



武汉大学学报(信息科学版)

Geomatics and Information Science of Wuhan University

ISSN 1671-8860, CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：一种融合多维关系的地理环境时空主题发现方法
作者：朱杰, 张宏军, 廖湘琳, 徐有为
DOI: 10.13203/j.whugis20210326
收稿日期: 2021-06-21
网络首发日期: 2022-07-18
引用格式: 朱杰, 张宏军, 廖湘琳, 徐有为. 一种融合多维关系的地理环境时空主题发现方法[J/OL]. 武汉大学学报(信息科学版),
<https://doi.org/10.13203/j.whugis20210326>



网络首发: 在编辑部工作流程中, 稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定, 且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式(包括网络呈现版式)排版后的稿件, 可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定; 学术研究成果具有创新性、科学性和先进性, 符合编辑部对刊文的录用要求, 不存在学术不端行为及其他侵权行为; 稿件内容应基本符合国家有关书刊编辑、出版的技术标准, 正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性, 录用定稿一经发布, 不得修改论文题目、作者、机构名称和学术内容, 只可基于编辑规范进行少量文字的修改。

出版确认: 纸质期刊编辑部通过与《中国学术期刊(光盘版)》电子杂志社有限公司签约, 在《中国学术期刊(网络版)》出版传播平台上创办与纸质期刊内容一致的网络版, 以单篇或整期出版形式, 在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊(网络版)》是国家新闻出版广电总局批准的网络连续型出版物(ISSN 2096-4188, CN 11-6037/Z), 所以签约期刊的网络版上网络首发论文视为正式出版。

DOI:10.13203/j.whugis20210326

引用格式：朱杰，张宏军，廖湘琳，等. 一种融合多维关系的地理环境时空主题发现方法[J]. 武汉大学学报·信息科学版，2022，DOI: 10.13203/j.whugis20210326 (ZHU Jie, ZHANG Hongjun, LIAO Xianglin, et al. A Method for Geographical Environment Spatiotemporal Topic Discovery of Multi-dimensional Relationships[J]. *Geomatics and Information Science of Wuhan University*, 2022, DOI: 10.13203/j.whugis20210326)

一种融合多维关系的地理环境时空主题发现方法

朱杰^{1,2} 张宏军¹ 廖湘琳¹ 徐有为¹

¹ 陆军工程大学指挥控制工程学院，江苏 南京，210002

² 73021 部队，浙江 杭州，315023

摘要：对战场文本数据的深入挖掘，可以高质量和高效率地发现时空主题结构，从而有效揭示战场事件发展的时空规律。针对现有的主题发现方法无法有效适用于具有多维异构关系的时空主题发现，提出了一种融合多维关系联合聚类的时空主题发现方法，首先构建以地理环境实体、地理位置与事件主题为节点的主题关系网络，其次以张量模型的 Tucker 分解建立主题关系的完全表达式作为主题分类的目标函数，最后运用块值矩阵分解方法进行联合聚类计算，获取主题分类结果和内聚结构。实验结果表明，该方法能够有效发现具有时空语义关系特征的主题结构，较好地体现出地理环境要素与时空主题之间的关联性，以及时空主题在地理位置与事件主题标签上的内聚性，进一步反映出主题的演化过程。

关键词：地理环境；多维关系；时空主题发现；块值矩阵分解；联合聚类

中图法分类号：P208

文献标识码：A

近年来，通过大数据技术，从大规模、复杂程度高的多维数据中获取最关键的主题信息成为研究热点之一，而时空数据作为反映人类活动规律的核心数据源，与时空相关的主题信息也可以反映现实世界的各种社会活动概貌。战场文本中蕴含丰富的时空语义信息，并结合位置信息的分析，成为发现战场态势变化、部队行动特征及应急事件规律的重要依据。因而，利用战场文本数据进行时空主题分析，从微观到宏观、由分散到聚集地获取地理环境要素的空间分布模式及其对战场活动的影响特征，挖掘战场事件的空间分布与发展规律，由此进一步丰富了战场环境感知方式，提供了新的战场环境效能分析手段，对于深入理解战场环境知识具有重要的意义和价值。

主题发现是利用一系列自然语言处理和分析的方法，从规模性的文本数据中自动抽取主题性和关键性信息，并加以聚类形成结构化主题的过程^[1]。时空主题发现侧重于研究与位置、时间相关主题或者文档主题，

相较于传统的主题发现，在考虑不同空间区域对主题的影响，和连续时间上的主题分布，能够引入时空因素扩展主题模型，提高文档主题发现能力^[2]。现有的面向社交媒体数据的时空主题发现研究已经取得了一定的成果，主要集中在基于文本聚类的主题分类^{[3][4]}和基于主题模型的主题提取^{[5][6]}两类计算方法，并在灾害事件影响分析、城市旅游功能识别、商业设施资源规划等方面得到了广泛的应用^[7-9]。前者主要是利用文本词向量位置距离或者相似性进行归类形成不同主题的簇或类别，其典型方法如共词分析法、划分法、层次法等^[10]，这些方法优点在于能够快速筛选关键词构建共词网络，但是仅通过词频统计方法挖掘主题信息，容易忽略词项的语义描述，存在向量结构单一和特征稀疏的问题，造成主题词之间的语义关系不能被发现；后者是一种降维能力较出色的方式，采用概率主题模型将关键词的语义特征量化，结合语义分析方法，对模型处理后的结果进行聚类并识别主题，其优点在于

收稿日期：2021-06-21

项目资助：中国博士后科学基金资助项目（2019M664028）。

第一作者：朱杰，博士，工程师，主要从事战场环境认知及战场位置服务的理论和研究方法研究。zjsoldierlee@163.com

能够提取文本潜在语义信息,比较适合时间主题信息发现。地理环境时空主题结构是融合多维异构关系的复杂主题结构,不仅包括地理环境实体之间的复合关系,还包括实体与事件、实体与位置等多种异构关系,然而上述方法均没有从实体关系的角度考虑时空主题之间的语义关系,难以描述由其关键词构成的主题关系网络。由此,如何有效构建实体复合关系网络并融合多维异构关系进行主题聚类是提高地理环境时空主题发现质量的关键技术。

基于此,本文提出一种多维关系联合聚类的时空主题发现方法,基于关键词共现关系构建以地理环境实体、地理位置与事件主题为节点的关系网络,从复杂网络的角度结合关系聚类思想发现时空主题结构。首先构建融合多维关系的时空主题张量模型,利用主题张量模型的 Tucker 分解给出主题关系的完全表达式;然后构建多维关系聚类的特征向量空间,以此作为主题分类的目标函数,并运用块值矩阵分解技术对此进行联合聚类计算,并利用核心张量矩阵解决数据稀疏问题;最后,通过仿真实验验证该方法的有效性,所得到的结果能够较好地体现出地理环境要素与时空主题间的关联性,以及时空主题在事件主题与地理位置标签上的内聚性,从而提高发现应急事件时空主题的质量和效率,进一步揭示事件的发展态势。

1 相关研究工作

目前,对于研究时空主题发现的方法主要集中在共词分析、隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)主题模型以及相关聚类算法等方面。文献[11]采用LDA模型建立词、文档与主题结构之间的概率相关统计,并结合支持向量机(Support Vector Machine, SVM),分析其中的主题及其相互联系,并能够输出主题演变的时间分布;文献[12-13]通过词频-逆文档频率分析自动筛选出与主题相关的关键词,以此构建基于关键词共现关系的共词网络,实现主题的挖掘与聚合;由于共词分析存在词与词之间位置关系不确定,且LDA主题模型易受词频影响,难以从词汇之间的关联来定义主

题之间的语义关系,因而,文献[14]利用语义分析方法,如潜在语义分析(Latent Semantic Analysis, LSA)方法、潜在语义索引(Latent Semantic Indexing, LSI)等发现数据中隐藏的语义关系,在此基础上进行位置主题提取。从上述词共现方法或主题模型可以看出,利用传统的主题聚类方法仅定义了与位置特征相关的词汇主题关系,而仅使用单维关系容易忽略主题间的语义关联,无法全面准确把握战场地理环境实体之间存在的多维异构关系的主题结构。

近年来,基于网络结构的社区主题发现算法能够融合用户、位置、媒体等多模实体及其多维关系,在挖掘复杂社区结构上表现出准确、高效的应用特征。文献[15]以图结构定义所包含的实体及其相互关系,并从语义角度赋予边权重,采用边聚类方式挖掘具有重叠结构的社区主题,但是容易造成因多维关系数据稀疏而降低计算效率的问题;文献[16]在矩阵分解方法能有效处理单维数据的基础上^[17],引入基于块值分解的联合聚类方法^[18]解决用户、位置与地理签到三者之间关联关系等多维度关系数据的聚类,并在关系数据稀疏问题上得到较好的处理,有效地提高了社区主题发现的质量和效率。

地理环境实体是指在战场及其周围对作战活动和作战效果产生影响的各种时空对象,按照要素类型划分,包括自然地理、人文地理、电磁空间以及核生化等要素组合构成实体对象^[19]。地理环境时空主题关系是由时空对象之间关联关系作用形成,实体对象关系研究由传统地理信息系统中地理对象的空间关系向全空间信息系统背景下的多粒度时空对象的关联关系研究发展^[20-21]。时空对象的关联关系可以用时间关联、空间关联、属性关联和综合关联等关系进行全方位的描述和表达,通过网络图方式将时空对象、关系类型和关系强度分别赋予节点、边和边权重的语义信息,从而实现对象关系逻辑计算的全方位表达^[22-24]。

基于此,本文以多粒度时空对象关联关系为基础,构建融合地理环境实体、地理位置、事件主题及其之间关联关系的复合关系网络,将基于矩阵分解的联合聚类方法改进

和扩展,使之适用于融合多维异构关系的时空主题挖掘,以有效提高时空主题发现的质量和效率,使得到的主题结构能够更好地表现出地理环境要素、地理位置与事件主题之间内聚的关联性,能够反映出主题的演化过程。

2 方法

2.1 地理环境时空主题复合关系网络模型

地理环境时空主题是一种由地理环境实体及其关联关系构成的复合关系网络,主要包括地理环境实体之间构成的实体-实体关联关系,实体在地理空间上分布构成的实体-地理位置关系,以及地理环境实体与事件主题特征分布构成的地理环境实体-主题相关关系,旨在以统一的方式对上述相互作用的关系进行建模,并从语义特征词中确定随时间和空间区域变化的主题划分。借鉴共词网络思想,构建以表示地理环境实体、位置和事件主题特征的关键词为节点、融合多维关系为网络边、关键词概率分布为边权重的复合关系网络,以此分析实体与主题、主题与主题之间的关系,进而有效挖掘时空主题结构。

如图 1 所示,该网络包含如下三类层次关系:

(1) 地理环境实体-实体关系层。该层次是由地理环境实体及其关联关系构成的加权有向图,表示实体之间的时间、空间以及属性关联关系,记为 $S=(E,O,W)$ 。其中, E 表示网络节点,对应文本集中提取的地理环境实体集,即 $E=\{e_1,\dots,e_n\}$; O 表示实体之间的关联关系组成的网络边集,即 $O=\{\langle e_i,e_j \rangle | e_i,e_j \in E\}$, $\langle e_i,e_j \rangle$ 是由节点 e_i 指向节点 e_j 的有向边; w 表示实体关系边权重集合,即 $W=\{w(e_i,e_j) | e_i,e_j \in E\}$,边权重 $w(e_i,e_j)$ 采用 Overlap^[25] 计算方法由实体关系关键词与实体关键词在数据集中的共现

频率表示,即

$$w(e_i,e_j)=\frac{|K'_{o_{ij}} \cap K'_{e_i} K'_{e_j}|}{\min(|K'_{o_{ij}}|, |K'_{e_i} K'_{e_j}|)},$$

其中 $K'_{o_{ij}}$ 表示实体关系关键词在语料数据集中出现的次数, K'_{e_i} 、 K'_{e_j} 分别表示发生关系的两个实体关键词出现的次数。

(2) 地理环境实体-地理位置关系层。该层次是由地理环境实体在地理空间上的位置分布形成的关系网络,表示实体与地理位置的紧密关系,可抽象成一个加权无向图,记为 $P=(E,L,H,C)$,且位置分布可转换为坐标的高斯分布,从空间域上得到不同的区域位置关系。其中, E 表示地理环境实体的网络节点集; L 表示地理位置的网络节点集,即 $L=\{l_1,\dots,l_n\}$; H 表示两者联接关系组成的网络边集,即

$$H=\{\langle e_i,l_j \rangle | e_i \in E, l_j \in L, E \cap L = \emptyset\};$$

边权重 C 表示两者关系边权重集合,即

$$C=\{c(e_i,l_j) | e_i \in E, l_j \in L\},$$

其中 $c(e_i,l_j)$ 是通过采用 Overlap 计算方法由地理环境实体关键词与地理位置关键词在数据集中的共现

$$c(e_i,l_j)=\frac{|K'_{e_i} \cap K'_{l_j}|}{\min(|K'_{e_i}|, |K'_{l_j}|)},$$

频率表示,即 $c(e_i,l_j)$ 分别表示发生关系的地理环境实体和地理位置关键词出现的次数。

(3) 地理环境实体-事件主题关系层。该层次是由带有事件主题特征的地理环境实体分布构成的关系网络,可抽象成一个加权无向图,记为 $Q=(E,U,V,D)$ 。其中, E 表示地理环境实体的网络节点集; U 表示事件主题的网络节点集,即 $U=\{u_1,\dots,u_n\}$,且事件主题具有时间性,按照事件从产生到发展直至消亡的周期,一般情况下不同时间阶段的事件主题由不同词汇表示,如果同一词汇

在不同阶段的重复表示事件主题, 则存在网络节点可重叠性, 从时间域上得到主题连续分布; v 表示两者带有主题标签的关系组成的网络边集, 即 $V = \{(e_i, u_j) | e_i \in L, u_j \in U\}$,

任意边 $v_{ij} = (e_i, u_j)$ 表示为无序对; 边权重 D 表示两者关系边权重集合, 即 $D = \{d(e_i, u_j) | e_i \in E, u_j \in U\}$, 其中 $d(e_i, u_j)$ 是通过采用 Overlap 计算方法由事件主题关键词与地理环境实体关键词在数据集中的共

现频率表示, 即 $d(e_i, u_j) = \frac{|K'_{e_i} \cap K'_{u_j}|}{\min(|K'_{e_i}|, |K'_{u_j}|)}$,

K'_{e_i} 、 K'_{u_j} 分别表示发生关系的地理环境实体和事件主题关键词出现的次数。

将上述3层关系融合构建实体复合关系网络模型 FR , 即:

$$FR = S \times P \times Q = (E, O, W) \times (E, L, H, C) \times (E, U, V, D) = ((E, L, U), O, H, V, O_{(w)}, H_{(c)}, V_{(d)})$$

上述包含3类节点: 地理环境实体集 E 、地理位置集 L 和事件主题集 U , 3类维度关系: 带权重的地理环境实体关联关系集 $O_{(w)}$ 、带权重的地理环境实体与地理位置联接关系集 $H_{(c)}$ 、带权重的地理环境实体与事件主题联接关系集 $V_{(d)}$ 。

原始的文本数据集经过分词处理、实体及实体关系抽取等过程形成“文本-词汇-主题关键词”的形式, 通过上述各层次网络的融合, 转变为由关键词作为节点, 关键词之间的语义关系作为网络边, 关键词的概率分布作为边权重构成的“网络-节点-边”的形式来表示“网络-时空主题, 实体/位置/事件-节点, 关联关系-边”的复合关系网络形式, 因此, 可通过网络中关联关系的聚类划分主题, 用所包含的关系紧密度解释主题结构, 通过核密度分析得到主题的空间分布, 以及不同时间阶段的主题变化趋势。

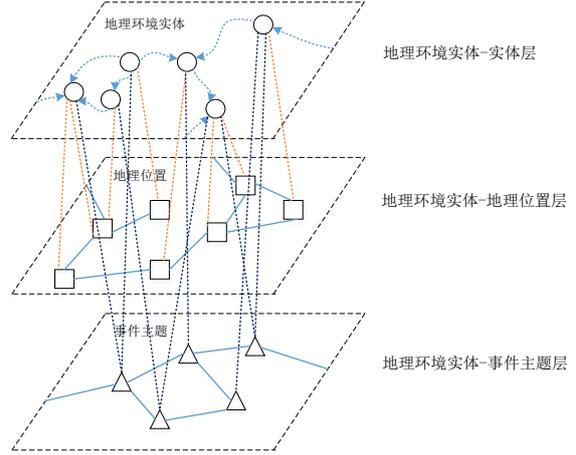


图1 地理环境时空主题复合关系网络

Fig.1 Geographical Environment Spatiotemporal Topic Compound Relation Network

2.2 时空主题复合关系张量 Tucker 分解

张量在数据结构上用于定义多维数组, 使向量和矩阵在输入输出及变换上应用于更高的维度, 能够以统一的方式来处理多线性映射, 因而在人工智能领域得到大量的使用。张量 Tucker 分解采用高阶主成分分析, 将一个张量分解表示成由一个核心张量沿每一个维度 (mode) 与一个因子矩阵相乘的形式, 从而完美解决张量的稀疏特征填补问题^[26]。以一个三阶张量 $\chi \in \mathcal{R}^{I \times J \times K}$ 为例, 将其 Tucker 分解表示成由一个核心张量 $Z \in \mathcal{R}^{P \times Q \times R}$ 沿每一个 mode 分别与相应的因子矩阵 $A \in \mathcal{R}^{I \times P}$, $B \in \mathcal{R}^{J \times Q}$, $C \in \mathcal{R}^{K \times R}$ 相乘, 即:

$$\chi \approx [Z; A, B, C] = Z \times_1 A \times_2 B \times_3 C \quad (1)$$

其中, \times_n 表示沿每一个 mode 的张量积模式, 因子矩阵 $A \in \mathcal{R}^{I \times P}$, $B \in \mathcal{R}^{J \times Q}$, $C \in \mathcal{R}^{K \times R}$ 通常是正交的, 可以视为沿相应 mode 的主成分。

由于实体复合关系网络模型 FR 从结构上可以表示为一个三阶二元张量, 其中每个元素对应一个二元张量, 网络边的权值用以表示张量的特征值, 因而, 为了能够用矩

阵完整表示时空主题复合关系网络模型，使其在高维空间统一形式进行计算，首先将上述每层的关系网络图结构转换成相应的邻接矩阵形式，分别表示为地理环境实体-实体关系、地理环境实体-地理位置关系和地理位置-事件主题关系等多维关系的邻接矩阵，然后对应时空主题结构，将其表示为三阶二元张量，如图 2 所示，引入三阶张量的 Tucker 分解方法将复合关系网络模型转换成由矩阵向量构成的二元张量积模式，即：

$$\phi[G;S,P,Q]=G \times_1 S \times_2 P \times_3 Q \quad (2)$$

其中， $S=A \in \mathfrak{R}^{M \times M}$ 表示地理环境实体-实体层网络关系的 M 阶方阵， $P=B \in \mathfrak{R}^{M \times N}$ 表示地理环境实体-地理位置层网络关系的 $M \times N$ 阶矩阵， $Q=C \in \mathfrak{R}^{M \times K}$ 表示地理环境实体-事件主题层网络关系的 $M \times K$ 阶矩阵， $G=Z \in \mathfrak{R}^{M \times N \times K}$ 为 $M \times N \times K$ 阶核心张量， M 、 N 和 K 分别为相应的地理环境实体、地理位置和事件主题的数量。

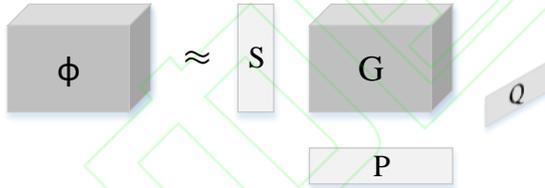


图 2 复合关系网络模型 Tucker 分解结构

Fig.2 Tucker Decomposition Structure of Composite Relational Network Model

(1) 地理环境实体-实体关系矩阵 S 。

令矩阵 $S = \{w_{ij}\}_{M \times M}$ ，元素 w_{ij} 表示地理环境实体 e_i 与 e_j 之间的关系值，是由实体间存在的时间关联 TR 、空间关联 SR 、属性关联 AR 等关联关系的综合权重表达，即：

$$w_{ij} = \begin{cases} w(sr)_{ij} \times w(tr)_{ij} \times w(ar)_{ij}, & w(sr)_{ij}, w(tr)_{ij}, w(ar)_{ij} \in w(e_i, e_j) \\ 0, & \text{otherwise} \end{cases}$$

(3)

其中， $w(sr)_{ij}$ 为实体 e_i 与 e_j 的空间关系

边权重， $w(tr)_{ij}$ 为实体 e_i 与 e_j 的时间关系边权重， $w(ar)_{ij}$ 为实体 e_i 与 e_j 的属性关系边权重。

(2) 地理环境实体-地理位置关系矩阵 P 。令矩阵 $P = \{c_{ij}\}_{M \times N}$ ，元素 c_{ij} 表示地理环境实体 e_i 在地理位置 l_j 上的分布概率，

可由地理环境实体 e_i 与地理位置 l_j 关系边权重得到，且满足 $\sum_j c_{ij} = 1$ 。若将矩阵 P 列

向量视为带有位置标签的地理位置分布，则矩阵 P 行向量可以看作是各地理实体在每个空间区域的位置分布，即 $P_i = [c_{i1}, c_{i2}, \dots, c_{ij}]$ 为实体 e_i 的区域位置向量。

(3) 地理环境实体-事件主题关系矩阵 Q 。令矩阵 $Q = \{d_{ij}\}_{M \times K}$ ，元素 d_{ij} 表示地理

环境实体 l_i 在事件主题 u_j 上的分布概率，可由地理环境实体 l_j 与事件主题 u_j 关系边权重得到，且满足 $\sum_j d_{ij} = 1$ 。若将矩阵 Q 列

向量视为带有主题标签的地理环境实体分布，则矩阵 Q 行向量可以看作是各地理实体在每个事件主题上的分布，即 $Q_i = [d_{i1}, d_{i2}, \dots, d_{ij}]$ 为实体 e_i 的事件主题兴趣向量。

(4) 核心张量 G ，是一个三阶张量的矩阵，用于在拟合过程中解决上述因子矩阵维度不一致的问题，经 Tucker 分解后每一个切片可以分别表示如下：

$$\begin{aligned} \phi_{(1)} &= SG_{(1)}(P \otimes Q)^T \\ \phi_{(2)} &= PG_{(2)}(Q \otimes S)^T \\ \phi_{(3)} &= QG_{(3)}(P \otimes S)^T \end{aligned} \quad (4)$$

由上可以看出, 关系矩阵 \mathbf{S} 、 \mathbf{P} 和 \mathbf{Q} 分别以 M 、 N 和 K 维的二元向量进行表示, 假设将每个关系矩阵对应的向量 w_{ij} 、 c_{ij} 和 d_{ij} 中的第 i 、 j 和 k 个元素分别设为 1, 其它元素设为 0, 则核心张量 \mathbf{G} 在第 ijk 元素上为 1, 因此, 三个关系矩阵向量表示与核心张量的乘积, 可通过利用逻辑函数准确地表示原始正交张量, 证明 Tucker 分解能够捕获所有向量数据中的潜在信息, 使得线性模型简单、直观, 并对所有嵌入维度的边界数据全面表达^[27]。

基于此, 利用张量分解能够将时空主题复合关系网络模型转换成矩阵和向量形式进行完整的表达, 张量的维度与地理环境实体、地理位置和事件主题的数量相关, 从而能够利用因子矩阵的多维向量联合聚类方法挖掘时空主题结构。

2.3 基于块值分解的多维关系联合聚类方法

多维关系联合聚类方法是在时空主题复合关系网络下, 通过其分解后形成的多维关系矩阵的行聚类和列聚类两个步骤进行迭代循环, 直至收敛, 确定地理环境实体的时空主题划分。由于时空主题复合关系张量是一个三阶张量矩阵, 将其看作由一个行向量与两个列向量矩阵构成, 因而, 行聚类是指在行向量上实现实体关系的主题划分, 即地理环境实体所关联时空关系的聚类, 具体方法是, 将实体关系划分为 R_1 个主题, 令

矩阵 $\mathbf{X} = \{\alpha_{ij}\}_{M \times R_1}$ 表示为 M 行实体关系与

R_1 个时空主题之间的关系值, 其中 α_{ij} 满足

$$\begin{cases} \text{若实体关系 } i \text{ 只属于主题 } j, \text{ 则 } \alpha_{ij} = 1 \\ \text{若实体关系 } i \text{ 不属于任何主题, 则 } \alpha_{ij} = 0 \\ \text{若实体关系 } i \text{ 分布于主题 } j, \text{ 则 } \alpha_{ij} \in [0, 1] \text{ 且 } \sum_j^{R_1} \alpha_{ij} = 1 \end{cases}$$

。列聚类是指在列向量上分别实现地理环境实体在区域空间上的聚类, 以及事件主题上的聚类, 具体方法是, 将地理位置空间区域划分为 R_2 个簇, 令矩阵 $\mathbf{Y} = \{\beta_{ij}\}_{R_2 \times N}$ 表示

N 列地理位置与 R_2 个簇区域的关系, 并满

足 $\beta_{ij} \in [0, 1]$ 且 $\sum_i^{R_2} \beta_{ij} = 1 (j \in N)$; 将事件

主题标签聚类设为 R_3 个簇, 令矩阵

$\mathbf{Z} = \{\gamma_{ij}\}_{R_3 \times K}$ 表示 K 列主题标签与 R_3 个簇

主题域的关系, 并满足 $\gamma_{ij} \in [0, 1]$ 且

$\sum_i^{R_3} \gamma_{ij} = 1 (j \in K)$ 。上述矩阵 \mathbf{X} 、 \mathbf{Y} 和 \mathbf{Z} 分

别表示了时空主题复合关系网络下, 地理环境实体-实体关系层、地理环境实体-地理位置关系层与地理环境实体-事件主题关系层对应聚类关系的特征向量空间。

由于时空主题复合关系每一阶张量都与上述特征向量空间相关, 为了融合聚类生成的向量空间, 利用复合关系矩阵建立特征向量矩阵之间的关联关系, 以及利用核心张量矩阵解决拟合过程中矩阵维度不一致的问题。块值分解的核心思想就是将二维数据矩阵分解为三个分量, 即行系数矩阵、列系数矩阵和块值矩阵, 系数矩阵表示其相应的簇作为维度的矩阵, 块值矩阵用于调和数据结构的紧凑和显式, 将行聚类与列聚类交织迭代, 解决矩阵降维和稀疏问题。

由此, 对应地理环境实体-实体关系矩阵 \mathbf{S} , 引入核心张量矩阵 $\mathbf{A} = \{\varepsilon_{ij}\}_{R_1 \times R_1}$ 作为

块值矩阵, 表示实体时空主题 i 与时空主题 j 之间的关系紧密程度, 并满足 $\varepsilon_{ij} \in [0, 1]$,

其中 ε_{ij} 越大, 表示两者关系越紧密; 同理,

对应地理环境实体-地理位置关系矩阵 \mathbf{P} ,

引入核心张量矩阵 $\mathbf{B} = \{\eta_{ij}\}_{R_2 \times R_2}$ 作为块值

矩阵，表示时空主题 i 与空间区域簇 j 之间的关系紧密程度，并满足 $\eta_{ij} \in [0,1]$ ，其中 η_{ij} 越大，表示两者关系越紧密；对应地理环境实体-事件主题关系矩阵 Q ，引入核心张量矩阵 $C = \{\mu_{ij}\}_{R_1 \times R_3}$ 作为块值矩阵，表示时空主题 i 与事件主题簇 j 之间的关系紧密程度，并满足 $\mu_{ij} \in [0,1]$ ，其中 μ_{ij} 越大，表示两者关系越紧密。

综合上述矩阵，基于块值分解的多维关系联合聚类方法构成的目标函数 F 最小代价形式表示为：

$$\min F = \min_{A,B,C,X,Y,Z} \left(\begin{aligned} &\|S - XAX^T\|^2 + \|P - XBY\|^2 + \\ &\|Q - XCZ\|^2 \end{aligned} \right) \quad (5)$$

其中， $A, B, C, X, Y, Z \geq 0$ 且 $R_1 \ll M, R_2 \ll N, R_3 \ll K$ ， $\|\bullet\|$ 表示 F-范数。

目标函数 F 每一项分别关联时空主题复合关系分解后的矩阵，第 1 项关联地理环境实体-实体关系矩阵，其中矩阵 X 表达了实体时空关系的主题划分；第 2 项关联地理环境实体-地理位置关系矩阵，其中矩阵 Y 表达

了实体与地理位置空间区域簇的关联性；第 3 项关联地理位置-事件主题关系矩阵，其中矩阵 Z 表达了实体与事件主题域簇的关联性，后两项关系矩阵聚类结果均受到第 1 项关系矩阵聚类结果的约束。

为了求得该目标函数最小解，通过使用一组乘法更新规则对函数进行交互优化迭代更新求解来达到收敛的目的，从而获取其局部最小值。设 $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$

分别为约束矩阵 A, B, C, X, Y, Z 的拉格朗日乘子，且满足：

$$\lambda_1 \in \mathfrak{R}^{R_1 \times R_1}, \lambda_2 \in \mathfrak{R}^{R_1 \times R_2}, \lambda_3 \in \mathfrak{R}^{R_1 \times R_3}, \lambda_4 \in \mathfrak{R}^{M \times R_1}, \\ \lambda_5 \in \mathfrak{R}^{R_2 \times N}, \lambda_6 \in \mathfrak{R}^{R_3 \times K}$$

由此，可将目标函数转换成拉格朗日函数进行求解：

$$La = F - \text{tr}(\lambda_1 A^T) - \text{tr}(\lambda_2 B^T) - \text{tr}(\lambda_3 C^T) - \text{tr}(\lambda_4 X^T) \\ - \text{tr}(\lambda_5 Y^T) - \text{tr}(\lambda_6 Z^T) \quad (6)$$

其中， $\text{tr}(\bullet)$ 为矩阵的迹求解。

然后，对式 (6) 进行求偏导并按照 Kuhn-Tucker 条件分别对 A, B, C, X, Y, Z 求 Hadamard 积 (\odot)，即：

$$\left. \begin{aligned} \frac{\partial La}{\partial A} &= 0 \\ \frac{\partial La}{\partial B} &= 0 \\ \frac{\partial La}{\partial C} &= 0 \\ \frac{\partial La}{\partial X} &= 0 \\ \frac{\partial La}{\partial Y} &= 0 \\ \frac{\partial La}{\partial Z} &= 0 \end{aligned} \right\} \quad (7)$$

$$\left. \begin{aligned} \lambda_1 \odot A &= 0 \\ \lambda_2 \odot B &= 0 \\ \lambda_3 \odot C &= 0 \\ \lambda_4 \odot X &= 0 \\ \lambda_5 \odot Y &= 0 \\ \lambda_6 \odot Z &= 0 \end{aligned} \right\} \quad (8)$$

由此求得：

$$\left. \begin{aligned} (\mathbf{X}^T \mathbf{S} \mathbf{X}) \odot \mathbf{A} - (\mathbf{X}^T \mathbf{X} \mathbf{A} \mathbf{X}^T \mathbf{X}) \odot \mathbf{A} &= 0 \\ (\mathbf{X}^T \mathbf{P} \mathbf{Y}^T) \odot \mathbf{B} - (\mathbf{X}^T \mathbf{X} \mathbf{B} \mathbf{Y} \mathbf{Y}^T) \odot \mathbf{B} &= 0 \\ (\mathbf{X}^T \mathbf{Q} \mathbf{Z}^T) \odot \mathbf{C} - (\mathbf{X}^T \mathbf{X} \mathbf{C} \mathbf{Z} \mathbf{Z}^T) \odot \mathbf{C} &= 0 \\ (\mathbf{S}^T \mathbf{X} \mathbf{A} + \mathbf{S} \mathbf{X} \mathbf{A}^T + \mathbf{P} \mathbf{Y}^T \mathbf{B}^T + \mathbf{Q} \mathbf{Z}^T \mathbf{C}^T) \odot \mathbf{X} - (\mathbf{X} \mathbf{A}^T \mathbf{X}^T \mathbf{X} \mathbf{A} + \mathbf{X} \mathbf{A} \mathbf{X}^T \mathbf{X} \mathbf{A}^T + \mathbf{X} \mathbf{B} \mathbf{Y} \mathbf{Y}^T \mathbf{B}^T + \mathbf{X} \mathbf{C} \mathbf{Z} \mathbf{Z}^T \mathbf{C}^T) \odot \mathbf{X} &= 0 \\ (\mathbf{B}^T \mathbf{X}^T \mathbf{P}) \odot \mathbf{Y} - (\mathbf{B}^T \mathbf{X}^T \mathbf{X} \mathbf{B} \mathbf{Y}) \odot \mathbf{Y} &= 0 \\ (\mathbf{C}^T \mathbf{X}^T \mathbf{Q}) \odot \mathbf{Z} - (\mathbf{C}^T \mathbf{X}^T \mathbf{X} \mathbf{C} \mathbf{Z}) \odot \mathbf{Z} &= 0 \end{aligned} \right\} \quad (9)$$

在此基础上，得到以下 A, B, C, X, Y, Z 的迭代更新规则：

$$A_{ij} \leftarrow A_{ij} \frac{(\mathbf{X}^T \mathbf{S} \mathbf{X})_{ij}}{(\mathbf{X}^T \mathbf{X} \mathbf{A} \mathbf{X}^T \mathbf{X})_{ij}} \quad (10)$$

$$B_{ij} \leftarrow B_{ij} \frac{(\mathbf{X}^T \mathbf{P} \mathbf{Y}^T)_{ij}}{(\mathbf{X}^T \mathbf{X} \mathbf{B} \mathbf{Y} \mathbf{Y}^T)_{ij}} \quad (11)$$

$$C_{ij} \leftarrow C_{ij} \frac{(\mathbf{X}^T \mathbf{Q} \mathbf{Z}^T)_{ij}}{(\mathbf{X}^T \mathbf{X} \mathbf{C} \mathbf{Z} \mathbf{Z}^T)_{ij}} \quad (12)$$

$$X_{ij} \leftarrow X_{ij} \frac{(\mathbf{S}^T \mathbf{X} \mathbf{A} + \mathbf{S} \mathbf{X} \mathbf{A}^T + \mathbf{P} \mathbf{Y}^T \mathbf{B}^T + \mathbf{Q} \mathbf{Z}^T \mathbf{C}^T)_{ij}}{(\mathbf{X} \mathbf{A}^T \mathbf{X}^T \mathbf{X} \mathbf{A} + \mathbf{X} \mathbf{A} \mathbf{X}^T \mathbf{X} \mathbf{A}^T + \mathbf{X} \mathbf{B} \mathbf{Y} \mathbf{Y}^T \mathbf{B}^T + \mathbf{X} \mathbf{C} \mathbf{Z} \mathbf{Z}^T \mathbf{C}^T)_{ij}} \quad (13)$$

$$Y_{ij} \leftarrow Y_{ij} \frac{(\mathbf{B}^T \mathbf{X}^T \mathbf{P})_{ij}}{(\mathbf{B}^T \mathbf{X}^T \mathbf{X} \mathbf{B} \mathbf{Y})_{ij}} \quad (14)$$

$$Z_{ij} \leftarrow Z_{ij} \frac{(\mathbf{C}^T \mathbf{X}^T \mathbf{Q})_{ij}}{(\mathbf{C}^T \mathbf{X}^T \mathbf{X} \mathbf{C} \mathbf{Z})_{ij}} \quad (15)$$

由式 (10) - (15) 可以看出，A, B, C, X, Y, Z 之间的矩阵计算是相互依赖的，因而可设定目标函数 La 收敛阈值为 ρ ，

通过上述矩阵迭代更新规则计算直至 La 收敛，求得联合聚类的结果矩阵 X, Y, Z，然后分别对其行向量或者列向量求取

$\max_{j=1, \dots, R_1} (\alpha_{ij})$, $\max_{i=1, \dots, R_2} (\beta_{ij})$, $\max_{i=1, \dots, R_3} (\gamma_{ij})$ ，从而

确定时空关系主题划分。其具体算法描述如下：

算法 1：基于块值分解的多维关系联合聚类方法

输入：构成时空主题复合关系的 3 类维度关系矩阵 S 、 P 、 Q ，以及时空主题数目 R_1 、空间区域簇数 R_2 、

事件主题簇数 R_3

输出：联合聚类的结果矩阵 X 、 Y 、 Z ，块值矩阵 A 、 B 、 C ，及其构成的时空关系主题归属结构 $\Omega(\alpha, \beta, \gamma)$

1. 初始化参数：

矩阵 X ， Y ， Z 初始化，使之服从均匀分布；矩阵 A ， B ， C 初始化元素均为 1 的矩阵； $F = +\infty$ 。

2. 利用交互优化策略迭代更新各个矩阵的值：

While True:

$F' = F$;

按照公式 (10) - (15)，迭代更新计算 A, B, C, X, Y, Z 的值；

按照公式 (5)，计算 F ；

If $|F' - F| \leq \rho$;

收敛且达到阈值，获得当前各个矩阵的值；

break

3. 基于聚类结果矩阵 X ， Y ， Z 分别计算行向量 $\max_{j=1, \dots, R_1}(\alpha_{ij})$ ，列向量 $\max_{i=1, \dots, R_2}(\beta_{ij})$ ，列向量 $\max_{i=1, \dots, R_3}(\gamma_{ij})$ ，

并分别将隶属度最大的值划分到相应的主题域中， $add(\alpha_i, \beta_j, \gamma_k) \rightarrow \Omega[i, j, k]$ ，从而构成时空关系主题归属结构 $\Omega(\alpha, \beta, \gamma)$ 。

3 实验分析

3.1 实验数据说明

为了验证算法在战场地理环境时空主题发现的有效性，本文选取某战术训练场近

年演习导调中所获得的 4670 篇文档作为文本数据集，如表 1 所示，主要包含了该地区的地理环境实体信息、任务事件信息、实体关系信息以及地理位置信息等数据。

表 1 实验部分数据集片段

Tab.1 Experimental Partial Data Set Fragment

| 文档数据集 | 实体 | 位置 | 事件 | 时间 | 关系三元组 |
|---|-------------------|-----|----------|-----------|--|
| <ent:org>机步 1 连</ent:org>在<ent:loc>公路桥(21, 04)</ent:loc>以<rel:dir>西</rel:dir>选择有利地形，<rel:act>构筑</rel:act>正面宽约 400 米、纵深约 200 米的<ent:fea>反坦克网状阵地</ent:fea>，坚决于<oth:time>5 日凌晨</oth:time><oth:event>阻止敌人沿尖岗、和平里方向突破</oth:event>。 | 机步 1 连 反坦克网状阵地 | 公路桥 | 阻敌 突破 | 5 日 凌晨 | (机步 1 连, 构筑, 反坦克网状阵地) (公路桥, 西, 反坦克网状阵地) |

实验前，上述表 1 数据集经过中文分词、语料标注、实体识别、实体关系抽取等一系列预处理步骤，将文本数据集简化为包含实体与实体关系三元组数据集、地理位置标签

和事件主题标签数据集，结果形成如表 2 所示的数据集统计，并采用 Hash 函数将标签数据集分类映射为区间 $[1, 400]$ 内的地理格

网标识号，去除重复信息后，最终共包含 7144 个地理环境实体，20771 个地理环境实体关系三元组，6214 条地理位置标签和 1167

条事件主题标签，本文后续实验将在上述数据进行预处理后得到的数据集上开展。

表 2 实验文本数据集的参数

Tab.2 Parameters of Experimental Text Data Set

| 数据集 | 地理环境实体数 | 地理环境实体关系三元组数 | 位置标签数 | 事件主题标签数 | 采集时间 |
|-----|---------|--------------|-------|---------|------------|
| A1 | 7095 | 17029 | 5682 | 1156 | 2015 年 3 月 |
| A2 | 7229 | 30363 | 7341 | 1197 | 2015 年 4 月 |
| A3 | 7201 | 20885 | 6750 | 1179 | 2015 年 7 月 |
| A4 | 7051 | 14808 | 5083 | 1138 | 2015 年 9 月 |

3.2 实验设计与验证

本文以时空主题复合关系网络模型为基础进行主题挖掘，从共词网络构建、主题标签聚类以及主题结构特征关联等方面设计实验流程，并给出验证方法，具体技术路线如下：

(1) 共词网络构建

对文本数据集进行预处理后得到的一系列实体关系三元组数据集、地理位置标签和事件主题标签数据集，按照 Overlap 计算方法技术相关关键词在数据集中共现频率，得到以表示地理环境实体、位置和事件主题特征的关键词为节点共词关系网络，如图 3 所示。

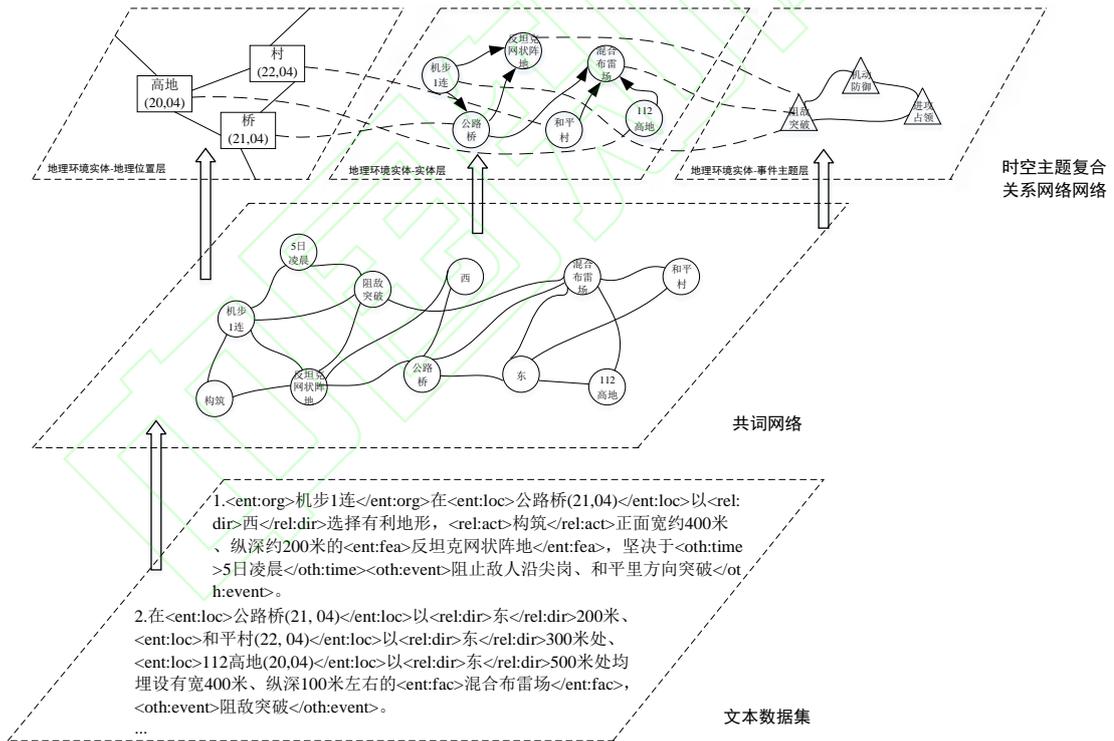


图 3 基于共词网络的时空主题复合关系网络示意图

Fig.3 Scheme Diagram of Spatiotemporal Topic Compound Relationship Network based on CO-word Network

(2) 主题标签聚类

利用基于块值分解的多维关系联合聚类方法在数据集上获得不同规模的时空关系主题划分，按照对其中所包含的关键词频率分布，将其采用词云图的方式，词汇的大

小表示不同时空主题与地理位置及事件之间关联的紧密度，如图 4 所示。

$F1=2PR/(P+R)$ 。

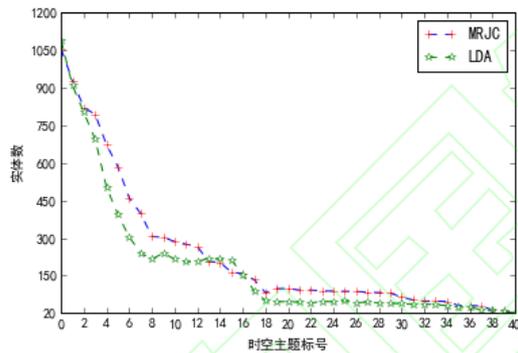
测试结果如表 3 所示，训练集中 4546 个地理环境实体，12606 条实体关系被正确聚类成时空主题结构，准确率分别为 88.4% 和 86.9%；测试集中 1105 个地理环境实体，3083 条实体关系被正确聚类成时空主题结构，准确率分别为 87.3% 和 85.8%。如图 5 所示，不同主题结构下所聚类的实体和标签数量，随着主题规模的减小而平缓下降，统

计结果得到最大时空主题中共有 1054 个地理环境实体、317 条地理位置标签和 97 条事件主题标签，出现最多的主题标签是机动、进攻和阻击，最多的位置标签是高地、公路和村庄。同时，多维关系联合聚类方法（MRJC）与 LDA 算法相比较，可以看出本文算法挖掘到的实体数与标签数总体上高于 LDA 算法，从而可以看出融合了多维关系后能够得到准确、清晰的时空主题结构。

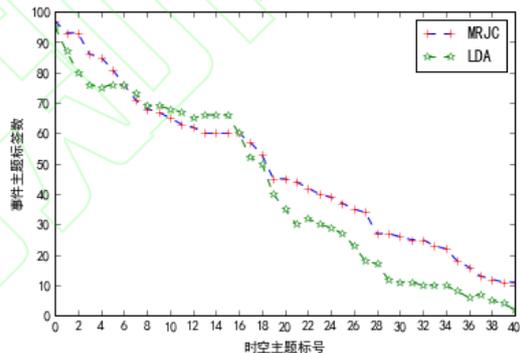
表 3 基于多维关系的时空主题分类精度

Tab.3 Accuracy of Spatiotemporal Topic Classification Based on Multi-Dimensional Relationship

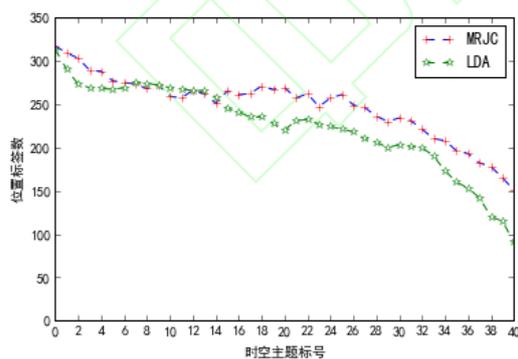
| 聚类类型 | 训练样本 | | | | 测试样本 | | | |
|----------|-------|-------|-------|-------|------|-------|-------|-------|
| | 正确数量 | 准确率 | 召回率 | F1 值 | 正确数量 | 准确率 | 召回率 | F1 值 |
| 地理环境实体 | 4546 | 88.4% | 79.5% | 83.7% | 1105 | 87.3% | 77.3% | 82.0% |
| 地理环境实体关系 | 12606 | 86.9% | 75.9% | 81.0% | 3083 | 85.8% | 74.2% | 79.6% |



(a) 实体数统计



(c) 事件主题标签统计



(b) 位置标签统计

图 5 不同时空主题规模下实体和标签数量统计
Fig.5 Statistics of The Number of Entities and Tags in Different Spatiotemporal Topic Scale

3.3.2 时空主题结构特征分析

多维关系聚类的块值矩阵 A ， B ， C 反映了地理环境时空主题结构的内部特征关系。如图 6 所示，时空主题之间的关联关系由块值矩阵 A 得到，横轴与纵轴均为主题标号，坐标的灰度表示两者关系的紧密程度，颜色越深对应关系越紧密，可以看出大多数色块分布以坐标系对角线成对称分布，表明时空主题内部具有较强的内聚性，时空主题之间关联较为紧密；如图 7 所示，时空主题与地理空间区域之间的关联关系由块值矩阵 B 得到，横轴表示空间区域簇号，纵轴为时空主题标号，坐标网格的灰度表示两者关

系的紧密程度，颜色越深对应关系越紧密，可以看出两者关系既包含一对一的关系，还包含一对多的关系，表明同一时空主题在不同空间区域体现不同的关联程度，同一空间区域对不同时空主题的影响程度不同；如图8所示，时空主题与事件主题之间的关联关系由块值矩阵C得到，横轴表示事件主题簇号，纵轴为时空主题标号，坐标网格的灰度表示两者关系的紧密程度，颜色越深对应关系越紧密，表明时空主题结构具有多事件主题标签特征，反映了与事件关联的紧密性。

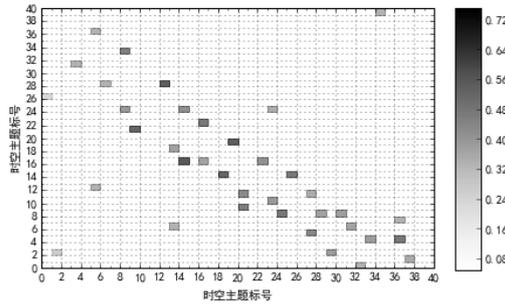


图6 时空主题之间的关联

Fig.6 The Relationships between Different Spatiotemporal Topics

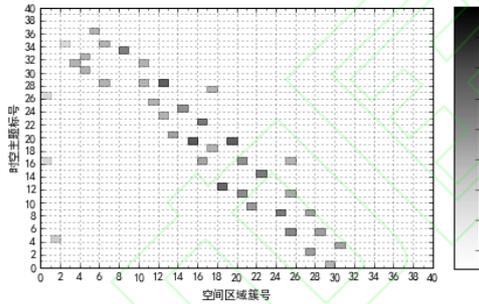


图7 时空主题与地理空间区域之间的关联

Fig.7 The Relationships between Spatiotemporal Topics and Geospatial Regions

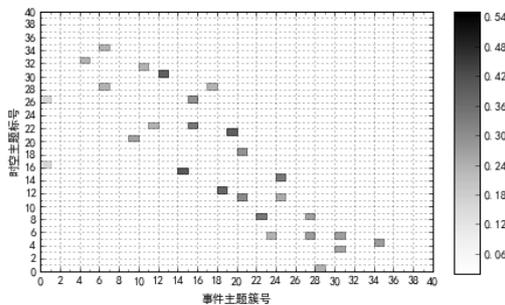


图8 时空主题与事件主题之间的关联

Fig.8 The Relationships between Spatiotemporal Topics and Event Topics

利用时空主题结构特征分析主题空间分布情况，可以有效发现事件发展的时空规律，有助于分析战场地理环境对作战行动的效能影响规律。本文选取某一个时间片段，获得当前文本数据集的时空主题结构相关联的地理环境实体、位置标签和事件主题标签等信息，如图9所示，采用OPTICS聚类方法对实体关联的位置信息进行聚类分析，每个实体以点状符号表示，可以看出在区域1、2、3出现热点区域，位置标签信息、事件主题标签信息集中区域与所关注的实体空间分布聚集区域相一致，这表明该区域的作战行动受到地理环境实体的影响程度较大，为进一步挖掘事件发展规律提供基础。

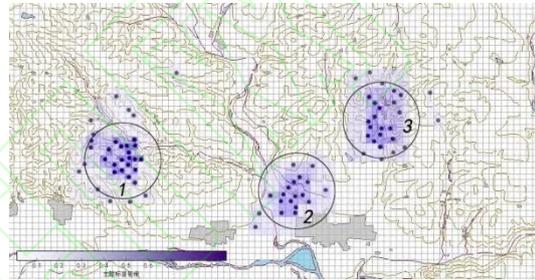


图9 某时间片段时空主题结构空间分布

Fig.9 Spatial Distribution of Spatiotemporal Topics Structure in A Certain Time Segment

4 结论

战场文本作为战场信息交互的重要媒介，包含了以时间、空间信息为基础的文本语义信息，对其进行深入挖掘，可以用来分析战场环境对作战行动的影响效能，从而揭示战场事件发展的时空规律，为指挥决策机构提供重要参考依据。本文在实体及实体关系抽取的基础上，构建以地理环境实体、地理位置与事件主题为节点的主题结构关系网络，利用张量模型的 Tucker 分解建立主题结构关系的完全表达式，并运用块值矩阵分解方法对多维关系进行联合聚类计算，以此获取时空主题分类及其内聚结构。实验结果表明，该方法能够有效提高时空主题发现的质量和效率，使得到的主题能够更好地表现出地理环境要素、地理位置与事件主题之间内聚的关联性，为能够清晰地反映时空主题的演化过程提供基础，并为掌握事件发展态势提供支持。由于本文在关系矩阵构建上

仅根据实体词共现频率作为权重值,使得数据分析存在一定偏差,下一步将结合注意力机制深挖多源文本数据,提高数据分析的效率与精度,在有效发现具备内聚性的热点区域的基础上,能够建立不同热点区域间的时空关联性,推断事件变化过程,为战场环境的动态推演提供重要参考价值。

参考文献

[1] Jo Y, Oh A. Aspect and Sentiment Unification Model for On-line review analysis[C]. ACM International Conference on Web Search and Data Mining. ACM,2011:815-824.

[2] Duan Lian, Guo Wei, Zhu Xinyan, et al. Constructing Spatio-Temporal Topic Model for Microblog Topic Retrieving[J]. *Geomatics and Information Science of Wuhan University*, 2014,39(2):210-213,243. (段炼, 吕维, 朱欣焰, 等. 基于时空主题模型的微博主题提取[J]. 武汉大学学报·信息科学版, 2014,39(2):210-213,243.)

[3] Li Luping, Zhao Xiaobing. Review on Topic Discovery Methods Based on Text Clustering[J]. *Information Research*,2020,(11):121-127. (李璐萍, 赵小兵. 基于文本聚类的主题发现方法研究综述[J]. 情报探索,2020,(11):121-127.)

[4] Saracoglu R, Tutuncu K, Allahverdi N. A Fuzzy Clustering Approach for Finding Similar Documents Using a Novel Similarity Measure[J]. *Information Processing and Management*,2006,6(2):600-605.

[5] Zhang Huibing, Zhong Hao, Hu Xiaoli. User Reviews Clustering Method Based on Topic Analysis[J]. *Computer Science*,2019,46(8):50-55. (张会兵, 钟昊, 胡晓丽. 基于主题分析的用户评论聚类方法[J]. 计算机科学,2019,46(8):50-55.)

[6] Chen You, Cheng Xueqi, Yang Sen. Finding High Quality Threads in Web Forums[J]. *Journal of Software*, 2011,22(8):1785-1804. (陈友, 程学旗, 杨森. 面向网络论坛的高质量主题发现[J]. 软件学

报,2011,22(8):1785-1804.)

[7] Resch B, Florian U, Havas C. Combining Machine-Learning Topic Models and Spatiotemporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment[J]. *Cartography and Geographic Information Science*,2018,45(4):362-376.

[8] Wen Congcong, Peng Ling, Yang Lina, et al. Topic Model Combined with the SVM for Small Scale Land Use Classification[J]. *Journal of Geo-information Science*,2018,20(2):167-175. (文聪聪, 彭玲, 杨丽娜, 等. 主题模型与 SVM 组合的小尺度街区用地分类方法[J]. 地球信息科学学报, 2018,20(2):167-175.)

[9] Liu Jingjing, Liu Yusi, Yi Disheng, et al. Extracting Mixed Topic Patterns within Downtown Beijing at the Block Level[J]. *Journal of Geo-information Science*,2020,22(6):1370-1382. (刘菁菁, 刘雨思, 伊迪升, 等. 北京市四环内街区尺度下的主题混合模式挖掘[J]. 地球信息科学学报,2020,22(6):1370-1382.)

[10] Han J, Kamber M. Data Mining Concepts and Techniques[M]. San Francisco: Morgan Kaufmann Publisher,2006.

[11] Li Hailin, Wu Xianli. Research on Topic Discovery and Evolution Based on Time Series Clustering[J]. *Journal of the China Society for Scientific and Technical Information*,2019,38(10):1041-1050. (李海林, 邬先利. 基于时间序列聚类的主题发现与演化分析研究[J]. 情报学报,2019,38(10):1041-1050.)

[12] Wang Yandong, Li Mengmeng, Fu Xiaokang, et al. A New Method to Detect the Development Situation of Disasters Based on Social Media Co-word Network[J]. *Geomatics and Information Science of Wuhan University*, 2020,45(5): 691-699. (王艳东, 李萌萌, 付小康, 等. 基于社交媒体共词网络的灾情发展态势探测方法[J]. 武汉大学学报·信息科学版,2020,45(5):691-699.)

[13] Wang Yandong, Fu Xiaokang, Li

- Mengmeng. A New Social Media Topic Mining Method Based on Co-word Network[J]. *Geomatic and Information Science of Wuhan University*, 2018,43(12):2287-2294. (王艳东, 付小康, 李萌萌.一种基于共词网络的社交媒体数据主题挖掘方法[J]. 武汉大学学报·信息科学版,2018,43(12):2287-2294.)
- [14] Shao Qi, Mu Dongmei, Wang Ping, et al. Identifying Subjects of Online Opinion from Public Health Emergencies[J]. *Data Analysis and Knowledge Discovery*, 2020, 4(9): 68-80. (邵琦, 牟冬梅, 王萍, 等.基于语义的突发公共卫生事件网络舆情主题发现研究[J]. 数据分析与知识发现,2020,4(9):68-80.)
- [15] Zhou Xiaoping, Liang Xun, Zhang Haiyan. User Community Detection on Micro-blog Using R-C model[J]. *Journal of Software*, 2014,25(12):2808-2823. (周小平, 梁循, 张海燕.基于 R-C 模型的微博用户社区发现 [J]. 软件学报,2014,25(12):2808-2823.)
- [16] Gong Weihua, Chen Yanqiang, Pei Xiaobing, et al. Community Detection of Multi-dimensional Relationships in Location-based Social Networks[J]. *Journal of Software*, 2018,29(4):1163-1176. (龚卫华, 陈彦强, 裴小兵, 等.LBSN 中融合多维关系的社区发现方法 [J]. 软件学报,2018,29(4):1163-1176.)
- [17] Yang L, Cao X, Jin D, Wang X, Meng D. A Unified Semi-supervised Community Detection Framework Using Latent Space Graph Regularization. *IEEE Trans. on Cybernetics*, 2015,45(11):2585-2598.
- [18] Long B, Zhang Z, Yu P S. Co-Clustering by block value decomposition[C]. *Proceeding of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining*. New York: ACM Press, 2005, 635-640.
- [19] Gao Jun, Wan Gang. *Theory and Technology of Battlefield Environment Engineering*[M]. Beijing: PLA Press,2016. (高俊, 万刚 等. 战场环境工程理论和技术[M]. 北京: 解放军出版社, 2016.)
- [20] Zhou Chenghu. Prospects on Pan-spatial Geographic Information System[J]. *Process in Geography*, 2015,34(2):129-131. (周成虎.全空间地理信息系统展望[J].地理科学进展,2015,34(2):129-131.)
- [21] Hua Yixin. The Core Problems and Key Technologies of Pan-spatial Information System[J]. *Journal of Geomatics Science and Technology*, 2016,33(4):331-335. (华一新.全空间信息系统的核心问题和关键技术[J].测绘科学技术学报, 2016,33(4):331-335.)
- [22] Wang Jianjian, Wang Yannan, Zhou Liangchen, et al. The Classification System and Expression Model of the Relationship of Spatio-temporal Object of Multi-granularity[J]. *Journal of Geo-information Science*, 2017,19(9):1164-1170. (王健健, 王艳楠, 周良辰, 等.多粒度时空对象关联关系的分类体系与表达模型[J].地球信息科学学报,2017,19(9):1164-1170.)
- [23] Zhang Zheng, Hua Yixin, Zhang Xiaonan, et al. The Basic Issues of Associative Relationship of Spatial-temporal Objects of Multi-granularity[J]. *Journal of Geo-information Science*, 19(9):1158-1163. (张政, 华一新, 张晓楠, 等.多粒度时空对象关联关系基本问题初探[J].地球信息科学学报,2017,19(9):1158-1163.)
- [24] Li Rui, Shi Jiahao, Dong Guangsheng, et al. Research on Expression of Multi-granularity Spatio-temporal Object Composition Structure[J]. *Journal of Geo-information Science*, 2021,23(1):113-123. (李锐, 石佳豪, 董广胜, 等.多粒度时空对象组成结构表达研究[J].地球信息科学学报,2021,23(1):113-123.)
- [25] Zhang Xueying, Zhang Chunju, Du Chaoli. Semantic Relation Between Spatial Relation Terms and Feature Types of Geographical Entities[J]. *Geomatics and Information Science of Wuhan University*, 2012,37(11):1266-1270. (张雪英, 张春菊, 杜超利.空间关系词汇与地理实体要素类型

的语义约束关系构建方法[J].武汉大学学报·信息科学版,2012,37(11):1266-1270.)

[26] Tamara G K, Brett W B. Tensor Decompositions and Applications[J]. SIAM review,2009,51(3):455-500.

[27] Balazevic I, Allen C, Hospedales T. Tucker: Tensor Factorization for Knowledge Graph Completion[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

Language Processing (EMNLP-IJCNLP). 2019: 5188-5197.

[28] Xing Hanfa, Meng Yuan, Hou Dongyang, et al. A Land-Cover Classification Method Using Point of Interest[J]. *Geomatics and Information Science of Wuhan University*,2019,44(5):758-764. (邢汉发, 孟媛, 侯东阳, 等.一种应用兴趣点数据进行地表覆盖分类的方法[J].武汉大学学报·信息科学版,2019,44(5):758-764.)

A Method for Geographical Environment Spatiotemporal Topic Discovery of Multi-dimensional Relationships

ZHU Jie^{1,2} ZHANG Hongjun¹ LIAO Xianglin¹ XU Youwei¹

¹ College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210002, China

² 73021 Troops, Hangzhou 315023, China

Abstract: Objectives: Using battlefield text data for spatiotemporal topic analysis, we can obtain the spatial distribution pattern of geographical environment elements and their impact characteristics on battlefield activities from micro to macro and from scattered to gathering places, and mine the spatial distribution and development law of battlefield events, which further enriches the perception mode of battlefield environment and provides a new means of battlefield environment efficiency analysis. It is of great significance and value for in-depth understanding of battlefield environmental knowledge. **Methods:** The key technology to improve the quality of spatiotemporal topic discovery in geographical environment is to effectively construct entity composite relationship network and integrate multi-dimensional heterogeneous relationships for topic clustering. Firstly, a spatiotemporal topic tensor model integrating multi-dimensional relationships is constructed, and the complete expression of topic relationship is given by using the Tucker decomposition of topic tensor model; then, the feature vector space of multi-dimensional relational clustering is constructed as the objective function of topic classification, and the block valued matrix decomposition technology is used for joint clustering calculation, and the core tensor matrix is used to solve the problem of data sparsity; finally, the block value matrix obtained by multi-dimensional relational clustering is used to obtain the association value between geographical environment elements and spatiotemporal topics. **Results:** The results show that: (1) The geographical environment entities and entity relationships are correctly clustered into spatiotemporal topic structure. The accuracy rates in the training set were 88.4% and 86.9% respectively, and in the test set were 87.3% and 85.8% respectively. (2) The number of entities and tags clustered under different subject structures decreases gently with the

reduction of subject scale. The statistical results show that the most subject tags are maneuver, attack and interception, and the most location tags are Highlands, roads and villages. (3) Compared with LDA Algorithm, the multi-dimensional relationship joint clustering method can be seen that the number of entities and labels mined by this algorithm is generally higher than LDA Algorithm, so it can be seen that the accurate and clear space-time topic structure can be obtained after integrating multi-dimensional relationships. (4) The block value matrix obtained by multi-dimensional relationship clustering reflects the internal characteristic relationship of the spatiotemporal topic structure of the geographical environment, indicating that the spatial-temporal theme has strong cohesion. **Conclusions:** This method can effectively improve the quality and efficiency of spatiotemporal topic discovery, making the obtained topics better show the cohesive correlation between geographical environment elements, geographical location and event topics, providing a basis for clearly reflecting the evolution process of spatiotemporal topics, and supporting for mastering the development trend of events. Since this paper only takes the co-occurrence frequency of entity words as the weight value in the construction of relationship matrix, there is a certain deviation in data analysis. In the future, we will combine the attention mechanism to dig deep into multi-source text data, improve the efficiency and accuracy of data analysis, and establish the temporal and spatial correlation between different hot spots on the basis of effectively discovering cohesive hot spots, for inferring the event change process and providing an important reference value for the dynamic deduction of battlefield environment.

Key words: geographic environment; multi-dimensional relationships; spatiotemporal topic discovery; block valued matrix decomposition; joint clustering

First author: ZHU Jie, Ph.D., specializes in the theories and methods of battlefield environment cognition and battlefield location service. E-mail: zjsoldierlee@163.com

Foundation support: China Postdoctoral Science Foundation Funded Project, No. 2019M664028.

网络首发:

标题: 一种融合多维关系的地理环境时空主题发现方法

作者: 朱杰, 张宏军, 廖湘琳, 徐有为

DOI: 10.13203/j.whugis20210326

收稿日期: 2021-06-21

引用格式:

朱杰, 张宏军, 廖湘琳, 等. 一种融合多维关系的地理环境时空主题发现方法[J]. 武汉大学学报·信息科学版, 2022, DOI: 10.13203/j.whugis20210326 (ZHU Jie, ZHANG Hongjun, LIAO Xianglin, et al. A Method for Geographical Environment Spatiotemporal Topic Discovery of Multi-dimensional Relationships[J]. *Geomatics and Information Science of Wuhan University*, 2022, DOI: 10.13203/j.whugis20210326)

网络首发文章内容和格式与正式出版会有细微差别, 请以正式出版文件为准!

您感兴趣的其他相关论文:

面向仿真事件的战场地理环境时空过程建模

朱杰, 张宏军

武汉大学学报·信息科学版, 2020, 45(9): 1367-1377, 1437

<http://ch.whu.edu.cn/cn/article/doi/10.13203/j.whugis20200175>

中国知网