# 附录 A：博士学位论文研究方法章节标注流程与规范

## A.1 数据标注背景

研究方法是学者进行研究过程中使用到的方法、工具或者手段，充当着实践与理论之间的桥梁，也是人们逐渐认识世界并改造这个世界所需的工具。在过往研究中，笔者发现学者们多采用理论分析为主的研究方法定性的研究本学科研究方法运用现状，缺乏定量角度的描述。那么我国人文社科研究方法的发展现状究竟如何？规范化程度如何？研究方法的运用存在着哪些特点？如果仅从定性角度进行描述，很难准确的给出这些问题的答案。因此，本课题将从定量的研究角度对研究方法进行计量分析，以期获得有价值的研究结论。

与此同时，学术文献的数量在当今这个大数据时代背景下正飞速增长，对于研究方法的运用也参差不齐。相对于其他文献，博士学位论文具有专业性以及内容结构规范等等优点，具有一定的代表性，其使用的研究方法也在一定程度上折射出该学科研究方法的运用现状。另外，博士学位论文中关于研究方法的界定也更加的规范，更便于提取分析，特别是人文社科的学位论文，一般都有研究方法专门的小节，这也为研究方法实体的抽取提供了可能。

然而，目前鲜有学者以博士学位论文为对象对整个人文社科领域研究方法应用现状进行相关研究，因此本课题将对人文社科领域所有学科博士学位论文数据进行检索，以此开展相应研究。但是由于博士学位论文数据全文中具有大量的与研究方法不相关的文本内容，例如研究创新点、研究背景、研究意义等等，因此在进行研究方法实体自动提取之前，需要进行研究方法的章节定位工作。研究方法章节定位用到了两种方法，基于人工标注的方法以及基于规则匹配的方法。

正是由于标题的不确定性与 OCR 识别的误差，导致规则匹配方法得到的章节匹配结果召回率与准确率较低，为了为下一步研究方法实体抽取获得更加规范的训练预料，在此处需要对研究方法章节进行人工标注与审核，以提高章节识别的语料质量，此数据标注任务正是基于以上背景。

## A.2 概念界定

### A.2.1 研究方法

研究方法，是指在研究中发现新现象、新事物，或提出新理论、新观点，揭示事物内在规律的工具和手段，在人类的学术活动中发挥着重要作用，例如“文献研究法”、“问卷调查法”、“比较研究法”等等，这些都是人文社科领域博士学位论文中常用的研究方法。对于一门学科而言，研究方法的发展以及规范程度也在很大程度上与其自身的发展成熟程度相关联。当一个研究方法实体出现在博士学位论文中时，则表示该研究方法实体被该博士学位论文使用了，或者是该博士学位论文引用该研究方法实体由于相关内容的分析或者比较。因此梳理博士学位论文中使用的研究方法实体，可以帮助相关研究者快速的了解文献的主旨内容。同时，通过研究学科领域研究方法的运用情况如演化情况等，能够了解人文社科领域学科的发展规律，为相关学者进行研究提供参考等。因此对博士学位论文中的研究方法进行研究具有重要意义。

### A.2.2 研究方法章节

研究方法章节就是研究方法所在的小节。学术文献的数量在当今这个大数据时代背景正

飞速增长，对于研究方法的运用也参差不齐。与一般的期刊论文不同的是，博士学位论文具有专业性以及内容结构规范等等优点，关于研究方法的界定也更加的规范，更便于提取分析，特别是人文社科的学位论文，一般都有研究方法专门的小节。如图 A1 所示为某情报学学科的博士学位论文，其中 1.4.1 小节就是研究方法章节。


**图 A1 研究方法章节实例（标题中）**

在我们的标注任务中，由于需要从正文中找出研究方法章节，因此我们将研究方法章节定义为从研究方法标题开始到下一节开始位置的中间部分，如图 A2 所示为正文中研究方法章节的例子，其中“1.3.2 研究方法”标题与“1.4 研究意义与创新点”标题之间的内容就是研究方法章节。

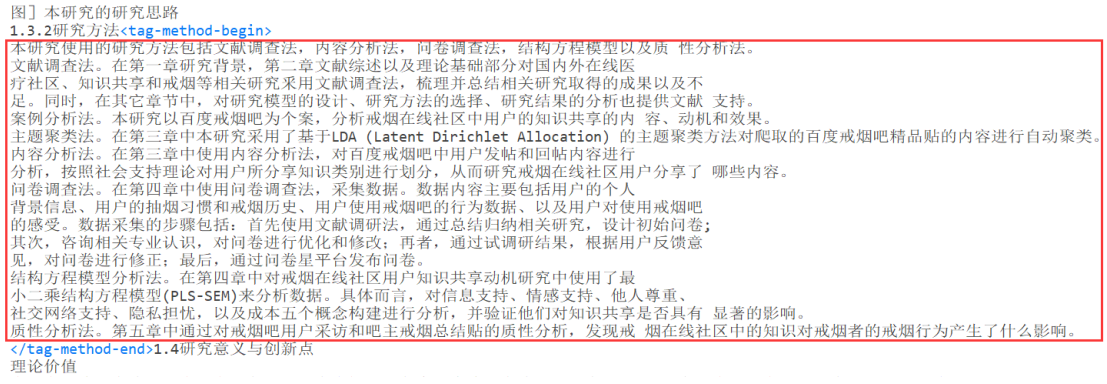
图］本研究的研究思路
1.3.2研究方法<tag-method-begin>
本研究使用的研究方法包括文献调查法，内容分析法，问卷调查法，结构方程模型以及质 性分析法。
文献调查法。在第一章研究背景，第二章文献综述以及理论基础部分对国内外在线医
疗社区、知识共享和戒烟等相关研究采用文献调查法，梳理并总结相关研究取得的成果以及不
足。同时，在其它章节中，对研究模型的设计、研究方法的选择、研究结果的分析也提供文献 支持。
案例分析法。本研究以百度戒烟吧为个案，分析戒烟在线社区中用户的知识共享的内 容、动机和效果。
主题聚类法。在第三章中本研究采用了基于LDA (Latent Dirichlet Allocation) 的主题聚类方法对爬取的百度戒烟吧精品贴的内容进行自动聚类。
内容分析法。在第三章中使用内容分析法，对百度戒烟吧中用户发帖和回帖内容进行
分析，按照社会支持理论对用户所分享知识类别进行划分，从而研究戒烟在线社区用户分享了 哪些内容。
问卷调查法。在第四章中使用问卷调查法，采集数据。数据内容主要包括用户的个人
背景信息、用户的抽烟习惯和戒烟历史、用户使用戒烟吧的行为数据、以及用户对使用戒烟吧
的感受。数据采集的步骤包括：首先使用文献调研法，通过总结归纳相关研究，设计初始问卷；
其次，咨询相关专业认识，对问卷进行优化和修改；再者，通过试调研结果，根据用户反馈意
见，对问卷进行修正；最后，通过问卷星平台发布问卷。
结构方程模型分析法。在第四章中对戒烟在线社区用户知识共享动机研究中使用了最
小二乘结构方程模型(PLS-SEM)来分析数据。具体而言，对信息支持、情感支持、他人尊重、
社交网络支持、隐私担忧，以及成本五个概念构建进行分析，并验证他们对知识共享是否具有 显著的影响。
质性分析法。第五章中通过对戒烟吧用户采访和吧主戒烟总结贴的质性分析，发现戒 烟在线社区中的知识对戒烟者的戒烟行为产生了什么影响。
</tag-method-end>1.4研究意义与创新点
理论价值

**图 A2 研究方法章节实例（正文中）**

### A.2.3 研究方法章节定位

在对研究方法与研究方法章节进行介绍以后，研究方法章节定位的概念就很容易理解。所谓研究方法章节定位就是使用特殊的区块标记将研究方法章节从整个文章中标记出来，以此将其与论文其他章节内容区分开来。学位论文大多存在研究方法专门小节，可以使用规则匹配的方法进行自动标注与定位，但是由于 CNKI 数据库中下载的博士学位论文为 can 格

式，需要使用 OCR 识别将其转换成 word 文档进行处理，然而 OCR 识别效果与文献原始质量相关，部分文献识别效果较差，无法通过规则匹配的方式提取专门章节，因此需要结合人工标注的手段进行章节定位工作。如图 A2 中“<tag-method-begin>”以及“</tag-method-end>”就是定位标记。提取学位论文的研究方法是进行跨学科研究的基础，在人工标注基础上，本研究将继续探究研究方法实体自动抽取以及方法的自动分类。

## A.3 标注过程说明

本节将对标注过程进行说明，分为自动识别、人工标注与人工审核三个部分。

### A.3.1 基于规则匹配方法的自动识别

在进行人工标注与审核之前，为了减少工作量，首先对数据及性预处理与自动标注，即自动识别过程，此过程可以理解为数据预处理过程，由笔者自行处理，无需数据标注志愿者参与，因其作为整个数据标注过程的一部分，因此在此进行简单介绍。首先，经过论文内容中研究方法章节的标题进行小规模调研，根据标题规律构建正则表达式用于标题匹配，构建例如“((([使运应]用)|研究).*[方范][法式])”的正则表达式对所有文献中含有研究方法小节的文献进行粗筛选，即首先将“不含”（机器无法识别）研究方法章节的文献剔除，提升模型的性能。其次是制定章节定位规则，即章节起始标志。在本课题中采取的规则是首先提取文献标题数据，对研究方法章节起始标题（上边界）进行识别，然后提取标题序号，如从“1.2.1 研究方法”中识别“1.2.1”。在得到序号之后，计算下一标题序号内容，如“1.2.2”或者“1.3”等等，由于程序比较复杂，在此不做赘述。得到下一标题序号或结构后根据正则表达式对后续标题字段进行识别，确定结束标题（下边界），在开始与结束位置进行标记即完成自动定位过程。

### A.3.2 人工标注与人工审核

经过自动识别过程后，由于上文中提到的 OCR 识别过程中的乱码以及其他问题，基于规则的匹配结果不是很理想，所以还有一部分需要标注员对其进行手动标注，另外，对于已识别的文献需要对其进行审核，以确保所有文献的研究方法章节均被准确标注，构建研究方法章节数据集，完成章节定位工作，为后续研究方法实体识别提供高质量预料。具体而言，人工标注过程为：首先对标注人员进行培训，确定标注员熟悉标注任务、标注过程以及一些注意事项，同时安装相关的软件。其次，管理员根据上一步中机器识别结果确定需要标注的文档与需要审核的文档，向标注员分发需要标注的数据进行，审核过程由管理员完成。最后标注员在指定日期内完成所有标注任务后将数据提交给管理员汇总，完成整个章节标注任务。

## A.4 标注步骤

### A.4.1 打开 Navicat 以及数据表

首先标注员打开 Navicat 软件，如图 A3 所示，分别双击①②③所示的“章节定位”（即设置的连接名）、“main”以及“表”，在标注员的界面中应该只存在 content 这一张表，双击 content 以后即打开了数据表。打开以后继续操作，首先点击需要标注的文档的第二列即“content”下的内容，选择右下角表单视图，如图中⑤所示，然后点击⑥所示文本选项，即可直接查看内容（否则直接查看会很麻烦）。请注意，在第一次使用时，如⑦所示，内容会被

压缩在软件靠下位置，需要将⑦所示的横杠向上拉到合适的位置进行标注。

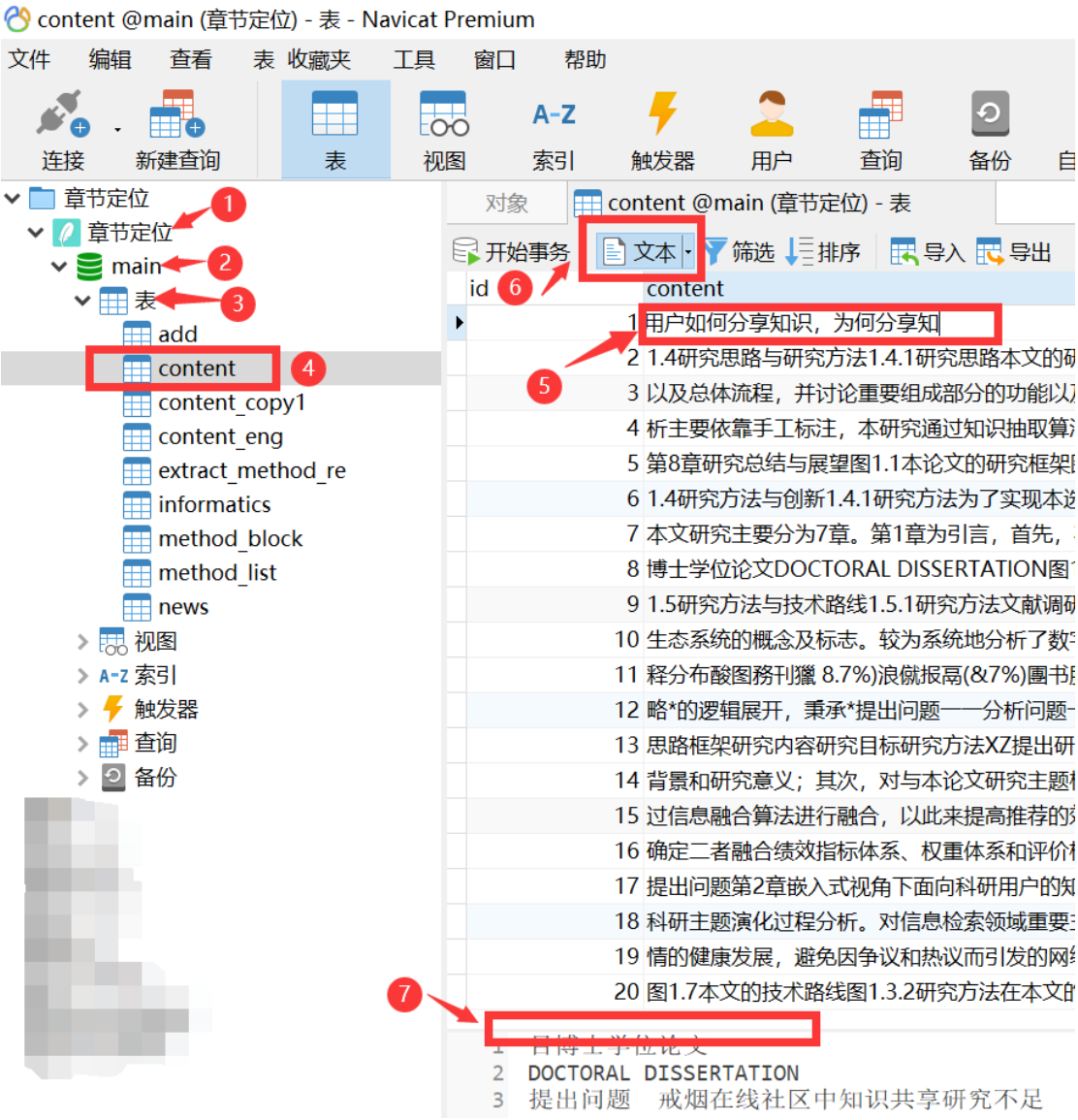


**图 A3 数据标注前的基本操作**

## A.4.2 文档标注

标注员需要对如下图 A4 所示文档进行标注，快速翻阅内容对于研究方法所在章节进行判断。

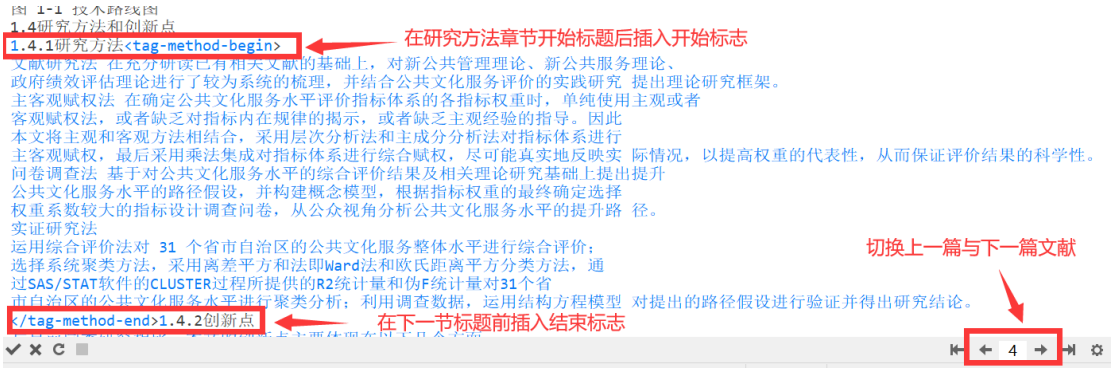


**图 A4 文档标注示例**

如图所示，可判断研究方法章节上下边界位置如图中紫色框所示，因此需要在“1.3.2 研究方法”后插入“<tag-method-begin>”，在 “1.4 研究意义与创新点”前插入“</tag-method-end>”，标注完成软件会自动保存，点击右下角左右按钮切换文档。无研究方法章节时，在当前文档任意位置插入“<no-method-article> ”即可。另外，在标注 nb 表时，有明显研究方法边界标记的文献使用“<clear-method-begin> </clear-method-end>”进行标记。为了方便标注，不方便观察上下界标题时，可直接在第一个出现的研究方法前插入开始标志，在最后一个研究方法介绍结尾处插入结束标志。另外，由于标记颜色和文本颜色在有些文本中是同色的，可以右键将语言改变为 HTML（默认为 CSS），如图 A5，能更好的观察到标记与文本之间的差异。

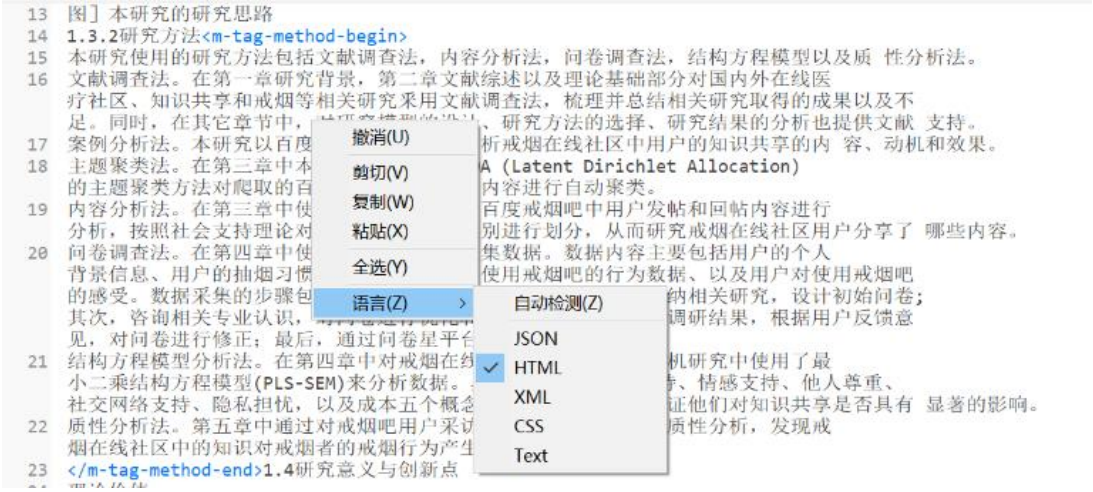


**图 A5 软件中设置显示的语言**

## A.5 标注相关软件工具使用介绍

本研究需要对研究方法所在章节进行标注，在边界上下位置插入标签进行界定，标注过程需要使用到的软件工具包括 Navicat、文本快捷插入插件。其中 Navicat 负责存储待标注文本，文本快捷插入插件负责设置标注标签的快捷键。

### A.5.1 Navicat 下载与使用介绍

软件可在官网下载并安装。安装完成以后，打开 Navicat 软件，使用 Navicat 创建新连接，并选择 SQLite（图 A6），设置连接名（图 A7，可设置为“章节标注”），类型选择现有数据库文件（即我这边导出的数据），选择文件位置，用户名密码可不设置，处理完成。


**图 A6 选择连接数据库类型—SQLite**


**图 A7 新建连接**

### A.5.2 文本快捷插入插件使用介绍

文本快捷插入插件使用较为简单，双击打开“快捷输入文本.exe”，软件截图如图 A8 所示，设置 F1-F10，点击启用即可开始使用，由于只需要标注出上下边界，此处 F1-F10 可根据个人喜好设置，将上边界标志设置为“<tag-method-begin>”，下边界设置为“</tag-method-end>”，未发现研究方法章节的在内容任意位置插入“<no-method-article>”。如图 A8 所示，分别使用 F1、F2、F3 表示三种文本的快捷键。值得一提的是，使用笔记本自带键盘时需要使用 fn+F1 等来使用。

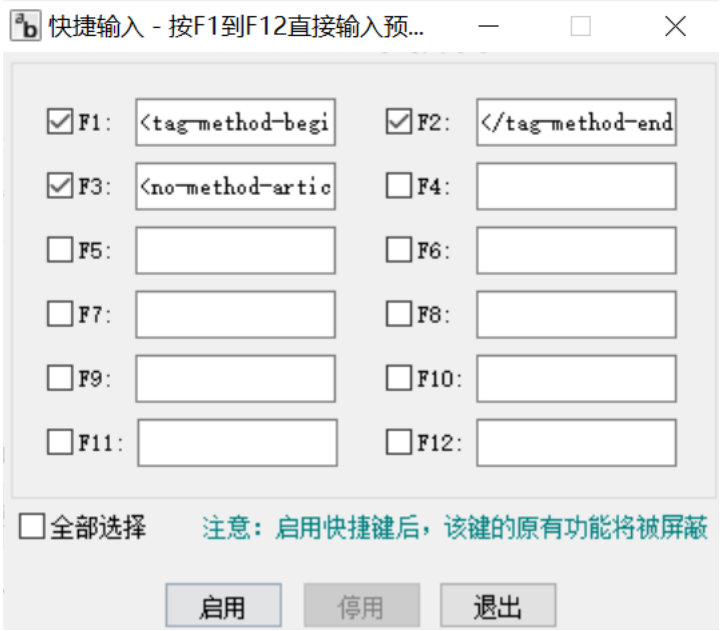


**图 A8 文本快捷插入软件截图**

## A.6 标注注意事项

在进行研究方法章节标注过程中可能会出现一些问题，本章将对常见问题以及解决思路进行描述。

### A.6.1 注意区分研究方法小节是否是本文的研究方法

部分文献中“研究方法”小节是文献综述中对现有文献的总结，不属于该文献使用的研究方法，真正使用的研究方法可能在后面才会介绍，要注意甄别。例如图 A9 所示这篇文献的研究方法小节，正是文献综述小节的一部分，这种研究方法小节是对以往某个主题研究中的研究方法综述，不是该学位论文所用到研究方法，不可在此进行标注。

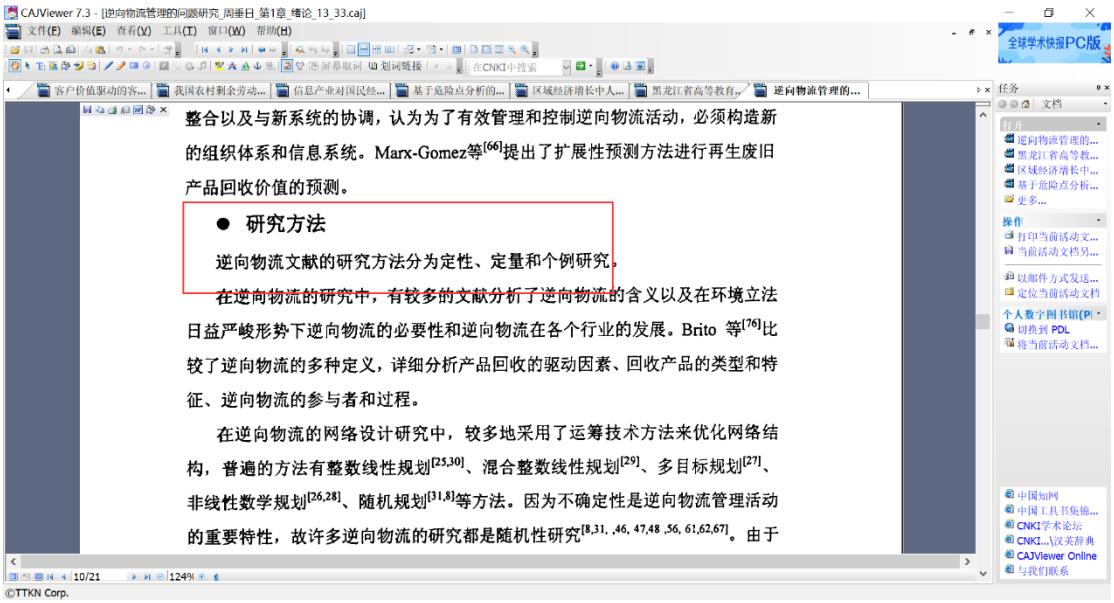


图 A9 文献综述中的研究方法小节举例

### A.6.2 根据 url 对内容缺失文献进行补充

```
1.3 研究方法与研究内容
研究方法

研究内容




17个
17
1.1
1.1
Fig 1.1 The Basic Structure of the Study
```

图 A10 未识别出内容的文献举例

在标注过程中，可能会遇到图 A10 所示的情况，即研究方法小节下没有提取到内容或者出现奇怪字符，这是由于部分文献年代久远字符不清等问题导致 OCR 无法识别出内容造成的；另一种情况是内容缺失，如图 A11 所示，由于文件大小与页数等原因，导致 OCR

过程提前结束，所以未识别全部内容。

```
63 1.3论文的研究思路和研究方法<m-tag-method-begin>
64 （论文的研究思路
65 论文围绕中石油自主创新机制的构建和自主创新能力的评价开展系统研
   究，首先总结评述了国内外研究的现状，阐述了论文研究所需的理论基础；
   然后在分析中石油自主创新现状的基础上，明确了中石油自主创新存在的问
   题；论文进而分析了中石油自主创新机制的内涵、特征和作用机理，为构建
   中石油自主创新机制奠定了坚实基础；以此为基础，系统构建了中石油自主
   创新机制；之后，论文构建了中石油自主创新能力评价的指标体系及模糊层
   次评价和网络层次评价模型，并进行实证研究；最后提出了提高中石油自主 创新能力的对策和建议。论文大致可以分为6个部分：
66
```

**图 A11 OCR 识别提前结束举例**

对于上述两类文献，需要使用标注数据中提供的 url 字段，在校园网的环境下（或者使用校外访问）根据 url 地址找到指定文献，找到指定文献后，点击分章下载或全文下载，找到指定研究方法章节进行下载（如果没有研究方法小节，可以直接下载绪论小节进行查看），如图 A12 所示，再使用阅读器自带的文字复制功能或者相关软件的图片转文字功能（如 QQ 的“Ctrl+Alt+A”快捷键），将研究方法介绍相关的文字内容复制后保存到数据库中，再进行标注（不需要全文复制，只需添加研究方法小节内容）。


**图 A12 文献下载示例**

### A.6.3 不同语种问题

在标注过程中还存在一些非中文文献，即使用的语言为其他国家语言，如日语、俄语等等（英文文献已剔除），对于这类文献，请使用标记“<Non-Chinese Documents>”进行标记，不需要再进行其他标注。

### A.6.4 繁体字、错别字问题

在标注过程中，如果遇到繁体字、或者错别字等问题，在保证标注进度的情况下可以下载原文进行修改，或者只对标题进行简单修改，例如“调查研究法”被识别成了“调查研究法”等可以直接进行修改。同样，仅对研究方法章节文字进行修改即可。

### A.6.5 其他注意事项

另外，在标注过程中可能还存在着诸多其他问题，例如句中分段、页码、参考文献出现在句中等等，标注人员需要注意审查，保证上下界标记之间为研究方法小节内容，无其它干扰内容，确保标注质量。