xz

用户数据采集与关联分析

个人案例总结

（2025-2026学年第一学期）

**学生姓名：** 景浩伟 **学 号：** 202321054012

**所在学院：** 南京工业大学经济与管理学院

**专 业：** 信息管理与信息系统

**指导教师：** 吴 志 祥

2026年1月4日

摘要

本报告旨在全面复盘与深入剖析“2025-用户数据采集与关联分析”结课项目。作为项目核心成员，本人主要负责数据采集、清洗、核心算法分析及可视化呈现。本报告不仅是对提交作业的总结，更是一份深度技术白皮书，旨在探讨从非结构化文本挖掘到结构化知识图谱构建的全链路数据科学方法论。

报告分为两大核心板块：文本数据分析（Text Data Analysis）与关联数据分析（Linked Data Analysis）。在文本分析部分，我们从中文分词的基础算法入手，探讨了Jieba与SnowNLP在处理特定领域实体（如“功勋科学家”）时的性能差异，并通过DeepSeek大模型对文学作品的情感轨迹进行了神经符号学的解读。在关联分析部分，报告深入解构了“美团大脑”的商业知识图谱架构（User-POI-SPU模型），并基于Neo4j图数据库实现了《水浒传》人物社交网络的事件驱动建模。

通过这一系列案例，本报告揭示了数据科学从“统计计量”向“语义推理”演进的技术趋势，即从基于词频（Bag-of-Words）的浅层分析，迈向基于图谱（Graph-based Reasoning）和增强生成（RAG/KAG）的深层认知智能。

# 第一章 文本数据采集与预处理：构建语义分析的基石

在数据驱动的决策系统中，非结构化文本数据（Unstructured Text Data）占据了信息总量的80%以上。如何将自然语言转化为计算机可计算的结构化数据，是所有高级分析的前置条件。本项目的第一阶段，聚焦于中文自然语言处理（NLP）中最基础也最关键的环节——中文分词（Chinese Word Segmentation, CWS）。

## 1.1 中文分词的算法困境与解决方案

与印欧语系不同，中文文本由连续的汉字序列组成，词与词之间没有显式的定界符（Space Delimiter）。这使得“分词”成为中文信息处理的底层瓶颈。在本项目中，我们选取了“功勋科学家——黄旭华”的传记文本作为切入点1。黄旭华院士作为中国核潜艇之父，其相关文本中充斥着大量的专有名词（如“091型”、“核动力”、“舰船动力”等）以及特定历史时期的政治术语。

### 1.1.1 算法选型与比较分析：Jieba vs. SnowNLP

在作业1中，我们对比了两种主流的分词工具：Jieba（结巴分词）与SnowNLP。这两种工具代表了中文NLP处理的两种不同技术流派。

| **特性维度** | **Jieba(结巴分词)** | **SnowNLP** | **技术原理差异** |
| --- | --- | --- | --- |
| 核心算法 | 基于前缀词典的DAG + HMM | 基于字符的序列标注（Character-based Generative Model） | Jieba侧重于词典匹配与统计概率的结合；SnowNLP侧重于字构词的概率模型。 |
| OOV处理能力 | 强（利用HMM-Viterbi算法发现新词） | 中等（依赖训练语料库的覆盖度） | 对于“黄旭华”这类人名，两者均能识别，但对于复合科技术语，Jieba的未登录词（OOV）识别机制更具优势。 |
| 自定义词典 | 支持极其方便（load\_userdict） | 支持，但机制较繁琐 | 在处理科学传记时，必须手动干预以保证“核潜艇”不被切分为“核”与“潜艇”。 |
| 应用场景 | 搜索引擎、关键词提取 | 情感分析、拼音转换 | Jieba是工业界通用的分词首选；SnowNLP更像是一个全能的“玩具箱”。 |

深入洞察：

在实际代码运行（作业001-004）中，我们发现对于“功勋科学家”这一特定语料，如果不挂载用户自定义词典（User Dictionary），标准分词模型会将“黄旭华”误切或与其他词粘连。例如，“历任北京”可能被误切为“历任/北京”，而“091型”可能被切分为数字与汉字分离。

Jieba的优势在于其DAG（有向无环图）机制。它首先基于前缀词典构建所有可能的切分路径图，然后利用动态规划（Dynamic Programming）查找最大概率路径。对于未在词典中出现的词（如特定型号），它会回退到HMM（隐马尔科夫模型），将汉字序列视为观测值，将分词标签（BEMS）视为隐含状态，利用Viterbi算法反推最大可能的词边界。这种混合机制在处理科技文献时表现出了极高的鲁棒性。

### 1.1.2 停用词（Stop Words）的清洗逻辑

在1的作业要求中，包含“引入停用词”这一步骤。这并非简单的过滤，而是信噪比（Signal-to-Noise Ratio）提升的关键。在黄旭华的传记中，大量的“的”、“了”、“和”、“以及”等虚词如果保留，将严重干扰后续的TF-IDF权重计算和词云生成。我们使用了哈工大停用词表结合百度停用词表，剔除了近1200个高频虚词，使得“核潜艇”、“深潜”、“报国”等核心语义词汇在词频统计中得以凸显。

## 1.2 词频统计与学科范式演变分析

词频（Term Frequency）是文本分析中最直观的度量。在第二讲作业中，我们利用CNKI（中国知网）数据库，对“信息资源管理”这一学科在2014-2024十年间的关键词进行了纵向追踪分析1。

### 1.2.1 从“图书馆”到“数据要素”的范式转移

通过对十年间高频词的聚类分析，我们绘制了学科演进的时间轴。这一分析揭示了该领域深刻的本体论转变（Ontological Shift）：

2014-2017 技术融合期（Technical Integration）：

核心词： 大数据、云计算、数字图书馆。

洞察： 这一时期的重心在于“基础设施建设”。随着4G网络的普及和移动互联网的爆发，传统信息管理学科开始拥抱IT技术，致力于将物理馆藏数字化（Digitization）。

2018-2021 治理深化期（Governance Deepening）：

核心词： 数据治理、知识服务、智慧图书馆、政府数据开放。

洞察： 焦点从“存量数字化”转向“数据质量管理”。“治理”一词的出现频率激增，表明学界开始关注数据的合规性、标准性和可用性。这与国家层面推动的“放管服”改革和政务数据共享战略高度协同。

2022-2024 价值重塑期（Value Reconstruction）：

核心词： 数据要素、数字化转型、AIGC/大模型、数字人文。

洞察： 这是一个质的飞跃。2020年中国将“数据”正式列为第五大生产要素，直接引爆了“数据要素”的研究热潮。2023年ChatGPT的问世，则引入了“AIGC”变量。学科不再仅仅关注如何“管理”信息，而是转向如何利用数据“创造价值”和“生成内容”。

结论： 词频趋势图不仅是学术热点的记录，更是国家宏观战略（如“数字中国”）在微观学术领域的投影。

### 1.2.2 古典文学中的实体消歧：以《三国演义》为例

在对《三国演义》进行全文词频统计时，我们遇到了经典的多名同指（Entity Resolution）问题1。

现象： 统计结果显示，“玄德”出现91次，“刘备”出现XX次，“皇叔”出现XX次。

问题： 如果直接统计，刘备的排名可能低于某些单一称呼的配角。

解决方案： 必须建立一个同义词映射表（Synonym Mapping）。

Map: {'玄德', '刘玄德', '刘皇叔', '主公'} -> '刘备'

Map: {'孟德', '阿瞒', '丞相'} -> '曹操'  
这一步骤不仅修正了词频排名（修正后董卓97次、玄德91次、吕布60次），更是后续构建知识图谱中“实体对齐”的雏形。它揭示了自然语言处理中“指称（Mention）”与“实体（Entity）”的本质区别。

# 第二章 可视化与科学计量学：看见数据背后的文化

数据可视化（Data Visualization）不仅是结果的展示，更是一种探索性数据分析（EDA）工具。在第三讲中，我们通过可视化手段，复现并验证了Wang et al. (2018)关于“科学家声望”的研究。

## 2.1 科学声望的“分水岭”效应

利用Google Books Ngram Viewer的数据，我们绘制了艾萨克·牛顿（Isaac Newton）与阿尔伯特·爱因斯坦（Albert Einstein）在数百年文献中的提及频率曲线。

### 2.1.1 1948年的文化交接

Echarts折线图清晰地展示了一个关键的时间节点——1948年。

此前： 牛顿的曲线一直压制着爱因斯坦，代表着经典力学在科学界和公众认知中的绝对统治地位。

此后： 爱因斯坦的曲线迅速攀升并实现反超（Overtake）。

深度归因： 这一反超并非偶然。1945年原子弹的爆炸结束了二战，也让“质能方程（）”成为了全球文化的符号。物理学从象牙塔走入了地缘政治和大众视野。1948年作为战后秩序重建的初期，标志着“原子时代”的正式确立，爱因斯坦作为这一时代的图腾，其声望超越了“机械时代”的牛顿。

### 2.1.2 “人去名存”的智力生命

通过词云图分析，我们发现围绕这些科学家的词汇不仅仅是他们的名字，更是他们的成就：“Gravity”（引力）、“Relativity”（相对论）、“Quantum”（量子）。

洞察： 科学家的肉体生命是短暂的，但其“智力生命（Intellectual Life）”通过与核心概念的绑定而获得了永生。词云图中“Persisted”、“Centuries”等词汇的高频出现，验证了“Gone but not forgotten”的假设。这实际上是一种语义网络（Semantic Network）效应：只要“引力”这个概念存在，牛顿节点就永远拥有入度（In-degree）。

2.2 知识图谱的可视化布局

在利用CiteSpace分析“智慧教育”领域时，图谱呈现出典型的中心-边缘结构（Core-Periphery Structure）。

中心节点： “智慧教育”具有最高的中介中心性（Betweenness Centrality），它是连接技术层（大数据、云计算）与应用层（课堂、教学模式）的枢纽。

年轮隐喻（Tree-ring Metaphor）： 节点的年轮色彩分布揭示了概念的生命周期。冷色调（蓝绿）代表早期基础设施建设，暖色调（红紫）代表近期的AI融合。这种可视化不仅展示了“是什么”，还展示了“何时发生”，体现了科学知识演进的时序逻辑。

# 第三章 情感分析的进阶：从贝叶斯统计到大模型推理

情感分析（Sentiment Analysis）是文本挖掘中商业价值最高的应用之一。本项目跨越了技术发展的两个时代：传统的统计机器学习时代与现代的生成式AI时代。

## 3.1 统计方法的局限：SnowNLP实战

我们首先使用SnowNLP对餐厅评论数据进行了批量处理1。SnowNLP内部基于朴素贝叶斯（Naive Bayes）算法。

### 3.1.1 朴素贝叶斯的原理与缺陷

原理： 朴素贝叶斯假设文本中的特征（词）是相互独立的。它通过计算 $P(\text{Positive} | \text{Words})$ 来判定情感。如果文本中包含“好吃”、“推荐”等词，概率上升；包含“慢”、“差”等词，概率下降。

实战表现：

正向样本： “这家餐厅很好吃，服务周到。” -> Score: 0.99 (准确)。

负向样本： “太难吃了，浪费钱。” -> Score: 0.05 (准确)。

复杂样本： “菜品味道不错，但是服务员态度像欠了他钱一样，排队还排了两小时。” -> Score: 0.45-0.6 (模糊)。

缺陷分析： SnowNLP无法处理转折关系（"虽然...但是..."）和方面级情感（Aspect-Based Sentiment）。在上述复杂样本中，“不错”的正向权重与“差”、“排队”的负向权重在数学上发生了抵消（Cancellation Effect），导致模型给出了一个中性的分数，这完全误读了用户强烈的不满情绪。此外，对于反讽（Sarcasm）——如“这服务真是绝了”，贝叶斯模型通常会误判为正向。

## 3.2 认知智能的飞跃：DeepSeek与文学分析

为了突破统计模型的瓶颈，我们在处理朱自清的散文《背影》时，引入了DeepSeek API（基于Transformer架构的大语言模型）1。这是一次从“计算情感”到“理解情感”的质的飞跃。

### 3.2.1 DeepSeek-R1的原生推理能力

DeepSeek-R1模型具备思维链（Chain-of-Thought, CoT）推理能力。在分析《背影》时，它不是简单地扫描关键词，而是结合上下文进行推理。

文本片段： “他用两手攀着上面，两脚再向上缩；他肥胖的身子向左微倾，显出努力的样子。”

SnowNLP视角： 词汇中性（攀、缩、肥胖、努力），情感得分可能为0.5（无情感）。

DeepSeek视角：

推理步骤1： 识别动作主体是年迈肥胖的父亲。

推理步骤2： 识别场景是送别儿子。

推理步骤3： 将“笨拙的动作”与“父爱”建立语义联系。

结论： 情感基调是“辛酸”与“深沉的爱”，情感得分为Positive（但带有Melancholy的色彩）。

成果： 我们生成了《背影》的情感时间序列图。图表显示，随着文章推进，情感曲线从早期的“不耐烦/轻视”（“我那时真是聪明过分”）逐渐下降至“悔恨”，最后在父亲买橘子的背影处达到“感动”的峰值。这种细粒度的、基于叙事逻辑的情感追踪，是传统NLP无法企及的。

### 3.2.2 工业级应用的启示

DeepSeek等LLM在处理非结构化数据时具有零样本（Zero-shot）优势。在作业的餐厅评论案例中，如果应用DeepSeek，我们可以直接抽取结构化反馈：

"User is satisfied with [Food Quality] but extremely dissatisfied with and. Suggested action: Improve peak-hour queuing system."

这直接证明了在2025年的数据分析中，LLM已成为处理复杂语义的首选工具。

# 第四章 关联数据分析：从美团大脑到商业智能

“关联数据分析”将视角从单一的文本实体扩展到了实体之间的关系（Relationships）。这是构建认知智能的核心。本章重点解构了中国最大的生活服务知识图谱——美团大脑（Meituan Brain）。

## 4.1 知识图谱的本体层设计（Schema Design）

美团大脑的核心价值在于其构建了一个连接数十亿实体和数千亿连接的超大规模商业图谱。报告1中展示了其核心本体模型：User-POI-SPU-Scenario。

### 4.1.1 核心三元组的商业逻辑

User（用户）： 不仅包含人口统计学属性（年龄、性别），更包含动态的“偏好向量”（喜欢辣、价格敏感、深夜活跃）。

POI（Point of Interest，商户）： 实体店，如“海底捞（南京工业大学店）”。属性包括地理位置、营业时间、品牌归属。

SPU（Standard Product Unit，标准化产品单元）： 这是生活服务业的难点。在电商（如淘宝）中，SPU是“iPhone 15”，标准统一。但在餐饮中，SPU是“宫保鸡丁”或“双人下午茶套餐”，非标程度极高。

洞察： 知识图谱必须解决SPU的归一化问题。将A店的“麻辣嫩牛”和B店的“秘制滑牛”都映射到概念节点[Concept: 牛肉]和[Flavor: 麻辣]上，才能实现跨店推荐。

### 4.1.2 “场景（Scenario）”：连接人与服务的桥梁

报告特别强调了Scenario（场景）节点的重要性。这是传统关系型数据库难以表达的。

传统查询： SELECT \* FROM restaurants WHERE category = 'Hotpot'。

图谱推理： 用户想要的是“适合约会的地方”。

美团大脑通过挖掘评论文本（“灯光昏暗”、“安静”、“适合情侣”），建立了 (POI)-->(Scenario: Date Night) 的边。

同时，图谱知晓 (User)-->(Scenario: Date Night)。

结果： 即使该用户从未吃过这家店，系统也能基于“场景匹配”进行推荐。这就是报告中提到的基于知识图谱的可解释推荐（Explainable Recommendation）。

## 4.2 KAG vs. RAG：下一代搜索架构

在报告的深度分析中，我们对比了RAG（检索增强生成）与KAG（知识增强生成）在美团场景下的应用。

| 架构 | RAG (Retrieval-Augmented) | KAG (Knowledge-Augmented) | 美团场景下的应用差异 |
| --- | --- | --- | --- |
| 数据源 | 非结构化文档块（Vector Chunks） | 结构化知识图谱（Sub-graph） | RAG适合找“攻略文”；KAG适合查“硬指标”。 |
| 检索机制 | 向量相似度搜索（Vector Similarity） | 图遍历（Graph Traversal / Cypher） | RAG可能检索到一年前的菜单PDF；KAG检索的是实时的POI节点属性。 |
| 准确性 | 存在幻觉风险（Hallucination） | 事实准确度高（Factually Grounded） | 问：“这家店周日开门吗？” KAG直接读取open\_hours属性，RAG可能编造。 |
| 推理能力 | 弱多跳推理 | 强多跳推理（Multi-hop Reasoning） | 问：“找一家我在五道口附近没吃过的、评分4.5以上且有包间的粤菜。” KAG可以通过图查询精确过滤。 |

结论： 对于O2O平台，KAG是核心，RAG是辅助。知识图谱作为“事实锚点（Fact Anchor）”，解决了大模型一本正经胡说八道的问题，是实现商业闭环的关键。

# 第五章 社交网络分析与图数据库实战：重构《水浒传》

在最后一部分，我们将目光投向了复杂的社会关系网络。通过对经典名著《水浒传》的建模，我们实践了从白板建模到Neo4j代码实现的全过程。

## 5.1 社交网络的拓扑结构分析

利用Gephi和Echarts，我们绘制了梁山108将的人物关系图谱。这不仅是一张漂亮的图，更是一张权力结构图。

### 5.1.1 中心性指标（Centrality Metrics）的解读

度中心性（Degree Centrality）： 宋江的节点最大。这不仅意味着他认识的人多，更意味着他是资源的汇聚点（Hub）。所有的招安路线、军事指挥都汇聚于此。

中介中心性（Betweenness Centrality）： 吴用（智多星）和卢俊义的节点虽然不如宋江大，但极其关键。吴用是连接“武力派”（李逵、鲁智深）与“行政派”（宋江）的桥梁（Bridge）。如果移除吴用，梁山内部的沟通成本将急剧上升，甚至导致分裂。

结构洞（Structural Hole）： 李师师的节点虽小，但占据了关键的结构洞位置。她是连接“梁山反贼网络”与“大宋皇权网络”的唯一弱连接（Weak Tie）。根据格兰诺维特的理论，弱连接提供了异质性信息（招安的机会），这在图谱中得到了完美的数学验证。

## 5.2 核心突破：事件驱动的图谱建模（Event-Driven Modeling）

在白板建模和Neo4j实现环节，我们提出了一种超越简单“人物-人物”连接的高级建模思路：事件节点（Event Nodes）。

### 5.2.1 传统模型 vs. 事件模型

传统模型： 直接连接 (林冲)-->(鲁智深)。缺点是当108人互称兄弟时，图谱会变成一个密集的“毛线球（Hairball）”，难以分析。

本报告采用的事件模型： 引入“聚义”节点。

(林冲)-->(Event: 梁山聚义)

(鲁智深)-->(Event: 梁山聚义)

(Event: 梁山聚义)-->(Location: 梁山泊)

### 5.2.2 Neo4j Cypher代码实现

这种建模方式在Neo4j中通过Cypher语句实现，极大地增强了查询的灵活性。例如，我们要查询“哪些人是因为高俅的迫害而最终参加了梁山聚义？”

Cypher

MATCH (villain:Person {name: '高俅'})-->(hero:Person)  
MATCH (hero)-->(e:Event {name: '梁山聚义'})  
RETURN hero.name  
 这条查询链 (高俅) -> [迫害] -> (好汉) -> [参与] -> (聚义) 揭示了《水浒传》的因果逻辑链（Causal Chain）。它证明了梁山不仅是一个社交圈，更是一个由特定外部压力（高俅）和内部事件（聚义）共同驱动的组织（Organization）。这种建模思维对于现实世界中的犯罪团伙挖掘或金融欺诈检测具有直接的借鉴意义。

结语：迈向神经符号人工智能

我的这篇报告详细梳理了从“文本数据采集”到“关联数据分析”的完整技术路径。

在文本端： 我们见证了从Jieba的统计分词到DeepSeek的深度语义推理的演进。数据不再是冰冷的字符，而是带有情感温度和逻辑深度的载体。

在图谱端： 我们实践了从静态的实体链接到动态的事件驱动建模的跨越。美团大脑和水浒图谱证明，只有将数据连接成网，才能涌现出“推荐”、“推理”和“洞察”等高级智能。

未来展望：

随着GraphRAG技术的成熟，未来的数据分析将走向神经符号AI（Neuro-Symbolic AI）的融合之路。大模型（DeepSeek）提供通用的推理能力，而知识图谱（Neo4j/Meituan Brain）提供精确的领域知识。本项目所涉及的“文本清洗 -> 实体抽取 -> 图谱构建 -> 语义推理”这一pipeline，正是构建下一代企业级AI应用的标准范式。

主要参考资料：

1 原始结课作业文档：'2025-用户数据采集与关联分析-结课作业.pdf'

2 SnowNLP与LDA模型在微博情感分析中的应用

3 DeepSeek-R1的可解释性情感分析研究

4.基于知识图谱与图神经网络的推荐系统

5. 多模态学科知识增强生成（MDKAG）框架

6. 使用Neo4j GraphRAG构建知识图谱

7. Neo4j官方文档：事件驱动建模与图分析