年鉴数字千分空自动化处理研究

周丹丹 *

摘　要　利用计算机技术对年鉴数字信息进行处理是年鉴编纂工作值得探讨的话题。成都年鉴社利用计算机技术对年鉴表格的数字千分空编辑处理，实现了提高编辑效率，降低人工编辑出错率，减轻编辑人员工作负担的目的。该程序依据有关法规要求，以 Word 文档或者 WPS 文档为处理素材，使用 C#语言编程实现数字处理功能，具有处理速度快、过程自动化、无差错等优势，可以较大程度减轻人工处理的压力，为后期校对工作提供很好的支持，也为年鉴出版的信息化智能化应用场景提供思考借鉴。

关键词　年鉴　计算机辅助技术　自动化处理

年鉴编纂工作对年鉴出版质量至关重要，需要编纂人员付出艰辛劳动。如何减轻编纂人员的负担，避免重复单一的劳动，进而提高工作效率就显得格外重要。笔者开发一款千分空数字化处理程序并试用于成都年鉴社年鉴编纂工作中，提高了效率，减轻了编辑负担，保证了工作质量。所谓千分空，是阿拉伯数字的一种规范用法，指一个数字，从小数点起，向左和向右每三位数字一组，组间空四分之一个汉字，即二分之一个阿拉伯数字的位置。① 此种数字用法常见于审计报告、法律文书等写作中，如“支付金额 42 300 000 元”，以达到醒目的效果。下面，笔者试就此话题展开论述。

一、数字千分空处理的依据及优势

《图书编辑工作基本规程》指出，加工整理是编辑工作不可缺少的环节。经过审稿决定采用的书稿，在内容、体例、引用材料、语言文字、逻辑推理等方面难免存在一些问题，需要进行加工整理，使内容更完善，体例更严谨，材料更准确，语言文字更通达，逻辑更严密，消除一般技术性、常识性差错，防止出现原则性错误，并符合排版和校对要求。② 数据量

---

*　周丹丹，女，内蒙古自治区开鲁县人，成都年鉴社管理岗位七级职员，主要研究方向为年鉴学。

① 中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会：《出版物上数字用法》（GB/T 15835－2011），2011 年 7 月 29 日。

② 新闻出版署图书司：《关于转发〈图书编辑工作基本规程〉的通知》，中国出版年鉴社编：《中国出版年鉴（1999）》，中国出版年鉴社，1999 年，第 302～303 页。

较大的时候,稍不注意就可能发生体例格式不统一的问题。在年鉴编辑加工中,关于户籍人口、国民经济和社会发展、固定资产等内容的表格,会大量涉及对数字的处理。以《成都年鉴(2021)》为例,全书 220 万字,其中数字数量非常大,传统的处理方式,需编辑人员对图表中的每个数字进行手工添加空格,费时费力,并且容易发生错漏,给编辑、校对和后续出版工作带来一定困难。年鉴作为工具书,不仅要求数字要准确、严谨,同时要保证数字体例格式的统一。因此研究千分空自动化处理的技术手段具有一定的必要性。

国家标准《出版物上数字用法》(GB/T 15835 - 2011)规定:"本标准出版物上汉字数字和阿拉伯数字的用法。本标准适用于各类出版物(文艺类出版物和重排古籍除外)"。第五部分"数字形式的使用"中"多位数"的规定:"为便于阅读,四位以上的整数或小数,可采用以下两种方式分节:第一种方式:千分撇。整数部分每三位一组,以',' 分节,小数部分不分节。四位以内的整数可以不分节。示例 1:624,000;92,300,000;19,351,235.235767;1256。第二种方式:千分空。即从小数点起,向左和向右每三位数字一组,组间空四分之一个汉字,即二分之一个阿拉伯数字的位置。四位以内的整数可以不加千分空。示例 2:55 235 367.346 23;98 235 358.238 368"。① 《出版物上数字用法》(GB/T 15835 - 2011)在促进中文出版物中数字表达形式的规范使用方面起到重要作用。同时,《有关量、单位和符号的一般原则》(GB3101—93)第三部分规定:"数一般应当用正体印刷。为使多位数便于阅读,可将数字分成一组,从小数点起,向左和向右每三位分成一组,组间留一空隙,但不得用逗号、圆点和其他方式。"②

《成都年鉴》一直坚持按照《出版物上数字用法》(GB/T 15835 - 2011)的要求,对全书的数字部分进行千分空处理。据笔者了解,目前《四川交通年鉴》《成都市教育年鉴》等也在文稿中对数字进行千分空处理。千分空通过把数字分节,使其更便于阅读和使用,对于包含很多统计数字的工具书,优势更为明显。此外,千分空作为数字处理的一种标准,可以使年鉴表格数字方面的文稿更加规范,尤其是对于超过 5 位数的数字,比如 2020 年粮食产量 2278585 吨,如果读者想要查阅这一条统计资料表格里面的数字,很显然"2 278 585 吨"的表述方式比"2278585 吨"的表述方式更为清晰。再比如,如果要比较 2019 年和 2020 年的粮食产量,2019 年粮食产量为"2 259 048 吨"和 2020 年粮食产量"2 278 585 吨",通过千分空方式处理过的表述可以让比较的结果一目了然。

二、计算机程序化处理分析

在实践中,需要处理的对象为表示数量的数字,期待实现的目标为每个数字从右到左每隔 3 位中间添加一个空格,处理对象、处理方式和输出结果明确。计算机程序具有处理速度快、过程自动化、无差错等优势,能够为我们解决电子信息领域各种问题,通过编程,还可满足适用性、方便性、可扩展性等需求,是解决图表数字千分空全自动处理的最优选

① 中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会:《出版物上数字用法》(GB/T 15835 - 2011),2011 年 7 月 29 日。

② 国家技术监督局:《有关量、单位和符号的一般原则》(GB3101 - 93),1993 年 12 月 27 日。

择。下面对计算机程序处理过程进行分析。

(一)程序处理流程分析

参照人工处理数字千分空的过程,使用计算机处理可以将处理步骤分为以下几步。

1. 打开待编辑处理的文档,将文档信息化数字化后写入内存。
2. 提取其中待处理的数字。
3. 对所有数字依此进行千分空处理。
4. 处理完毕,保存文档。

以上步骤均可使用现有计算机技术进行编程实现,下面将详细分析实现算法和使用的具体技术。

(二)处理目标及对象分析

目前,在《成都年鉴》编辑实践中,所用素材基本为 Word 文档或者 WPS 文档。其中,以 Word 文档居多,WPS 为国产金山软件公司开发的办公软件,可兼容 Word 文档。Word 办公软件目前有“.doc”和“.docx”两种文件后缀格式。其中“.docx”格式的文档是 Microsoft Office 2007 办公套件及以后的版本默认的文档格式,该文档支持 Open XML 技术标准。Open XML 是一种基于 XML 的 Office 文档格式,包括处理文档、电子表格、演示文稿以及图表、形状和其他图形材料。该标准由微软公司开发,并于 2006 年被 ECMA International① 采用为 ECMA-376。第二个版本于 2008 年 12 月发布,第三个版本于 2011 年 6 月发布。该标准已被 ISO 和 IEC 采用为 ISO/IEC 29500。该格式改善文件和数据管理、数据恢复以及与软件系统的互操作性,扩展以前版本的二进制文件的功能,任何支持 XML 的应用程序都可以访问和处理其数据,最重要的是这种格式基于开放打包约定(Open Packaging Conventions),可以使用第三方开发工具 OpenXML SDK 对文档进行二次开发以对文档的内容进行快速解析和处理,解析后的文档为一系列特定结构和特定标记的 XML 文档,为本程序的处理提供技术基础。为提高程序兼容性,后续可以考虑直接支持 WPS 文档的解析,这里可手工转存,利用 WPS 程序自带功能直接将文档另存为“.docx”格式。为方便处理,本文选择的处理目标统一为“.docx”文件,其他格式文件可以通过各种方式进行转换。

需要注意的是,千分空处理对象为阿拉伯数字,但是这里的数字还需要加以区分,即分为表示数量的数字和表示年份或特定称谓的数字,但处理特定称谓的数字就需要在提取数字过程中进行筛选,用计算机编程实现,需要用到文本分析技术。文本分析技术是人工智能的一个细分领域,本文处理的场景相对简单,仅需判断一段文本或表格中的数字是表示数量还是年份或某些特定称谓,可采用上下文特定关键字匹配的方式进行处理。

(三)核心算法及实现过程

根据上文分析,我们使用 C#语言编程实现所需功能。C#是一个现代的、通用的、面向对象的编程语言,开发人员利用 C#能够生成在 .NET Framework 运行库中运行的多种安全

① Ecma International 是一家国际性会员制度的信息和电信标准组织。

可靠的应用程序。其核心算法和实现过程如下：

用户选择文件后，使用 OpenXML SDK Tool 开发包中的方法①打开并获取文件中的所有文本数据及超文本标记语言标识的特定符号，依赖于 OpenXML SDK Tool 工具包，可以很容易将待处理文件的主体文档部分提取为结构化对象，并通过其自带的函数，遍历文档中所有结构，找出整个文件中的所有数字，并将其存入一个缓存池。缓存池中的数据还需要加以区分，一类为需要处理的数字，一类为不需要处理的数字。这里有两种处理方式。方式一：先将所有数字全部放入缓存池，再逐一进行判断和处理；方式二：先对数字进行判断，需要处理的数字放入缓存池。这里我们采用的是第二种方式，需要处理的标准为整数部分超过 3 位，且该数字不是表示年份或特定称谓。对后者的判断可采用上下文文本分析法，寻找该数字所在文本段落前后是否有特定的关键字，如“年”“月”“日”“地名”“代号”一类，如有则判断该数字不作处理，否则需进行处理，找出所有需要处理的数字后，对缓存池中的数字依此进行千分空处理并替换原有数字。为方便校对，可以提示处理的数字个数并为所有被处理的数字添加批注或进行颜色标记。

所有数字处理完成后，将处理后的文档另存一份，并将已处理和未处理的文档同时打开，方便编辑人员对文档进行人工校验。

（四）实现的效果分析

下面以某部年鉴中不同时期的航空运输、邮电业务、城市规模和建设用地、市政设施水平等行业相关数字图表处理的单数据处理时长及效果进行分析。

表 1—表 4 为处理前的数据表格。

表 1　航空及公路运输情况

	单位	2016 年	2017 年	2018 年	2019 年	2020 年
航空运输						
民用航空线路条数	条	270	315	335	358	367
飞机架数	架	242	271	294	334	349
旅客吞吐量	万人	4603.9	4980.2	5295.1	5585.9	4074.2
货邮吞吐量	万吨	61.16	64.29	66.50	67.20	61.85
公路运输						
公路通车里程	公里	26037	26294	27731	28260	29627
轨道运输						
轨道交通运营里程	公里	108.55	179.59	239.90	341.54	557.84
全社会各种机动车辆	万辆	466.74	494.18	548.44	577.24	603.81
#载货汽车	万辆	22.97	24.86	28.64	31.91	36.04
载客汽车	万辆	388.23	425.1	457.33	484.45	496.13
摩托车	万辆	51.13	59.09	56.67	55.43	55.43
#私人汽车	万辆	369.87	398.24	420.28	438.83	441.39
电动自行车	万辆	442.63	450.67	457.78	461.32	465.00

① 这里的方法是指计算机编程技术中的方法相当于一个函数，并非通常意义中的方法。

表 2　邮电业务基本情况

年份	邮电业务总量(万件)	函件(万件)	移动电话用户(万户)	固定电话用户(户)
1978 年	2768	2140	—	9396
1980 年	3066	2655	—	11140
1990 年	15520	7538	—	50562
2000 年	716428	9481	99. 1	1481152
2010 年	4412654	7980	1732	3959092
2019 年	19333530	1843	2616. 96	6505639
2020 年	25468033	1511	2876. 20	6070524

注:①从 2011 年起,电信相关资料均按国家新政策进行统计;②从 2016 年起,邮电业务总量中电信业务总量按 15 年不变价计算。

表 3　城市规模和建设用地情况

	单位	2019 年	2020 年		单位	2019 年	2020 年
市区人口	万人	876. 47	895. 73	公共管理与公共服务用地	平方公里	82. 18	83. 01
市区面积	平方公里	3639. 81	3639. 81	工业用地	平方公里	137. 95	141. 25
建成区面积	平方公里	949. 58	977. 12	道路与交通设施用地	平方公里	158. 32	168. 00
城市建设用地面积	平方公里	879. 19	909. 65	物流仓储用地	平方公里	16. 88	19. 74
居住用地	平方公里	300. 7	306. 58				

表 4　市政设施水平

	单位	2019 年	2020 年		单位	2019 年	2020 年
用水普及率	%	99. 81	99. 64	人均公园绿地面积	平方米	14. 58	14. 51
用气普及率	%	98. 82	99. 35	建成区绿地率	%	36. 98	37. 68
人均拥有道路面积	平方米	15. 87	18. 70	建成区绿化覆盖率	%	43. 46	43. 90

表 5—表 8 为处理后的数据表格。

表 5　航空及公路运输情况

	单位	2016 年	2017 年	2018 年	2019 年	2020 年
航空运输						
民用航空线路条数	条	270	315	335	358	367
飞机架数	架	242	271	294	334	349
旅客吞吐量	万人	4 603. 9	4 980. 2	5 295. 1	5 585. 9	4 074. 2
货邮吞吐量	万吨	61. 16	64. 29	66. 50	67. 20	61. 85
公路运输	—	—	—	—	—	—
公路通车里程	公里	26 037	26 294	27 731	28 260	29 627
轨道运输	—	—	—	—	—	—

（续表）

	单位	2016 年	2017 年	2018 年	2019 年	2020 年
轨道交通运营里程	公里	108. 55	179. 59	239. 90	341. 54	557. 84
全社会各种机动车辆	万辆	466. 74	494. 18	548. 44	577. 24	603. 81
#载货汽车	万辆	22. 97	24. 86	28. 64	31. 91	36. 04
载客汽车	万辆	388. 23	425. 1	457. 33	484. 45	496. 13
摩托车	万辆	51. 13	59. 09	56. 67	55. 43	55. 43
#私人汽车	万辆	369. 87	398. 24	420. 28	438. 83	441. 39
电动自行车	万辆	442. 63	450. 67	457. 78	461. 32	465. 00

表 6　邮电业务基本情况

年份	邮电业务总量（万件）	函件（万件）	移动电话用户（万户）	固定电话用户（户）
1978 年	2 768	2 140	—	9 396
1980 年	3 066	2 655	—	11 140
1990 年	15 520	7 538	—	50 562
2000 年	716 428	9 481	99. 1	1 481 152
2010 年	4 412 654	7 980	1 732	3 959 092
2019 年	19 333 530	1 843	2 616. 96	6 505 639
2020 年	25 468 033	1 511	2 876. 20	6 070 524

注：①从 2011 年起，电信相关资料均按国家新政策进行统计；②从 2016 年起，邮电业务总量中电信业务总量按 15 年不变价计算。

表 7　城市规模和建设用地情况

	单位	2019 年	2020 年		单位	2019 年	2020 年
市区人口	万人	876. 47	895. 73	公共管理与公共服务用地	平方公里	82. 18	83. 01
市区面积	平方公里	3 639. 81	3 639. 81	工业用地	平方公里	137. 95	141. 25
建成区面积	平方公里	949. 58	977. 12	道路与交通设施用地	平方公里	158. 32	168. 00
城市建设用地面积	平方公里	879. 19	909. 65	物流仓储用地	平方公里	16. 88	19. 74
居住用地	平方公里	300. 7	306. 58				

表 8　市政设施水平

	单位	2019 年	2020 年		单位	2019 年	2020 年
用水普及率	%	99. 81	99. 64	人均公园绿地面积	平方米	14. 58	14. 51
用气普及率	%	98. 82	99. 35	建成区绿地率	%	36. 98	37. 68
人均拥有道路面积	平方米	15. 87	18. 70	建成区绿化覆盖率	%	43. 46	43. 90

以表 5—表 8 为例，处理数字 31 个，表示年份的数字没有被处理，处理准确率 100%，处理时长约 1 秒。采用传统人工编辑处理方式，平均需花费约 3 分钟。计算机程序处理效率与人工相比提高近 180 倍。在实际工作中，编辑人员每天都需要面对各种稿件及事

务性工作,工作状态很难长时间维持在较高水平,而且处理的数字图表越多,花费的时间越多,且人工处理时长同数字量成指数级增长,错误率也会增高。而程序处理时长几乎不会因数据量的增长而增加,程序的处理时间主要与电脑硬件配置相关,如 CPU 主频、芯片核心数、内存容量及速率、硬盘转速及读写速度等。以 Intel i7 8 核处理器,2.6GHz 主频为例,处理的数据为 1 万个时,所需的时间约 2 秒左右。程序算法还可以不断进行优化,只要算法和代码设计合理,程序处理的数据准确率可以达到 99% 以上,很大程度减轻人工处理的压力,为后期校对工作提供很好的支持。

然而,该程序也存在一定的优化改进空间。比如兼容性方面,目前仅支持“.docx”格式的文件处理,后期可以扩展兼容 EXCEL 等各类文件,因金山 WPS 文档也支持 Open XML 标准,通过对程序的改进可以直接支持 WPS 文档的处理,同时还可以加入选项,支持用户自行选择使用千分空或千分撇进行处理,使程序兼容性更加广泛;易用性方面,为方便文稿校验,可利用 COM 组件,将处理前后的文件在一个窗体中分左右屏同时打开,同步浏览,滑动其中一个窗体的滚动条,另一个自动同步滚动,便于校验;程序效率方面,可以优化数字判断和处理算法,例如采用语义分析法等,更精准地判断待处理的数字,提高数字识别的精准性,使程序更加完善高效。

三、结　语

千分空通过把数字分节,使其更便于阅读和使用,可以使年鉴表格数字方面的文稿更加规范。笔者着重讨论了使用计算机程序辅助年鉴编辑加工中数字千分空的自动化处理的应用场景,可以实现在提高编辑效率的同时,有效降低错误率,减轻校对压力,可以为实现年鉴素材的高效、快速处理提供数字化支撑。此外,对编辑加工环节中数字数据的自动化处理应用进行探讨,为年鉴出版的信息化智能化应用场景提供了广阔的思考空间。在实际工作中,素材收集整理、文稿编辑校对、数据存储备份、出版发行宣传、档案归档管理、网络信息安全等,都可以借助信息化手段提高工作效率,减少人工成本。例如,在素材收集整理方面,可利用网络爬虫技术在媒体和政府网站等渠道自动获取各类年鉴图片、文字素材,利用智能化算法实现素材自动获取,自动分类存储和智能化编目管理,利用人工智能技术实现年鉴关键字索引自动生成和管理;在编辑校对方面,虽然目前市面上已有黑马、方正等专业校对软件,但其智能化程度还不够,未来可以在基于中文语义分析为技术框架的智能校对和纠错方面进行探索;在年鉴数字化对外宣传方面,依托互联网短视频平台、大数据分析平台等各类新媒体平台,实现出版物精准投放和宣传;在数字方志馆建设方面同样可期。综上所述,利用现代信息技术,为年鉴编纂出版赋予非常大的潜能。思考、创新应是广大年鉴工作者具备的一种基本工作思维。打造精品年鉴,以信息化助力年鉴事业更快更好发展,我们永远在路上。

责任编辑:李一龙