

文本内容分析法训练营

不用懂编程、超快速上手内容分析方法和文本挖掘工具

DiVoMiner[®]
Data in Value out

让研究更容易

课程简介

- 内容分析法是一种对文本内容进行分类、编码、语义判断及形成可供统计分析之用的量化分析方法，它广泛应用在社会科学研究领域。近年来，随着研究技术的进步和研究场景的演变，传统内容分析法也不断与时俱进，发展出了结合人工智能算法和大数据技术的全新的内容分析法。
- 全新的内容分析法既是一种机遇，又是一种挑战，尤其是对于那些缺少计算机编程经验的社科研究者而言，就算有再好的研究思路，如果不会写代码、没有好用的研究工具，面对海量的文本资料的时候也是一筹莫展。
- 本次内容分析方法训练营，恰恰正是为了解决那些不懂编程的社科领域广大师生的内容分析和文本挖掘需求而组织的。
- 本次训练营由新传学苑和DiVoMiner[®]研发团队联合定制、联袂推出，旨在帮助广大人文社会科学领域的本硕博学生和教师群体，在最短的时间内、以最快的速度、用最优的方法学会最好用的内容分析方法和文本挖掘工具。
- 通过1整天6小时的训练营课程，报名学员将会学习到，即使不会编程，也可以借助现成的大数据技术平台DiVoMiner[®]，进行文本内容分析，省却80%处理数据的时间，协助产出高质量的论文。

张荣显博士

澳门互联网研究学会会长
前澳门大学传播学系教授



港澳地区的互联网及大数据研究先锋及实践者，网络传播学会副会长，亚太区互联网研究联盟(APIRA)现任主席。专长于互联网研究、民意研究、网络挖掘、电子政务及大数据项目策划与落地。于2009年创办澳门易研方案研究咨询机构及易研信息科技，其后在珠海横琴创立博易数据技术有限公司，独家研发DiVoMiner®文本大数据挖掘与分析平台。创业前，任教于澳门大学传播系十多年，曾获「世界民意研究学会」2003年伊利莎伯·尼尔森最佳论文奖。

主讲嘉宾介绍**丁奕 香港浸会大学传播学博士**

香港浸会大学传播学博士，应用语言学硕士，曾赴密苏里大学新闻学院以及传播学院交流。研究方向与兴趣为新闻学、内容分析、框架分析、政治传、媒介效果等。博士论文课题为《中美贸易战美国精英媒体框架分析及其影响》，在论文中借助DiVoMiner®作为内容分析/框架分析的研究工具，将质化和量化分析方法相结合，获得了答辩评审的高度称赞，正在利用论文中的多个实证研究和发现开展后续研究，拟发表三篇海外顶级期刊论文。曾多次参加各类国际传播会议，分享阶段性以及总结性研究成果，发表多篇关于头条新闻倒U型螺旋效果分析，以及华人对于海外大众媒介的观感分析等的科研论文。

曹文鸳 博易数据首席研究顾问

澳门大学文学硕士，有着十余年的科研经验，张荣显博士研发团队核心成员，深度参与DiVoMiner®文本大数据挖掘与分析平台的搭建和服务。基于严谨的社会科学研究方法，结合大数据技术，经过技术沉淀和优化研究方法流程，逐步受到学术界和业界的青睐和肯定。多年来积极参与学术活动，在国内外学术期刊和国际研讨会发表多篇论文。

真正让研究更容易的文本大数据挖掘与分析工具

DiVoMiner®是一个助力研究和教学的在线数据挖掘与分析平台，以内容分析法为设计核心，利用机器学习编码，人机结合的操作流程，在线完成内容分析法的全部流程，并提供灵活而强大的研究执行及团队协作管理功能，是市场上唯一一个兼具实用性和学术性要求的文本内容挖掘和分析平台。

DiVoMiner®基于云计算、大数据技术架构，以SaaS订购账号的服务模式，用户无需安装软件在本机，只需通过浏览器就可登入平台，线上可处理定量与定性数据的编码与统计分析，也可自行上传文本或定量数据进行编码与统计。

DiVoMiner®运用自然语言处理、机器学习等技术，内置多个大数据文本挖掘算法模型，例如自动化情绪分析、语义网络分析、主题提取等等，也可按需定制化模型。

传统方法，创新执行！

让80%时间用在研究设计及分析上，20%执行劳务性工作

DiVoMiner®
Research Done Right

文本大数据挖掘及分析平台

大大缩短研究时间 高质高效完成学术报告

- ✓ 定量和定性分析
- ✓ 在线编码
- ✓ 编码员间的信度测试
- ✓ 随机抽样
- ✓ 统计分析
- ✓ 情绪分析
- ✓ 自动化主题提取
- ✓ 词关系分析
- ✓ 社交网络分析
- ✓ 数据探索及预处理
- ✓ 团队在线协作管理
- ✓ 语义网络分析

请马上关注公众号
获取第一手研究干货

DiVoMiner



使用DiVoMiner®平台完成的成果案例（部分）

已经有很多学者利用DiVoMiner®写作并发表了多篇核心期刊、C刊论文乃至SSCI期刊论文，想要了解这些论文的详细情况，可以扫描二维码关注“文本数据挖掘与分析”公众号进行查看。

这些论文和研究成果包括新闻传播、公共管理、政策分析等多方面题材。本次课程将邀请高水平的学术论文作者，分享研究经验和心得。



危机管理

- Mak, A. K. Y., & Song, A. (2021). **Regenerative crisis, social media and Internet trolling: A cultural discourse approach**. *Public Relations Review*, 47(4).
- Mak, A. K. Y., & Song, A.. (2019). **Revisiting social-mediated communication model: The Lancome regenerative crisis after the Hong Kong Umbrella Movement**. *Public Relations Review*, 45(10), 10812.

直播研究

- 王建磊.(2018). **如何满足受众：日常化网络直播的技术与内容考察**. *国际新闻界*, 40(12), 21-33.

企业形象

- 赵莹, 林坤燕 & 张荣显.(2018). **中国企业在海外新闻媒体中的形象研究**. 中国报告2018. 北京: 中国社会科学出版社. 189-210.

国际传播

- 马诗远 & 郑承军.(2021). **新信息环境下海外社交媒体中的北京形象研究**. *现代传播(中国传媒大学学报)*(07), 150-157.
- 赵欣.(2021). **国际传播视野中的中国故事叙事之道——“第一主讲人”人类命运共同体意涵的国际分享**. *新闻与传播研究*(01), 5-25+126.
- 王丹 & 郭中实.(2020). **整合框架与解释水平: 海内外报纸对“一带一路”报道的对比分析**. *新闻与传播研究*(03), 5-20+126.
- 万晓红, 周榕 & 周晓宇.(2018). **“中国体育形象”的媒介呈现-基于国外主流媒体里约奥运会涉华报道文本分析**. *中国国家形象传播报告(2017~2018)*, 243-267.

管理学

- 单学鹏 & 罗哲.(2021). **成渝地区双城经济圈协同治理的结构特征与演进逻辑——基于制度性集体行动的社会网络分析**. *重庆大学学报(社会科学版)*(02), 55-66.
- 曹文鹭, 赵莹, 林坤燕 & 张荣显.(2018). **基于大数据的公共服务研究：对海峡两岸暨香港、澳门公共自行车的新闻报道的内容分析**. *新媒体与社会*(01), 239-263.

舆情研究

- 曹文鹭, 赵莹, 姚欣妤 & 张荣显.(2017). **“一带一路”倡议在澳门舆论环境中的传播**. 舆情蓝皮书: 中国社会舆情与危机管理报告(2017), 北京: 社会科学文献出版社. 291.
- 张荣显 & 曹文鹭.(2016). **网络舆情研究新路径: 大数据技术辅助网络内容挖掘与分析**. *汕头大学学报: 人文社会科学版*, 8(1), 111-121.

健康传播

- Chang, A. Schulz, P. J. & Cheong, A.. (2020). **Online Newspaper Framing of Non-communicable Diseases: Comparison of Mainland China, Taiwan, Hong Kong and Macao**. *International Journal of Environmental Research and Public Health*.
- 程萧潇, 金兼斌, 张荣显 & 赵莹.(2020). **抗疫背景下中医媒介形象之变化**. *西安交通大学学报(社会科学版)*(04), 61-70. doi:10.15896/j.xjtuskxb.202004007.

危机管理

- Mak, A. K. Y., & Song, A. (2021). **Regenerative crisis, social media publics and Internet trolling: A cultural discourse approach**. *Public Relations Review*, 47(4).
- Mak, A. K. Y., & Song, A.. (2019). **Revisiting social-mediated crisis communication model: The Lancome regenerative crisis after the Hong Kong Umbrella Movement**. *Public Relations Review*, 45(4), 10812.

本次课程以传统内容分析法为开篇，明晰其概念、背景、要素。随科技进步，内容分析法也有了新的发展。尤其是到了大数据时代，产生了对文本大数据的研究需求。在传统方法和创新执行的碰撞下，大数据技术辅助在线内容分析法应运而生。我们将围绕研究方法，介绍DiVoMiner®平台的设计理念，并详细拆解流程图的每个步骤。学员可在该体系下，理解内容分析法的来龙去脉，以及前沿的研究执行方法。

目录

CONTENTS

01

**理论基础：
传统内容分析法**

02

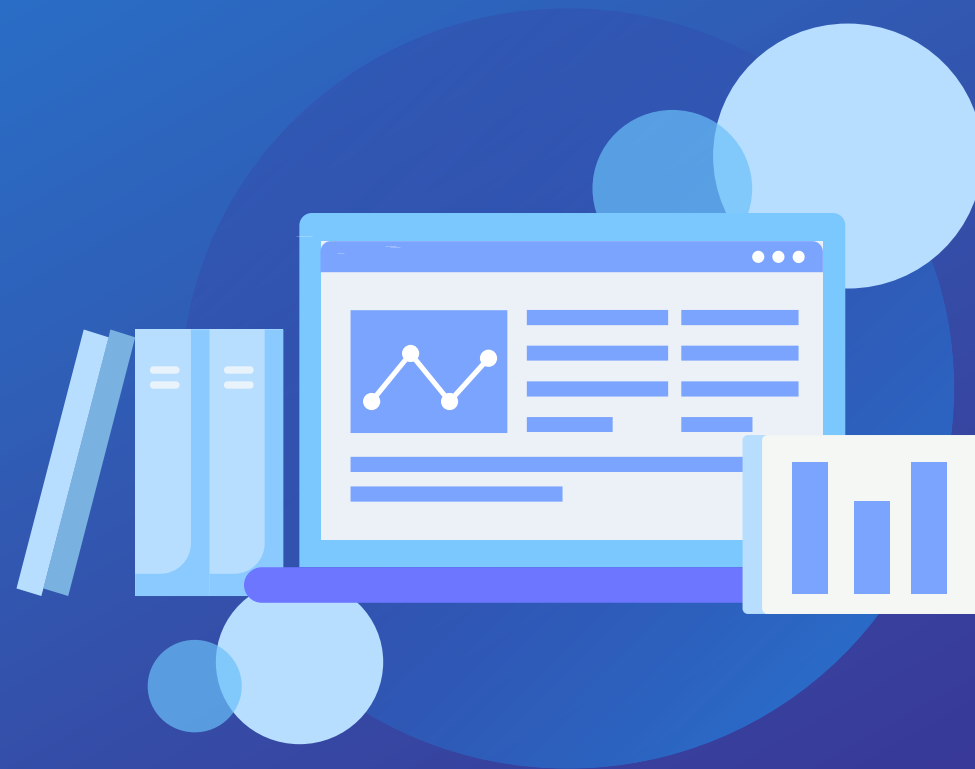
**内容分析法新路径：
传统方法与创新执行**

03

**DiVoMiner®
大数据技术辅助在线内容分析法
操作流程及优势**

01 理论基础

传统内容分析法



传统定量问卷调查 (小) 数据 vs. 文本 (大) 数据

传统意义上的小数据

- 研究人员主导，结构化问卷，对象是人
 - 结构化
 - 抽样方式
 - 抽样偏差
 - 代表母体
- 时间周期长
- 细节模糊
- 敏感度低

如何从人来理解人的行为和想法

大数据时代的文本数据

- 研究人员被动，开放式数据，对象是文本
 - 非结构化
 - 普查方式
 - 数据源偏差
 - 部分全覆盖
 - 实时
 - 细节清晰
 - 敏感度高

如何从文本来理解人的行为和想法

内容分析法(Content Analysis)

- 是社会科学研究方法中的一种对文本内容进行编码、分类、语义判断及形成可供统计分析之用的量化分析方法。
- 以学术的说法, 它是指一种以系统、客观与量化的方式, 来研究与分析传播内容, 藉以测量及解读内容的研究方法 (Kerlinger, 1973)。

系统的方法

- 随机样本、系统的类目建构与编码程式

客观的程序

- 遵守明确的标准与规则

量化的分析

- 为所有的变量下操作性定义, 确定测量尺规, 进行内容编码及统计分析

Lasswell's better known statement which succinctly encapsulates what media content analysis is about, published in 1948, (as cited in Shoemaker and Reese, 1996), describes it as:

Who says what through which channel to whom with what effect

常用的文本分析研究方法

(详细可参阅“文本数据挖掘与分析”公众号文章)



	内容分析法(Content Analysis)	扎根理论(Grounded Theory)	文本分析(Textual Analysis)	话语分析(Discourse Analysis)
研究类型	定量研究	定性研究	定性/定量研究	定性研究
应用领域	社科研究	社科研究	计算机、语言文学领域	社科研究
哲学基础	实证主义	建构主义	诠释主义	诠释主义
研究对象	任何一种可被传播的信息，包括文字、图像、符号、视频、音频等记录在案的资料	访谈、反思、文本、文献、观察、问卷、备忘录等	文字、图像、符号、视频、音频等记录在案的资料	口头+书面话语
代表人物	贝雷尔森等	格拉塞、斯特劳斯等	不明确	费尔克拉夫等
分析逻辑	演绎法	归纳法，从经验资料基础建立理论	归纳法	归纳法
研究目的	1.描述、解释文本内容特征和性质 2.判断趋势效果及差异 3.验证理论	1.强调发现理论 2.分析数据作为一种解释 3.理论是解释性分析，是建构的	1.找寻文本意义功能；2.描述文本内容结构；3.理解文本相关变量； 4.评述资料事实结果； 5.挖掘文本背后意识形态和权力结构	1.分析传播立场语境；2.揭示背后观念受众；3.解释传播内容语言； 4.考察华语实践作用； 5.挖掘文本背后意识形态和权力结构
研究步骤	1.确定研究问题或假设；2.抽取样本； 3.定义分析单位；4.类目建构； 5.编码员信度测试；6.内容编码； 7.编码结果检查；8.解释与分析结论；	1. 开放式和选择性编码；2. 不断比较；3. 理论性采样；4. 理论性报告； 5. 理论性编码（三级编码）；6.写备忘录和手工整理备忘录	1.文本查阅 2.鉴别评价 3.归类整理	1.确定理论；2.选择研究问题； 3.样本选择和收集； 4.资料整理；5.资料分析； 6.总结和得出研究结果
信度检验	编码员间信度	无	无	无
流程标准化	标准化，要求系统化、客观性、可量化	无标准化，经验证据	无标准化，间接主观	无标准化
研究团队配置	研究员、编码员、研究督导（常由研究员兼任）	研究员	研究员	研究员
适用工具	DiVoMiner®、DiscoverText、PRAM（计算信度）	Nvivo、ATLAS.ti、MAXQDA	DiVoMiner®、Nvivo、ATLAS.ti、MAXQDA、LIWC、CiteSpace、UCINET、KH Coder、WordStat、WordSmith、BICOMB、AntConc、Python、R	常与文本分析工具混合使用

历史背景

- 内容分析(content analysis)”一词最早见于1961年的韦伯字典(Krippendorff, 2013) ，它的历史可追溯到17世纪教会倡导的圣经注解的文本分析。
- 哈罗德·D. 拉斯韦尔(Harold D. Lasswell)针对第一次世界大战宣传技巧进行了内容分析，成为传播学史上首次对宣传进行分类的实证研究。
- 第二次世界大战期间，拉斯韦尔在“战时传播研究”(the Study of Wartime Communications)的工作中，对过对德国公开发行的新闻媒体进行内容分析和研究，着力探讨了抽样、测量、类目及信度和效度等一系列内容分析的核心问题，将内容分析发展的更为成熟。
- 战后用于研究大众传播媒介的宣传方法、描述传播内容、比较媒介真实与社会真实、评估特殊社会团体的形象、建立媒介效果研究的起点等。
- 20世纪60年代后，内容分析法进入美国各大高校的课堂。1971年，内容分析法被列为从1900年至1965年62项“**社会科学的重大进展**”之一。随着相关研究的出现，内容分析法也逐渐成型并广泛的应用于各领域。

哪些资料可以用来做内容分析？

- 内容分析法的研究对象，可以是任意一种可被传播的信息，包括单词、意义、图像、符号、思想、主题等，可用于内容分析的资料包括书面的、视觉的或是口头表达的，例如书籍、文章、采访内容、讨论内容、报纸标题和报道、历史资料、演讲、谈话、广告、戏剧、非正式交谈或者任何交流性的语言。在社会科学领域，可以用来做内容分析研究的资料有新闻报道、社交媒体内容、文学作品、历史档案、访谈、学术文献、政策文本、发言稿、图片和视频等。
- 内容分析法的优势主要体现在它作为**非介入性研究** (Unobtrusive Research)，不受变量的类型以及信息生产或呈现的背景限制，是一种总结性的、依赖科学研究过程的分析方法，也是一种可复制的且可进行有效推论的研究方法。互联网时代，使用便利、成本低廉的数据库和搜索引擎使得内容分析法更加适合于资源匮乏而有志于实证研究的学生和青年教师。
- 即便内容分析法有这些优势，但烦琐的劳务性工作，例如文本处理和数据分析过程，仍会消耗大量的人力、物力和时间，因而，计算机在内容分析的过程中扮演着辅助的角色。

内容分析实例

- 美国911事件发生后，研究人员考察了CNN、ABC、CBS和NBC在最初5个小时突发性新闻的报道，展示了这一次恐怖袭击如何让新闻记者感到巨大震撼，以至于他们放弃了传统的角色观念和专业主义标准，报导起个人的想法和听到的传言，并且使用匿名消息来源(Reynolds & Barnett, 2003)
- 每五年一次对《生活》、《新闻周刊》和《时代杂志》刊登的非洲裔美国人的照片进行分析，以编年史的方式反映出了从20世纪初期到后期，就美国社会各个层面来说，黑人的角色日益重要(Lester & Smith, 1990)。
- 心理学上，四代心理学家对丹尼尔·斯瑞博的《一个神经症患者的回忆录》进行内容分析，将其中的词语划分为“奇异的”、“明显精神病的”和“本质上的妄想症”等类型，再和健康人的自传中的词语统计进行对照分析，也把作者正常期和急性妄想症时期写下的东西进行对照，最终寻求心理疾病归因。
- 比较不同报章的社论对某某事件的评论。
- 研究中西方电视广告呈现的价值观。
- 分析某某两候选人演说的修辞技巧。
- 。 。 。 。 。 。

内容分析实用功能

- 揭发社会问题
- 探讨社会趋势
- 评估媒介表现
- 商业策略分析
- 竞选活动
-

DiVoMiner®

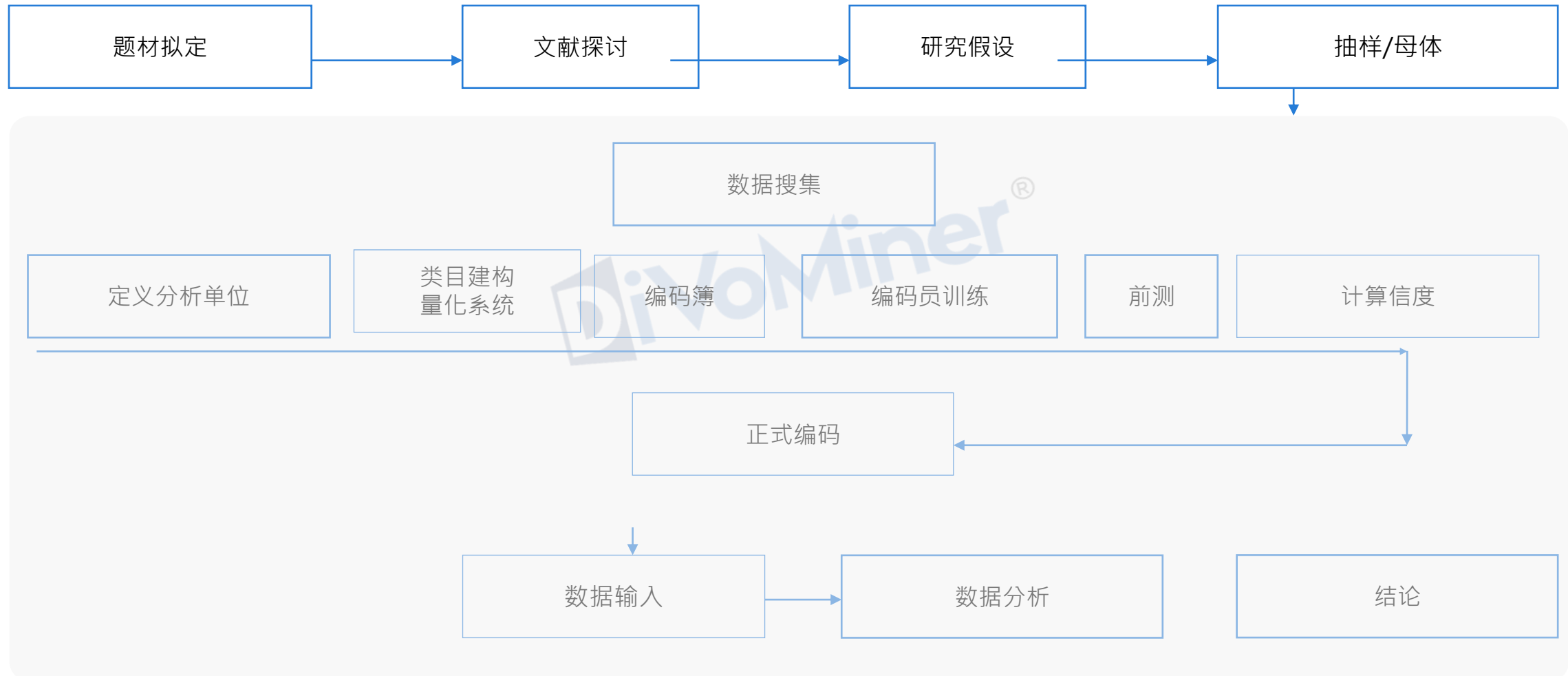
研究设计模式

- 比较传播内容的趋势：分析同一传播来源的内容，在不同时期或阶段的变化。例如：分析过去10年来新闻媒体的环保新闻报导主题的变化。
- 评估情势对传播内容的影响：探讨同一传播来源的内容，在不同的历史、政治、文化情势下，会有哪些变化。例如：分析广告在互联网前后时代的设计手段变化。
- 评估读者对传播内容的影响：探讨同一传播来源，面对不同的读者，是否会生产不同的内容。例如：分析《人民日报》国内版和海外版有关国际贸易议题报导的差异；比较政治人物针对不同群众发表演讲的内容。
- 分析传播内容变项间的关系：分析同一传播来源中，不同内容的关联性。例如：分析某自媒体所发表的大量内容之间，是否有关联性；分析同一电视台的不同节目所呈现的价值观念，是否有关联性。
- 比较传播者之间的差异：比较不同的传播来源的内容，藉以推论传播者之间的差异。例如：比较精英与大众新闻媒体的社论立场，以探讨不同读者导向的新闻媒体之社论立场，是否有所不同。
- 评估传播者的表现：在采取某一特定的标准，来评估传播者的表现。例如：以警方的记录为标准，和报章对「暴力案件」的报导作一比较，以评估新闻报导是否正确；以某国的民意代表在选举前的政见为标准，和当选后的质询内容或言论作一比较，藉以评估民意代表的表现。

内容分析法研究程序

- 选择研究题材
- 熟悉研究题材（文献探讨）
- 提出研究问题或假设
- 从母体中进行适当的抽样
- 定义分析单位
- 类目建构
- 建立量化系统
- 训练编码人员及进行前测
- 信度分析
- 按照先前建立的定义，对内容进行编码
- 资料分析与解释
- 结论

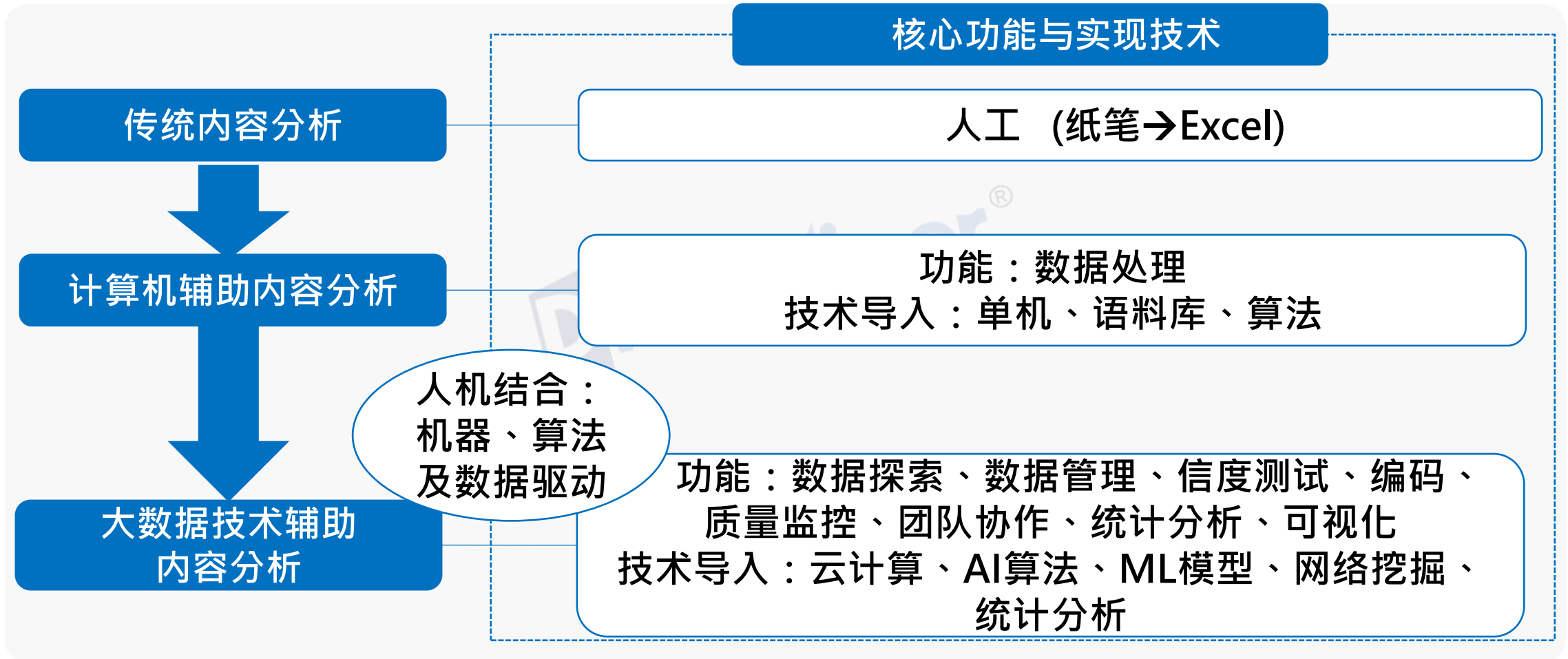
一个传统的内容分析流程图



02 内容分析法新路径： 传统方法与创新执行



内容分析法的技术演变



——以皮尤研究中心(Pew Research Centre)为例

- 皮尤研究中心于2016年进行一项研究项目，目的为研究在美国总统竞选中，立法者公开表达的政治敌意的情况。该研究采用一种新型的混合研究方法，结合机器学习技术和传统内容分析法。研究者采集了自2016年1月1日至2016年4月30日、超过20万条相关新闻报道和Facebook讨论帖内容。在整体数据中，按规则抽取7,000条新闻报道和Facebook讨论帖，并针对这部分抽样样本进行人工编码，编码结果作为机器学习的训练数据集(Training data)，使用电脑分析后形成关系逻辑判断，利用该机器学习结果再自动完成剩下样本的分类与编码。
- 研究结果发现，在新闻报道和Facebook表达渠道中，极端的负面表达非常之少，大约有10%的新闻报道和9%的Facebook讨论帖表达出了反对其他党派的意见，其中包括愤怒、怨恨和烦恼等情绪。
- 局限：1)类目设置仅能用简单的二分法；2)操作难度大，需要针对单独项目专门编程已完成机器学习。

- Does the document mention a specific benefit for people in the district?
- Does the document discuss a foreign policy or international issue?
- Does the document mention one or more members of both parties agreeing on something or working together in a bipartisan manner?
- Does the document describe the U.S. political system/process or government as broken, corrupt, wasteful, or dysfunctional?
- Does the document express any indignation or anger?
- Does the document express opposition toward or disagreement with any of the following?
(Select all that apply)

什么是文本大数据？

文本类数据 (Text data)

新闻、社交网络、访谈文字记录、历史档案、文献文档、政策文本、文学作品、领导发言稿，包括文字、图片、视频等等

感应数据 (Sensor data)

• 雷达、GPS、CCTV摄像、读卡器、RFID等

网络日志 (Web logs)

• 用户浏览行为、选择行为等

交易数据 (Big transaction data)

• 行业/企业：电商、通讯、连锁超市、信用卡等

参考数据 (Big reference data)

• 国家资料、工商黄页、个人资料、社经指数等

高
非结构化程度
低

大数据的特征



文本大数据研究中的难题

关于自动化人工智能应用于文本大数据研究中的问题

感知→认知→判断

弱人工智能 (AI, Artificial Intelligence)

- 利用机器来模拟人的某些特定技能的智能，来处理一些特定场景和应用问题
- 例子：常见的语音/图像/人脸识别、自然语言处理、信息检索、自动驾驶、智能控制机器人等等
- 他们比较偏向于感知（识别）层面的水平

强人工智能

(AGI, Artificial General Intelligence)

- 指达到或超越人类水平的、能够自适对外界环境挑战的、具有自我意识的人工智能系统
- 例子：尚未可见
- 他们比较偏向于认知和判断层面的水平

文本数据的分析中涉及**认知**及**判断**层面时遇到的问题

- 中文语境·上文下理（场景、常识）
- 文本中的具体指向物（尤其出现多个时）
- 反讽、暗语
- 价值判断
- 多变量关系

人工智能 / 机器学习的“能”与“不能”



从感知、认知到判断:

- 重要的不是我们所看到的可视化结果（冷冰冰的图表），而是这些结果能给我们做出判断的信心和确定性。
- 这就需要从一开始的数据库建立开始，到设定分析框架到测量到分析，由我们“人”来掌握！

“你能穿多少就穿多少”，这句话在冬天的时候跟夏天的时候，含义是完全不一样的。

另外一个就是说北漂，老家的妈妈经常催他结婚，然后他回答“我原来喜欢一个人，现在喜欢一个人”

~引自网络~



当前最需要关注的处理文本大数据的3大挑战

覆盖 (Coverage)

数据是否齐全/代表性?

测量 (Measurement)

可以测量什么? 如何测量?

解释 (Interpretation)

如何分析、挖掘及解释发现?

归根到底，还是3个社会科学研究中永恒的问题：
信度！效度！变量之间的差异及关系！

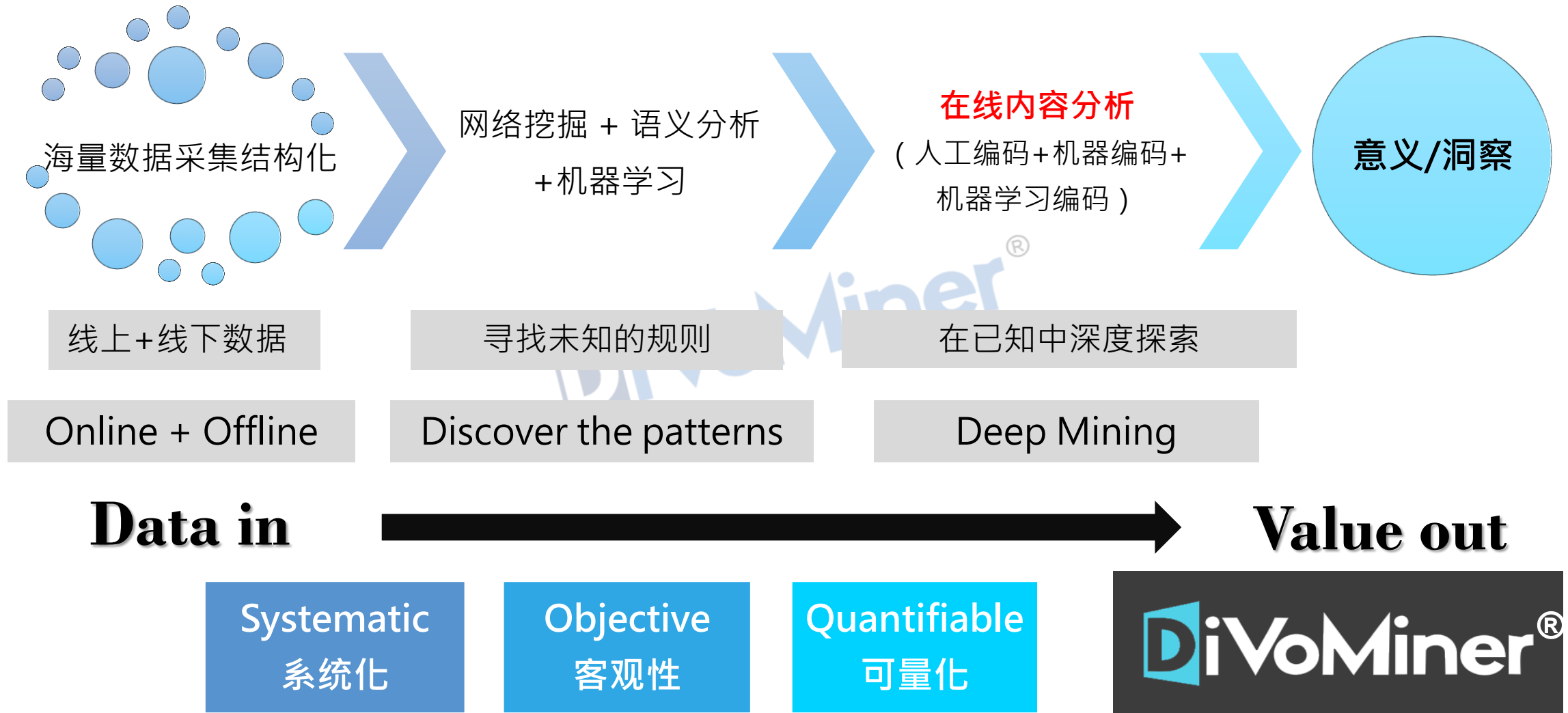
怎样迎接3大挑战？

我们提出新路径：

大数据技术辅助在线内容分析法

Big-data-tech-aided Online Content Analysis (BACA)

独创：大数据技术辅助在线内容分析法 [兼具量化与质化分析]





不会计算机编程的救星

两大组成部分

文本大数据挖掘及分析系统

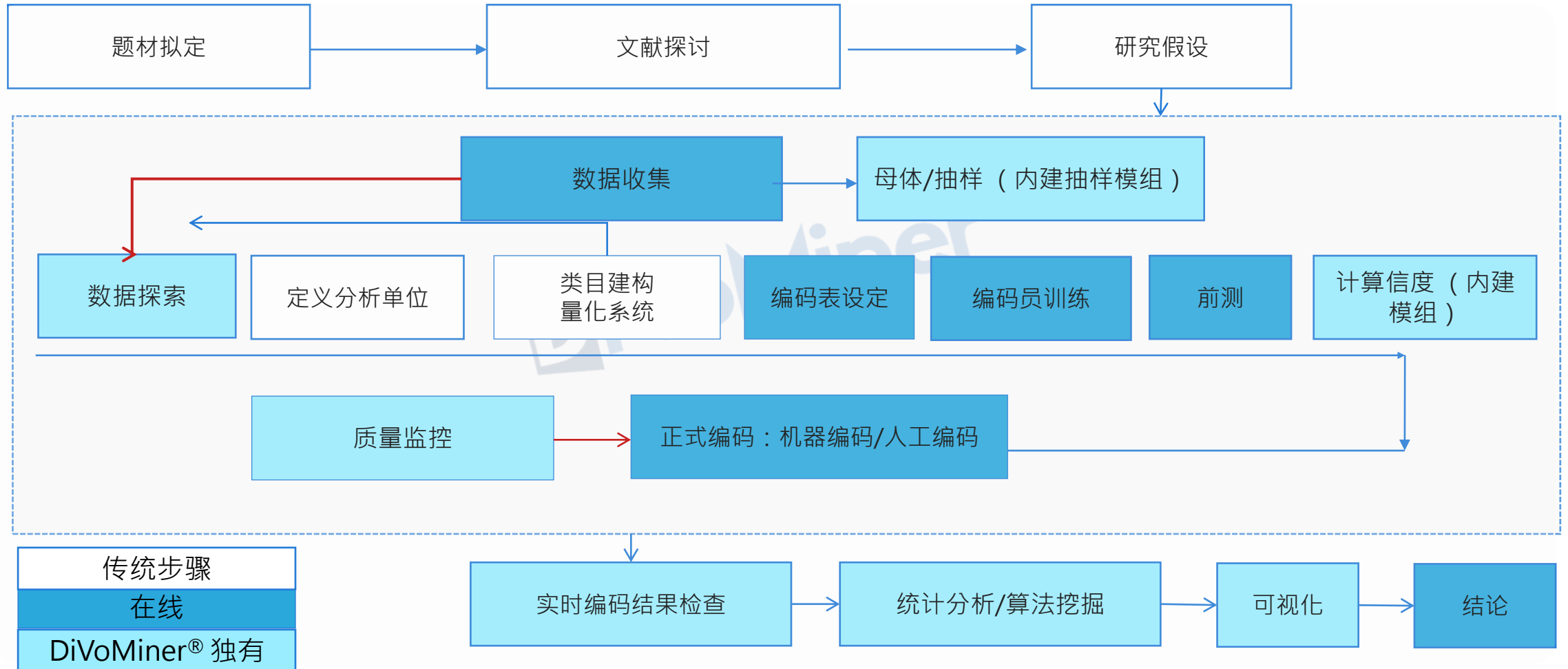
来自全球网络可爬取的数据及自行上载的数据
(线上+线下)

DiVoMiner®利用大数据技术与自然语言处理技术建构的 数据管理及挖掘与分析模组



此外，可根据深度用户需求，在人工内容编码后，插入并调用用户自建之机器学习模型

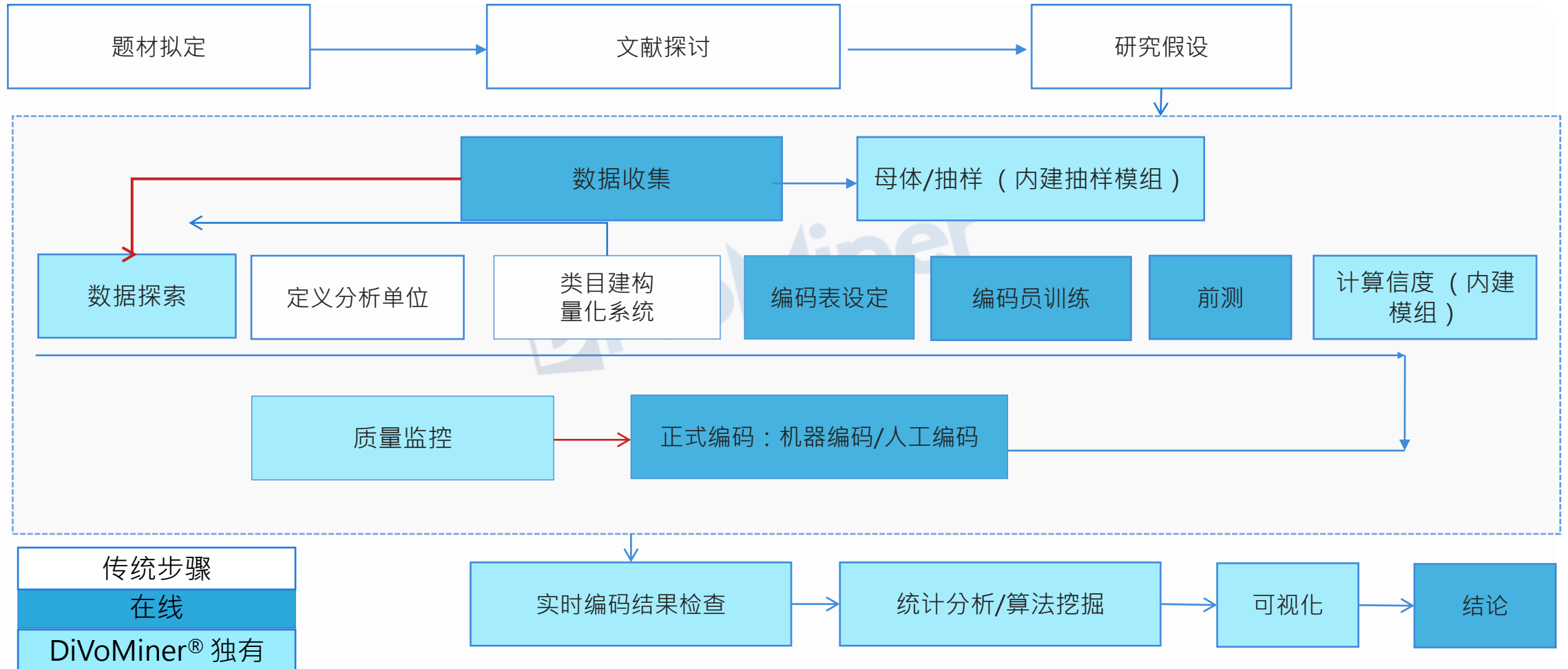
DiVoMiner[®] 大数据技术辅助在线内容分析法流程图



03 DiVoMiner[®]大数据技术 辅助在线内容分析法 操作流程及优势



DiVoMiner[®] 大数据技术辅助在线内容分析法流程图



从理论到数据的落地：概念->维度->指标

- 在以理论验证或回答研究问题的研究中，会涉及到由理论概念化，概念测量等要求。研究者从一个研究话题相关的观点出发到测量的过程，是一个概念化、操作化和测量的过程（Babbie, 2007）。概念本身难以直接测量，这是因为在概念到文本数据之间存在对应理解落差，因此在文本分析的研究中，通常会有一个将概念具体化的步骤，即是将概念落实到维度与测量指标。从理论过渡到数据的解读，落实到操作层面，则是以编码类目为桥梁。
- **拟定研究问题或提出研究假设**
- **设计指标及类目**
- 提出研究问题后，很多时候难以直接用于文本分析，即是在测量层面上有难度。因此，需要给涉及的每个概念设定维度和指标，其中指标层级就是概念的具体化结果，可用于制作内容分析的编码类目，提供给编码员使用的编码簿。

操作流程 – 定义分析单位

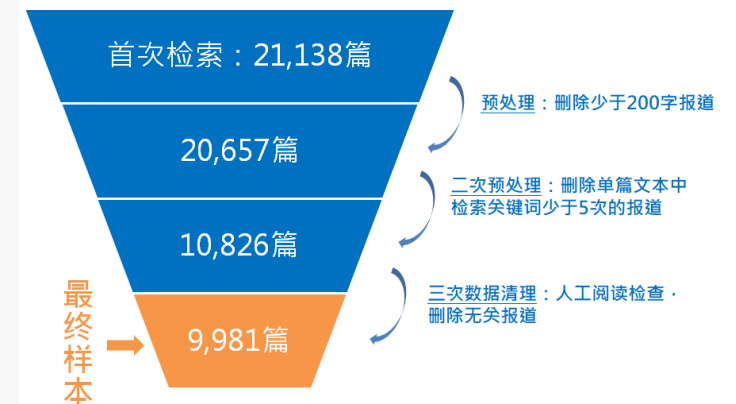
- 分析单位是研究过程中被分析的主体，主要是“**什么**”或“**谁**”被分析。就像化学和物理学一样，其中所有物质都可以基于原子或分子单位（忽略更深层次的亚原子粒子物理学），任何其他科学或其他科学理想的领域应该是以元素的不可约的基本组合为基础（Ernest, 2016）。
- 内容分析将分析单位定义为一段明确的内容，特点是将其归入一已设定的类目中（Holsti, 1969）。亦即是说，当编码员针对一段传播内容编码进行一个类目编码时，这一段传播内容就是一个记录单位。记录单位是搜集抽样单位中的资料，提供分析的基础。

分析单元的操作定义和实例

研究者	主题	总体	样本	分析单位
Bramlett-Solomon和Subramanian(1999)	杂志广告中的老年人形象	1990-1997年《生活和乌木树》中的全部广告	除分类广告外的所有广告	65岁以上的人物，或者有灰白头发，面部多皱纹，使用拐杖
龚金红，彭家敏，谢礼珊(2011)	了解顾客不公平行为的具体表现及其背后的认知因素	在服务过程中顾客对员工(或企业)不公平的行为	68名MBA学员写在服务过程中顾客对员工不公平的行为文本	文本中描述或罗列的行为要点
Signorielli和Kahlenberg (2001)	黄金时段电视节目	1990-1998年黄金时段的广播电视	每年秋季或春季一个星期的节目	主角、配角及其他辅助角色从事的活动
潘玉青(2014)	来了解目前幼儿在幼儿园中接触到的是怎样的儿歌，是否有助于我们的幼儿更好地发展。	温州市幼儿园教材中的儿歌	《幼儿园体验·探究·交往课程》，《教育活动设计》，《幼儿园建构式课程》（三套十八册）中的所有儿歌	教材中的每首儿歌

数据准备：需要考虑覆盖度及数据质量

- 在社会科学领域，可以用来做内容分析研究的资料包含新闻报道、社交媒体内容、文学作品、历史档案、访谈、学术文献、政策文本、发言稿、图片和视频等。到了大数据时代，网络数据大而全，同样需要考虑数据齐全、具有代表性，保证数据质量等问题，否则数据样本失焦，难以满足研究要求。
- 根据研究主题确定数据范围
- 架构概念化逻辑，检索获取数据样本
- 在研究实践中，检索条件的设计优劣，会影响到数据样本结果的数据量和准确度。使用概念化逻辑检索的思路，利用多元检索关键词搭配，配置一定的逻辑语言，建立一套检索概念（有点像建立检索概念数据库），利用检索词在概念上的相关性，检索获得同属一类概念的结果。
- 清理数据
- 由语言表达的复杂性和网络信息的杂乱无章，即便使用概念化逻辑条件检索文本数据，依然无法保证数据恰如其分刚好是研究范畴内的数据。所以获取初步数据样本之后，需要清理数据。在实践中，清理数据的方法也有多种做法。
- 具体方法：使用关键词排除、使用条件排除（如文本长度、提及关键词次数等）、人工清理。



数据抽样：文本大数据研究中，产生了新的抽样需求

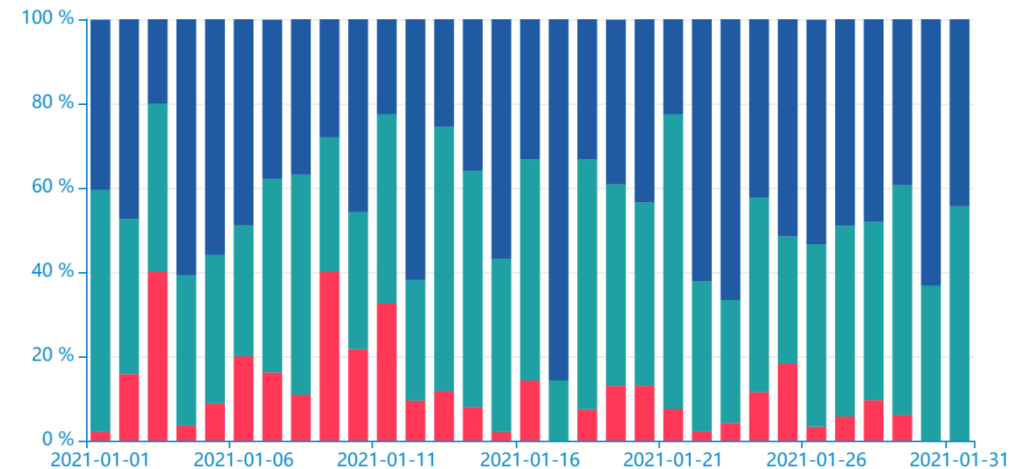
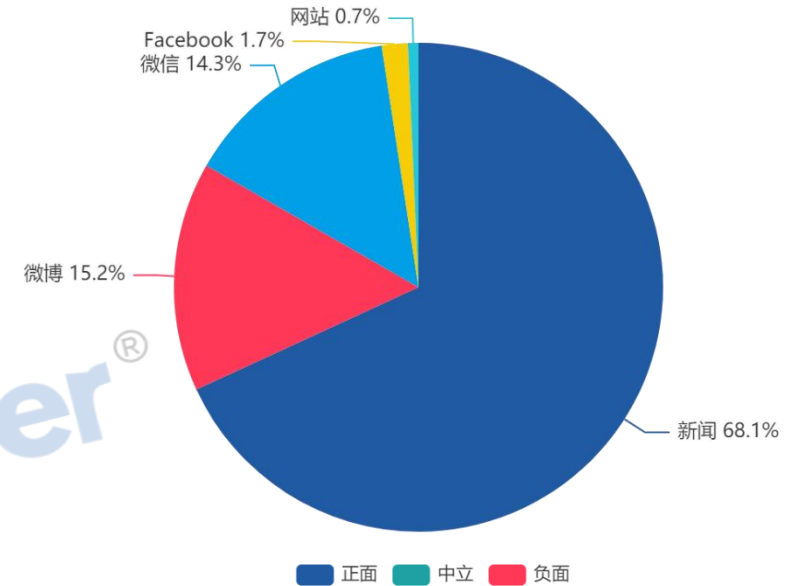
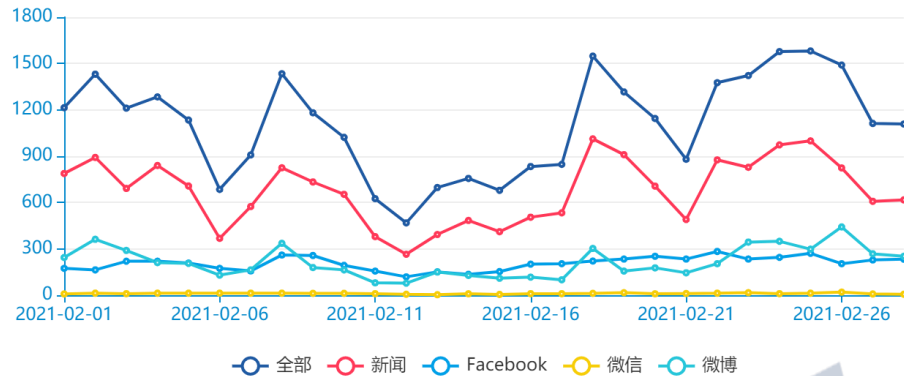
- 以大体量数据为对象的内容分析法，在样本的选取上，通常两种选择方向：
- 第一类，分析**全体样本（母体）**，比如，利用技术，快速分析文本数据的客观性信息（来源分布、时间序列、主题、人物、表达词等）；
- 第二类，文本相对主观性内容，例如态度的分化等，技术所不能达，需要人工介入，而由于人工内容编码需要时间和人力，当需要应对海量文本数据的时候，如果人工编码的压力过大，时效性会大打折扣。这种场景下就可能要考虑**抽样**，选取部分样本，在编码的时效性、样本的覆盖广度和编码后的分析深度上做出一个平衡。

The screenshot displays the DiVoMiner interface for data sampling. On the left, there are two panels:

- 新闻数据库 (News Database):** 编码库 (Encoding Library) 36,674, 回收库 (Recycling Library) 2, 已人工编码 (Manually Encoded) 2. A '自定义' (Customize) button is visible.
- 新闻抽样库 (News Sampling Library):** 编码库 (Encoding Library) 7,335, 回收库 (Recycling Library) 0, 已人工编码 (Manually Encoded) 0. A '抽样库' (Sampling Library) button is visible.

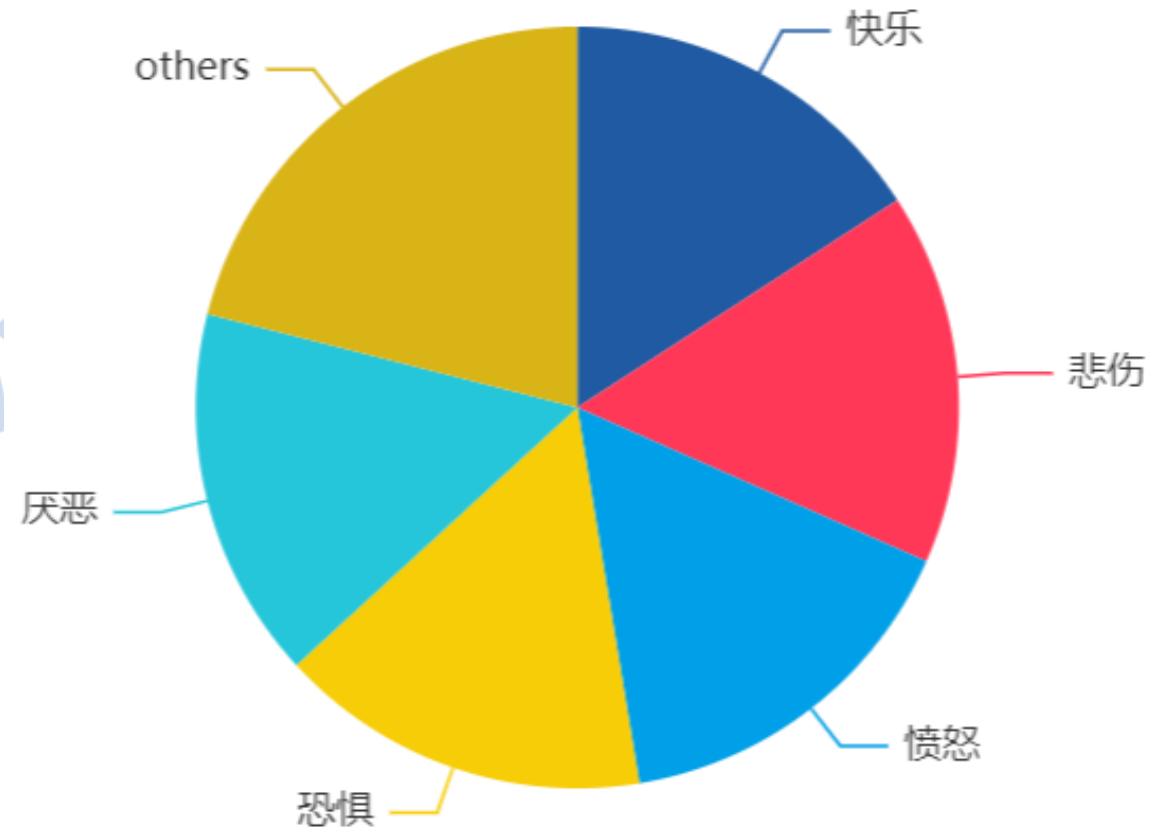
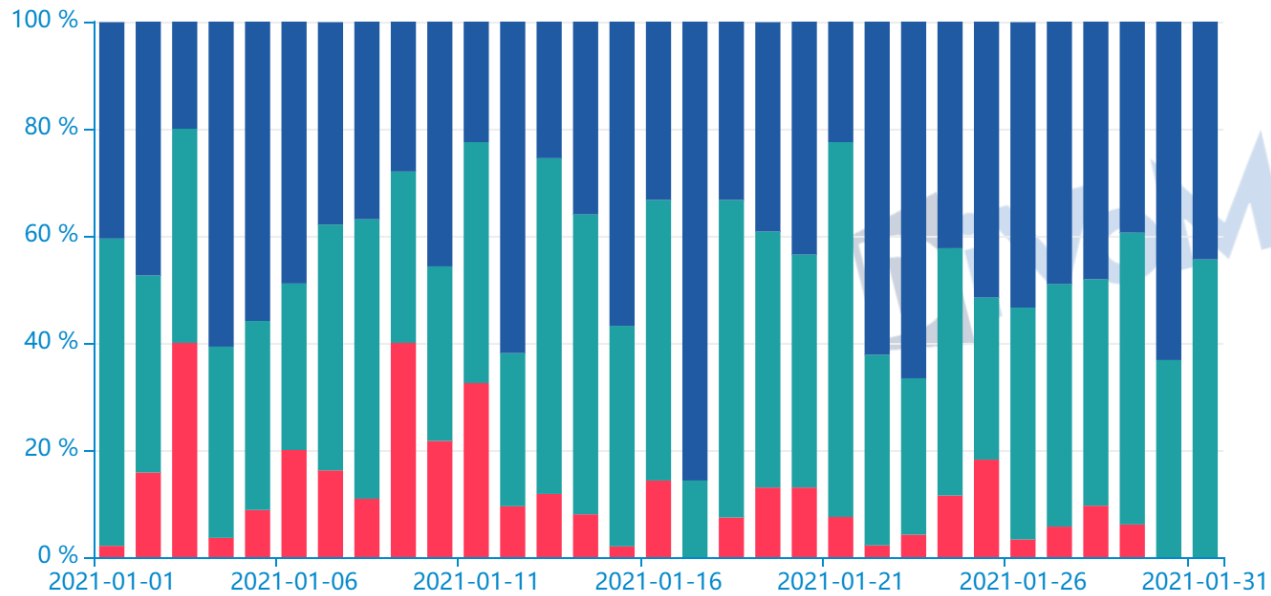
On the right, the '选择抽样库' (Select Sampling Library) step is active, showing options for '随机' (Random) and '设置抽样范围' (Set Sampling Range). A '+ 添加' (Add) button is also present.

运用自动化分析初步探索数据：网络挖掘

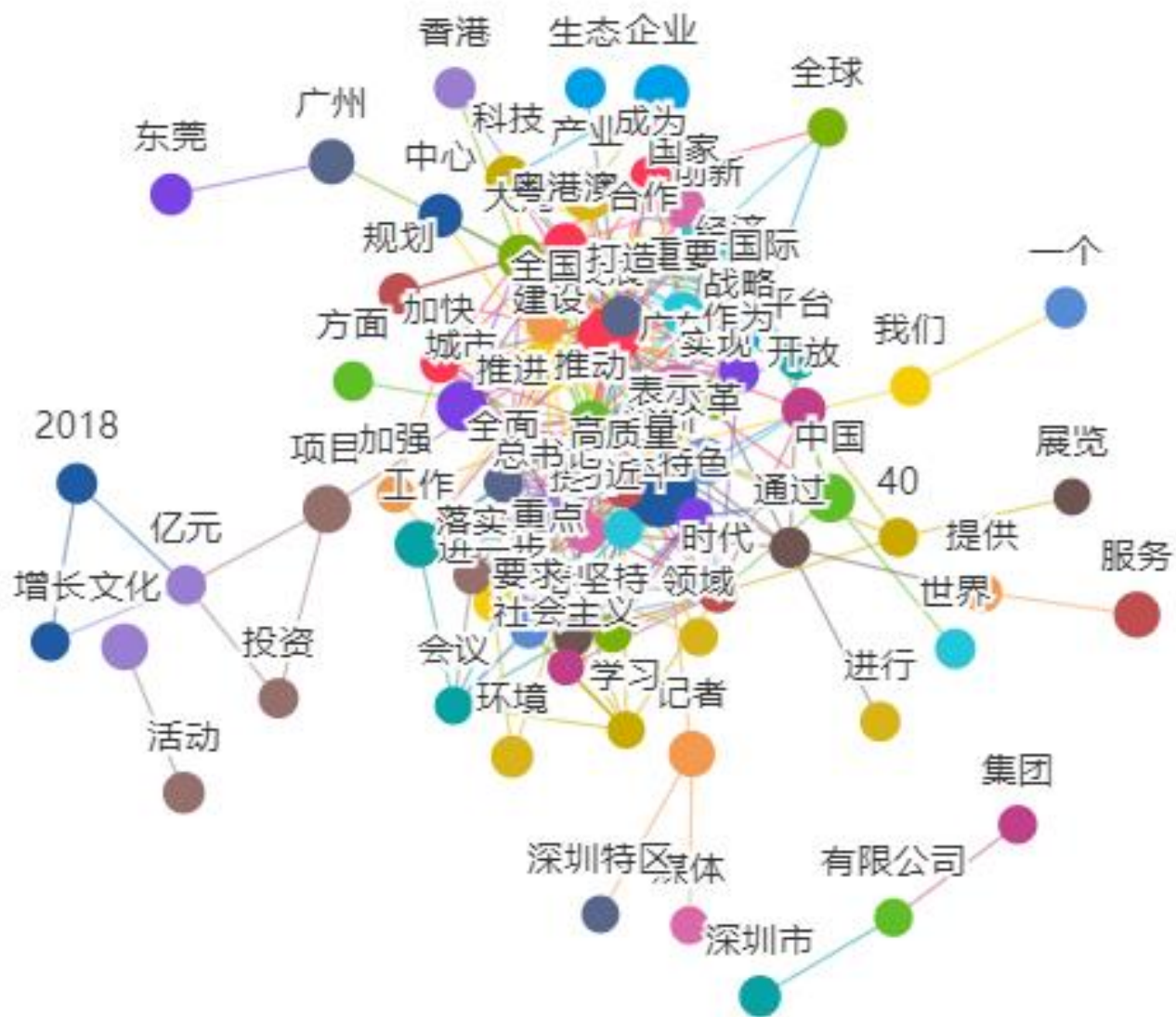


运用自动化分析初步探索数据：情感分析

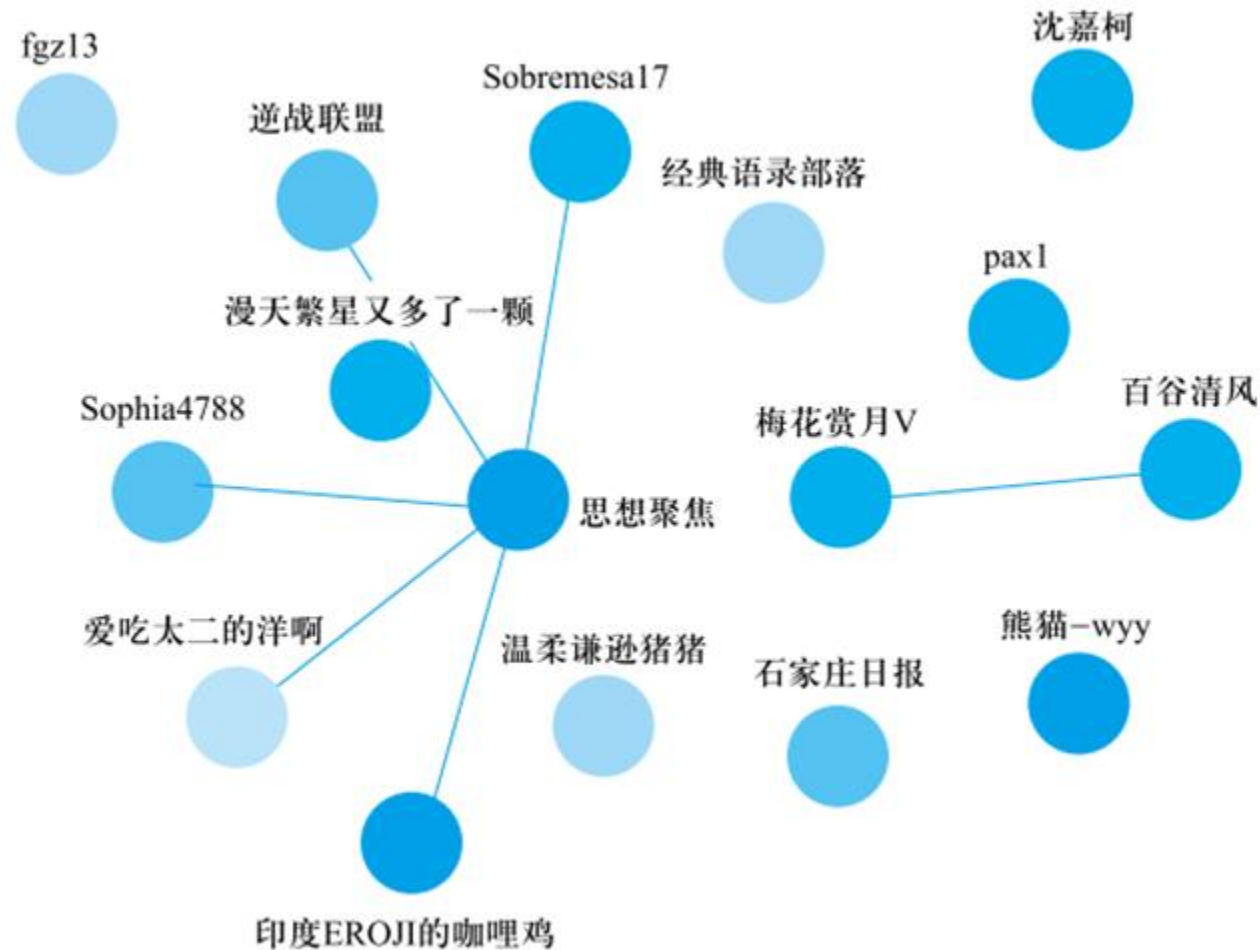
■ 正面 ■ 中立 ■ 负面



运用自动化分析初步探索数据：共现词分析



运用自动化分析初步探索数据：社交网络分析



类目建构 (Classification)原则和方法

■ 类目建构的原则

- 互斥 (类目不要重复)
- 穷尽 (涵盖所有的类别 , 其中 “其他” 类别不宜超过10%)
- 可靠 (不同的编码员的编码结果应大致相同)

■ 类目建构的方法

- 根据理论或是过去的研究结果来建构类目 ;
- 由研究者根据常识、经验和研究目的而自行建构 ;
- 参考其他研究的类目 , 依据当地的习惯或一些特定的标准 , 研究者将之改良后适合于本地的类目

类目的种类：说什么、如何说

■ 说什么 what is said :

1. 主题项目 (subject) : 依据传播内容的主题加以分类, 例如报章社论的主题可分为: 政府官员、当地政府、立法机关、司法机关、解放军、国内问题、国际事务、政党问题、商业经济问题、社会民生问题、地区治安问题、文化活动、宣扬爱国主义、其他等14项。
2. 方法类目 (methods) : 指达到目的的手段, 例如分析小说的故事情节时, 把小说中人物成功的方法, 分成: 运气、个人努力、家庭背景、投机取巧、其他等5项。
3. 特性类目 (traits) : 指在传播内容中的人物具有那些特征, 例如: 年龄、性别、种族、职业、婚姻状况、社会阶级等6项。又如: 把消息来源人物依职业分为: 政府官员、意见领袖、学者专家、人大代表、社会工作者、商人、一般民众等7项。
4. 主角类目 (actor) : 指在传播内容中, 出现的代表性人物或领导人物, 例如分析两会开会的报导, 藉以找出提出提案最多的代表。
5. 权威项目 (authority) : 指传播内容中, 以人或团体的名义来发表声明或谈话, 例如分析某地区各界对疫情的反应的报导, 把该地区各界分成: 政府、商界、医护界、压力团体、社区组织、大V等6项。
6. 来源类目 (origin) : 指传播活动的发生地点, 例如分析新闻时, 依据新闻事件发生的地点, 把新闻分为: 国际新闻、国内新闻、台湾地区新闻、香港地区新闻、澳门地区新闻等5项。

类目的种类：说什么、如何说

■说什么 what is said :

7. 目标类目 (target) : 指传播的对象，传播常以某些特定的人或团体为诉求对象，例如分析某候选人的政见，把候选人政见的诉求对象分为：工人、公务员、白领、青年学生、家庭主妇等5项。
8. 标准类目 (standard) : 以传播内容特性的方法为标准，而加以分类。例如分析电视广告中妇女的形象，依性格类型分为乐观和悲观；独立和依赖；主动和被动；积极和消极等。
9. 方向类目 (direction) : 指传播内容所显示的态度与立场，常见有：赞成和反对；有利和不利；同意和不同意。例如分析报纸社论，把社论对政府的立场分为：有利、中立、不利等3项。

类目的种类：说什么、怎么说

■ 怎么说 how is said :

1. 传播形式或类型 (form or type of communication) : 指传播内容以何种形式或类型来表达 , 例如分析电视节目的形态 , 把电视节目分为 : 新闻、教育、娱乐、公共服务、其他等5项。
2. 叙述形式 (form of statement) : 指传播内容的文法或造句形式 , 例如分析某篇新闻报导时 , 把新闻报导的结构分为 : 报导、推论、意见等3项。
3. 强度类目 (intensity) : 指传播内容中表达态度、感觉、立场的强烈程度 , 例如分析报纸社论的立场 , 把社论对某社会行为的批评或赞扬的程度分为 : 强烈批评、中度批评、稍为批评、中立、稍为赞扬、中度赞扬、强烈赞扬等7项。强度类目与方向类目相似 , 但方向类目只显示方向 , 没有指出态度、感觉或立场的强度 , 而强度类目两者兼备。
4. 策略类目 (device) : 指传播内容中所用的修辞或宣传手法 , 例如分析电视广告的策略 , 把广告所采取的诉求方式分为 : 感性诉求、理性诉求、恐惧诉求等3项。

建立量化系统：测量标尺

- 上述的类目建构过程，实际上是为概念或构念赋予数值，使之成为变项的过程。根据不同的研究要求，研究者决定使用不同的测量标尺。这个跟问卷调查的问卷选项的测量标尺类似。
- 不同编码员对分析单位所属类目，意见应该一致。这种一致性在内容分析中是以量化方式呈现，称之为“**编码员间的信度 (intercoder reliability)**”。
- 一般规则是，类目多比类目少要好，因为分析单位确定之后，将几个小类目变成一个大类目，比将一个大类目分成数个小类目要容易多了。

编码 Coding

- 在内容分析中，把类目建构后，接着是把文本内容归类，这个过程称为编码，负责编码的人员称为**编码员**。
- 编码过程的功能是把一堆杂乱无章的文本内容转化为有组织的资料档案，最后的成果通常是数字而不是文字，亦即是**量化**。
- 编码员必须经过训练，目的是确保所有的编码员熟悉研究的目的、研究问题、研究假设、变项的操作性定义、编码的规则及类目的定义。
- 在训练过程中，研究者可能会不断地修正类目，制作

编码簿 (code book) :
包括拟定编码说明 (coding instruction) 及编码表 (coding sheet) 。

编码簿的设计

- 编码表是记录研究者遵循类目定义，按照分析单位界限，遵守编码规则，从研究素材（传播内容）转化成符合研究目的、能回应研究问题的原始资料，其所描述的现象最能代表研究素材的内容。

- 编码表的内容应包括三种资讯
 1. 行政性资讯：例如内容分析计划、所应用的单位、区分资料的一般类型、指明编码表的状况、资料识别编号、编码员姓名、编码员编号、编码时间等。
 2. 资料组织资讯：可记录某些资料是如何结合而得来的，以便作其他分析目的之用。
 3. 分析资料：有关所有要编码的现象和资料的资讯，是整个编码过程的核心。

编码簿范例

1. Name of anthology and edition number: *American Tradition in Literature* 10th ed.
2. Editor: George Perkins and Barbara Perkins
3. Number of volumes: 2
4. Year of publication: 2002
5. Publisher and place of publication: McGraw-Hill Co.: New York
6. Number of pages on which materials by or about male authors appear: 3,095
7. Number of pages on which materials by or about female authors appear: 619
8. Number of pages on which materials by or about an author of unknown gender appear: 43
9. Number of pages on which materials by or about male Native American authors appear: 0
10. Number of pages on which materials by or about female Native American authors appear: 5
11. Number of pages on which materials by or about Native American authors of unknown gender appear: 27

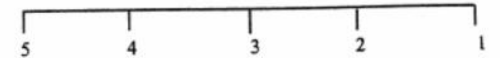
FIGURE 13.1 Partial Sample Coding Sheet from “Reloading the Canon” Study

Exhibit - 1

1. Which of the following newspapers do you read?
 - a) *The Times of India*
 - b) *The Hindustan Times*
 - c) *Indian Express*
 - d) *Pioneer*
 - e) *Any other (Please specify)*

If the respondent does not read Times of India, move to question number 6.
2. I have been subscribing to TOI for (please tick as applicable)
 - a) *Less than 6 months* _____
 - b) *6 months - less than 1 year* _____
 - c) *1 year - less than 2 years* _____
 - d) *Two years or more* _____
3. Please rate the following features on a scale of 1 to 5, as per the interest you have in each one of them. (5 - Very interested, 4 - Interested, 3 - Not particularly, 2 - Uninterested, 1 - Very uninterested)
 - a) *Political News* _____
 - b) *International News* _____
 - c) *City News* _____
 - d) *Corporate and Business News* _____
 - e) *Sports News* _____
 - f) *People and Lifestyles News* _____
 - g) *Leisure, Art and Entertainment News* _____
4. Please rate TOI on the following attributes along the scale marked below
 - a) *News Content*

Very Adequate *Very Inadequate*



编码簿范例

■ Neuendorf著作《The content analysis Guidebook》2017第二版第七章《Content Analysis in the Interactive Media Age》中的案例论文，以ebay上网络游戏的拍卖网站页面（截图）为样本，下图为编码簿（部分）和编码指引。

Annotations:

- #5 and #6 Picture(s)
- #15-17 Subtitle
- #9 Item condition
- General product information provided by eBay
- #4 Seller's certification
- #8 Payment options
- #25 Shipping cost
- #26 Delivery time
- #27 Shipping range
- #18-24 Description provided by the

Annotations:

- #10 - 14 Title of auction listing
- #2 Seller's feedback rating
- #3 Feedback percentage
- #30 Time left
- #28 Starting price, click here.
- #31 Current bid number.
- #29 Current bid
- #7 Return Guarantees

III. Coding Scheme

1. Coder ID. _____
- A. **Credibility**
 2. What is current seller's feedback rating score? (Open ended question) _____
 3. What is current seller's feedback percentage? (Open ended question. Just record the numerator, for example: -94% is recorded as -94, 85% is recorded as 85) _____
 4. Does the current seller have any seller certification? (For example: Top-rated seller certification, or Power seller)
 0. No
 1. Yes
 5. Including the title picture, how many pictures of the auctioning item does the seller provide? (Open ended question) _____
 6. Among these provided pictures, how many of them are the representation of the exact auctioning item.
 0. None of them.
 1. Just 1 picture
 2. 2 pictures
 3. 3 pictures
 4. More than 3 pictures
 7. Does the seller provide return guarantee for the current auctioning item?
 0. No
 1. Yes
 9. Not specified
 8. What payment options does the seller provide? (For this question, "other payment options" is defined as all the payment options except PayPal, such as check, money order, and credit/debit card)
 1. Only PayPal
 2. PayPal and other payment options
 3. Only other payment options
 9. What is the item condition indicated by the seller?
 1. Brand new
 2. Like new
 3. Very good
 4. Acceptable
 5. Not specified
- B. **Attractiveness**
 - Questions 10 – 14, Comparing with the default title of auction listing of the item, the current title is: (The Default Title: The default title of the game on eBay auction list is: "Call of Duty: Modern Warfare 2 (Xbox 360, 2009)")
 10. Exactly same with the default title
 0. No
 1. Yes
 11. Differently capitalizing letters than the default title (Examples: "CALL OF DUTY: MODERN WARFARE 2 (XBOX 360, 2009)", or "SEALED Call of Duty: Modern Warfare 2")
 0. No

DiVoMiner® 编码簿 (Code Book) 与编码说明

编码簿

内容分类[必填]

- UGC
别名:1
- 转载新闻
别名:2
- 两者皆有
别名:3

地区分类[必填]

- 澳门
别名:1
关键字:澳門 OR 澳门 OR mac...
- 内地
别名:2
关键字: 中國 OR 内地 OR 大陆...
- 香港
别名:3
关键字:香港 OR 本港 OR 九龍 ...
- 台湾
别名:4
关键字:台灣 OR 台中 OR 台北 ...
- 国外
别名:5
关键字:英國 OR 美國 OR 朝鮮 ...

二. 内容分类

内容分类

- UGC 转载新闻 两者皆有

UGC 为原创;

转载新闻为转载网站、报纸等其他媒体的新闻报导、评论等;

两者皆有为转载新闻报导、他人言论之后, 另作评论(一般从标题、开头、结尾可判断)。

DiVoMiner®大数据技术辅助内容分析机制

东莞阳光网讯 按照市委、市政府机构改革要求，今天（26日），**东莞**市自然资源局东城分局、**东莞**市应急管理局东城分局等部门分别举行揭牌仪式。

东莞市自然资源局东城分局揭牌成立

原文

人工编码+机器编码

1.大湾区城市

香港

澳门

广州

深圳

东莞 (机器建议)

珠海

编码簿

依据左侧内容，进行编码

1.大湾区城市

1.1香港-定位

1.2澳门-定位

1.3广州-定位

1.4深圳-定位

1.5珠海-定位

1.6佛山-定位

1.7惠州-定位

1.8东莞-定位

1.9中山-定位

1.10江门-定位

1.11肇庆-定位

2.品牌个性

3.城市旅游资源

3.1人文资源 51

DiVoMiner®平台支持复杂的关键词逻辑条件设置

研究员设置类目选项关键词
平台自动匹配选项

关键词:国际金融中心

X

[(香港 OR "HK" OR "Hong Kong" OR "HongKong")] AND (金融中心 OR 商贸中心)

香港特区政府财政司陈茂波司长则就内地与**香港**合作现状及前景分享了自己的看法。他表示，**香港**是国际**金融中心**，期待**香港**能进一步助力国家金融改革。陈茂波赞扬“金融青年汇”北京暑期实习团帮助**香港**提升金融人才素质。他也寄语实习团团员们：“粤港澳大湾区是一个很重要的机遇，大湾区拥有金融、**科技**实力雄厚的城市，城市之间的合作与成果也充满了可能性。青年学子应该去大湾区了解，或许会寻到未来事业发展的机会。”

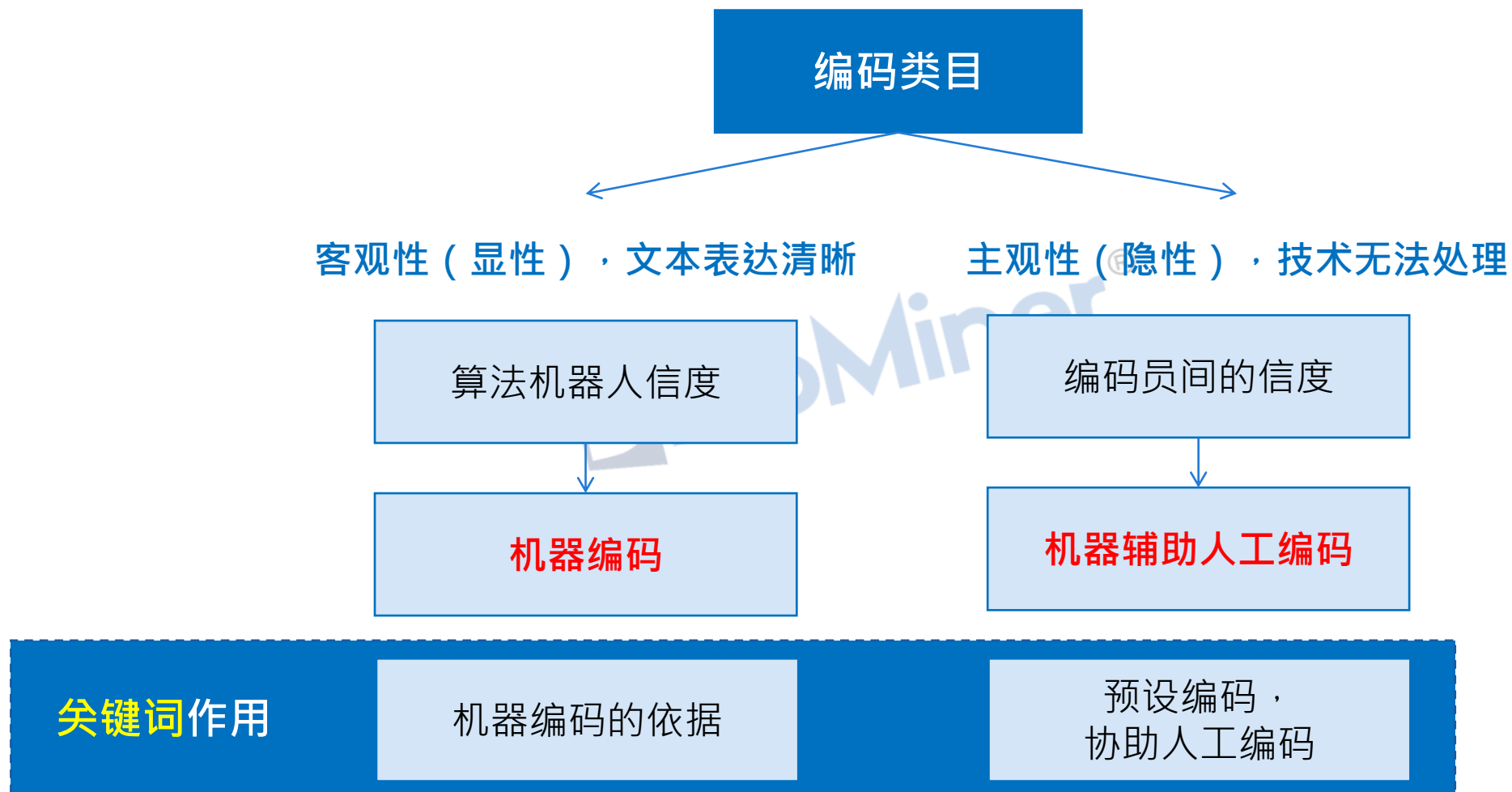
随后，与会的资深金融从业人员与学生们进行了对话。学生们就热

1.1香港-定位

- 国际金融中心 (机器建议)
- 国际大都会
- 区域商贸枢纽
- 生产力强创意无限的城市

2.品牌个性

编码类目与两种内容编码处理方法



编码员间的信度 (Inter-coder Reliability)

- 在内容分析中，需要多于一个的编码员来进行编码工作，这些独立的编码员对一段讯息/记录内容的特征（也就是记录单位）作出判断，并且达致一致的结论。这种一致性以量化方式呈现，称之为“编码员间的信度”。
- 不同的编码员应该对每一个分析的对象给予相同的评分（对等距或者等比标尺而言，即使不是完全相同的数值，也应该是相近的值），这种实质的同意程度是检验“编码员间的信度”的基础（Tinsley & Weiss, 2000, p98）。

为什么我们要重视编码员间之信度？

- 通常我们研究的讯息有明显的内容 (manifest content) 和隐藏的内容 (latent content)。对于明显的内容，例如版面面积或者消息来源，很容易以客观的判断来达致高度一致性。但是，对于隐藏的内容来说，例如报导态度或者价值观，编码员必须根据他们自己的思维系统作出主观的诠释。这样的话，编码员之间的相互主观判断变得更加重要，因为当这些主观判断由所有编码员共享的时候，也就是它们更有可能让读者产生相同的意义 (Potter and Levine-Donnerstein, 1997, p.266)。
- Neuendorf (2002, p141)：既然内容分析的其中一个目标是相对客观地界定及记录信息的特征，那么信度攸关重要。没有建立信度，内容分析的测量只是空谈。
- Rust and Cooil (1994, p11)指出，从现实的角度来看，编码员间的信度对于市场研究者来说至为重要，因为高信度意味者决策者作出错误决策的机会相对减少。
- 总而言之，编码员间的信度乃衡量一个内容分析研究效度的必要条件（虽然不是充分条件），没有信度，那么，该研究的结论便值得怀疑，甚至显得毫无意义。

如何评估和报告编码员间之信度？

1. 训练中的非正式的信度评估：在编码员训练期间，依照编码指引进行少量样本的信度评估、编码调整，直至这种非正式的评估达致适当的同意度为止。注意此阶段的编码样本可以是方便样本。
2. 测试中的正式的信度评估：以随机方式抽出具代表性的少量样本，一般为30篇，此时各编码员应该独立进行编码，不能互相讨论或指导。如果信度未达效果，则需再训练或改善编码指引和程序。在逼不得已的情况下，唯有更换编码员。
3. 总样本的正式信度评估：根据测试的信度结果，判断是否可以进行正式的编码工作。之后，在总样本中随机抽出一有代表性的样本，一般样本数不少于50篇或约占总样本的10%，进行编码并计算其信度，此信度为最终在报告中出现的正式信度。
4. 结合总样本：各编码员独立进行编码工作之后，以适当的步骤选择上述（6）中的其中一个编码员的编码结果结合至总样本中。因为编码员之间始终有若干差异存在，选择的准则由研究员决定，可以根据多数编码员的共识或其他专家的意见。
5. 报告编码员间之信度

信度报告应该包括以下资讯:

1. 信度分析的样本数及理由
2. 信度样本与总样本的关系：是总样本的一部分还是额外样本
3. 编码员资料：人数（须为2或更多）、背景，研究员是否也是编码员
4. 每名编码员的编码数量
5. 信度指标的选择和理由
6. 每个变项的编码员间之信度
7. 编码员的训练时间
8. 在总样本的编码过程中遇到不同意见时的处理方式
9. 读者可以在哪里得到详细的编码指引、程序和编码表
10. 要报告每一个变项的信度水平，不要只报告所有变项的整体信度。

编码员间之信度计算

目前，大概有39种不同的同意度指标（[Popping, 1988](#)），常用的有以下几种：

2名编码员：	测量水准
Percent agreement	名目 (nominal)
Holsti's Coefficient Reliability	名目
Cohen's kappa(κ)	名目
Scott's pi(π)	名目
多名编码员：	
Cohen's kappa (κ)	名目
Krippendorff's alpha(α)	适用于各测量水准

Holsti's Reliability

R: 相互同意度

$$R = \frac{2M}{N1 + N2}$$

M: 两位编码员编码结果相同的次数

N1: 第一位编码员编码的次数

N2: 第二位编码员编码的次数

选择适当的最低接受信度程度：.90或以上最佳，.80可接受，.70在试探性研究中可接受。对于宽松的指标（如percent agreement），对于保守的指标（如Cohen's kappa, Scott's pi, and Krippendorff's alpha）则接受程度相对可降低。

如何计算编码员间的信度

- 手动 (By hand) : 笔、计算机、公式 (参考<http://matthewlombard.com/reliability>)
- 专门软件 (Specialized software)
 1. AGREE. [Popping's \(1984\)](https://www.agreetrust.org/)设计, 计算 Cohen's kappa , <https://www.agreetrust.org/>
 2. Krippendorff's alpha 3.12a. 计算 Krippendorff's alpha , 新版本还在设计中 , 可到Professor Krippendorff 网站了解。【找遍互联网, 找不到, 很可惜】
(<http://www.asc.upenn.edu/usr/krippendorff/>) 。
 3. PRAM. 信度评估软件, 可以处理多个编码员【免费, 实用, 推荐使用】PRAM可以计算的指标有: Percent agreement, Holsti' s reliability, Scott' s pi, Cohen' s kappa, Krippendorff' s alpha

使用DiVoMiner信度测试模组, 只需几个按键就搞定

如何计算编码员间的信度

- 统计软件包含信度分析 (Statistical software packages that include reliability calculations)
 1. Simstat. 可以从Provalis Research下载试用版 (<http://www.simstat.com/>) 可以计算 :Percent agreement, Scott's pi, Cohen's kappa, Free marginal (nominal), Krippendorf's r-bar, Krippendorf's R (for ordinal data), Free marginal (ordinal) 。
 2. SPSS. 可以计算Cohen 's kappa 【不建议使用，只能对2X2的表格才有效】 。
 3. QDA Miner v1.1.可以计算percentage of agreement, correction for chances (Free Marginal, Scott's Pi, and Krippendorff's alpha), list of disagreements等 。参考：
<http://www.kovcomp.co.uk/QDAMiner/qdambroc.html> 【需购买，未有机会使用，不知效果如何】

- 在SPSS或SAS中使用 (适合对电脑程序熟悉者)
 - 可以参考以下网址：<http://www.temple.edu/mmc/reliability/#Macros>

使用DiVoMiner信度测试模组，只需几个按键就搞定

大数据背景下DiVoMiner®信度测试的操作流程

- 根据编码类目的属性不同，决定了后续执行内容编码的方式有两大类，机器编码及机器辅助人工编码两种方法在信度测试方面也有不同的要求。
- **人工编码**要求进行编码员之间的信度，在信度达到接受的一致性后，再执行正式编码；
- **机器编码**则要求检验人工和机器之间的准确性。

在DiVoMiner上，编码员间的信度测试的操作流程



机器编码同样需要检验信度水平的例子

- 对于文本大数据研究来说，借助大数据技术辅助机器编码，在人工少量介入的情况下，可提升编码效率同时确保编码质量。但是机器编码模式下，由研究人员设定机器编码条件，指引机器对文本进行关键词的自动化标注和自动化填答选项，准确性依然依赖于人工介入的部分。
- 在操作实践中，有研究者采用人机比对的方法，检验机器编码是否可以达到可接受水平。
- 程潇潇的研究团队随机抽取1%的样本作为检验样本（107篇），执行人工编码，对照大数据算法自动编码结果，计算信度，所有类目准确度均在0.87以上，显示大数据自动编码结果良好，可采用该结果作为解读依据。这一步骤是为确保机器编码的准确程度。
- 张文瑜教授（2020）在一项健康传播的研究中，同样随机抽取了1%的数据作为比对样本，由4位经过训练的研究助理完成人工编码的部分，且编码员间的信度达到Cohen' s Kappa = 0.78 ($p < 0.001$), 95% CI (0.604, 0.948)。最终人机对比一致性达到80%，证明算法编码结果可接受（Chang, A. Schulz, P. J. & Cheong, A. 2020）。

DiVoMiner[®]信度测试功能界面

编码员: blair × Wenny Cao × ∨ 信度指标: 霍尔斯蒂系数 ∧ 计算

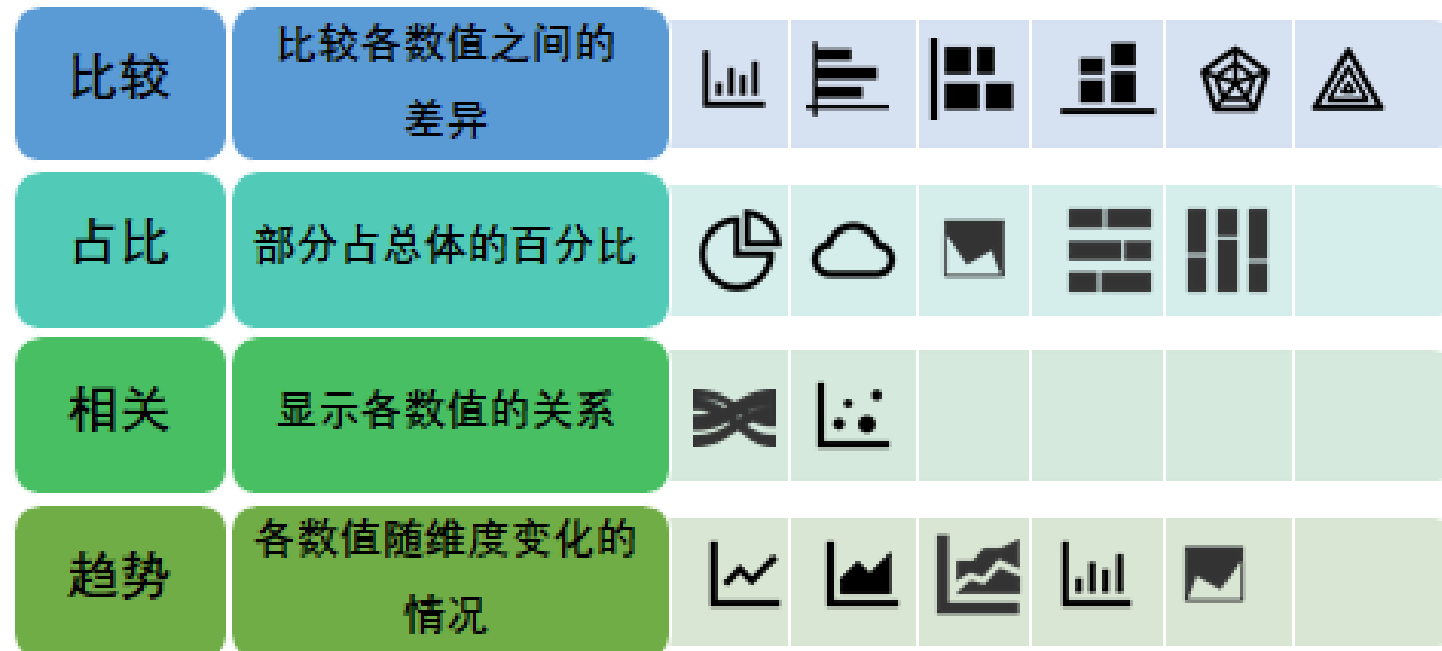
获取机器编码信度测试结果 下载数据

#		
	复合信度	
+	1.研究方法	0.74
+	1.1.内容分析法是否汇报编码员之间信度结果	0.95
+	2.分析数据类型 (对象)	0.61
+	3.数据处理工具/软件	0.74
+	4.数据分析方法/统计方法	0.90

霍尔斯蒂系数
科恩系数 (Cohen's kappa ...
斯科特系数 (Scott's Pi)
克里本多夫系数

统计分析结果的可视化及解释

- 选择何种统计方法和研究主题有直接关系。
- 最常用的数据描述方法，单变量频次的呈现和多变量交叉分析、时间序列表达，在各类研究主题中均有运用。在可视化表达方面，图表的应用同样丰富，图表应用目的以比较、占比、相关和趋势四大类为主，也可以借助LDA模型进行主题分析、社交网络或语义网络分析等。



使用DiVoMiner统计及可视化模组，只需几个按键就搞定

流程	传统定量内容分析法	DiVoMiner®
研究主题	研究者线下处理	
文献探讨		
研究假设		
确定样本(母体/抽样)	手动抽取样本，利用excel随机或以其他方式抽取样本	一键抽样，可使用随机或多种抽样方式
数据搜集	翻旧新闻媒体，复印逐个样本形成样本库；不同类型数据难以汇整，无法实时更新，操作易造成混乱	快速提取不同类型数据，结构化后形成方便管理的数据库
建构类目和制作编码簿	将编码簿写在卡片上，每个样本对应一张卡片；一旦定稿，编码过程中，难以修改编码类目	编码簿电子化，随时更新和调整
前测与信度计算	抽取少量样本，各编码员在编码卡片上做前测，需手动录入和统计前测编码结果，计算编码员间的信度，如信度不理想，需要重复多次前测。	一键随机抽取前测样本，各编码员在平台上做前测编码，督导即时查看前测结果和信度计算结果。重复过程简单。
正式编码	在卡片上或excel里做编码，需手动对照样本和做类目编码	样本及编码类目呈现在同一个页面，有关键字高亮提醒和预设编码功能，同时利用机器算法自动将类似内容分派给同一位编码员，大为提高编码效率。
数据输入	手动输入	无需人工输入，机器自动统计
检查编码数据	难以检查编码数据，需逐一检视，耗时耗力，如修改编码，需手动处理后续数据输入问题	可随时查看和修正编码结果，轻松修改
数据分析	手动统计	机器自动统计，提供交互式可视化统计分析



登录网址 me.divominer.cn

或扫描二维码体验！



预告：
关于文本数据的量化内容分析法，将于今年内出版的《传播研究方法》一书中有专门篇章详细阐述，敬请垂注！

DiVoMiner[®]
Data in Value out

让研究更容易

文本大数据挖掘与分析平台