规则在识别地名时,其正确率是不同的。为了表示规则的可信程度,引入规则频度这一概念。规则频度的定义如下:

$$F(R)=\frac{CorrectTime(R)}{AllTime(R)}$$

其中,CorrectTime(R)表示规则R识别地名正确的个数;AllTime(R)表示规则R识别地名总数。

(3)根据匹配结果和规则频度的反馈信息,通过增加奖惩规则,对规则库进行修正和完善。增加的规则如:

奖励规则:

①若候选地名中出现"国"、"府"、"州"、"县"、"郡"等地名特征词,(见表2)。

惩罚规则:

②若候选地名长度大于5个汉字长。

③若候选地名右侧2个汉字内出现"记"、"志"、"丛话"等表示书名的字词。

④若候选地名左侧1个汉字是"按"、"见"、"案"等表示引用文献的字词。

⑤若候选地名中出现"一"、"二"……"十"等数词。

⑥若候选地名中出现"上"、"中"、"底"、"边"等方位词。

⑦若规则字和前后汉字组成固定词语,如"蔓生"、"野生"、"飞出"、"出入"、"土产"、"水产"等。

表2 物产地名特征词

地名特征词	地名特征词	地名特征词	地名特征词
国	海	田	番
府	洋	塘	楼
州	江	池	谷
郡	湖	园	岗
县	溪	林	墙
乡	山	地	坑
村	峰	陇	路
……	……	……	……

2.2.3 物产地名识别

(1)运用规则库匹配物产解释,产生候选地名;

(2)通过奖惩规则对不同类型的规则产生的候选地名做相应的过滤处理;

(3)用地名标引词典扫描经过处理的候选地名,进一步修正通过规则识别的地名。

具体识别算法如下:

①读入一条物产解释;

②判断物产解释是否为空;

③是 执行空地名信息插入,转①;

④否 遍历规则信息;

⑤根据规则类型,获取地名信息;

⑥判断地名信息是否为空;

⑦是 地名置空,转④;

⑧否 遍历惩罚规则;

⑨根据当前匹配规则的类型,执行相应的

惩罚规则过滤处理;

⑩判断惩罚规则遍历是否结束;

⑪否　转⑧;

⑫是　判断地名信息是否为空;

⑬是　地名置空,转④;

⑭否　遍历奖励规则,过滤地名信息;

⑮用地名表扫描经过处理的候选地名,修正识别地名;

⑯地名信息插入;

⑰判断规则信息遍历是否结束;

⑱否　转④;

⑲是　判断物产解释遍历是否结束;

⑳否　转①;

㉑是　结束退出。

方志物产地名识别流程见图1。

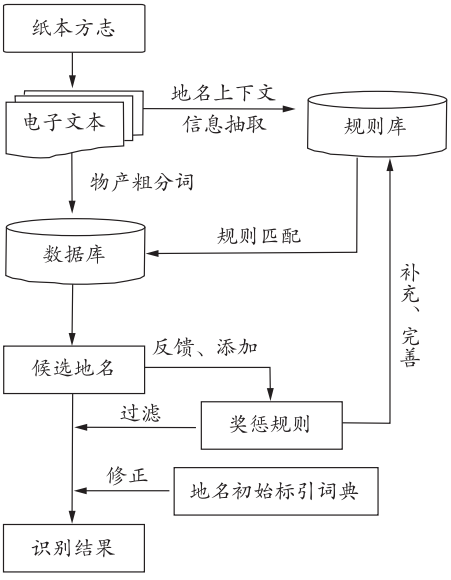

图1　方志物产地名识别流程

3　物产地名识别系统的实现

3.1　系统开发软件的选择

系统开发运行的环境是 Microsoft. NET Framework。项目的类型为 ASP. NET, ASP. NET 是由 ASP(Active Server Pages)发展而来,它是完全基于对象的,每个对象都有自己的属性、方法和事件,开发人员只要选用相应的控件并调整其属性,就可以建立业务解决方案[14],这种结构为 Web 应用的开发提供了一种面向对象的方法,使得 Web 应用的开发更为简易、方便和灵活。系统开发的语言选用 C#,开发的工具选用 Visual Studio 2008。后台数据库选用 Microsoft SQL Server2005, SQL Server2005 是微软基于客户端/服务器模式的数据库系统,适用于大型数据库管理和电子商务,能确保数据的完整性和安全性,可为方志内容挖掘和知识发现提供支持,且 SQL Server2005 支持 Unicode,适合古籍特殊数据存储,因而选作后台数据库。数据访问采用 ADO. NET 技术, ADO. NET 是由 ADO(ActiveX Data Objects, ActiveX 数据对象)发展而来[15-16],它是一种无连接、基于消息的数据访问模型。数据源上的数据可作为 XML 文档进行传输和存储,这样,只要应用能够解析 XML,就能够实现数据访问。基于 B/S 模式的设计思想,便于系统扩充应用和更新维护,用 Internet 访问 Web 页面,实现文件管理、规则库管理、地名识别、信息统计等功能。

3.2　系统实现

系统主要包括文件管理、信息抓取入库、规则和奖惩规则管理、地名识别、信息查询与统计等功能。

文件管理:浏览、上传需要进行地名识别的文件,支持多文件上传。已上传的文件可以通过文件名查询,文件的详细内容可以点击查看,也可以随时删除文件。

信息抓取入库:将已上传的文件按照标注的代码转入数据库,同时完成物产粗分词的过程,界面如图2。

规则和奖惩规则管理:查询、添加、编辑和删除地名识别规则及奖惩规则。编辑奖惩规则时,选定一条规则,从判断方向、字符长度、规则状态、过滤信息等选项进行编辑。

地名识别:对已抓取入库的文件进行物产地名的识别,识别结果分"已编辑"和"未编辑"两类分别显示。"已编辑"是抓取到地名信息的结果,可逐页浏览每个物产的物产名、地名、规

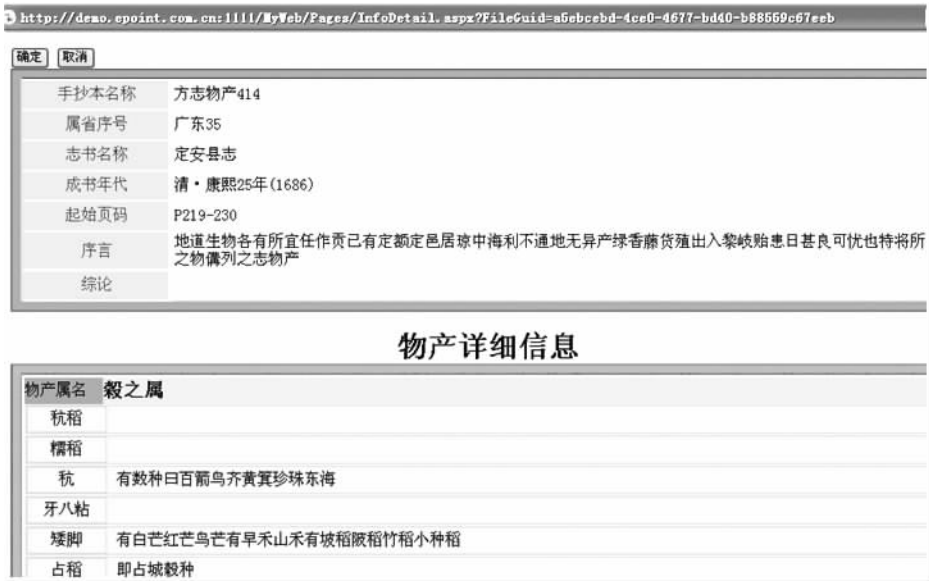

图 2　信息抓取入库界面

编号	物产品	地名信息	规则信息
1	稷	大城皆货及	出
2	棕榈	岭南西川	出
3	紫榆	海舶	来自
4	紫檀	扶南	出
5	紫菜	万州	出
6	紫背	海	产于
7	梓	四会	产者
8	梓	香山	出
9	梓	岭南	惟有之
10	维粟	文昌	出
11	烛竹	南岭	生于
12	烛竹	燃此竹火光	生
13	竹叶青	连州	出
14	竹席	黄岭角	出者佳

图 3　物产地名识别结果界面

则等信息，系统实现了同一物产相关信息的集中显示；“未编辑”是未能抓取到地名信息的结果。识别结果的显示界面中，用户均可点击查看物产的详细信息，必要时可根据物产的解释，人工添加、修改地名和规则信息，界面如图 3。

信息查询与统计：显示全部的地名识别结果，提供物产名、物产属名、物产地名、规则信息、志书名称、成书年代等检索入口和排序依据，可分类统计信息，并具有去重及筛选的功能。

3.3　系统测评

3.3.1 测评指标

为衡量系统的识别效果,采用三个评估指标对系统进行测评,分别是准确率 P、召回率 R 和综合指标 F。它们的定义如下:

$$P=\frac{\text{计算机识别出的正确物产地名数}}{\text{计算机识别出的物产地名总数}}\times 100\%$$

$$R=\frac{\text{计算机识别出的正确物产地名数}}{\text{人工识别出的正确物产地名总数}}\times 100\%$$

$$F=\frac{(\alpha^2+1)\times P\times R}{\alpha^2 P+R}$$

其中,α 是准确率 P 和召回率 R 之间的权衡因子,这里我们认为 P 和 R 同等重要,因此,α 取 1,此时综合指标称为 F－1 值。

3.3.2 测评方法

随机抽取 10 个文件作为测试集,请相关专家仔细研读后人工识别出正确的物产地名,同时标出地名对应的规则信息。由于一条物产解释中有可能涉及多个规则和地名,例如:

芒果 种传外国实大如鹅子状生则酸熟则甜惟新会香山有之

此物产解释中包含的地名信息:芒果　种传外国　惟新会香山有之

为方便测试,将这一条解释中的地名信息作为两对识别结果来加以记录:

芒果　种传外国

芒果　惟新会香山有之

这样,专家人工识别出的正确的物产地名 643 对,计算机识别出的物产地名 841 对。测试时,把计算机识别出的物产地名和人工识别出的正确的物产地名逐一对比,找出相同的对数,结果见表 3。

表 3　物产地名识别测试结果

人工识别的正确物产地名数	643
计算机识别出的物产地名总数	841
计算机识别出的正确物产地名数	533
计算机识别出的错误物产地名数	308

3.3.3 测评结果及错误原因分析

计算得出,准确率为 63.38%,召回率为 82.89%,综合指标为 71.83%。通过对识别结果的分析,笔者认为导致系统误识别和漏识别的原因主要有三点:一是规则库的覆盖面有限,识别规则不能涵盖所有的地名信息,导致漏识;惩罚规则也不可能穷举所有可能导致地名误识别的情况,导致误识。二是规则匹配对,规则之间存在冲突,导致重复识别。三是原始方志资料数字化处理过程中存在生字、错字,当利用地名初始标引词典对候选地名做最后的修正时,计算机无法识别、修正错误的地名信息,降低了识别结果的召回率和准确率。

4　结语

本文尝试将命名实体识别技术应用到方志类古籍的内容挖掘中,一方面,为方志类古籍的整理和开发利用提供了一种新方法、新技术;另一方面,也为命名实体识别技术的应用研究开辟了新领域。从实证研究的效果看是可行的,要达到实际应用的程度,有待通过提高电子文本质量、增加规则和优化算法等途径进一步提高地名识别的准确率和召回率。

参考文献:

[1] Grishman R, Sundheim B. Message understanding conference－6: A brief history[C]// Proceedings of the 16th International Conference on Computational Linguistics COLING－96, 1996－08.

[2] 乔永波. 规则与统计相结合的中文命名实体识别[D]. 济南:山东大学,2007.

[3] 周波,杨国纬. 基于贝叶斯算法的中国人名识别[J]. 计算机应用,2006(4):998－1000.

[4] 李丽双,黄德根,陈春荣,等. SVM 与规则相结合的中文地名自动识别[J]. 中文信息学报,2006(5):51－57.

[5] 王浩畅,赵铁军,李艳. 生物医学命名实体识别的特征选取与评价[C]//孙茂松,陈群秀. 内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集. 北京:清华大学出版社,2007.

方正阿帕比助力海峡两岸图书馆特色资源建设
——阿帕比亮相第三届海峡两岸大学图书馆合作发展论坛

2011 年 3 月 31 日—4 月 3 日,“第三届海峡两岸大学图书馆合作发展论坛”在四川大学举办。论坛的主题为“特色资源数据的收集与建设”。

中国高等教育文献保障系统(CALIS)于 2011 年 3 月对全国高校图书馆发出了 CALIS 第三期专题特色数据库子项目申报的邀请,要求挖掘整理未开发利用的资源,重点建设一批定向专题数据库,补充 CALIS 资源体系。本次论坛就是对一些高校图书馆提出的专题项目进行研讨。

方正阿帕比受邀出席会议并带来其建设特色资源数据库的技术解决方案——方正阿帕比德赛系统。德赛系统是特色资源数字化加工与安全发布系统,能够将 DOC、DOCX、PDF、EPS、JPG、TIF、TXT、PS、PS2、S72、S92、S10 以及扫描文件等各种格式的文献资源,统一成符合国际标准格式的电子资源,进行深度数据加工和加密处理后在网络上安全发布,或以光盘介质出版,供读者使用。

德赛系统支持图书馆特色资源的版权保护;具有先进的曲线显示技术,物理、数学公式高保真原版原式;支持全文检索和加工 CEBX 文件,可将自有资源进行移动阅读应用;支持图片加工发布;未来还将支持条目数据库加工制作,并可承建 CALIS 三期建设特色库服务平台 1.0 版本。

目前德赛系统已成功应用于北京大学图书馆人物库、云南理工大学图书馆外文期刊库、黑龙江工程大学图书馆工程图片库等。该系统是国家级火炬计划项目,而且拥有军用信息安全产品军 C + 级证书,为各图书馆珍贵资源版权保护提供有力的保障。

[6] 邹涛. 一种电子产品领域命名实体识别方法研究[D]. 西安:西安电子科技大学,2010.

[7] 付瑞吉. 音乐命名实体识别技术研究[D]. 哈尔滨:哈尔滨工业大学,2009.

[8] 王铮. 基于 CRF 的古籍地名自动识别研究[D]. 南宁:广西民族大学,2008.

[9] 王世昆. 中医症状病机实体识别及其关系挖掘研究[D]. 厦门:厦门大学,2009.

[10] 来新夏. 中国地方志的史料价值及其利用[J]. 国家图书馆学刊,2005(1):5 - 8.

[11] 牟力科. Web 中文信息抽取技术与命名实体识别方法的研究[D]. 西安:西北大学,2008.

[12] 李诺,张全. 利用地名用字分析的中文地名识别处理[J]. 计算机工程与应用,2009(28):230 - 232.

[13] 衡中青. 地方志知识组织及内容挖掘研究[D]. 南京:南京农业大学,2007.

[14] Donny M, Doug S. ASP. NET 数据驱动 Web 开发[M]. 林琪,张伶,朱涛江,译. 北京:中国电力出版社,2003.

[15] 周治平. ADO 数据存取技术[J]. 计算机应用,1999(7):93 - 95.

[16] 叶德谦,马勤勇. 使用 ADO. NET 对关系数据库的访问[J]. 微型电脑应用,2001(8):39 - 42.

朱锁玲 南京农业大学人文社会科学学院博士研究生。通讯地址:南京农业大学图书馆。邮编:210095。

包 平 南京农业大学图书馆馆长,教授,博士生导师。通讯地址同上。

(收稿日期:2010-11-18)