2020 年第七届中国可视化与可视分析大会

数据可视分析挑战赛

（ChinaVis Data Challenge 2020）

作品说明文档

参赛队名称：　　西南科技大学-张建军

作品名称：　疫.谣言与新闻

作品主题关键词：　　疫情与舆情的时空变化趋势 舆情之间的关联关系 特定主题舆情的态势变化

团队成员：　张建军，西南科技大学，1482901238@qq.com，队长

周阳，西南科技大学，320235617@qq.com

吴毅，西南科技大学，1347419360@qq.com

薛炜，西南科技大学，1165021785@qq.com

杨甜，西南科技大学，1513552824@qq.com

彭莉娟，西南科技大学，12811924@qq.com，指导老师

张庆明，西南科技大学，29932627@qq.com，指导老师

团队成员是否与报名表一致（是或否）：是

是否学生队（是或否）：　是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：　Excel、Echarts、Python、D3

共计耗费时间（人天）：　60 人天

本次比赛结束后，我们是否可以在网络上公布该文档与相关视频（是或否）：是

一、作品简介：请围绕作品主题、要解决的问题\场景、目标用户\读者、应用价值等方面简要介绍作品（建议参赛者描述本部分内容不多于 500 字，图表不多于 1 个）

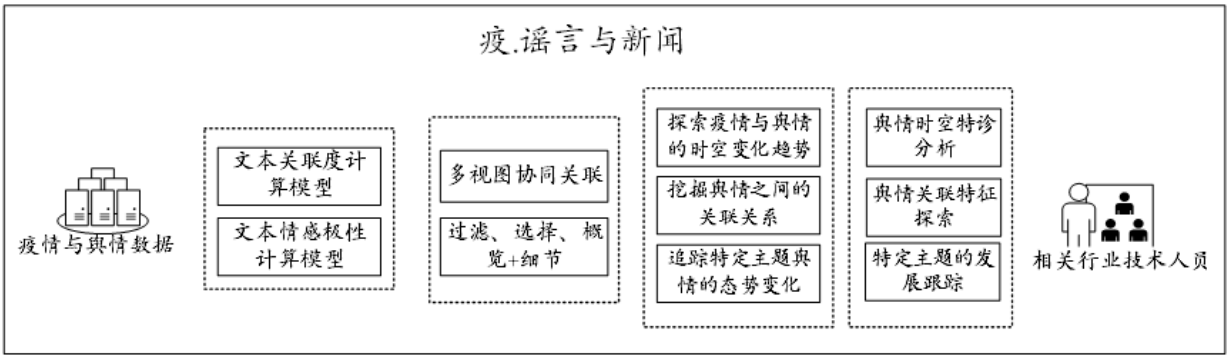

图 1-1 系统总体结构概览图

为探索疫情与舆情的时空变化趋势、挖掘舆情之间的关联关系以及追踪特定主题舆情的态势变化，针对全球“抗疫”过程中产生的疫情、新闻和谣言数据，利用文本关联度计算模型与文本情感极性计算模型，结合多视图协同关联可视化技术与过滤、选择、概览+细节等交互方式，我们设计并实现了“疫.新闻与谣言”可视分析系统。为相关行业技术人员提供有效的分析和探索工具，辅助其在疫情舆情领域开展深入研究，实现对抗击疫情过程中舆情的时空特征分析、关联特征探索与特定主题的发展跟踪。利用系统对疫情期间的实际案例进行探索分析，验证了该系统的有效性与易用性。

二、数据介绍：请围绕数据来源、数据格式、数据严谨性、数据清洗等方面简要介绍（建议参赛者描述本部分内容不多于 500 字,图表不多于 3 个）

本作品使用的数据包括疫情和舆情两类数据。疫情数据来自国家卫健委的官方播报。舆情数据分为谣言和新闻数据，谣言数据来自今日头条疫情防控专栏下防控辟谣专区，新闻数据则来自南方传媒“记疫”专栏，舆情转发、点赞、评论和发布该舆情的媒体数量从微博中获取，复工复学专题数据从已有舆情数据中提取。疫情数据时间为 2019 年 12 月 31 日至 2020 年 5 月 27 日，截止到 2020 年 6 月 1 日 0 时，共获取新闻数据 1289 条，谣言数据 920 条，提取复工复学专题数据 101 条。上述数据每周进行一次更新。

● **疫情数据**

疫情数据字段包括国家、省或城市的中文和英文名及日期编号。当日未公布数据记为空。

表 2-1 疫情数据

REGION_CN	REGION_EN	T20200204	T20200205	T20200206	...
...	...	...	...	...	...

● **新闻数据**

删除原新闻数据中起止时间、链接、图片、视频与重要程度等字段，最终字段为事件标题、类别、报道时间、详细内容、主题、转发量、点赞量、评论量及发布该舆情的媒体数量。

表 2-2 新闻数据

EVENT	CATEGORY	TIME_POINT	REMARK	THEME	TRANSMIT	THUMBSUP	COMMENT	MEDIUM
...	...	...	...	...	...	,..	...	...

● **谣言数据**

删除原谣言数据中的图片、链接和视频等字段，最终字段为内容、发布时间、类被、来源、转发量、点赞量、评论量及发布该舆情的媒体数量。在数据清洗过程中删除了 23 条无效内容的谣言条目。

表 2-3 谣言数据

TITLE	PUBTIME	KEYWORD	SOURCENAME	TRANSMIT	THUMBSUP	COMMENT	MEDIUM
...	...	...	...	...	...	...	...

三、分析任务与可视分析总体流程（建议参赛者描述本部分内容不多于 500 字，图表不多于 3 个）

1、分析任务

● 疫情、新闻和谣言时序特征分析

在疫情发展过程中，新闻和谣言同样也在发生着变化，通过比较这三者时序特征，能够从宏观上理解舆情与疫情在整个时间层面的进程变化。

● 疫情、新闻和谣言时序相关性探索

舆情的走向与疫情的发展有着千丝万缕的关系，探索疫情、新闻与谣言的时序相关性，能够从微观上解析舆情和疫情在时间层面的相关性。

● 舆情关联特征探索

新闻和谣言的产生不仅与疫情进程相关，二者也是相互影响的。通过探索二者在特定时间段上的关联特征，可以更好地了解新闻与谣言的产生过程，起到溯源的作用。

● 舆情数量地理分布特征分析

舆情的变化不止体现在时间层面，在空间上即地区分布上同样体现着差异性。分析舆情数量的地理分布特征，能够更迅速的感知社会舆情的变化。

● 特定时空下的舆情特征探索

在感知到社会舆情的变化后，对时空进行界定，探索特定时空下的舆情特征。

● 特定主题舆情的态势分析

对特定主题舆情的态势进行监测，帮助系统使用者从整体上观测舆情的发展，分析舆情的情感倾向，把握舆情的整体状态，从多个方面剖析主题，能够更全面的掌握特定主题的舆情的发展及状态。

2、可视分析总体流程

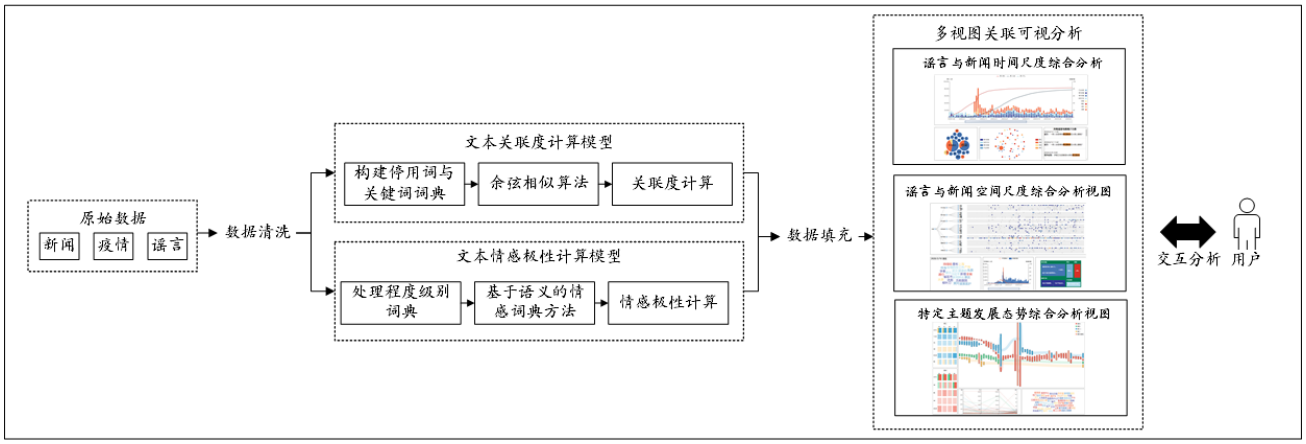

图 3-1 可视分析总体流程图

四、数据处理与算法模型（建议参赛者描述本部分内容不多于 1000 字，图表不多于 5 个）

1、舆情文本关联度计算模型

● 文本相似度计算

舆情文本关联度计算模型的核心算法是余弦相似度算法，通过计算两个向量的夹角余弦值来评估它们的相似度。在使用余弦相似度算法计算两段文本的关联度时，以文本中单个字为向量，取每个字在文本中出现的次数作为此字向量的值。假设文本 1 中出现的字为: $Z1c1$、$Z1c2$、$Z1c3.....Z1cn$；它们在章节中的个数为: $Z1n1$、$Z1n2$、$Z1n3.....Z1nm$；文本 2 中出现的字为:、$Z2c1$ Z2c2、Z2c3..... $Z2cn$；它们在章节中的个数为: $Z2n1$、$Z2n2$、$Z2n3.....Z2nm$；其中 $Z1c1$ 和 $Z2c1$ 表示两个文本中同一个字，$Z1n1$ 和 $Z2n1$ 是它们分别对应的个数，那么两个文本的相似度便可以进行如下计算：

$$\mathrm{Sim}\textit{ilaryValue}=\frac{(\mathrm{Z1n1}\times\mathrm{Z2n1})+(Z1n2\times Z2n2)+(Z1n3\times Z2n3).....\times(Z1nn\times Z2nn)}{\sqrt{Z1n1^2+Z1n2^2+Z1n3^2.....+Z1nn^2}\times\sqrt{Z2n1^2+Z2n2^2+Z2n3^2.....+Z2nn^2}}$$

● 构建停用词与关键词词典

为排除 jieba 分词后的一些并不能体现文本特征且对关联度有极大干扰的词语，例如数字、程度副词等，新增停用词词库。为更好突出文本中关于疫情的特征，结合新冠肺炎爆发以来的新闻热点及社会舆论焦点，在原有的 jieba 分词基础之上新增关键词词库。

● 关联度计算

将经过数据预处理的新闻谣言数据作为计算文本带入关联度计算模型，每一条新闻数据为文本 1，谣言数据为文本 2。

在对文本进行分词的过程中，根据 jieba 库与新增的停用词词典和关键词词典进行分词操作，对原始算法进行改进，选取文本中的每个分词作为向量，取每个词在文本中出现的次数作为此词向量的值，此时已将新闻谣言文章特征进行数字化处理，便可根据余弦相似度进行新闻谣言关联度计算。

2、舆情文本情感极性计算模型

● 情感极性计算

舆情文本情感极性计算模型是利用基于语义的情感词典方法对文本情感做极性计算，即利用情感词典及否定词典及程度级别词典，给情感强度不同的情感词赋予不同权值，结合否定词的情感反转以及程度副词的强弱值，采用权值算法通过情感打分的方式进行文本情感极性判断。

假设一个文本中有 $\mathrm{N_p}$ 个正面情感词汇，$\mathrm{N_n}$ 个负面情感词汇，正面情感词汇权值为 wp_{i}，每个正面情感词汇前有 dp_{i} 个否定词，该情感词汇前的程度副词数值为 $\mathrm{vp_i}$，负面情感词汇权值为 $\mathrm{wn_j}$，每个负面情感词汇前有 $d\mathrm{n_j}$ 个否定词，该情感词汇前的程度副词数值为 $\mathrm{vn_j}$，那么该文本的情感极性数值为

$$\overline{\mathrm{E}}=\sum_{i=1}^{N_p}(-1)^{\mathrm{dp_i}}\times v\mathrm{p}_i\times wp_i+\sum_{j=1}^{N_n}(-1)^{\mathrm{dn_j}}\times v\mathrm{n}_j\times wn_j$$

最后确定阈值判断文本倾向性,结果为正是正面倾向，结果为负是负面倾向，结果为零无倾向。

● 处理程度级别词典

对程度级别词典进行程度强弱分类，根据词语语义设置程度数值处理效果如下。

表 4-1 程度级别词典

程度副词	程度数值
百分之百	4
多么	3
那么	2
有点	0.5

● 情感极性计算

将已处理过的词典以及需要计算的舆论文本带入舆情文本情感分析模型，在进行舆论文本分词过程之前，将新增的停用词词典与关键词词典引入，减少干扰因素，所得数值则为该舆论的情感极性数值。

五、可视化与交互设计（建议参赛者描述本部分内容不多于 1500 字，图表不多于 10 个）

1、谣言与新闻时间尺度综合分析视图

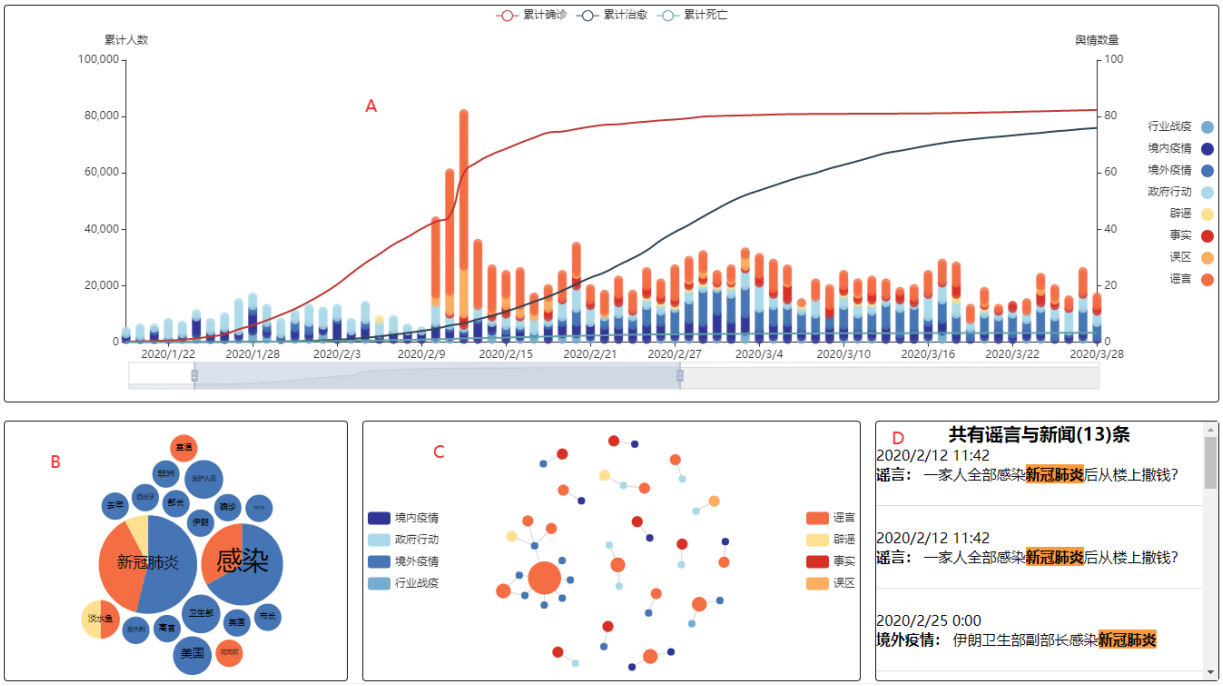

图 5-1 谣言与新闻时间尺度综合分析视图

● 舆情与疫情概况分析

为分析疫情、新闻和谣言在时间上的发展趋势与关联，我们设计了舆情与疫情概况组合图（图 5-1A）。三条曲线分别代表累计确诊、累计治愈和累计死亡；不同颜色的散点代表不同类型的舆情，新闻与谣言分别是蓝色系与红色系；左侧竖轴代表人数，右侧竖轴则代表舆情数；横轴代表时间，并提供选框用于对时间的筛选。图例提供点击交互用于疫情以及舆情数据的类别筛选；点击某个散点，对当天的谣言与新闻进行展示（图 5-1D）；如图 5-2，对时间进行框选时，系统会自动生成框选时间段的舆情关系图（图 5-1C）。

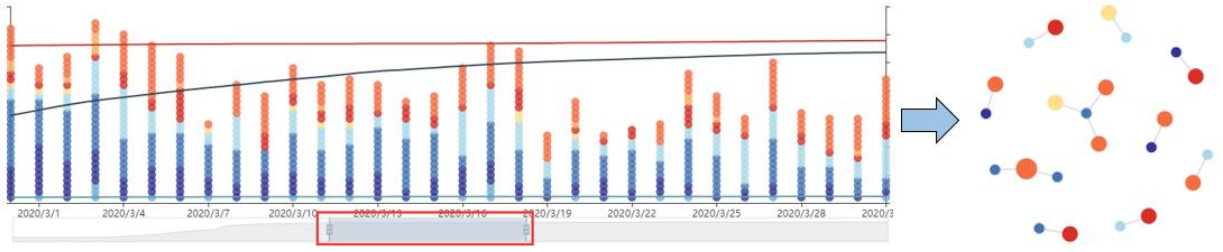

图 5-2 选框交互示意图

● 谣言与新闻关联性分析

为探索谣言与新闻的关联特征，我们设计了舆情关系图（图 5-1C）与关键词气泡图（图 5-1B）。舆情关系图中，不同的散点代表不同的舆情，其颜色通过舆情的类别来映射；散点的大小代表了与之相关的

其他舆情的数目；散点之间的连线则代表它们对应的舆情存在关联关系。关键词气泡图中，不同的气泡代表了不同的关键词；气泡的大小由关键词在所选类簇中出现的次数来映射；气泡的颜色是该关键词所处舆情的类别。

图例提供点击交互用于针对不同类别谣言和新闻的筛选；如图 5-3，点击某个散点，将该类簇中舆情的具体信息进行展示，并对共有关键词进行标注，同时更新该类簇的关键词气泡图。

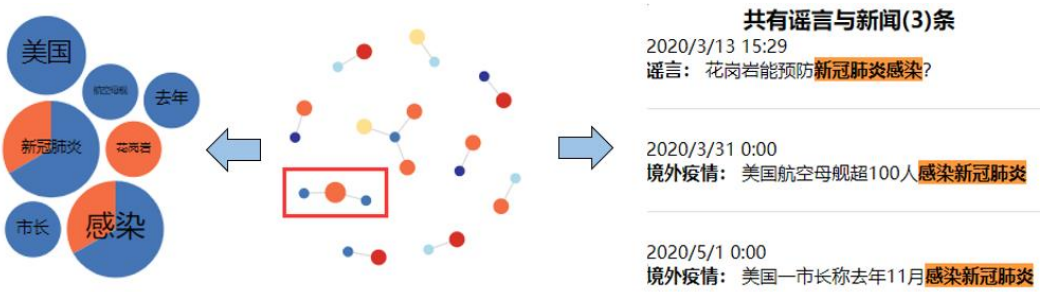

图 5-3 散点点击交互示意图

2、谣言与新闻空间尺度综合分析视图

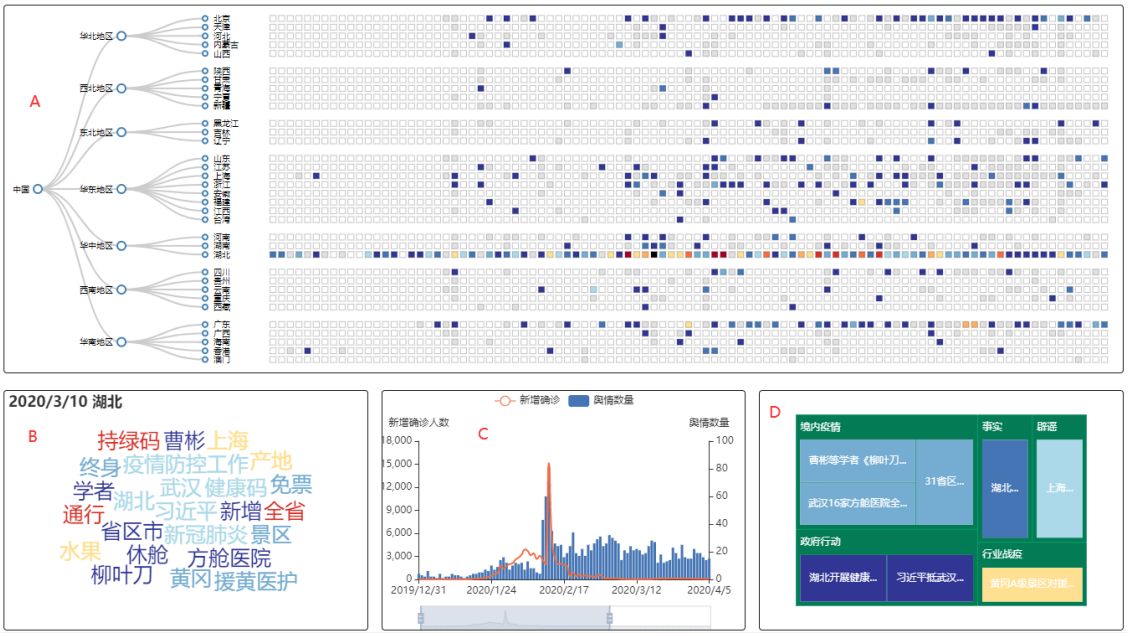

图 5-4 谣言与新闻空间尺度综合分析视图

● 舆情与地区关联性分析

为了分析各地区各省的舆情状况，我们设计了舆情与地区关联性分析视图(图 5-4A)。树图展示了我国地区的结构，使用热力图展示了各地区各省每天在新闻与谣言中被提及的次数。不同的颜色代表热度的大小，颜色越明亮热度越大。

如图 5-5，在树图中点击第二层的节点将各地区中的省份合并以查看整个地区的热度情况。点击热力图中的区域，舆情类别矩形树图与舆情关键词云图会进行更新，以便更细粒度的查看。

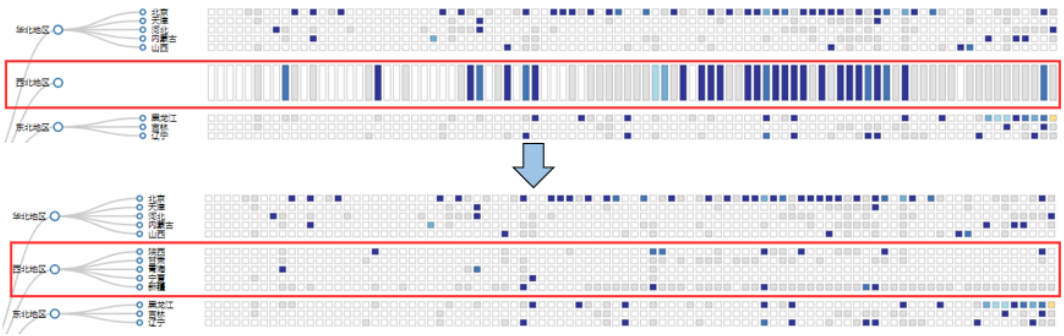

图 5-5 图像合并交互示意图

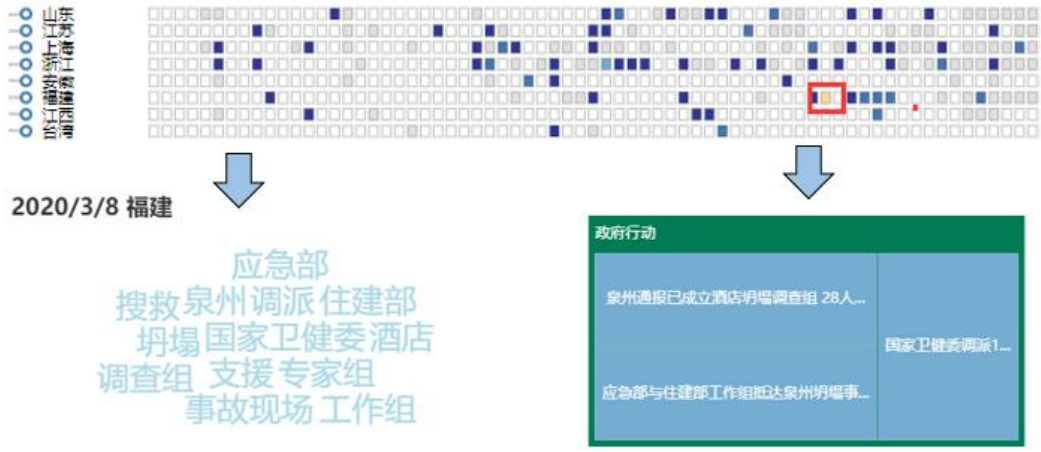

图 5-6 细节探索交互示意图

● 舆情与疫情趋势分析

为探索疫情与舆情在时间变化中的趋势关系，我们设计了舆情与疫情走势组合图（图 5-4C）。横轴代表时间，左纵轴代表人数，右纵轴代表舆情数；折线代表每日新增的国内确诊人数，矩形块代表每日的舆情数目。通过横轴提供的选框可以界定时间范围，同时更新舆情与地区关联组合图。

● 特定舆情特征分析

为分析特定舆情中的类别概况与关键词信息，我们设计了舆情类别矩形树图（图 5-4D）与舆情关键词云图（图 5-4B）。舆情类别矩形树图中，矩形的大小由舆情出现的次数映射，颜色则代表舆情的类别；舆情关键词云中，词的大小由关键词出现的次数映射，颜色则代表该关键词对应舆情的类别。

3、特定主题发展态势综合分析视图

● 特定主题舆情态势分析

为把控舆情的情感概况，我们设计了舆情整体情感概览视图（图 5-7A）。结合舆情的微博热度情况与舆情的类别，对微博热度数据分组，例如复学的第一个矩形所代表的数据为类别为“政府行动”的舆情转发的数量，并用透明度来映射数值大小。使用不同的颜色来区别舆情所属主题与情感。

如图 5-7B，根据计算得到的情感极性以及相关新闻谣言的时间属性，形成了特定主题舆情发展图。每一个矩形的高度反应了舆情在微博的讨论热度。在微博热度平行坐标图中（图 5-7C），展示了相关舆情的点赞、评论、转发以及发布的媒体的数量，并提供选框用于筛选热度高的舆情。通过特定主题舆情发展图的点击以及平行坐标的筛选操作，将选中的舆情关键词进行展示，如图 5-7D 所示。

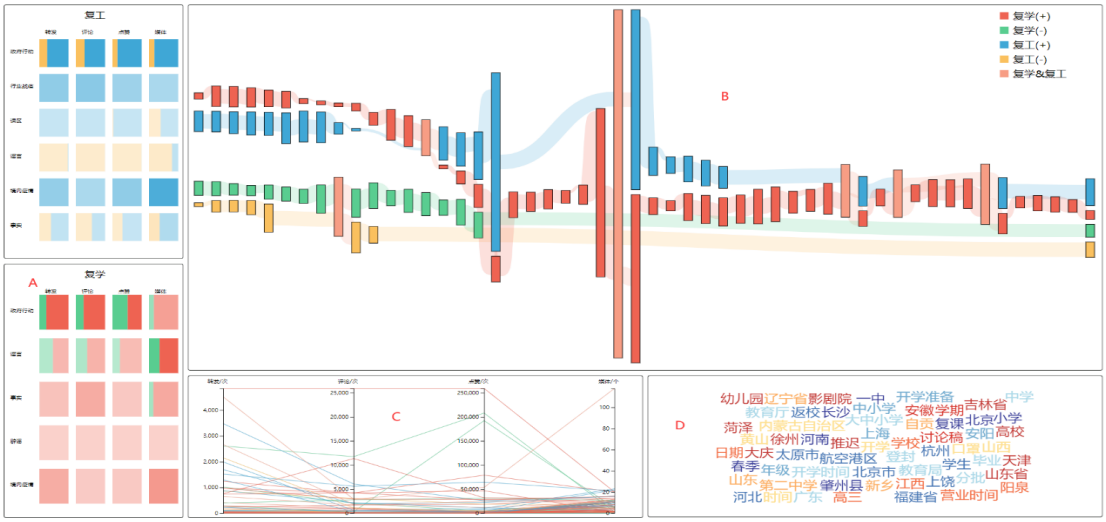

图 5-7 特定主题发展态势综合分析视图

如图 5-8，通过与主题发展桑基图结合，可以进一步查看某一类别舆情的发展，从各个方面对主题的发展进行分析。

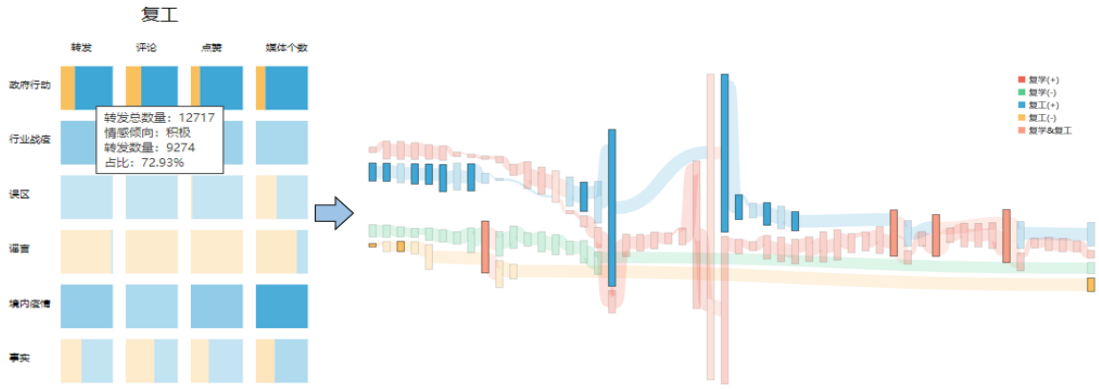

图 5-8 主题整体概览交互示意图

六、实验\案例\场景分析（建议参赛者描述本部分内容不多于 2000 字，图表不多于 15 个）

1、疫情、新闻和谣言时序特征分析结果

● 疫情初期政府行动与境内疫情类新闻数据较多，谣言数据较少

如图 6-1，可以很明显的发现红色系的散点几乎没有，即谣言类舆情在疫情初期传播较少。疫情初期，国内正处于“谣言”敏感阶段，因此少有传播较广的谣言出现，且这段时间对谣言的收集较少，许多媒体平台对谣言的整理在这个阶段都没有启动。

通过交互发现，新闻类数据中占比较高的是政府行动（图 6-1A）与境内疫情（图 6-1B）。说明在政府层面对疫情的重视程度较高，这也是我国抗疫成绩优异的原因；境内疫情新闻较多也与政府的“疫情透明”相关政策有关，向民众持续的通报疫情概况。

图 6-1 疫情早期舆情与疫情概况示意图

● 疫情高峰期谣言类数据较多，新闻数据较少

如图 6-2，可以发现在红色的线（累计确诊）增长率最大时，谣言类数据不论是占比还是本身的数量都是远高于其他时期的，这也符合谣言的产生特征。

通过交互发现，在谣言数据中，随着谣言不断增加的还有误区类数据（图 6-2）。不仅是谣言的快速传播，误区即民众对知识的错误认知也在不断上升，说明混乱的舆论环境同样会影响人的判断与认知

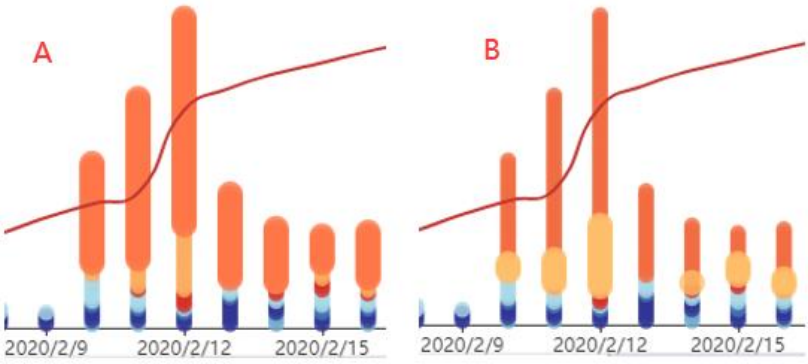

图 6-2 疫情高峰期舆情与疫情概况示意图

2、疫情、新闻和谣言时序相关性探索结果

● 谣言与累计确诊人数增长率相关

如图 6-3，谣言数量在达到高峰后迅速下降，后续趋于平稳；累计确诊数量在达到峰值后同样趋于平稳。从图中框选部分可以看出，二者趋势发生突变的时间几乎一致，可以说谣言与确诊人数的增长率呈正相关。

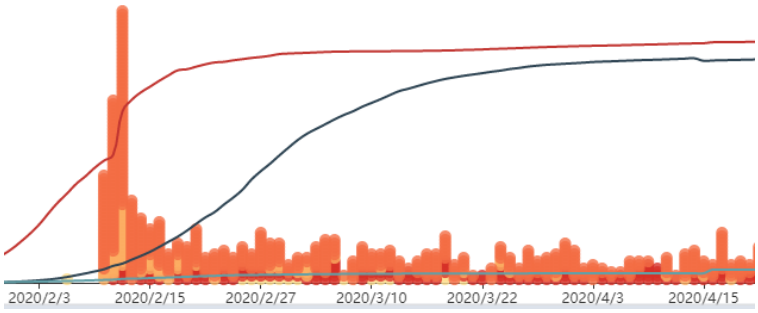

图 6-3 谣言与疫情趋势示意图

● 新闻与疫情是否得到有效控制相关

如图 6-4A，新闻数量在整个时间范围上都一直处于波动状态，从图中框选部分可以发现，累计确诊人数趋于平稳初期，新闻数量达到峰值。

出现该结果的原因可能是在这期间境外已经出现较大范围的疫情传播，大量新闻是和境外疫情相关。通过交互筛选掉新闻中境外疫情的部分，如图 6-4B，框选部分是新闻最密集的时期，可以看到，红色曲线与蓝色曲线逐渐靠近，即累计治愈人数逐渐与累计确诊人数相近，也就是疫情得到有效控制的过程。疫情是否得到有效控制是决定新闻数量的关键因素。

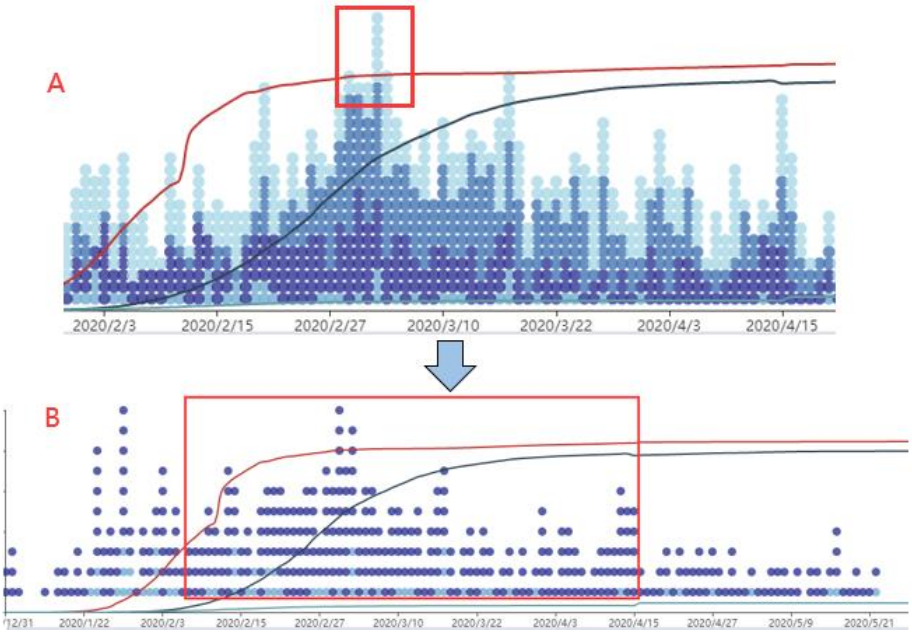

图 6-4 新闻趋势示意图

3、舆情事件关联特征探索结果

● “郭某鹏”事件相关新闻溯源

如图 6-5，可以看到关系图中自动将郭某鹏相关的舆论关联到了一起。观察此类舆情的特征可以发现，主要集中在“郭某鹏”以及“密切接触者”两个话题，且时间上谣言传播较新闻更早。此类谣言先行产生后，引起了较大的社会关注，逐渐也成为了新闻的热点话题。

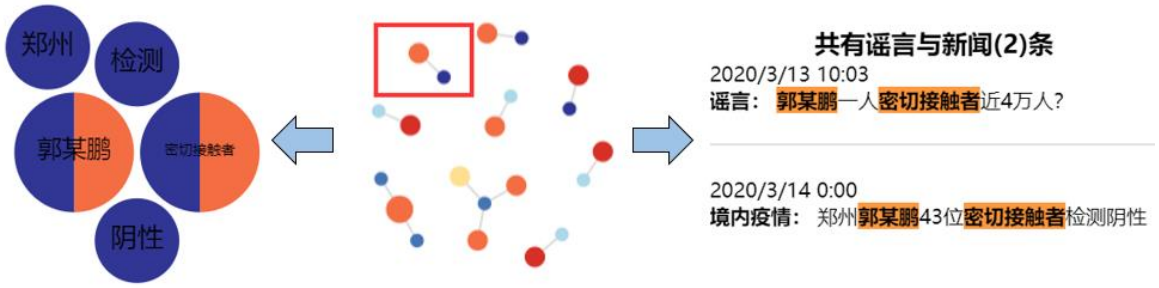

图 6-5 “郭某鹏”舆情关联图

● “武汉在院新冠肺炎清零”误区谣言溯源

如图 6-6，在点击框中节点后，观察该类关联舆情的特征，发现是与“武汉在院新冠肺炎清零”相关，主要的关键词是“武汉”、“清零”与“在院”，且新闻发布时间较误区传播更早。一则新闻发布后，其中可能会存在与常规概念不同的词汇，对此类词汇的认知差异，则会导致误区认知的传播，进而形成谣言。

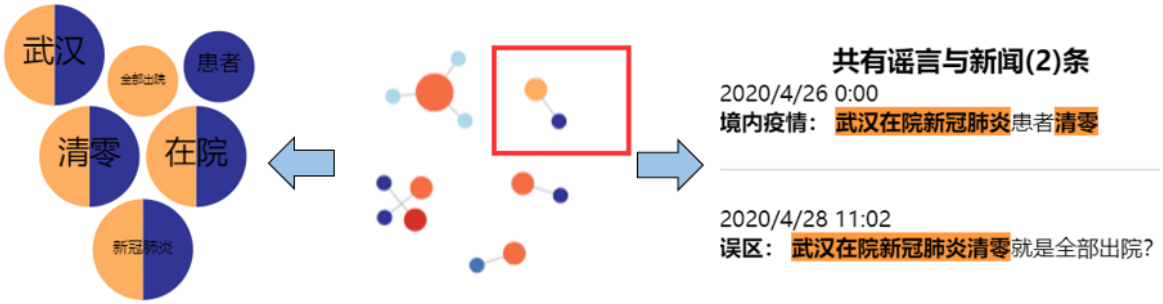

图 6-6 “武汉在院新冠肺炎清零”舆情关联图

4、舆情数量地理分布特征分析结果

● 疫情高风险地区舆情数量更多

如图 6-7，很容易可以看到北京、湖北、广东舆情数量分布密集且数量较多，随着疫情不断得到控制，这些省份的舆情数目也逐渐减少；同样也可以看到黑龙江和吉林在疫情后期舆情数量分布较密集，这与疫情后期该类地区未有效防范输入病例造成多人接触感染有关。

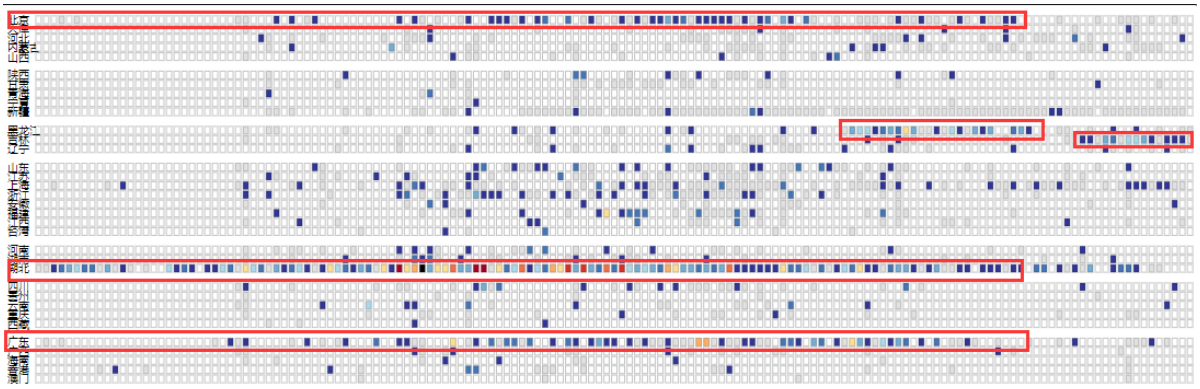

图 6-7 高风险地区舆情数量地理分布图

5、特定时空下的舆情特征探索结果

● **福建舆情突增异常**

如图 6-8，发现 2020/3/8 这天，福建的舆情数目发生了突增。通过交互对当日舆情下探可以得知，当天福建泉州一隔离酒店发生了坍塌，造成了舆论的突增。

图 6-8 福建异常舆情下探示意图

● **黑龙江舆情持续增长异常**

如图 6-9，发现 2020/4/13 日起，黑龙江的舆情数目突然持续增多。通过交互查看这几日舆情关键词发现，自 4/13 日起，就不断的有境外输入病例相关的舆论，甚至不断有“封城”，“封路”等关键词，可以看到民众对于输入病例的恐惧。

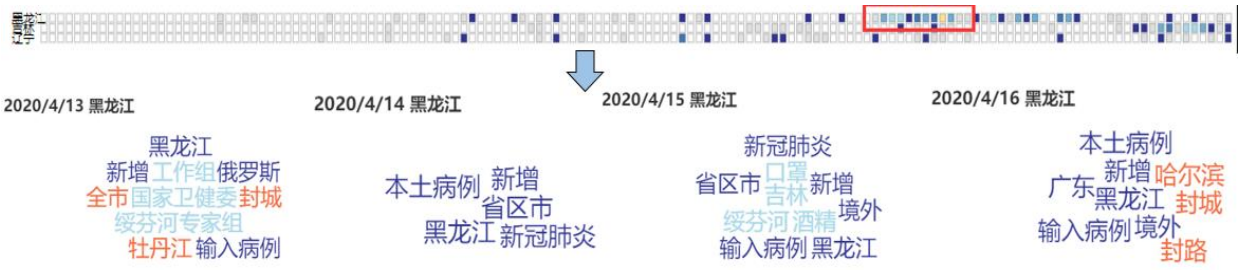

图 6-9 黑龙江异常舆情下探示意图

6、特定主题舆情的态势分析结果

● 复学相关谣言情感偏向积极

如图 6-10 所示，复学中谣言的情感倾向积极的比消极的多，到 4/19 之后，基本没有负面情绪的谣言出现。而偏向积极的谣言依然存在且大都与开学相关，说明对于舆情中对复学的期待值是越来越高的。

图 6-10 复学谣言情感倾向视图

● 高热度舆论概况分析

针对部分在微博上被广泛讨论的舆论，需要进行额外的关注，防止有害舆论的进一步传播。如图 6-11，我们筛选了转发数超过 2000 次的新闻与谣言。

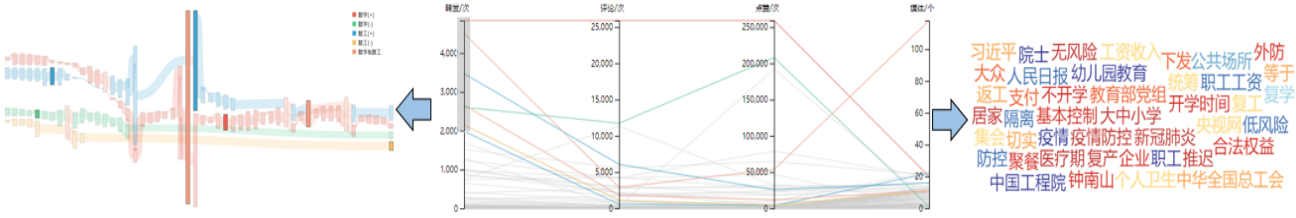

图 6-11 微博转发数较高的舆论视图

点击桑基图中的矩形，可以查看舆论的具体内容，如图 6-12 所示。我们查看了第一个热度较高的关于复工的新闻，可以看到，人们在关注在家办公人员的工资问题；钟南山院士呼吁人们不要懈怠，继续保持距离的新闻则是报道媒体做多的；而对于复学，关注最多的是开学时间，关于“已有 17 个省份明确高校开学时间”是讨论的最多的话题。

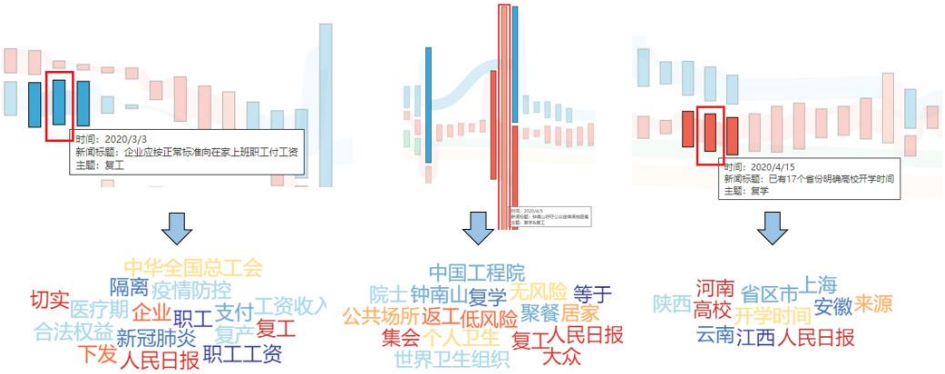

图 6-12 舆论具体内容图

七、讨论与总结（建议参赛者描述本部分内容不多于 500 字）

本次挑战赛中，我们设计并实现了“疫.新闻与谣言”可视分析系统，该系统构建了谣言与新闻时间尺度综合分析视图、谣言与新闻空间尺度综合分析视图以及特定主题发展态势综合分析视图三个页面，并提供多种交互手段来支持具体任务的灵活探索分析。

通过该系统，我们对实际案例进行了疫情初期和高峰期新闻和谣言的时序特征分析、新闻和谣言在时序上与疫情的相关性分析、舆情事件的关联性特征探索并溯源、疫情高风险地区舆情地理分布特征分析、异常舆情探索以及“复学”主题相关的舆情情感与热度分析，最终分析结果也验证了系统的有效性。后期可进一步对舆情评论与转发的文本进行挖掘，观测具体舆情在社会中的变化过程，探索舆情的传播与兴衰过程。