大规模接警日志语义时空模式交互探索

贺一桐¹⁾, 李杰^{1)*}, 陶文源¹⁾, 高星¹⁾, 张康²⁾

¹⁾ (天津大学智能与计算学部软件学院 天津 300350)

²⁾ (Department of Computer Science, Texas University Dallas 75080)
(jie.li@tju.edu.cn)

摘要: 接警日志包含时间、空间和案件描述信息,属于非结构时空数据。与时空社交媒体相比,接警日志的数据项之间存在较少的联系,数据项之间不能形成复杂网络关系,在挖掘其数据模式时难以提供有价值的线索,因此,其分析更加依赖于其中的语义挖掘和语义时空模式探索。针对这一问题,提出了一个可视分析框架支持对大规模非结构接警日志时空模式的交互探索。首先,提出了一种基于主题模型集成的方法,实现从异构文本中抽取主题;其次,该框架包含一个数据立方体,实现快速响应用户的查询请求;第三,设计并实现了一个可视化交互系统,支持对数据立方体的可视化交互探索。最后,使用国内某城市真实接警日志进行实验,找到的丰富的模式和主题预测准确性证明了方法的有效性。

关键词: 接警日志; 语义时空模式; 可视分析; 主题模型集成; 数据立方体

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18813

Interactive Exploration of Semantic-Spatial-Temporal Patterns of Large-Scale Alarm Police Logs

He Yitong¹⁾, Li Jie^{1)*}, Tao Wenyuan¹⁾, Gao Xing¹⁾, and Zhang Kang²⁾

¹⁾ (School of Computer Software, Department of Intelligence and Computing, Tianjin University, Tianjin 300350)

²⁾ (Department of Computer Science, Texas University, Dallas 75080)

Abstract: The alarm log contains time, space, and case description information, which is unstructured spatiotemporal data. Compared with spatio-temporal social media, there are fewer connections between the data items in the alarm log, and complex network relationships cannot be formed between the data items. It is difficult to provide valuable clues in data patterns. Therefore, its analysis depends more on the semantic mining and the exploration of semantic spatiotemporal patterns. In response to this problem, a visual analysis framework is proposed to support large-scale unstructured alarm log spatiotemporal patterns interactive exploration. First, a method based on topic model integration to extract topics from heterogeneous texts is proposed; Second, the framework includes a data cube to quickly respond to user queries; Third, a visual interactive system is designed and implemented to support the visual interactive exploration of the data cube. Finally, we use a real alarm log of a city in China to evaluate the approach. The rich detected patterns and high topic prediction accuracy prove the effectiveness of the approach.

收稿日期: 2020-11-25; 修回日期: 2021-01-04. 基金项目: 国家重点研发计划(2018YFC0831700); 国家自然科学基金(61972278, 61602340). 贺一桐(1994—), 男, 硕士研究生, 主要研究方向为数据可视化; 李杰(1984—), 男, 博士, 副教授, 研究生导师, CCF 会员, 论文通讯作者, 主要研究方向为机器学习、大数据可视分析; 陶文源(1971—), 男, 博士, 教授, 博士生导师, 主要研究方向为数字媒体技术、物联网技术与应用; 高星(1987—), 男, 硕士研究生, 主要研究方向为数据可视化; 张康(1959—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算美学、信息可视化。

Key words: alarm log; semantic-space-time mode; visual analysis; topic model integration; data cube

接警日志,即人们所熟知的110报警记录,是一类常见的非结构时空数据。报案时间、事发地点与报案内容描述是接警日志中的核心的信息。接警日志呈现出一种时序线性结构,相比于通过“关注”和“转发”形成复杂动态网络的时空社交媒体^[1],其结构更加简单,并没有为数据分析提供更多有价值的线索,使分析接警日志必须依赖挖掘案件描述中包含的丰富的语义信息,促使语义时空模式探索成为接警日志分析^[2]的必要手段。

探索语义时空模式可以有效提高接警日志的分析效果,主要体现在3个方面。(1) 增加分析结果的准确性。例如,在分析某类案件的时空分布时,分析人员通常会选择某个报案类型的接警日志,分析其在空间和时间上的分布情况。一般情况下,一个案件只会属于一个案件类型,如果存在一个日志从属于多个案件类型的情况,如驾驶被盗车辆发生交通事故,则很难自动地获取其对应的全部类型,从而影响分析的准确性。(2) 提高分析内容的丰富性。语义挖掘是从自然语言表达的案件描述中自动地挖掘其中可能存在的各种语义,并不受案件类型等人为数据划分的影响,根据后期实验的结果,接警日志中存在大量的体现更加丰富场景的主题,如“黑窝点”等,这是传统基于数据属性分析无法得到的。(3) 探索接警日志的语义时空模式可以提高分析预测的时效性。预测某类案件的未来趋势是接警日志分析的重要需求,将大量文本转换为具体的单数值类型的语义流行度,可以避免文本的高维稀疏性影响预测准确度,并且由于语义流行度是一个数值,可以有效地提高预测速度,使其满足实时交互的需要。

实现交互探索海量非结构接警日志也存在较大挑战。首先,常见的主题模型对参数设置非常敏感^[3],轻微的参数变化可能产生完全不同的结果,在没有先验知识的条件下,很难合理地设置主题模型参数。其次,主题时空模式包含大量的交互探索任务^[4]。例如,用户不仅可能感兴趣主题的时间趋势或空间变化,也可能对比分析在某个时空坐标下具体的主题内容,需要构建一致的任务组织方式以更好地支持不同的分析应用场景。

为了解决以上挑战,本文提出了一个可视分析框架,以交互探索大规模接警日志的语义时空模式。首先,框架采用了一个基于模型集成的主题抽取方法,通过将不同参数的主题模型抽取的结

果投影在平面上,用户可以直观地了解不同主题的差异,并从中选择感兴趣的主体,从而解决无法准确设定主题参数的问题。其次,框架集成了一个基于DataCube^[5]的数据和任务组织结构,通过预先存储主题在不同时间和空间坐标下的2类指标,以及将不同的分析任务对应为数据体内的各类投影和切片操作,实现对各类交互探索任务的快速响应。

1 相关工作

1.1 文本可视化

在处理文本数据在空间和时间分布的可视分析^[6]中,工作重点是找到文本数据中的关键词。常见的方法是分析文本数据检测在特定时间和地点发生的相关联的事件,从具有相同地点和相同或重叠时间的文本组中识别事件^[7]。Marcus等^[8]通过推特数据中的峰值检测热门的事件,并使用推文中的关键词对其进行有语义的标记。Zhou等^[9]使用机器学习中检测突发词技术识别事件。Chae等^[10]使用季节性趋势分解识别异常事件。在提取事件后,使用地图和时间轴或基于日历的显示来可视地探索事件与空间和时间的关系,使用词云表示与事件相关的文本^[11]。ScatterBlogs系统^[12]根据在附近地点和时间发布的消息中多次出现的具有相同关键词的事件进行可视检测,将这些关键词按照使用与出现次数成比例的字体大小显示在出现位置的空间地图。ScatterBlogs2^[13]通过添加用于文本过滤的复杂工具扩展ScatterBlogs,以便将可视化分析应用于先前选择的可能相关的文本。Chae等^[14]提出了根据发布的推文位置和时间分析推特用户的时空分布的工具,分析师可以在检测到的推特用户的时空群集中发布的推文上进行主题建模,并查看由词云表示的主题内容。然而,这些方法只是在关键词的层面对文本内容进行展示,并未涉及文本语义分析。

随着主题模型的广泛应用,近年来在对文本数据可视化时会应用主题模型。Xu等^[15]提出了一个主题竞争模型,以表示公众对多个主题的关注。Sun等^[16]的工作对该模型进行了扩展,用于分析社交媒体中不同主题之间的竞争和合作关系。Wang等^[17]提出了一种可视化分析系统,用于分析不同社会群体之间的主题传播。Dou等^[18]的工作支持探

索主题之间的等级关系. Cui 等^[19]的工作侧重于不同时期主题之间的动态等级关系. 上述工作主要使用社交媒体的文本数据, 然而它们是预先将主题模型参数设定好, 使用训练好的主题模型对文本数据提取语义. 由于主题模型对参数很敏感, 这样做无法保证主题模型的质量, 因此会影响模型提取语义的能力.

本文对接警日志中的文本数据使用主题模型集成, 将不同参数生成的模型集成在一起, 可以减少参数对模型质量的影响; 同时设计可视化视图将集成的主题模型投影供用户交互选择, 使用用户选择的主题提取出的主题分布值存入数据立方体, 将具有语义信息的主值按照地区和时间聚类, 对查询结果进行可视探索分析.

1.2 数据立方体

数据立方体^[5]是用于支持在线分析处理(online analytical processing, OLAP)查询的多维数据的模型. 在构建数据立方体时, 需要选择一些数据字段作为维度, 在单元格的内存预先计算好的结果, 如总和以及平均值等, 通过输入查询条件, 返回数据立方体对应维度上单元格的信息. 数据立方体模型近年来已被使用在不同领域的数据集, 例如, 对社交媒体数据^[20]和交通数据^[21]的分析. 最近提出的几种方法, 如 Lins 等^[22]提出的用于实时探索时空数据集的纳米数据立方体; Pahins 等^[23]通过构建哈希数据立方体更简单高效地实现实时可视化探索大数据; Wang 等^[24]提出了对大型多维数据集实时可视化探索的高斯立方体; Miranda 等^[25]提出对长时间序列交互可视分析的数据结构 Time Lattice, 这些工作专注于使用立方体结构实现对交互操作的实时响应, 但上述数据立方体结构是面向结构化数据, 无法用于交互探索非结构时空数据.

近年来也有工作开展了面向非结构数据的数据体的研究. Text Cube^[26]旨在对具有文本字段的多维数据集提供 OLAP 查询. Zhang 等^[27]通过文本分析获得主题层次构建主题立方体, 主题是立方体维度之一, 数据立方体单元格包含 2 种度量: 主题的单词分布和文档的主题覆盖. 但这些工作专注于面向文本内的 OLAP 的支持, 本文的工作重点是在数据立方体中加入语义, 使用主题模型抽取警务数据中的文本的语义信息, 使用包含语义信息的主值存储在数据立方体中的单元格, 将警务数据中的时间、空间和语义属性作为立方体的 3 个维度, 使用数据立方体的形式满足用户探索这 3 个维度之间的联系, 提高查询速度, 实现文本语

义、空间和时间之间关系的实时可视化探索.

1.3 安全可视分析

安全问题已成为每个国家的重点问题, 引起了研究人员和从业人员的关注. 城市安全是应用可视化分析技术的重要领域, Cao 等^[28]将数据在时空立方体模型转换为张量时间序列, 并提出分析算法将数据的当前状态与历史状态进行比较以检测异常情况, 使用可视化方式显示数据, 帮助用户判断和探索异常情况. Malik 等^[29]的工作重点是探索地理空间和时间尺度问题的可视化, 以便领域专家可以将他们的决策过程调整到准确合适的时空分析空间, 并利用可视化方式分析报警数据进行城市治安分析. Andrienko 等^[30]使用可视化系统分析群体在空间上的移动轨迹. Ahmed 等^[31]通过可视化方式探索比特币被盗的扩散过程. 然而, 这些工作只停留在对事件的轨迹或时空分布可视探索, 没有对数据中的文本数据进行语义分析, 这些工作的可视化过程中只注重对时空分布的探索, 没有将事件中语义信息和时空角度结合. 本文使用主题模型提取警务数据中的语义信息, 并设计了可视化系统将警务数据中的语义信息与时间和空间属性共同展示, 实现对警务数据在语义时空层面上的探索, 便于用户发现数据中存在的安全问题.

2 问题定义

2.1 数据描述

接警日志数据包括时间、地点、内容接警类型等信息. 其中时间和地点信息可以反映报警数据在时间和空间的分布情况; 接警内容是对报警人讲话的记录, 可以有效地描述发生的时间, 通过对内容的挖掘可以对案件进行语义角度上的处理; 接警类型是对该条记录分类标注.

接警日志数据所使用字段如表 1 所示. 报警时间字段是记录警局接到报警电话的时间, 时间最小刻度为秒. 该样本数据中报警时间范围是 2017-07-25—2017-12-08, 持续 137 天. 报警地点字段是记录报警电话打出的位置, 数据中的地点以文字方式记录, 由于文字描述属于非结构化数据, 不利于分析处理, 因此, 通过百度经纬度转换应用程序接口(application programming interface, API)将文字描述的地点转变成经纬度, 然后通过经纬坐标对案件进行空间角度的聚类. 报警内容是对每个记录在语义角度上的描述, 通过该字段可以对记录进行类型划分. 现在的警用数据分析

方法大多使用报警类型字段作为数据的分类依据,由表1可以看出,报警类型字段只是记录最相关的类型,不能详细和准确地概括该记录的内容,所以使用该字段进行分类会对后续工作造成误差.本文使用主题模型对报警内容中的信息进行语义提取,根据主题分布值对数据分类.这样可以减少语义信息的丢失,提高分类的准确度,并丰富案件分类的类型.

表1 接警日志数据所使用字段

报警时间	报警地点	报警内容
2017-07-25 T22:34:00	A市某区×镇× 小区×号楼×号	报警人A.男,60岁,在其现住地用手机从网上下载的2个理财软件进行投资理财,发现其通过投资理财被诈骗2万5千元
2017-07-26 T01:43:11	C市某区×小 区×-×-×	B被网友骗到C市某区×庄村荒地的一个猪窝处的传销组织,期间传销组织内的人员限制B的人身自由,不让他离开
...	...	...
2017-09-26 T01:43:11	D市某区×市 场×门面内	我所民警工作中发现在D市某区×市场内一个门面有一名男子涉嫌贩卖濒危物种陆龟,后民警将该男子口头传唤至公安机关接受调查

2.2 主题抽取

本文使用主题模型(latent Dirichlet allocation, LDA)提取接警日志中文本信息的语义, LDA对参数设置非常敏感,轻微的参数变化会产生差异很大的结果,而训练模型时需要定义大量参数.表2展示了主题模型中不同参数条件下生成的主题模

表2 不同参数的主题模型

主题模型1(5个主题数)		主题模型2(6个主题数)	
编号	主题词	编号	主题词
1.1	单号,无人,讲话,骚扰,电话,接到,反打,挂断,警员,骚扰电话	2.1	应答,男子,停车,人员,疑似,醉酒,女子,言语,配合,办理
1.2	应答,派出所,电话,违章,黑窝点,赌博,男子,举报,保密,跑	2.2	骚扰,车辆,派出所,货车,车主,一辆,客人,物业,家庭,门口
1.3	妥善处置,工作,应答,敏感,男子,打架,依法,手机,停车,黑窝点	2.3	电话,出警,黑窝点,诈骗,赌博,举报,卖淫,斗殴,怀疑,出租车
1.4	汽车,被盗,车辆,撞,停放,电动车,停车,砸,交通事故,受伤	2.4	纠纷,咨询,堵塞,电话,现场,异地,交通,走,疏导,速去
1.5	求助,拨打,挪车,交通,堵塞,催促,堵塞,到场,道路,告知,人员	2.5	无人,讲话,电话,接到,两小客,错,反打,手机,警员,位置
		2.6	处置,派警,骚扰,现场,速到,扰民,告知,打架,工作,敏感

型的内容,其中的主题模型虽然使用相同的语料库训练,但仅是主题个数参数的变化便会导致生成模型的内容截然不同,所以在框架中需要解决参数设定对主题模型的影响.

在同一主题模型中会同时存在质量好与坏的主题,有的主题内容丰富,便于用户理解主题表达的内容(主题模型1.4, 1.5, 2.3, 2.4);有的主题内容比较混乱,无法清晰分辨主题的含义(主题模型1.2, 2.1, 2.2).所以需要在主题集成中加入对主题质量的评估,同时将主题的内容进行可视化展示,便于用户理解主题内容并从中选择需要的主题.

2.3 任务分析

通过与领域专家的沟通,本文详细了解领域专家在对警务数据时空探索分析时的需求,以及在分析数据时遇到的困难.根据领域专家的描述,对本系统进行了任务分析.

接警日志在时空语义层面上一共涉及3个维度,分别为时间 t 、空间 s 和主题 p ,通过对这3个维度设定条件,得到查询结果-主题值和关键词权重 M ,探索主题值在这3个维度上的变化情况,从而对接警数据进行语义时空模式的分析.然而,用户同时查看3个维度进行探索和做出判断是困难的,因此,可以将复杂的查询任务分解成简单的子任务.

任务1. $t/\sum t \rightarrow (s \rightarrow p):M$,即先确定时刻或时间范围,查看在指定时间条件内不同地区中哪个主题更加显著.

任务2. $t/\sum t \rightarrow (p \rightarrow s):M$,即先确定时刻或时间范围,查看在指定时间条件内不同主题在地区上的分布情况.

任务3. $s/\sum s \rightarrow (t \rightarrow p):M$,即先确定某地区或地区范围,查看在指定地区条件内主题热度随时间的变化情况.

任务4. $s/\sum s \rightarrow (p \rightarrow t):M$,即先确定某地区或地区范围,查看在指定地区条件内不同主题在何时热度高.

任务5. $p/\sum p \rightarrow (s \rightarrow t):M$,即先确定某主题或主题范围,查看在指定主题条件内不同地区该主题随时间的变化情况.

任务6. $p/\sum p \rightarrow (t \rightarrow s):M$,即先确定某主题或主题范围,查看在指定主题条件内某个时间内该主题的地区分布情况.

上述子任务中包含某一刻度和范围区间2种情况,同时查询结果也有主题值和关键词2种情

况,即共有 24 个子任务(6 个子任务 $\times$ 2 种查询类型 $\times$ 2 类返回结果)。如果为每个任务设置独立的接口会增加用户的操作和理解难度,则为了在框架中更好地组织和实现上述的子任务,便于用户对数据的探索和理解,需要为上述的所有任务设计一致的接口,将不同维度的查询条件转化为一致格式的输入,返回对应查询结果,提高框架的一致性和易用性。

2.4 系统目标

本文的目标是设计和构建一个对接警日志语义时空模式交互探索的通用可视化框架。该框架可以解决主题模型难以确定合理参数的问题,将用户的经验纳入对数据分析的过程。同时,基于上述对接警数据多维度查询任务,该框架需要设计一种数据结构,对上述子任务提供一致接口,降低用户探索时的操作难度。最后,为了构建对接警日志完整的分析流程,该框架还应有一个可视化交互界面。

2.5 系统构成

根据上面所描述的任务和目标,本文将系统划分成 3 个模块,如图 1 所示,分别是主题模型集成、主题时空体和可视交互与预测模块。

主题模型集成模块对接警数据中的文本数据训练主题模型并集成。具体操作包括主题生成、主题质量评估和主题降维投影。通过迭代不同参数生成多个主题模型,对其进行质量评估后选择优质主题进行集成,将集成的所有主题降维投影在二维平面,使用词云展示主题中的内容。

主题时空体模块将接警数据中的时间、空间和文本数据转换成立方体数据结构。首先设计立方体数据结构组织和支持上述的查询任务,再将离散化后的时间空间和主题作为数据立方体的 3 个维度,在单元格的存储语义信息,从而满足对接警

图 1 本文框架结构

日志在语义时空层面上的探索。

可视交互与预测模块对文本数据立方体提供可视化交互探索。该可视化系统包括主题投影、时空投影、模式对比 3 个视图。主题投影视图用于交互选取用户感兴趣的主体。时空投影视图包含数据投影和热门地区排名 2 个子视图,数据投影可以根据用户选择的时间、空间、主题查询条件将查询结果投影展示在地图上。模式对比视图便于用户对比发现不同查询条件下案件的发生模式。本文还提供一种语义层面的预测方法,将地区历史的主题值数据转化成时间序列,使用但不限于长短期记忆网络(long short term memory, LSTM)方法^[32]预测未来一段时间数值变化。

3 主题模型集成

3.1 主题模型生成

接警日志中报警内容字段是非结构化的文本数据,本文采用交互式主题模型集成的方式实现从异构文本中抽取主题,如图 2 所示。该方式可以有效地避免传统主题模型难以确定合理参数的问题,将集成后主题模型降维投影在可视化界面,供用户选择。

本文使用 LDA 模型^[33]对接警日志提取主题,这是一种概率主题建模方法,通常用于处理大量

图 2 主题模型集成

文本. 主题被定义为给定词汇表上的概率分布, 即一组关键词, 其中具有高概率的关键词表示主题的语义内容. 除了主题-关键词分布之外, LDA 模型还生成文档主题分布, 即文档在每个主题上的分布值, 它可以表示文档与主题的贴合程度. 本文将接警日志内的接警内容字段作为 LDA 模型的原始语料库, 每条记录的接警内容作为一个文档. 由于接警内容中包含大量干扰信息, 会导致生成的主题模型质量较低, 因此本文在训练模型前需要将经过分词和去停用词处理后的文本再进行加工, 主要包括去掉无用字符、同义词替换和统计词频. 无用字符是指没有实际语义却不在停用词表中的字符, 本文所使用的接警内容数据中的无用字符是数字、英文单词以及特定的地理名称和人物名称. 数字和英文单词可以通过指定范围 ASCII 码删除; 地理名称和人物名称则通过添加到特定停用词表中去除. 同义词替换是指将分词后的数据中意思相近的词语换成同一个词语表示, 通过该方法可以减少同义词对主题模型的干扰, 能有效地提升生成模型的质量. 统计词频是指统计每个单词在语料库中出现的次数, 该操作按照词频由大到小保留前一部分词语, 可以去掉在语料库中出现次数很小的词语, 如人物名称和物品名称, 有效地降低稀疏词语对主题模型的干扰, 提升每个主题的聚合效果.

原始语料库在经过上述方法处理后可以作为训练模型用的语料库. 为了摆脱主题模型质量对参数的依赖, 本文使用迭代的方法, 通过改变影响较大的参数生成多个主题模型, 例如, 主题模型数量、 α 和 β , 其中 α 和 β 参数通常使用默认值或根据训练数据自动调整. 虽然这样设定可以解决大多数情况, 但存在参数设置不合理的可能, 所以根据这些参数的默认值迭代几个相邻值, 对主题模型数量在合理的区间内迭代, 生成多个主题模型, 之后将生成的主题模型集成在一起, 用户可以查看并选择其中的主题用于提取接警文本中的语义信息.

本文使用 LDA 主题模型提取文本语义信息, 作为接口使数据可以保存在数据立方体中. 系统中集成了多个 LDA 主题模型, 其中包含众多主题, 既有表示大类的主题, 也有表示细分类型的主题, 用户可以根据自身需求选择. 当用户在使用时发现主题内容不符合要求时, 可以在可视化界面中调节主题模型的参数, 生成涵盖不同语义范围大小的主题, 系统会根据用户选择的主题提取文本

语义信息, 保存在数据立方体中.

3.2 主题模型投影

主题中包含一系列关键词和对应的权重值, 为了方便用户查看各主题内容和之间的差距, 将主题模型集成中的主题信息投影在二维平面上^[34]. 首先把每个主题转换成多维向量表示, 由主题模型中导出的主题内容包含主题词、主题词在词库中对应 ID 和主题词权重, 词库记录了训练 LDA 模型时为语料库中的词语提供与之对应的 ID, 为每个主题初始化一个词库长度的零值数组, 将主题词的权重存在数组中词语对应 ID 的位置, 从而将主题变成向量; 之后再通过这些多维向量计算每个主题之间在二维平面上的距离, 将主题信息投影在二维坐标系上.

每个主题使用圆形表示, 圆心的坐标代表主题的投影位置, 圆的半径代表该主题中关键词涵盖范围, 半径越长表明该主题中语义信息越丰富, 圆心会显示该主题中权重最高的关键词. 在投影平面的下方显示有主题内容的词云, 当用户点击视图中主题时, 会在这里显示该主题中包含的词语和权重信息, 主题中词语的权重对应词云中显示的大小和位置, 权重高的词语会被安排在中间位置, 词云中的颜色代表该主题的位置信息, 投影位置相近的主题使用相近的颜色, 从而便于用户从颜色上区别各主题间的差异.

根据投影的结果, 用户可以通过点击许多距离较近的主题, 查看主题对应的词云了解主题内容, 通过对比词云从中选择最满意的主题使用, 右击其他主题可以取消选择. 同时, 投影中会出现比较离群的主题, 此时用户可以选中查看该主题内容, 如果发现是由于主题中词语语义不明确导致的, 可以选择删除该主题. 用户通过点击视图中的词云选择主题, 系统会根据用户选择的主题在主题时空体中找到对应主题的单元格并返回其中存储的指标.

4 主题时空体

主题时空体是一种立方体数据结构, 通过输入 3 个维度的查询条件可以查看存储在对应单元格中的指标值. 将接警数据中的时间 t 、空间 s 和主题 p 模型中的主题作为立方体的 3 个维度, 将主题值和关键词权重 M 作为指标存储在单元格中. 首先, 设计立方体数据结构组织以支持上述的查询任务, 再将离散化后的时间、空间和主题作为数

据立方体的 3 个维度,在单元格内存储通过主题模型中获取的主题值和关键词权重,从而满足快速

对接警数据在语义时空层面上的探索.主题时空体的构建如图 3 所示.

图 3 主题时空体构建

4.1 任务与数据组织

本节对总结的任务进行统一组织,目的是为所有任务提供一致的接口,便于后续的结果可视化.数据立方体的三维结构不便于整体探索,所以在进行查询任务时引入了切片的方法^[35],先将立方体在其中一个维度上进行切片,该切片的厚度可以是单位长度,也可以是一段范围,在切片中存储的值是范围内的聚合值.例如,在任务 1 中,如果选定的是某一刻度值时间 t ,就将立方体在刻度 t 上切片,之后在该切片上探索指标值 M 在主题和空间上的变化关系.如果选定的是段时间范围 Σt ,就将立方体在时间范围 Σt 切片,之后将该切片中存储的指标沿着时间维度投影,将相同空间相同主题不同时间中存储的值加和平均,之后再对投影的切片上探索指标值 M 在主题和空间上的变化关系.最后对结果进行可视化,如此可以实现对立方体中任意维度上的查询.

在接警数据中,时间和空间属性是连续的,由于构建数据立方体需要离散的时间和空间信息,因此要对连续的数据进行离散处理.从时间角度,本文按照自然时间单位划分,根据警务数据数量选择合理的最小时间精度,经过测试选择时间精度为小时最合适.用户在对数据查询时,大于小时的查询范围将拆分成小时进行查询.从空间角度,常用 2 种方法进行空间离散,第 1 种方法是采用预先存在的规则,如按照行政区进行区域划分;第 2 种方法是按照固定的形状(如矩形、六边形等)将空间划分成更小的子空间,或者基于数据的分布特

点进行不规则形状划分^[36].由于第 1 种方法需要根据城市设定划分,不具备普遍性,因此本文选择第 2 种方法,即将某个城市划分成相同大小的矩形子地区,根据数据的密度情况,划分为 $n \times n$ 个子空间.主题维度用来表示用户在主题模型集成中选择的多个主题.通过上述操作,完成了语义时空体的 3 个维度,分别是以小时为单位的时间 t 、以经纬度为单位的地区 s 和主题 p .

4.2 指标计算

需要对接警数据中的文本数据进行语义提取,使用主题模型计算某个点 $P(p, s, t)$ 上的主题分布值(S_p)和关键词权重(K_w),将这 2 个指标存储到语义时空体对应点的单元格内.文本有时会属于不同主题,可能会有在多个主题上的分布值,本文将每个文本的所有主题分布值存到对应主题的单元格内,通过这些主题分布值共同表达文本中的语义构成情况.

4.2.1 主题分布值 $S_p(p, s, t)$ 计算

收集所有在空间 s 和时间 t 上的记录,得到文档子集 $D(s, t)$,然后通过主题模型导出该子集中每个文本在用户选择的多个主题 $p_1, p_2, \dots, p_n$ 上的主题值,所有记录在对应的单个主题 $p_x (x=1, 2, \dots)$ 的主题值相加,即为点 $P(p_x, s, t)$ 的主题分布值,如

$$S_p(p, s, t) = \sum_{d \in D(s, t)} v_{dp} \quad (1)$$

其中, d 为文档子集 D 中的记录; v_{dp} 为该记录在主题 p 上的主题值,该值可以在主题模型中文档-主题中获得,接警日志记录的文本信息输入模型中,

即可以得到该数据在每个主题上的 v_{dp} 值。

4.2.2 关键词权重 $K_w(p, s, t)$ 计算

本文用得到的 S_p 值乘以词语在主题模型中该主题的权重值 v_{pk} 作为该关键词在点 P 中的权重, v_{pk} 值可以在主题模型中主题-主题词中获得。如

$$K_w(p, s, t) = S_p(p, s, t) v_{pk} \quad (2)$$

如果某个主题在该点上具有较高的分布值, 并且某个主题词在该主题具有较高的权重, 则该关键词在点中的 K_w 值较高。

用户通过可视化界面交互选择主题和输入查询范围, 根据用户的查询条件在主题时空体中查找对应刻度上与存储的结果, 将结果输出以可视化方式展示(如图3中⑤所示)。使用该方法可以快速地完成用户的查询任务, 并返回包含语义信息的结果, 实现对接警数据在语义时空分布上的实时交互探索。

5 可视分析与预测

5.1 LDA 主题选择

用户在主题投影视图先查看并选择自己感兴趣的主体, 主题投影视图是将主题模型集成中的

主题进行降维投影。根据各个主题中词语的相似程度计算主题间距离, 将高维数据显示到二维平面, 在计算距离时加入防碰撞检测^[37], 避免在投影时出现重合的现象。生成的每个主题则对应某个类型的案件, 根据主题的选择就可以对案件进行分类。本文通过修改参数生成多个主题模型, 从中选取了主题内容明确的56个主题, 将这些主题进行集成。用户可以在这56个主题中任意选择其中感兴趣主题, 并将对应的主题值显示在其他视图, 用于将案件分类并计算对应类型的主题值。投影后主题圆圈大小代表主题词涵盖的范围, 主题间的距离表示他们间的差距。用户可以点击其中主题进行交互选择, 选择的主题及其包含的主题词将显示在下方的主题备选列表。

主题备选列表显示如图4a所示用户选择的任意个数主题的词云, 其中是用户从主题中选择的主题, 用户在投影视图中选中某个主题, 该主题的内容会以词云的方式显示。主题内容是由主题模型中的主题-主题词权重得到, 词云中字体大小代表该词语在主题中的权重值, 权重值大小反映该词语在主题中的重要性。在该列表中, 用户可以查看自己所选的多个主题中包含的主题词, 帮助用户判断是否要选择这些主题, 用户可以选择取消

图4 可视化系统

其中不满意的主题, 并选择其中某一个主题进行接下来的时空分布探索, 在使用中可以随时切换备选列表中其他主题并查看对应主题值的时空分布情况.

5.2 历史统计

5.2.1 时空分布

根据用户在主题备选列表中选中的主题, 如图 4b 所示将对应主题值的时空分布情况投影在时空投影视图, 该视图用于探索报警数据主题分布值在时空上的分布情况, 用户可以自定义选择查询时间查看. 时间方面, 用户可以通过调整下方的时间轴选择投影的时间范围(如图 4b2 所示), 选择起始终止时间则只统计选定时间范围内指定时间段的主题数据, 时间的最小刻度单位为小时. 空间方面, 本文根据经纬度将该城市外环以内地区平均分为 16×16 的网格(如图 4b1 所示), 每个网格代表一个子地区, 与主题时空体中地区划分对应. 本文将子地区编号为 A_{ij} , i 和 j 分别表示经纬度坐标位置, 地图视图中左下方为原点, 对应的地区编号为 A_{00} .

该系统将每个地区内在指定时间范围内的报警数据所对应用户选择的主题值求和, 得到该时空查询的主题值, 查询结果的数值按照大小排名显示在地区主题值排序图(如图 4b3 所示). 由于一些案件的发生地点可能跨越多个子地区, 原始数据的分布情况会受到地区划分宽度的影响而改变. 为了降低地区划分对数据分布的影响, 便于用户发现主题案件地区分布的热点, 对每个子地区主题值, 本系统使用核密度估计方法^[38]选择平滑的高斯核密度估计. 一个子地区周围有 8 个子地区与之相邻, 所以本系统中核密度估计使用的窗口范围为 8 个子地区.

将使用核密度估计后得到的每个地区主题值投影到对应位置, 透明度与主题值大小成正比. 通过投影结果, 用户可以看到在指定时间段内所选主题的主题值在地区上的分布情况. 在使用时将时间刻度选择到最后一天, 就可以查看未来一天预测结果的时空分布情况; 通过选择 LDA 主题视图中不同主题, 可以查看各类型案件的分布情况.

5.2.2 模式比较

用户如果有对多地区间对比需求, 如图 4c 所示可以将指定的时空查询条件生成模板保存到统计对比视图中. 不同类型案件分布会受到时间影响呈现不同的时空分布情况, 该视图通过 3 个不同时间精度的折线图分别展示时间分布的情况, 分别是每天不同时刻分布图、每周不同星期分布图和

每月不同日期分布图; 3 个折线图帮助用户从各种角度探索指定主题类型案件的时间分布情况. 图 4c 显示了使用本系统保存的 2 个不同时空范围产生的模板视图, 视图左侧显示该模板对应的时空查询条件, 视图右侧为 3 个不同时间刻度的 3 个折线统计图. 图 4c1 折线图横坐标为小时, 纵坐标为主题值, 表示主题值对应每天不同小时的分布情况; 图 4c2 折线图横坐标为星期几, 纵坐标为主题值表示主题值对应一周内不同星期的分布情况; 图 4c3 折线图横坐标为日期, 纵坐标为主题值, 表示主题值在全部时间内的分布情况.

用户可以通过该视图查看自己选择的主题在指定时空范围内的分布情况, 并与其他不同时空范围的分布对比分析, 方便用户对多地区多时间主题类型案件分布情况进行对比探索.

5.3 文本主题预测

本文提供了一个对具有时空属性的警用数据进行预测的方法. 通过将历史主题值数据整理成时间序列, 使用 LSTM 模型对主题值时间序列进行预测^[39], 既可以查询历史数据, 也可以查看未来一段时间的结果.

主题时空体中地区内对应主题值按照时间顺序排列, 可以得到主题值随时间变化的序列, 即主题值时间序列; 之后对提取得到的时间序列进行分析处理, 以便进行后续的预测操作. 本文对子地区单独处理, 对上述步骤得到的子地区以天为单位报警记录的在同一主题上分布值, 按照时间顺序排列, 便可以得到该地区每个主题的分布值时间序列.

通过对时间序列预测, 可以提高预测的准确度, 同时比起直接对文本数据预测, 该方法可以提高预测的速度, 更适用于对大规模警务数据的预测. 可以使用多种方法对时间序列预测, 在本文中使用 LSTM 模型对经过处理后的主题时间序列预测用于测试, 对时间序列预测模型的结果评估在本文第 6 节中.

6 案例分析与系统评估

6.1 案例 1. 交通事故分布情况探索

在本案例中, 用户通过使用本文的可视化系统对接警日志数据中某市区内交通事故时空分布情况进行探索, 以此证明本文工作的实用性. 交通事故的发生通常与车辆密集程度有关, 所以交通事故的发生会与时间和空间有很大关联, 本文通

过选择交通事故主题,对比该主题在工作日和节日的分布情况对该市交通事故案件进行探索。

首先在 LDA 主题视图主题投影中选择多个与交通事故有关的主题加入到下方,本次实验选择了主题编号 16, 3, 1, 28, 对比查看这 4 个主题,发现编号 16 主题最适合。于是在备选主题中选中该主题进一步查看,只查看与交通事故主题有关的案件。其次在时空分布视图中选择查询的时间范围,因为使用的数据时间范围内包含国庆节,所以对国庆节的 7 天假期的时间进行节日分布情况探索,使用国庆节前 2 个月时间的数据进行工作日分布情况探索。选择时间查询条件 2017-10-01—2017-10-07, 点击查询,在图 5c 所示视图中可以看到交通事故主题事件在该市区内的分布情况。再将时间查询条件调整到 2017-08-01—2017-09-30, 点击查询,在图 5b 所示视图中查看交通事故主题时间在工作日时的分布情况。图 5d 所示将热点较高地区的事件内容生成的词云,可以看出交通事故占有很大的比重。

通过对比 2 次查询结果发现,在空间分布上,交通事故类事件在工作日有多个热点地区,分布在市区的大部分位置,这些地区多为办公和上学的集中区域,车流量较大,导致交通事故容易发生。查看工作日时间全部地区的小时分布图(如图 6a 所示)可以看到,在上班高峰期有明显的峰值。在节假日期间,交通事故分布较为集中,该区域是旅游景点及商业街集中的位置。查看节日期间全部地区的小时分布图(如图 6b 所示)可以看到,节日期间交通事故没有明显的峰值,从上午 10 点到晚上 8 点一直保持较高的发生率。在周图中可以看到,节日期间 10 月 1 日和 2 日的日发生量远远大于工作日的发生量,且都集中发生在区域②(如图 6 所示)。因为在这 2 天会有大量游客前往该区域游玩,车辆较集中且多为外地车辆,导致该区域集中爆发交通事故;而在节假日中期会下降至平时值,节假日末期会有回升。车辆分布集中对交通事故的处理速度有更高的要求,在节假日前期需要加大该区域的交通警察部署,以防止交通事故的大规模发生。

图5 探索主题时空分布情况

图6 交通事故模式对比图

下面进行详细比较,选择无论节假日还是工作日都是热点的区域②与在节假日和工作日变化较大的区域①③进行区域比较.分别选中①②③区域添加到模式对比视图中,生成这3个区域的详细统计值(如图6所示).通过右侧的3组24小时分布折线图可以发现,对于区域②,节假日期间高峰期时间段事件数量会高于工作日,但时间分布无明显变化;对于区域①③,节假日期间的值远低于工作日期间的值,并且峰值由工作日时上午9点变成节日时下午3点.由此可以看出造成区域①③交通事故多发的原因主要是上班高峰期的拥堵,而区域②事故多发的原因是景区和商业街人流量大.这也验证了图5中的分布结果,在工作日多区域是由上班高峰造成的交通事故,而区域②是由游客导致的车辆拥堵.所以在节假日期间可以将交通警力资源多集中分布在火车站及周围景区分布,工作日期间则要在上下班高峰期多部署交通警力,以减少交通事故的发生.

6.2 案例 2. 特殊事件分布情况探索

首先,在 LDA 主题视图投影中选择多个质量高的主题,本次实验选择编号 34 主题,该主题中包含有“赌博、黑窝点、诈骗、举报”等主题词(如图7所示).其次,在时空分布视图中,对国庆节7天假期的时间进行节日分布情况探索,使用国庆节前2个月时间的数据进行工作日分布情况探索.分别选择时间查询条件 2017-10-01—2017-10-07 和 2017-08-01—2017-09-30,在地图视图中查看该主题事件在工作日时的分布情况,查询结果如图7所示.

通过对比2次查询结果可以发现,在空间分布上,特殊类型事件在工作日期间集中分布在和平区和火车站附近,这是因为火车站周围人流量较大,人员复杂,所以会有较多的赌博和贩毒的事件发生;而和平区是商业集中地,存在很多的地下窝

点涉及卖淫和赌博的事件.在节日期间,该主题事件比起工作日又多出几个热点区域,分别是区域②和区域③(如图7所示).在时间分布上,在周图中可以看到,节假日前期10月1日和2日的日发生量远大于工作日的发生量,而在节假日中期会下降至平时值,末期会有回升.

图7 主题内容及区域事件词云

为进一步探索变化的原因,分别选中区域①②③添加到模式对比视图中,生成这3个区域的详细统计值如图8所示.通过中间的3组24h分布折线图可以发现,对于区域①,节日期间高峰期时间段事件数量会略高于工作日,但时间分布无明显变化;对于区域②③,节日期间的值高于工作日期间的值,并且峰值的时间变化较大.区域②在工作日有明显的高峰时间,而在节假日会出现多个高峰,由此可以推断出该区域可能分布较多的卖淫事件,因为工作日时下班后才会有大量事件发生,而节假日全天都会有人前往.区域③的工作日峰值在上午10点,所以此地的事件不同于区域②,主要是由诈骗和赌博的事件构成,该地区位于离市区较远的位置,人口密度较低,较隐蔽,有利于从事违法犯罪活动.为了进一步探索这2个区域的案件构成,本文对原始数据中的接警文本生成词云(如图7所示),可以看到区域②中“卖淫”类词语

图8 特殊事件模式对比图

权重较大,区域③中“赌博、财产”类词语权重较大。这也验证了图7中的分布结果,在工作日主要集中在区域①,因为这里商业密集,位于火车站周边,造成该地区该主题事件发生大于其他地区。在节假日,由于不需要工作,因此有时间前往距离办公地较远的区域②嫖娼,而节假日期间人流量增多有利于区域③的赌博和贩毒进行。所以在节日期间可以将警力资源多分散分布在多个热点区域,并根据每个热点区域特点分工,工作日期间则要在市区中心多部署警力,以减少该主题事件的发生。

6.3 专家反馈

为进一步验证本文系统的有效性与实用性,邀请警务数据分析领域和城市治安领域的专家参与可视化分析系统进行实际操作。在实验开始前对本系统进行了详尽的讲解,之后专家们使用本系统对不同类型警务数据的时空分布情况进行探索,并在使用后对该系统提出了一些使用的评价和建议。

警务数据分析领域专家的评价是,传统的对接警数据处理是依靠数据中案件类型的字段进行案件分类,会存在字段缺失和分类不准的问题,并且案件类型不够丰富。传统方法对接警数据处理时,需要较长的响应时间,难以支持用户对数据的详细探索。本文提出的系统可以有效地解决上述问题。首先,用户可以通过 LDA 主题视图,比较选择最符合需求的主题,并且可以同时选择多个主题,一次操作可以查看多个类型的案件。其次,该系统在使用时响应迅速,当用户进行操作时,系统可以快速地将查询结果更新,满足用户的实时交互查询。

6.4 预测评估

本文对主题值时间序列的拟合效果,预测结果是否准确进行评估。本文使用均方根误差(root mean square error, RMSE)、平均绝对误差(mean absolute error, MAE)和准确率作为评价指标。评价使用的数据是时间长度为 137 天的某市警务报警数据,将某市划分成 16×16 个子地区,生成 6 个主题模型,挑选其中 56 个主题进行集成。由于主题数量众多,本文挑选其中 6 个具有代表性主题,计算方法为将这 6 个主题在全部子地区中的主题值时间序列作为模型的输入,时间序列的前 133 天作为训练数据,后 4 天为测试数据,通过对比该 6 个主题的预测数据与真实数据的误差评估模型的预测性能。考虑每个主题值时间序列的预测难度不

一致,单纯计算拟合误差并不能完全说明预测结果的准确性。本文为了对预测结果进行定性分析,对每个时间序列设置错误区间,如果预测值与真实值的误差在错误区间内,即认为预测结果准确。本文将错误区间的值设置成该时间序列的标准差,如

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (3)$$

其与真实值差值的绝对值小于标准差时,认为该预测为准确的。最后对全部地区 4 天的预测结果进行统计,计算最终预测准确数量 C_{corret} 占全部数量 C_{total} 的比例作为准确率 A 。

表 3 所示为本文为检验预测性能得出的结果,各主题在可以接受的误差范围的预测结果基本保持在 0.6 以上,基本满足用户对预测的要求。对于事件较多的地区,提取出的时间序列较稳定,预测的准确率较高。对于数据充足的主题(主题 1, 2, 6),预测效果会优于数据较少的主题(主题 3, 5),主题对应的事件较少时,预测的准确率相对较低,主要是由于从数据库中提取主题值时间序列稀疏,虽然通过地区融合方法可以缓解稀疏问题,但预测准确率也会低于平均值。

表 3 各主题预测结果

主题	RMSE	MAE	A
1	0.125	0.147	0.715
2	0.140	0.165	0.672
3	0.149	0.171	0.597
4	0.145	0.171	0.662
5	0.145	0.168	0.611
6	0.143	0.168	0.681

7 结 语

本文设计了一个对警务数据在语义时空模式层面交互探索的可视分析系统,该可视化系统包含主题模型集成、文本数据立方体和可视化界面。使用主题模型集成实现对文本数据的语义抽取,同时解决模型质量对参数的依赖,对比单纯通过事件类型分类,可以有更丰富的主题类型,更全面地提取语义信息。按照时间、空间和语义 3 个维度构建的数据立方体,可以实现对警务数据在这 3 个维度的快速查询,并返回其中存储的主题值信息。可视分析界面可以将查询返回的主题值信息投影在对应视图中,用户可以通过主题视图交互选择需要查看的案件类型,输入时间空间查询条件得

到需要的结果. 系统提供了一种基于主题的预测方法, 可以在时空投影视图查看通过模型得到的未来预测结果. 最后, 本文通过案例分析以及对预测结果的评估验证本文方法的可用性.

在未来工作中, 将尝试改进主题模型集成方法, 进一步提升生成主题模型的质量, 防止出现主题语义模糊的问题. 调整预测模型的参数和结构, 以获得更高的预测准确率. 未来也将改进可视化工具, 使用户对数据进行时空探索时更容易操作与理解, 并进一步提升可视化界面的美观程度和交互程度.

参考文献(References):

- [1] Swann T. Social media: a critical introduction[J]. *Consumption Markets & Culture*, 2017, 19(3): 1-3
- [2] Wang Haiyan, Hu Ting, Wang Jun. Application research of intelligent police system in municipal public security bureau[J]. *Computer Science*, 2018, 45(S1): 531-535(in Chinese)
(王海燕, 胡婷, 王俊. 智慧警务系统在地市公安局的应用研究[J]. *计算机科学*, 2018, 45(S1): 531-535)
- [3] Papanikolaou Y, Foulds J R, Rubin T N, *et al.* Dense distributions from sparse samples: improved Gibbs sampling parameter estimators for LDA[J]. *The Journal of Machine Learning Research*, 2017, 18(1): 2058-2115
- [4] Ibrahim Y. Temporality, space and technology: time-space discourses of call centres[J]. *New Technology, Work and Employment*, 2012, 27(1): 23-35
- [5] Gray J, Chaudhuri S, Bosworth A, *et al.* Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals[J]. *Data Mining and Knowledge Discovery*, 1997, 1(1): 29-53
- [6] Peng Yanni, Fan Xiaoping, Zhao Ying, *et al.* A survey of visualization of temporal event sequence data[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2019, 31(10): 1698-1710(in Chinese)
(彭燕妮, 樊晓平, 赵颖, 等. 时间事件序列数据可视化综述[J]. *计算机辅助设计与图形学学报*, 2019, 31(10): 1698-1710)
- [7] Wang X, Dou W, Ma Z, *et al.* I-SI: scalable architecture for analyzing latent topical-level information from social media data[J]. *Computer Graphics Forum*, 2012, 31(3pt4): 1275-1284
- [8] Marcus A, Bernstein M S, Badar O, *et al.* Twitinfo: aggregating and visualizing microblogs for event exploration[C] // *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 2011: 227-236
- [9] Zhou X L, Xu C. Tracing the spatial-temporal evolution of events based on social media data[J]. *ISPRS International Journal of Geo-Information*, 2017, 6(3): 88
- [10] Chae J, Thom D, Bosch H, *et al.* Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition[C] // *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Los Alamitos: IEEE Computer Society Press, 2012: 143-152
- [11] Dou W W, Wang X Y, Skau D, *et al.* LeadLine: interactive visual analysis of text data through event identification and exploration[C] // *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Los Alamitos: IEEE Computer Society Press, 2012: 93-102
- [12] Bosch H, Thom D, Wörner M, *et al.* ScatterBlogs: geo-spatial document analysis[C] // *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Los Alamitos: IEEE Computer Society Press, 2011: 309-310
- [13] Bosch H, Thom D, Heimerl F, *et al.* ScatterBlogs2: real-time monitoring of microblog messages through user-guided filtering[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12): 2022-2031
- [14] Chae J, Thom D, Jang Y, *et al.* Public behavior response analysis in disaster events utilizing visual analytics of microblog data[J]. *Computers & Graphics*, 2014, 38: 51-60
- [15] Xu P P, Wu Y C, Wei E X, *et al.* Visual analysis of topic competition on social media[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12): 2012-2021
- [16] Sun G D, Wu Y C, Liu S X, *et al.* EvoRiver: visual analysis of topic competition on social media[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 1753-1762
- [17] Wang X T, Liu S X, Chen Y, *et al.* How ideas flow across multiple social groups[C] // *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Los Alamitos: IEEE Computer Society Press, 2016: 51-60
- [18] Dou W W, Yu L, Wang X Y, *et al.* HierarchicalTopics: visually exploring large text collections using topic hierarchies[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12): 2002-2011
- [19] Cui W W, Liu S X, Wu Z F, *et al.* How hierarchical topics evolve in large text corpora[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2014, 20(12): 2281-2290
- [20] Cao G F, Wang S W, Hwang M, *et al.* A scalable framework for spatiotemporal analysis of location-based social media data[J]. *Computers, Environment and Urban Systems*, 2015, 51: 70-82
- [21] Shekhar S, Lu C T, Liu R, *et al.* CubeView: a system for traffic data visualization[C] // *Proceedings of the 5th IEEE International Conference on Intelligent Transportation Systems*. Los Alamitos: IEEE Computer Society Press, 2002: 674-678
- [22] Lins L, Klosowski J T, Scheidegger C. Nanocubes for real-time exploration of spatiotemporal datasets[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12): 2456-2465
- [23] Pahins C A L, Stephens S A, Scheidegger C, *et al.* Hashed-cubes: simple, low memory, real-time visual exploration of big data[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2017, 23(1): 671-680
- [24] Wang Z, Ferreira N, Wei Y H, *et al.* Gaussian cubes: real-time modeling for visual exploration of large multidimensional datasets[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2017, 23(1): 681-690
- [25] Miranda F, Lage M, Doraiswamy H, *et al.* Time lattice: a data structure for the interactive visual analysis of large time se-

- ries[J]. Computer Graphics Forum, 2018, 37(3): 23-35
- [26] Lin C X, Ding B L, Han J W, *et al.* Text Cube: computing IR measures for multidimensional text database analysis[C] //Proceedings of the 8th IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2008: 905-910
- [27] Zhang D, Zhai C X, Han J W, *et al.* Topic modeling for OLAP on multidimensional text databases: topic cube and its applications[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2009, 2(5/6): 378-395
- [28] Cao N, Lin C G, Zhu Q H, *et al.* Voila: visual anomaly detection and monitoring with streaming spatiotemporal data[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 23-33
- [29] Malik A, Maciejewski R, Towers S, *et al.* Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1863-1872
- [30] Andrienko N, Andrienko G, Barrett L, *et al.* Space transformation for understanding group movement[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2169-2178
- [31] Ahmed M, Shumailov I, Anderson R. Tendrils of crime: visualizing the diffusion of stolen bitcoins[C] //Proceedings of International Workshop on Graphical Models for Security. Heidelberg: Springer, 2018: 1-12
- [32] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471
- [33] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022
- [34] Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics[C] //Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore: ACL Press, 2014: 63-70
- [35] Li J, Chen S M, Chen W, *et al.* Semantics-space-time cube: a conceptual framework for systematic analysis of texts in space and time[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(4): 1789-1806
- [36] Andrienko N, Andrienko G. Visual analytics of movement: An overview of methods, tools and procedures[J]. Information Visualization, 2013, 12(1): 3-24
- [37] Fan Zhaowei, Wan Huagen, Gao Shuming. A fast collision detection algorithm in image space[J]. Journal of Computer Aided Design & Computer Graphics, 2002, 14(9): 805-809(in Chinese)
(范昭伟, 万华根, 高曙明. 基于图像的快速碰撞检测算法[J]. 计算机辅助设计与图形学学报, 2002, 14(9): 805-809)
- [38] Leibbrandt R, Günnemann S. Making kernel density estimation robust towards missing values in highly incomplete multivariate data without imputation[C] //Proceedings of the SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2018: 747-755
- [39] Xiao Fen, Gao Xieping. An approach for short-term prediction on time series from parameter-varying systems[J]. Journal of Software, 2006, 17(5): 1042-1050(in Chinese)
(肖芬, 高协平. 参数可变系统时间序列短期预测方法[J]. 软件学报, 2006, 17(5): 1042-1050)