# 司法判决文书数据中演化模式的挖掘和可视分析

王慧[1)], 孙国道[1)*], 丁伟杰[2)], 王潇[1)], 张颖[1)], 梁荣华[1)]

[1)] (浙江工业大学计算机科学与技术学院　杭州　310000)
[2)] (浙江警察学院大数据与网络安全研究院　杭州　310000)
(guodao@zjut.edu.cn)

**摘　要:** 现有的司法数据分析以传统的统计图表为主, 与案件数据的内部信息关联不够, 低效落后, 难以满足司法数据的分析需求. 因此, 针对司法判决文书数据, 提出一种通过用户交互的方式展示司法数据时序演化过程的方法. 首先提出基于分层结构的层次演化视图, 从整体上分析时序变化; 结合矩阵视图、词网络视图、主题聚类图和词云, 展开案件内部多元信息的精细化分析; 提取案情三元组信息, 并展开关键实体的约束布局, 帮助快速定位和掌握案件内容. 基于千万级规模司法判决文书数据, 展开案件纠纷类型演变和案件内部关系的案例分析, 并针对多个视图任务结合评分方式展开定量用户研究评估, 反馈得到所有问题平均评分均大于 3.5, 为后续司法文书数据的可视分析提供了一种解决思路.

**关键词:** 司法判决文书; 司法数据演化; 文本分析; 可视分析
**中图法分类号:** TP391.41　　**DOI:** 10.3724/SP.J.1089.2022.18916

# Visual Analysis of Evolutionary Patterns in Judicial Judgment Documents

Wang Hui[1)], Sun Guodao[1)*], Ding Weijie[2)], Wang Xiao[1)], Zhang Ying[1)], and Liang Ronghua[1)]

[1)] (*College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou*　310000)
[2)] (*College of Big Data & Network Security, Zhejiang Police College, Hangzhou*　310000)

**Abstract:** Existing analysis of judicial data is mainly based on traditional statistical charts, and can not better relate to the internal information of judicial document nor satisfy the requirement of judicial data analysis. A visual analysis approach is proposed to analyze the judicial judgment document data by interactively presenting the evolution of judicial data. First, the approach consists of a hierarchical structure based evolution view to analyze the overall changes of judicial judgment document data. Second, the approach combines matrix view, word network view, topic clustering graph, and word cloud to carry out a refined analysis of internal multiple information. The triplet information of the case is further extracted and the key entities are visualized with constraint layout to help locate and grasp the contents of the case quickly. Based on a large-scale judicial judgment documents, case analysis is conducted on the evolution of case dispute types and the internal relations of the case, and quantitative user research and evaluation for multiple view tasks combined with the scoring method are carried out. The average score of all questions is larger than 3.5, which provides a potential solution for the subsequent visual analysis of judicial document data.

收稿日期: 2020-01-29; 修回日期: 2021-06-15. **基金项目**: 国家自然科学基金(61972356, 62036009); 公安部科技强警基础工作专项(2020GABJC35); 浙江省公益技术研究计划(LGF19G010001). **王慧**(1995—), 女, 硕士研究生, 主要研究方向为信息可视化; **孙国道**(1988—), 男, 博士, 副教授, 论文通信作者, 主要研究方向为信息可视化与可视分析; **丁伟杰**(1980—), 男, 副教授, 主要研究方向为犯罪数据可视分析; **王潇**(1997—), 男, 硕士研究生, 主要研究方向为信息可视化; **张颖**(1997—), 女, 硕士研究生, 主要研究方向为信息可视化; **梁荣华**(1974—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究方向为可视化和计算机图形图像处理.

**Key words:** judicial judgment documents; evolution of judicial data; text analysis; visual analysis

随着人类生产生活的快速发展和社会活动的日益丰富，法律的规制和法律体系变得越来越庞杂；同时，得益于互联网大数据技术的迅猛发展，司法数据呈爆炸式增长[1]. 现有的司法数据处理系统以传统报表为主，对海量数据的处理方法相对落后，这导致了数据堆积，新旧数据未能建立有效联系，无法高效地利用新数据. 因此，研究司法大数据有助于解决司法判决文书数据的堆积问题；有助于掌握社情民意，全面感知社会矛盾风险，透视社会矛盾的热点、重点、难点和新点[2-3].

然而，在当前的司法大数据处理工作中，存在以下几个技术难点和问题：如何高效、系统地掌握案件详情，同时对新旧司法数据进行关联性的挖掘分析；如何处理数量巨大、类型多样且数据特征隐藏较深的数据；如何用可视化方式对案件主题的多特征信息进行呈现.

为了解决上述技术难点和问题，本文提出了对司法判决文书数据进行演化模式探索的可视分析方法. 采用文档图谱视图对个案进行可视化，对非结构化文本采用深度学习的方法抽取，结合结构化文本的关键信息，基于实体布局约束算法进行三元组排布展示，帮助用户快速、准确地掌握案件数据之间的联系. 通过层次演化视图进行不同层级案件类型的聚合和分解，同时使用时间控制器控制展示堆叠图中细粒度级别的时间段，并与词云进行交互探索呈现众多类别随时间的变化，以观察案件纠纷类型的演变，帮助用户发现热点纠纷类型和特殊时间节点. 矩阵视图基于主题强度和熵值，通过横向与纵向对比颜色插值的变化，探索案件文档主题的关注度与聚合程度，同时结合基于隐含狄利克雷分布(latent Dirichlet allocation, LDA)模型训练得到的主题聚类视图，帮助用户直观地了解不同案件主题之间的关系. 词网络视图使用关键词共现的方式将词频和共词相结合，帮助用户更好地识别案件的发展过程，找出不同类型的关键词随时间变化的关系，发现热点问题.

整个可视分析过程从宏观到微观分析司法数据的演变. 针对案件抽取出三元组形式构建新型的可视呈现方式，帮助更好地挖掘实体之间的关联；面向司法数据提供一种可视分析思路，结合文书内部关联、案情关系与时序进行可视分析；基于本文系统提供的司法案例分析和用户实验，检验了可视分析系统的可用性和有效性.

## 1　相关工作

### 1.1　时序可视化

近年来，可视化被广泛应用于分析面向时间的数据. Sultanum 等[4]利用时间线与 breadcrumbs 和直方图为医生提供了自动化审计和可定制性，同时也可链接到病人相关病史的原始文本. Galex[5]帮助用户探索任意学科的演化和交叉过程，允许用户通过时间刷选取时间片来探索某个领域迅速扩展到其他哪些领域，以及某些学科的研究近况. Flowstrates[6]从热力图中可以显著地看出随着年份的变化，难民从起源地流动到目的地的数量变化情况. CiteRivers[7]提供了与数据挖掘技术集成的多个协调视图，它利用光谱聚类技术，在用户可以交互调整的粒度级别上分层地聚合文档、期刊和会议. Cuenca 等[8]使用多种交互技术，根据各个时间序列间的邻近程度进行层级分组，方便用户的探索. 然而，对于层级的分组和内部的探索都是单独存在的，研究者很难提取更多的时序信息.

Yang 等[9]提出了一种交互式指导方法对大数据进行数据分析，利用公众知识和个人知识监督分层聚类，该方法有助于以有效的方式构建定制的聚类树. 通过对过去 20 年间可视化领域发表的论文进行处理，Isenberg 等[10]从主题和关键词变化这 2 个角度分别进行了分析. ArchiText[11]主要用于文本数据分层主题建模的交互，运用人在回路的主题建模方式，帮助用户探索和指导流程，更好地产生符合需求和更高质量的主题. TMvis[12]通过用户对数据和任务的理解调整主题模型，并可视化主题建模得到主题语义的信息，以帮助用户理解主题模型. El-Assady 等[13]设计了一种数据转换方式，用结构化的方式表示主题模型的输出结果，并支持用户在主题内和主题间进行疾病症状的探究. 主题随时间演变趋势的可视化支持对大型文档的理解，因此本文设计了交互的方式探索主题的演变.

### 1.2　文本分析可视化

文本可视化中涉及众多内容总结和相关信息提取的方式. VISPubComPAS[14]是一个需求驱动的分析系统，用于分析和比较电气与电子工程师协会(institute of electrical and electronics engineers, IEEE)可视化会议上出版物的作者和从属关系，以帮助用户找到适合自己需求的合作者. PyramidTags[15]在摘要文档集中将可视化标签的时间演变和语义关系结

合在一起, 利用悬停选择标签的方法探索语义关系并检索最相关的文档, 通过概览语料库的主要术语和短语方式掌握其主要思想. embComp[16]提出了一种用于比较 2 个嵌入的方法, 将局部信息与全局视图相关联, 可以捕获类似单词或文档对象之间的相似性. 有些文本可视分析系统允许用户对单篇文档进行可视化表示. 本文将通过构建模型的方式展示单篇文档中数据的内在联系.

文本可视化也适用于各个领域. VAiRoma[17]运用文本分析技术, 从时间、地点、事件和它们之间的关系等多个角度展示罗马的历史, 提供的树状和环状视图更加有利于用户探索主题信息. 张玮等[18]通过文本关联和时空角度对宋词进行可视分析, 使文学爱好者可以从宋词文本和词人本身的年代背景等多个角度进行关联性地探索分析宋词的诸多特性. Carvalho 等[19]设计了一个用于自动提取文档语义信息的模型, 用于提高法律文本的可读性. 而本文针对司法数据进行文本可视化, 将有价值的部分展示给用户.

### 1.3 智慧司法

“中国裁判文书网”上线以来, 裁判文书的公开已经取得了巨大的进步, 对司法从业人员和数据从业人员探索智慧司法具有重大价值. 谢小平等[20]通过计量学和可视化分析进展, 探索我国司法鉴定领域发展脉络, 刻画法医学工作热点和前沿动态.

司法数据内容繁杂, 信息较多, TextRank[21]的文本摘要已经无法满足快速预览案件内容的需求, 而面向司法领域知识图谱的工作可以满足上述需求. Chen 等[22]关注网络嵌入的问题, 在网络的结构信息外增加了网络中边的标签信息, 以更好地捕捉节点之间的关系. Ghiasnezhad 等[23]提出的规则挖掘模型, 通过对规则相关的子图进行采样的方式, 在减少算法的搜索空间和计算量的同时保存了必要信息. 本文同样涉及案情知识图谱, 但增加了对段和关键实体的绑定, 使图谱更加有章可循.

目前, 将智慧司法与可视化技术结合的工作较少, 现有工作形式单一, 缺乏可探索性. 本文通过对多个视图进行交互性分析, 对司法数据进行多角度展示, 具有一定开创性.

## 2 系统概览和数据处理

### 2.1 系统分析任务和设计目标

通过司法判决文书的数据特征并总结用户需求, 确定了以下分析任务.

**任务 1**. 在单个司法案件层面, 发现个案数据之间的联系.

**任务 2**. 在时空语义层面, 探索司法案件随时间地点的变化而演变的规律.

**任务 3**. 在主题分布及语义信息层面, 挖掘案件主题的演变规律.

针对上述任务, 本系统的设计目标如下.

**目标 1**. 抽取并可视化单个案件的多方面信息, 如事件起因、赔偿情况和法院判决等(任务 1).

**目标 2**. 可视化编码不同类型案件数量, 支持多层级类型案件数量之间的比较, 并设定区间分析案件内部关联, 以及可视化对比不同地域之间的案件数量(任务 2).

**目标 3**. 可视化司法案件主题, 基于文本层面直观地了解不同案件主题的被关注程度及聚合演化情况(任务 3).

### 2.2 系统流程介绍

本文系统流程如图 1 所示. 首先对数据进行预处理, 然后根据依赖语法分析方法, 将预处理后的文本进行三元组抽取, 对抽取出的结果进行约束布局, 用于构建文档图谱视图; 接着对数据进行分类, 并根据时间段进行模型训练与算法计算, 采用颜色渲染的方式展示; 最后, 结合文书内部关联、案情关系与时序通过可视分析方法对数据进行多角度展现.

### 2.3 数据处理

数据集是基于 2014—2018 年总共 10 012 571 篇司法判决文书, 包括序列号、案件名称、所属纠纷、原被告、审判庭、立案时间、结案时间和案件内容 8 个参数. 本文支持用户对司法关键词的搜索及可视化分析, 由于对搜索的快速性要求较高, 因此本文的数据集使用 Lucene 进行数据存储. Lucene 数据库是一个用于全文检索和搜索的开放源代码程序库, 它提供的关键词全文检索和相关度排序功能, 可以极大地提升文书检索的速度和判断的准确率.

司法判决文书中包含大量的数据信息且较为复杂, 很难对其进行有效的分析. 因此, 对原有数据进行了预处理, 包括案件内容的分词、数据采样、数据降维、词向量模型和文书分类等.

(1) 案件内容的分词. 对原始数据中的案件内容进行分词处理, 即去掉 340 个停用词(如“的”“吗”), 包括各种符号, 并且去掉 208 个无关词(如“甲方”“乙方”“民事判决书”).

(2) 数据采样. LDA 主题模型无法处理原有的

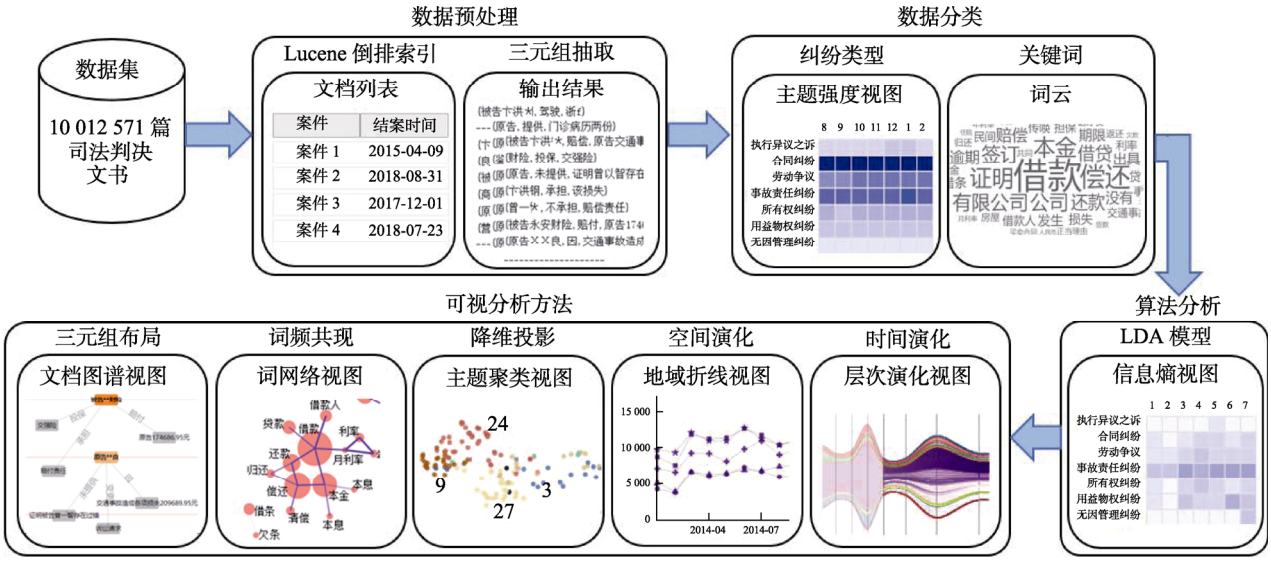



图 1　本文系统流程

数据集这么庞大的数据，因此需要对原数据进行采样. 这里不直接对原数据随机采样，因为某些类型的案件数量占比较少，随机采样后的数据量过小会影响该类型的熵值，所以选择对各类型数据随机采样 2 万条进行熵值计算.

(3) 数据降维. 采用非监督式算法 Doc2Vec 将文本数据进行向量表达. 在可视化质量方面，统一流形近似与投影(uniform manifold approximation and projection, UMAP)降维比主成分分析(principal component analysis, PCA)、多维尺度变换(multidimensional scaling, MDS), *t* 分布随机近邻嵌入(*t*-distributed stochastic neighbor embedding, *t*-SNE)降维保留了更多的全局结构，具有优越的性能. 最重要的是 UMAP 降维方法对嵌入维数没有计算限制，因此本文采用此方法进行数据向量降维.

(4) 词向量模型. 将原数据集中对案件内容进行分词处理结束后，以行为单位来处理文本. 所以每条数据占一行，由此组成语料库，再利用 Word2Vec 生成一个 100 维的词向量训练数据集.

(5) 文书分类. 根据最高人民法院印发《关于修改<民事案件案由规定>的决定(法[2011] 41 号)》文件中的案件所属类别划分可知，此文件将案件划分为了 43 个大类、424 个子类和 838 个细类.

## 3　可视分析与设计

本文结合可视化的方式，挖掘司法大数据的演变. 基于对案件三元组中实体的约束布局设计了文档图谱视图，用于直观地展示案件人物关系；采用层次演化视图对不同层级的案件纠纷在时序上进行比较分析，结合词云发现热点纠纷类型及案件前后的内容变化；矩阵视图和主题聚类视图从主题分布和语义层面挖掘数据的演变；运用词网络视图识别案件热点方向随时间的发展过程. 整个可视化系统如图 2 所示.

### 3.1　文档图谱视图

单篇司法文书之间的案件人物关系越来越复杂，用户很难快速掌握单一案件数据之间的联系，本文采用司法文档图谱视图对个案进行可视化. 通过规则梳理形成的司法文档图谱能够运用到个案，用于辅助用户对案件信息进行处理，帮助用户快速地检索相关法务内容，提高法院审判工作质量和效率. 文档图谱是由节点和链接组成的结构化视图，将原需阅读长篇案件内容转换为只需浏览关系型图解，对于快速掌握案件内容具有重要意义.

在司法文书案件中，构建案件的知识图谱需要结构化文本和非结构化文本. 结构化文本包括案件名称、所属纠纷、原被告、审判庭、立案和结案时间；非结构化文本需抽取案件内容. 本文采用深度学习的方法抽取出案件内容中的三元组(实体-关系-实体)，再结合结构化案件内容展示案件中的关键信息. 对于实体对的抽取，从案件内容的句法规则出发，利用实体在上下文中的语法关系进行抽取. 对于句法规则的处理，采用 Hanlp 中的依存句法分析模块，通过寻找“主谓关系”“定中关系”等语法关系，构建句中不同词之间的内在联系. 根据依存句法分析，得到句子的句法关系树，依据设定好的句法规则对该树进行深度搜索，以抽取出语义三元组.

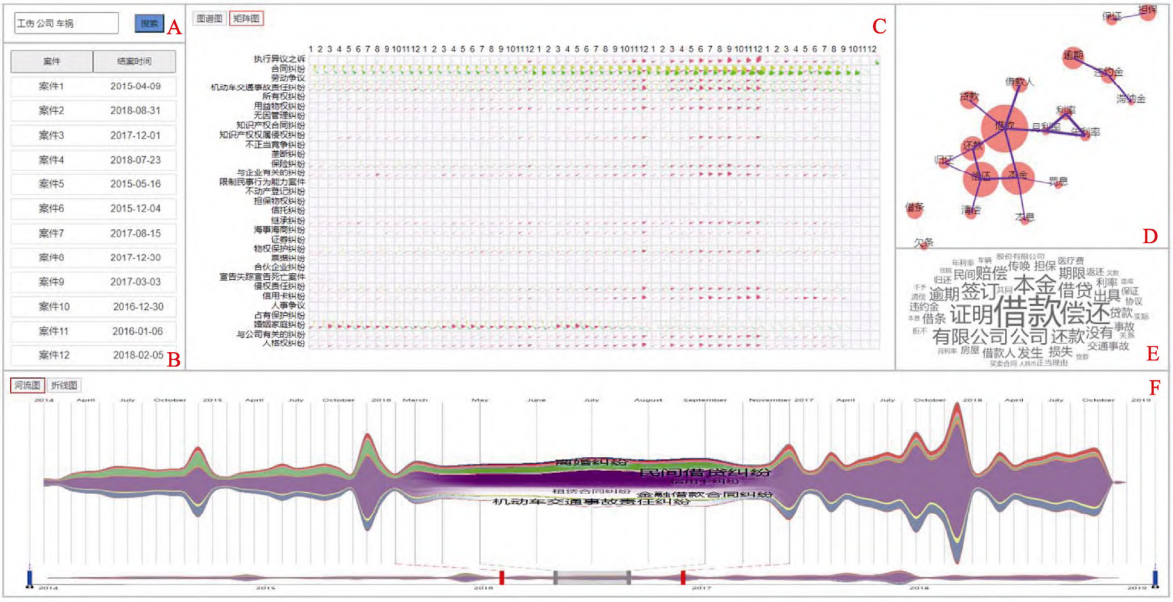



图 2 司法判决文书数据可视化系统

以“主谓宾”三元组为例抽取语义三元组. 首先确认句子的核心节点, 如图 3 中的“涉及”, 再通过主谓关系搜索到核心节点下的主语节点“合同纠纷”; 同理, 通过“动宾关系”搜索到宾语节点“关系”, 得到初步的主谓宾三元组{合同纠纷, 涉及, 关系}. 对于初步的三元组能有效表示该句子的核心含义, 但不难看出存在过于简略的缺陷, 对此又通过搜索“状中结构”“定中关系”等搜索到主谓宾的修饰词, 有选择性地扩充到初步的三元组中, 最终得到的主谓宾三元组为{合同纠纷, 往往涉及, 多方金融关系}. 其中图 3 底部字母(nz 专有名词、d 副词、v 动词、n 名词、u 助词)是按照北大标准/中国科学院标准的汉语词性对照表进行标注的, 底部阿拉伯数字是词在句中的序号.

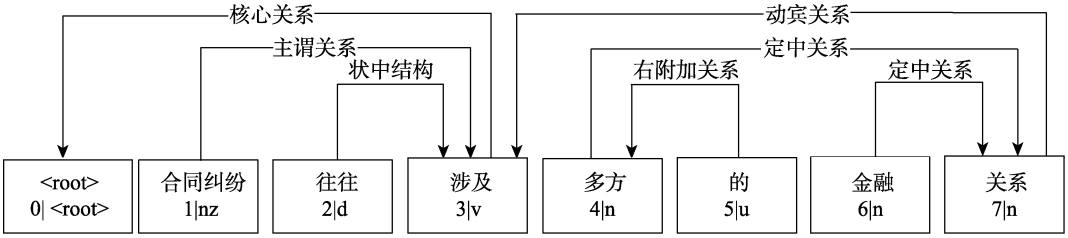



图 3 依存句法分析样例

对最终得到的三元组基于传统力引导算法进行布局, 显示如图 4 所示. 可以看到案件内容展示杂乱并无逻辑, 用户难以找到案件的分析入口. 因此, 为了增强阅读逻辑, 本文在段落时序逻辑上对三元组进行排布. 排布的重点是选好关键实体, 选择的标准是: 若某个实体在案件中的不同段落出现率高于所设阈值, 则将此实体作为案件内容的关键实体. 那么节点链的排布以关键实体为主体展开, 每列内的关键实体是按在原始案件文书内容的出现顺序排列的, 并把关键实体连接节点多的实体尽量放置到中间位置, 这样布局从美观度上也有很好的效果. 详细的实体布局计算过程如图 5 所示.

基于本文方法可视化出的文档图谱视图, 一列代表司法判决文书中的一段, 将文书从上到下的段落顺序用从左到右的方式可视化呈现出来, 以从左到右的段落时序逻辑顺序引导用户逐步了解整个案件的前因后果. 效果展示如图 6 所示. 其中, 灰色代表非关键实体, 其余颜色代表不同的关键实体. 通过分析大量的判决文书, 发现段落顺序即是逻辑顺序, 因此对每段均进行三元组的抽取, 并在视图中按段落顺序将三元组排布在不同区域中. 图 6 中显示有 3 列即为 3 段, 段与段之间用虚线分割, 节点链均被限定在固定区域内, 阅读顺序是箭头指向的顺序. 此外, 将不同段落的相同关键实体设置在一条线上(如图 6 中 1, 3 段中的“被告卞

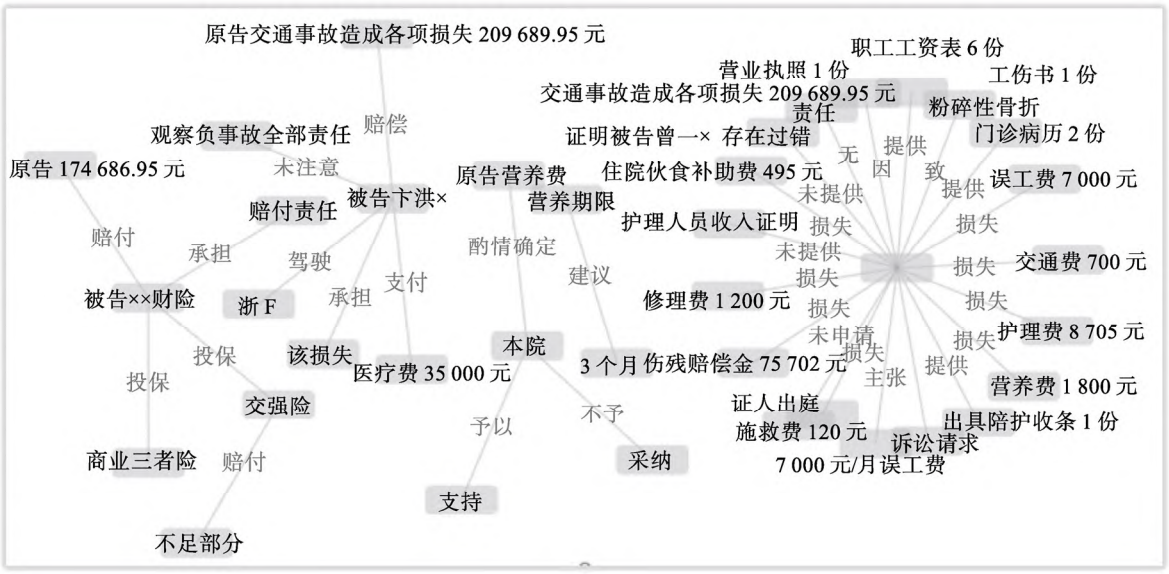



图 4　基于传统力引导算法的文档图谱视图

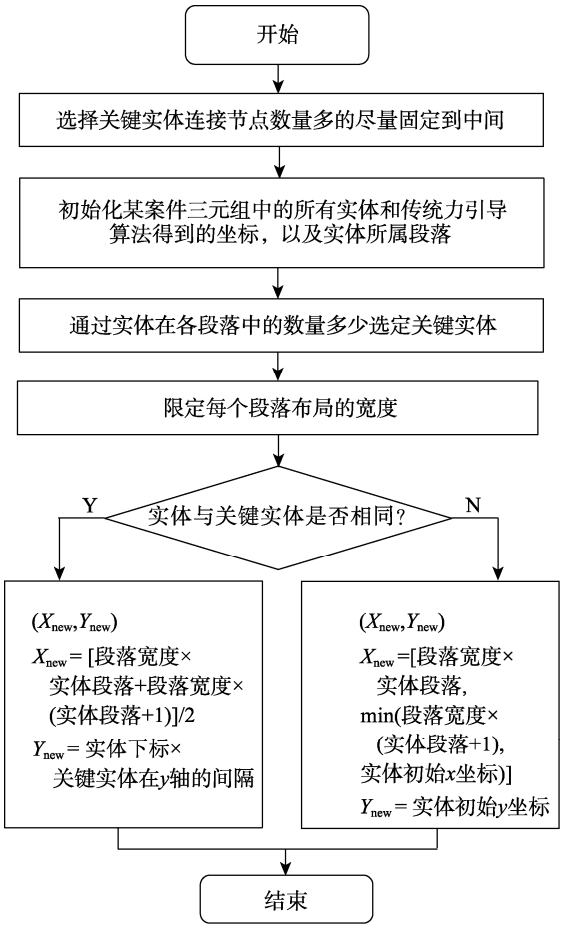



图 5　实体布局算法流程图

洪×”)，其余节点基于关键实体线分散排布．通过浏览关键实体和存在关系的实体可以快速掌握案件内容的逻辑重点，散落的非关键实体负责对案件内容进行额外的描述补充．

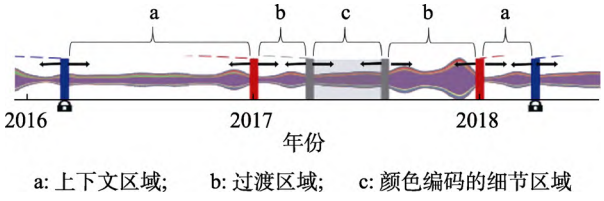
### 3.2　层次演化视图

司法判决文书数据是存在众多类别的时间序列数据，通过描述类别在一段时间内的变化观察案件纠纷类型的变化，发现热点纠纷类型或特殊时间节点．本文使用河流图的形式将文书数据的新旧案件串联起来，划分出不同类型的纠纷数据，沿着时间序列的方式，根据类型的层级进行时间流的聚合和分解．若将不同类型纠纷的数据可视化为无序堆叠的图形，会严重限制用户发现视图中有趣之处．MultiStream[8]是改进的堆叠图，它允许用户根据聚合和分解分层结构探索数据的演化过程．因此，本文依照上文数据处理部分的文书分类进行数量上的整合，采用 MultiStream 方法在时序上可视化呈现案件数量．在此基础上，为了帮助用户理解纠纷类型及发生案件的关键信息，还增加了词云视图．整个层次演化视图包含层次流图、时间控制器和词云 3 部分．

如图 7 所示，时间控制器是用于控制展示堆叠图中细粒度级别而非粗粒度级别的时间段，对应于时间轴上的灰色区域和 2 条垂直红线．基于时间刷，它允许移动或扩展时间控制器以更新堆叠区域，从而帮助用户更好地探索细粒度纠纷类型的演变信息．时间控制器显示的是 2014—2018 年的数据变化情况，它的设计是由灰色垂线、红色垂线和蓝色垂线各 2 条组成．

如图 8 所示，层级流图用于分层展示案件纠纷数量随时间的演变．其中，流的粗细表示在当前时间该类型纠纷的案件数量多少，密集的竖线用以


图 6 基于本文方法得到的文档图谱视图




a: 上下文区域; b: 过渡区域; c: 颜色编码的细节区域

图 7 时间控制器

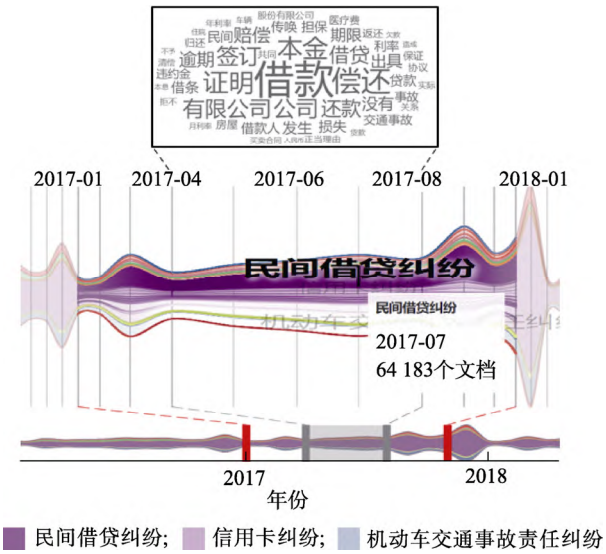



图 8 层次演化视图的交互展示

表示时间信息，根据所属类型进行了颜色编码(如民间借贷纠纷用紫色表示). 这有助于区分层次结构中的不同分支，保留层级演变并增强比较任务. 鼠标悬停位置相对应的纠纷类型颜色显示的饱和度会增强，其余部分饱和度降低，同时弹出一个文本框显示此时间和数量的信息. 要想了解此纠纷类型或对比前后发生案件的关键内容，用户可以通过对时间控制器的移动，在灰色区域选定时间段，双击流中某类纠纷，此类纠纷在该灰色时间段内关键字的重要性会显示出来，权重较大的关键词于中间位置.

### 3.3 矩阵视图

司法判决文书数据是存在时间序列的文本数据，适合从主题分布及语义信息层面挖掘演变，帮助用户直观地了解不同案件主题之间的关系. 使用矩阵视图挖掘分析案件文档主题的演化主要涉及 2 个方面: 一是用于度量在时序下各个主题集合的分布的主题强度; 二是通过时间轴的方式探索文档主题语义上的聚合程度的主题熵值.

(1) 主题强度是用于探索特定主题在特定时间内的被关注程度. 因此，计算出的数值越大，代表该主题的关注度越高，反之相反. 主题 $c$ 在时间窗口 $t$ 内的被关注程度为 $Q_c^t = D_c^t \Big/ \sum_{x=1}^{43} D_x^t$. 其中，$D_c^t$ 表示主题 $c$ 在时间窗口 $t$ 内的案件总数; 根据主题划分情况，将 $x$ 的取值设置为 1~43. 在得到 $Q_c^t$ 后，对其值进行归一化到 0~1，越接近 1 表示关注度越高. 由于案件时间分为立案时间和结案时间，因此对两者分别进行计算.

(2) 主题熵值是对 43 种主题类型纠纷在时序下混乱度的度量标准，其中 12 种因原数据集中数量极少，在此不对其进行计算分析，所以仅计算分

析原数据集中数量多的 31 种. 计算出的主题熵值越大，表示该主题文档内部越混乱；熵值越小，表示主题文档关联更紧密.

要计算主题熵值，首先需要对主题进行聚类投影. 采用 LDA 主题模型对 31 种主题类型进行内部主题划分. 尽管 LDA 是非监督的主题模型，但仍需根据案件内容人工确定需要划分的主题数. 因此，引入了 $P_{\mathrm{r}}$ 用于确定合适的主题数量. $P_{\mathrm{r}}$ 值越低，表明聚类效果越好. 计算公式为

$$P_{\mathrm{r}}(d)=\exp\left(-\left(\sum_{1}^{Z}\mathrm{lb}P(w)\right)\Big/\sum_{d=1}^{M}N_d\right).$$

其中，$Z$ 为当前主题个数；$M$ 为当前主题个数；$N$ 为数据集中的分词后的所有词(不去重)；$d$ 为数据集的每篇文档；$P(w)$ 为数据集中出现每个词的概率，计算公式为 $P(w)=\sum_{z=1}^{Z}\sum_{w=1}^{N}\sum_{d=1}^{M}P(z\mid d)P(w\mid z)$. 其中，$z$ 为训练过的主题；$P(z\mid d)$ 为一篇文档中每个主题出现的概率；$P(w\mid z)$ 为每个词对应某个主题出现的概率.

由 $P_{\mathrm{r}}$ 值确定主题数量，将数据处理中用于计算熵值采样出的数据，以 60 个月为基准时间单位，对立案时间进行划分；再通过 LDA 模型及 UMAP 降维，得到主题聚类图；最后进行混乱度的计算. 混乱度的计算公式为 $C_{\mathrm{f}}=\sum_{x=1}^{a}\sum_{y=1}^{b}P(x,y)H(x,y)$. 其中，将主题聚类图分割为 $a\times b$ 个小格($a$，$b$ 为自行确立的适宜常量)后；$P(x,y)$ 为第 $(x,y)$ 格包含点的数量；$H(x,y)=-\sum_{u=1}^{U}\sum_{v=1}^{V}\frac{P_{c(v)}}{P_{C(U)}}\mathrm{lb}\left(\frac{P_{c(v)}}{P_{C(u)}}\right)$. 其中，$C(u)$ 为数据集中某一大类；$U$ 为数据集汇总大类的总量；$c(v)$ 为 $C(u)$ 中某一小类；$V$ 为 $C(u)$ 中小类的总量；$P_{C(u)}$ 为第 $(x,y)$ 格中属于 $C(u)$ 类点的数量；$P_{c(v)}$ 为第 $(x,y)$ 格中属于 $c(v)$ 类点的数量. 对每格 $(x,y)$ 计算混乱度后再根据该区域的数据量对值进行加权并累加. 如果区域只包含一个类的数据，则值为零. 最后将计算得到的值进行归一化处理.

对于上述计算得到的熵值，采用矩阵热力图的可视方式，通过横向与纵向对比颜色插值的变化展示时序演变过程，以展示某个案件文档主题的混乱程度，横轴代表 31 类纠纷类型(仅部分标注)，纵轴代表 2014 年到 2018 年每年的 12 个月份，颜色越深，表示该文档主题语义上的聚合程度越差，通过颜色映射深浅可以直观地看出各个主题混乱度随月份的变化情况，展示图如图 9 所示.

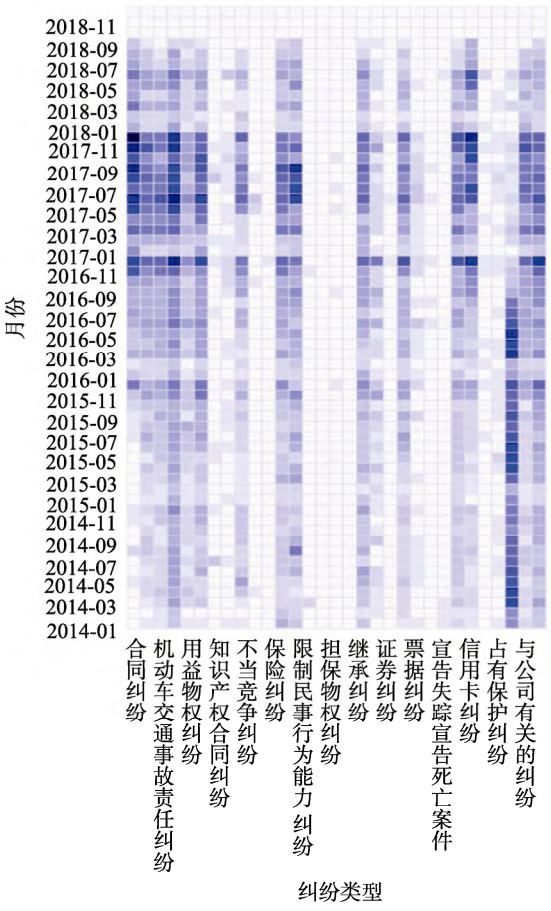



图 9　5 年每个月熵值热力图

为了更加形象地比较上述度量指标产生的演变情况，本文设计了一个可视化组件，揭示每个小方格所代表案件文档主题的关注度与聚合程度. 采用饼状图的方式，用不同的颜色来表示不同指标，绿色表示结案时间强度，黄色表示立案时间强度，红色表示熵值；使用圆环弧度的大小表示不同日期内用户对该主题的关注程度和该主题的聚合程度，如图 10 所示.

在图 2 中的 C 区域展示由饼状图+矩阵图组合成的多指标矩阵热力图. 其中，横轴代表 5×12 个月份，纵轴代表 31 类纠纷类型(由于其中 12 种计算出的主题强度极小，几乎为零，所以图中不再展示). 可以直观地看到主题在哪些时间窗口被关注度比较集中，在哪些时间窗口文档主题在语义上的混乱度明显提高，以及混乱度是否与文档主题的被关注程度存在一定的联系.

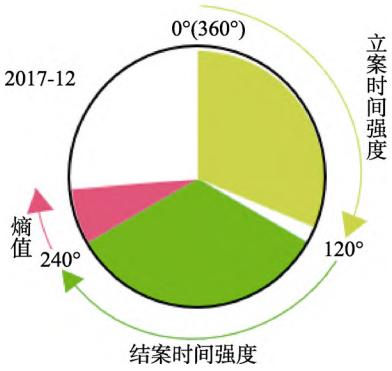
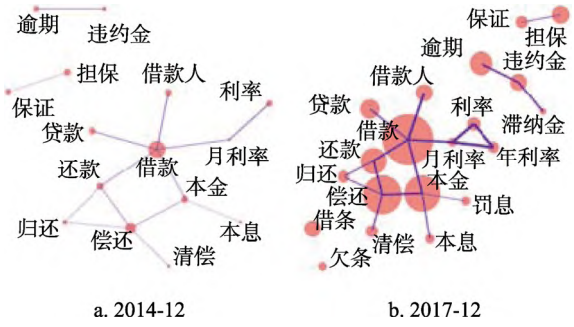



图 10 饼状图




a. 2014-12 b. 2017-12

图 11 不同年份 12 月词网络视图

### 3.4 词网络视图

为了有效地利用司法文书数据中各类型的特点，识别案件的发展过程，找出不同类型的关键词随时间变化关系，发现热点问题，本文使用关键词共现的方式设计词网络视图，包括词频分析和共词分析. 前者通过计算文本中的关键词频次，确定该文本数据中的关键实体，能够简单、直接地概括文本的核心内容；后者通过分析文本中关键词共同出现的情况，确定该文本中各关键词之间的关系. 本文采用这 2 种方法相结合并基于力导向布局的方式展示案件文档的主要关注内容，直观地了解该类型案件的热点问题及发展趋势.

词网络视图中每个圆代表一个关键实体，每个关键实体的确定是通过将某类型纠纷在当前时间窗口下、每篇文档用 Hanlp 取词频的前 20 个词后去掉单个字符，再依次加入词库取出词频位于前 50 的词作为视图中显示的关键实体. 词的共现是计算每篇文档 2 个不同关键实体是否共同出现，在视图中以连接线的方式体现.

为了防止出现视觉混乱，将上文数据处理部分生成的词向量模型，利用 Word2Vec 计算出词与词之间的语义相似度，只有相似度值大于阈值 0.65 的前提下才会建立连接，并且只保留存在有连接线的实体；然后对坐标点不断进行迭代，此处迭代 500 次. 图 11 展示的是不同年份 12 月合同纠纷的词网络视图，可以看出借款、本金和偿还等词的圆偏大，说明这些词的词频较高；借款、利率和月利率存在连接线且线条较粗，表明这 3 个词互相出现在某一篇文档的概率比较高.

## 4 案例分析

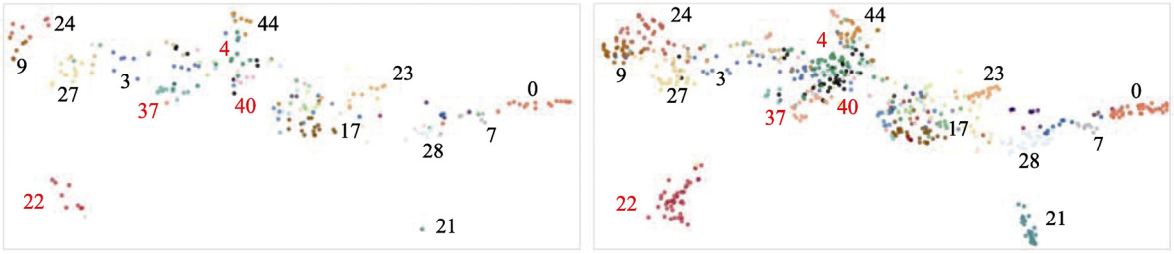
### 4.1 案件纠纷类型的演变分析

本案例通过系统展示热点纠纷在整个纠纷中的演变情况，根据规定将案件划分成 43 种类型后，观察分析某类纠纷在不同时序下的演变情况. 本案例选择对合同纠纷的 12 月份进行同比分析.

为了获取及分析合同纠纷随月份改变的主题信息，首先，用户点击图 2 中“矩阵图”按钮，此时，图 2 中的 C 区域切换为矩阵视图，横向比较小方格中的饼状图发现，合同纠纷的主题强度和熵值随月份呈上升态势.

然后，若要更细致地对比分析主题的演变规律，用户可以点击矩阵图中的小方格，在图 2 中的 F 区域弹出相应的主题聚类视图的同时，在图 2 中的 D 区域显示出词网络视图. 当用户同时选择多个小方格时，会同时弹出多个主题聚类视图(图 12a 和图 12b)以切片的可视化方式呈现出. 在图 12a 和图 12b 中用不同的颜色编码合同纠纷类型案件中包含的不同主题，可以看出主题 22 与主题 4、主题 37、主题 40 距离相对较远，说明主题 22 与另外 3 个主题之间的相关性较低. 另外，还发现主题 4、主题 37、主题 40 的散点数量在变多的同时，距离上也越来越近，这表明三者之间主题更近了. 结合表 1，通过图 12 中主题的颜色编码和标注序号，从表 1 中找到一致的主题序号对应的主题词，发现以上 3 个主题序号对应的主题词都是与民间借贷相关的，从侧面反映出民间借贷纠纷案件在不断增加的同时，案件之间的关联度也越来越高.

为了反映出合同纠纷当下较之前产生的热点问题，用户点击 2014-12 对应的合同纠纷小方格，图 2 中的 D 区域显示出词网络视图(图 11a)，用户再次点击 2014-12 对应的合同纠纷小方格时 D 区域切换为图 11b. 从图 11a 到图 11b，通过比较圆圈大小发现，借款、本金、偿还和贷款类关键词明显增多，说明此类案件数量较往年增长迅速. 从图 11b 中借款、利率、年利率和月利率之间的连线粗细可知，这些词同时出现在同一篇案件中的频率较高. 与此同时，也出现了借条、欠条、滞纳金以及年利




a. 2014-12　　b. 2017-12

图 12　不同年 12 月份合同纠纷的主题聚类图的效果对比

**表 1　合同纠纷部分主题划分**

| 主题序号 | 主题词 |
|---|---|
| 0 | 商品房 房地产 办理 开发 交付 登记 买卖合同 房屋 |
| 3 | 借款 偿还 本金 月利率 借据 本息 农村信用社 期限 |
| 4 | 借条 出具 借贷 偿还 民间 关系 现金 月息 |
| 9 | 抵押 担保 债权 财产 优先 登记 融资 拍卖 |
| 17 | 货款 付款 买卖合同 价款 购买 货物 产品 发票 |
| 21 | 服务 物业 业主 物业管理 小区 交纳 服务费 物业费 |
| 22 | 信用卡 透支 滞纳金 还款 本金 偿还 合约 欠款 |
| 23 | 工程 施工 工程款 结算 项目 建设工程 价款 建设 |
| 24 | 贷款 罚息 利率 本金 分行 复利 逾期 律师费 |
| 27 | 支行 贷款 股份有限公司 农村 签订 商业银行 本息 偿还 |
| 28 | 租赁 有效 租金 签订 使用 内容 解除 承租人 |
| 37 | 利率 借贷 年利率 逾期 民间 偿还 适用法律 超过 |
| 40 | 借款 本金 归还 还款 期限 借款人 到期 告谢 |
| 44 | 共同 夫妻 个人 夫妻关系 问题 解释 存续 证明 |

率等词，然而这些词大多出现在合同纠纷中的民间借贷纠纷类型中，说明民间借贷纠纷面临的是与图 11a 中关键词密切相关的问题，如民间借贷的利率问题和借条是否有效等.

另外，结合地理位置分析时序演变，对裁决文书中的序列号按省份简称进行匹配，若未匹配则根据案件内容句首包含的省份再次匹配，可得到案件所属省份. 统计并比较得到各时间段出现的案件数量最多的省份，分别有浙江省、广东省、江苏省、山东省和河南省，绘制出地域折线图(图 13)，横轴代表月份，纵轴代表案件总量. 用户点击图 2 中 F 区域的“折线图”按钮时，系统会将 F 区域视图切换为地域折线图. 从图 13 左图中发现在 2014-12 月份内，案件数量最多的是江苏省(★)；从图 13 右图中发现在 2017-12 月份内，案件数量最多的是广东省(▲). 前 4 个省都是沿海省份，经济发展水平高，推测我国的案件总体数量受经济发展水平差异的影响；河南省是人口大省，说明这也可能与人口数量相关.

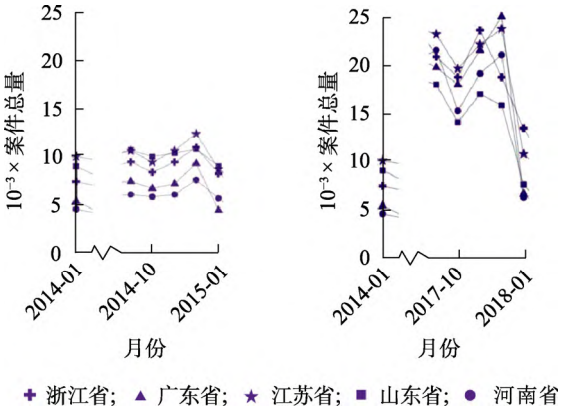



图 13　不同月份下地域折线图的效果对比

通过以上海量数据的分析，可以为今后法院对社会热点问题和案件提供较高的数据参考价值和决策支撑依据.

### 4.2　案件数据关系的展现

本案例将通过关键词搜索的方式实现个性化的类案推送，并展示与之相匹配的司法判决文书的数据内容. 如图 2 所示，首先，用户在系统界面 A 区域的输入框中输入“工伤 公司 车祸”多个关键词；然后，用户点击系统中的“搜索”按钮，此时，系统在图 2 中的 B 区域展示出了与工伤、公司、车祸关联度最高的 12 例案件，案件的顺序由 Lucene 数据库自带的倒排索引查询顺序的相关度决定，用户可以看到文档列表中的“案件 1 2015-04-09”排在第 1，这表明案件 1 对应的司法判决文书在出现工伤、公司、车祸时的相关度最高；最后，用户点击列表中各案件序号，如“案件 1 2015- 04-09”，此时，系统中以文档图谱视图的方式在图 2 中的 C 区域展示出与此案件对应的司法判决文书的详情内容(图 6).

由图 6 第 1 段中可知：本案件是由被告卞洪×驾驶的浙 F 车在未注意观察周围时造成原告××良粉碎性骨折，因此认定被告负此次事故全部责任，原告无责任. 图 6 第 2 和 3 段展示了具体赔付情况，可以看到原告××良在此次事故中提供的诸多费

用, 损失较多, 但同时原告未提供护理人员的收入证明; 也可以得知被告卞洪×需要赔付的费用款项, 以及被告卞洪×投保的车险公司××财险需要赔付的费用款项. 与案情相关的重要信息均可在可视化展现中快速、准确地获取.

## 5 用户实验

### 5.1 问 卷

为了评估本文设计的可视化系统, 邀请了 15 位计算机相关专业的研究生完成一份有 9 个问题的问卷, 针对本文方法对可视分析系统设计的实用性、有效性进行评估.

在每项任务开始之前, 先为每位参与者演示本文系统的每处细节, 之后针对主要任务要求参与者完成一系列任务, 任务罗列以问卷调查的形式, 如需要, 也会进一步对参与者讲解. 当所有参与者的任务完成后, 让参与者在李克特量表上对任务进行评级打分, 打分的标准是1~5分表示从非常不同意到非常同意.

从图 14 中的评级结果来看, 参与者普遍认为通过文档图谱视图能够快速定位到司法文书中的关键实体, 而实体关联动作事件的方式也对掌握主体内容有帮助(平均 4 分). 参与者认为, 层次演化视图通过流的宽度变化较好地展现了司法纠纷类型的变化, 一些特殊的变化如短时间内突然变宽等足以引起深入探索的兴趣, 此外, 词云的辅助展示能帮助用户掌握到某一类的主要内容(平均 4 分). 但在对高粒度到细粒度的分解上认可度不高, 这可能是由于颜色渐变不够显著, 需要花一定的时间进行理解(平均 3.5 分). 对于矩阵视图中展示的主题信息, 通过纵向与横向的饼状图细节信息对比, 可以帮助比较主题关注度在不同类型不同月份的演化规律(平均 4 分), 而主题聚类视图在一定程度上能帮助区分不同主题的分布, 但当主题数较多时使用了过多颜色, 对于发现某些主题之间的变化需要花费一定的时间代价(平均 3.5 分). 总体而言, 所有问题的平均评分都大于 3.5, 证明了可视分析方法的有效性.

### 5.2 用户反馈

本文设计的可视化系统面向的目标用户是普通用户和司法人员, 所以另外邀请了6名法学院研究生进行用户反馈, 结合计算机相关专业和其他领域法学相关专业的研究生对本文系统的反馈总结如下.

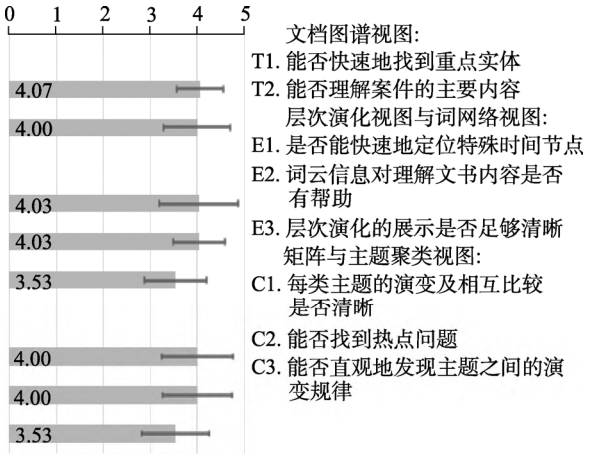



图 14 用户的任务设计和评分分析

在可视化设计方面, 所有用户一致认为该系统很直观, 对交互式印象深刻, 交互流畅, 对分析和探索司法数据很有帮助.

在可用性方面, 所有用户均认为本文系统对于司法文书的演化探索具有一定的实用性和有效性, 并且法学相关专业的用户希望在研究中使用该系统, 指出这个过程从宏观到微观的方式探索模式是合理的, 对司法数据分析有用, 并且易于交互, 方便掌握.

在建议方面, 对于纠纷种类的层次分解上的揭示, 由于细粒度层的颜色区分不够显著需要耗费一定的时间理解. 建议揭示更多的信息, 分析更多的内容, 如案件运用到的相关法律及最终诉讼的胜负. 这说明本文系统还有需要改进的地方.

## 6 结 语

基于司法判决文书数据, 本文设计了一套可视分析流程, 用于交互式可视化挖掘和分析, 以帮助挖掘案件的时序演化模式. 在可视化分析视图中, 针对案件内容抽取出三元组形式, 基于关键实体约束的布局算法与可视分析相结合, 实现段落在时序逻辑上的文档图谱视图, 更加有利于快速理清案情的数据关系. 通过层次演化视图提供分层展示不同数据类型的层级关系, 并利用时间控制器控制时间段对词云进行交互探索. 矩阵视图与词网络视图及主体聚类视图的交互, 更好地帮助用户从整体到局部挖掘演变. 通过分析用户需求, 构建了一个面向司法领域的可视分析系统, 允许用户从宏观到微观进行探索, 使其能够基于大量真实数据分析演变, 发现社会热点问题, 也可以推广到其他领域.

然而，本文工作也有不足之处．首先，在设计文档图谱视图中没有考虑案件本身在时间上的分布，不能很好地从时间上构建实体关系的关联；另外，不能满足多个类型案件之间的词关联度可视分析．因此，在后续的工作中，将针对单篇司法文书研究对案件过程进行时间上的梳理与排序，以及进行多类型案件之间的联合探索，便于结合新的可视化设计解决可扩展性的问题．当然也更希望其他学者能够从本文工作中得到启发，参与到可视化技术的工作中来，对司法领域的数据进行更深入、更具价值的探索．

## 参考文献(References):


[1] Jiang Huiling. What can judicial big data bring to us[J]. People's Tribune, 2017(36): 74-76(in Chinese)
(蒋惠岭. 司法大数据能为我们带来什么[J]. 人民论坛, 2017(36): 74-76)

[2] Liu Pinxin. Academic observation of big data judicature[J]. People's Procuratorial Semimonthly, 2017(23): 29-31(in Chinese)
(刘品新. 大数据司法的学术观察[J]. 人民检察, 2017(23): 29-31)

[3] Tang Weijian. New breakthrough, new prolongation and new finding: summary on civil procedural law research in 2018[J]. People's Procuratorial Semimonthly, 2019(3): 29-37(in Chinese)
(汤维建. 新突破 新拓展 新发现: 2018 年民事诉讼法学研究综述[J]. 人民检察, 2019(3): 29-37)

[4] Sultanum N, Singh D, Brudno M, *et al*. Doccurate: a curation-based approach for clinical text visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 142-151

[5] Li Z Y, Zhang C H, Jia S C, *et al*. Galex: exploring the evolution and intersection of disciplines[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1182-1192

[6] Boyandin I, Bertini E, Bak P, *et al*. Flowstrates: an approach for visual exploration of temporal origin-destination data[J]. Computer Graphics Forum, 2011, 30(3): 971-980

[7] Heimerl F, Han Q, Koch S, *et al*. CiteRivers: visual analytics of citation patterns[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 190-199

[8] Cuenca E, Sallaberry A, Wang F Y, *et al*. MultiStream: a multiresolution streamgraph approach to explore hierarchical time series[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(12): 3160-3173

[9] Yang W K, Wang X T, Lu J, *et al*. Interactive steering of hierarchical clustering[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(10): 3953-3967

[10] Isenberg P, Isenberg T, Sedlmair M, *et al*. Visualization as seen through its research paper keywords[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 771-780

[11] Kim H, Drake B, Endert A, *et al*. ArchiText: interactive hierarchical topic modeling[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(9): 3644-3655

[12] Tang Ying, Su Jianming, Tong Ning. TMvis: a visual analysis system for topic modeling based on LDA Topic Modelling[J]. Journal of Computer Aided Design & Computer Graphics, 2019, 31(10): 1728-1738(in Chinese)
(汤颖, 苏建明, 童宁. TMvis: 基于 LDA 的主题建模可视分析系统[J]. 计算机辅助设计与图形学学报, 2019, 31(10): 1728-1738)

[13] El-Assady M, Kehlbeck R, Collins C, *et al*. Semantic concept spaces: guided topic model refinement using word-embedding projections[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1001-1011

[14] Wang Y, Yu M Z, Shan G H, *et al*. VISPubComPAS: a comparative analytical system for visualization publication data[J]. Journal of Visualization, 2019, 22(5): 941-953

[15] Knittel J, Koch S, Ertl T. PyramidTags: context-, time-and word order-aware tag maps to explore large document collections[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(12): 4455-4468

[16] Heimerl F, Kralj C, Moller T, *et al*. embComp: visual interactive comparison of vector embeddings[J]. IEEE Transactions on Visualization and Computer Graphics, 2020: 1-16

[17] Cho I, Dou W, Wang D X, *et al*. VAiRoma: a visual analytics system for making sense of places, times, and events in roman history[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 210-219

[18] Zhang Wei, Tan Siwei, Liu Kai, *et al*. A new perspective on the study of literature(Songci): aext correlation and spatio-temporal visual analytics [J]. Journal of Computer Aided Design & Computer Graphics, 2019, 31(10): 1687-1697(in Chinese)
(张玮, 谭思危, 刘凯, 等. 宋词研究的新视角: 文本关联与时空可视分析[J]. 计算机辅助设计与图形学学报, 2019, 31(10): 1687-1697)

[19] Carvalho N R, Barbosa L S. Transforming legal documents for visualization and analysis[C] //Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance. New York: ACM Press, 2018: 23-26

[20] Xie Xiaoping, Pan Zhenjie, Wang Kang, *et al*. Analysis of the publication status of Chinese Journal of Forensic Sciences from 2001 to 2018 based on bibliometric indicators[J]. Chinese Journal of Forensic Sciences, 2019(4): 22-28(in Chinese)
(谢小平, 潘桢杰, 王康, 等. 基于文献计量学指标的《中国司法鉴定》2001—2018 年发文状况分析[J]. 中国司法鉴定, 2019(4): 22-28)

[21] Mihalcea R, Tarau P. TextRank: bringing order into text[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona: Association for Computational Linguistics, 2004: 404-411

[22] Chen H C, Sun X F, Tian Y T, *et al*. Enhanced network embeddings via exploiting edge labels[C] //Proceedings of the ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2018: 1579-1582

[23] Ghiasnezhad O P, Wang K W, Wang Z. Scalable rule learning via learning representation[C] //Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: International Joint Conferences on Artificial Intelligence Organization, 2018: 2149-2155