

基于 TF-IDF 和 SGTM 模型的标准技术地图构建及实证研究

修乙丹 郑素丽 王鹤媛

(中国计量大学经济与管理学院)

摘 要: 在关键领域技术研发加快和标准竞争加剧的背景下,亟需一种有效的工具对标准的技术内容进行深入分析,为标准创新活动提供决策支持。本文将文本挖掘方法和知识地图绘制思路引入标准化领域,提出一种基于TF-IDF和SGTM模型的标准技术地图构建方法,实现了对标准文本内容的挖掘、聚类与可视化呈现,可以全面细致地揭示一个技术领域的标准布局现状和特点。在此基础上以V2X技术为例进行实践应用,验证了该方法的可行性和有效性,也为V2X标准研究提供了理论指导和政策启示。

关键词: 标准技术地图, TF-IDF, SGTM, V2X

DOI编码: 10.3969/j.issn.1674-5698.2023.10.002

Standards Technological Map Construction and Empirical Research Based on TF-IDF and SGTM Models

XIU Yi-dan ZHENG Su-li WANG He-xuan

(School of Economics and Management, China Jiliang University)

Abstract: In the context of accelerated technical R&D in key areas and intensified standards competition, there is an urgent need for an effective tool to deeply analyze the technical content of standards and provide decision-making support for standards innovation activities. This paper introduces text mining methods and knowledge map drawing ideas into the field of standardization, and proposes a method of standards technological map construction based on TF-IDF and SGTM models. This method realizes the mining, clustering, and visual presentation of standards texts, comprehensively and meticulously revealing the current status and characteristics of the standards layout in a technical field. This paper also demonstrates the practical application using V2X technology as an example, verifies the feasibility and effectiveness of the method, and provides theoretical guidance and policy inspiration for V2X standards research.

Keywords: standards technological map, TF-IDF, SGTM, V2X

0 引言

随着第四次工业革命的推进和数字经济的加

速发展,标准作为产业技术基础和国家治理工具的作用更为突出,各国纷纷加快了关键领域的标准布局。2022年2月,欧盟委员会首次发布欧盟层面的

基金项目: 本文受国家社科基金一般项目(项目编号: 21BGL004),浙江省哲学社会科学重大项目(项目编号: 21QNRC15ZD)资助。

作者简介: 修乙丹,硕士研究生,研究方向为标准化与技术创新。

郑素丽,博士,教授,研究方向为技术创新、知识产权和标准化管理。

王鹤媛,硕士研究生,研究方向为标准化与技术创新。

标准化战略^[1], 重点关注弹性、绿色和数字化等领域的标准化战略需求, 试图重塑其标准“先行者”地位。2023年5月, 美国白宫发布《美国关键和新兴技术的国家标准战略》^[2], 旨在通过对关键和新兴技术国际标准的制定, 进一步强化美国在这些领域的创新和竞争力。在此背景下, 如何对关键领域的技术标准布局进行有效的跟踪、分析和研判, 以更好地应对日益激烈的标准竞争挑战成为一个重要的研究课题。

标准是通过标准化活动, 按照规定的程序协商一致制定, 为各种活动或其结果提供规则、指南或特性, 供共同使用和重复使用的文件^[3]。标准文献作为标准信息最有效的载体, 往往包含了名称、起草人、起草单位、术语定义、应用范围、技术指标等内容, 可以全面反映所在领域的标准布局和发展方向^[4]。因而, 近年来标准文献的分析成为标准化领域的一个研究热点。不少学者针对特定行业或领域的标准起草人^[5]、起草单位^[6]、技术领域^[7]、地域分布^[8]等展开分析, 部分学者利用社会网络分析方法对起草人合作网络进行了初步的探索^[9], 还有一些学者利用大数据方法对标准关键词等内容进行了可视化呈现^[10-12]。上述研究丰富了我们对于标准信息理解和利用, 然而以上研究主要是针对浅层次的结构化文献信息的分析, 并没有对包含核心技术内容的标准文件本身进行深入挖掘, 因此对标准所规定的技术内容的解释是极为有限的, 难以满足我们对国内外标准技术布局的分析需求。

事实上, 标准文本包含了丰富的技术信息, 然而这些信息本身不是以包装好的形式呈现的, 需要通过一定的技术手段挖掘后才能加以识别应用。近年来, 文本挖掘技术如: 文本分词、摘要提取、主题识别、信息地图绘制等快速发展^[13, 14]。这些基于文本的处理技术可以针对数量大、高维空间中的数据, 批量挖掘并概括大量隐藏在冗长文本内部的潜在信息^[15], 有效地进行技术趋势分析和技术领域的划分, 在科学计量、专利计量等领域得到了广泛的应用^[16]。然而, 目前利用文本挖掘方法来呈现技术标准信息的研究还很少, 笔者检索到的仅有 Jeong & Yoon和吴菲菲等的研究。Jeong & Yoon

在对标准和专利文本数据进行挖掘的基础上来识别潜在的标准必要专利^[17], 吴菲菲等则通过对标准文本的挖掘识别标准技术空白^[18]。上述研究将文本挖掘技术引入标准文本领域, 丰富了标准化领域的研究方法。然而, 上述研究的重点均聚焦在特定技术标准空白点的识别上, 而没有对整个技术标准构成和布局进行有效呈现。

基于上述背景, 本文尝试将文本挖掘方法和知识地图绘制的思路引入标准化领域, 提出一种基于TF-IDF关键词提取和SGTM模型的标准技术地图构建方法, 并以V2X技术为例进行实证应用。研究结果显示, 该方法不仅可以全面系统地呈现技术标准分布地图, 识别主要的技术标准研发方向, 还可以通过密度信息将标准研发的重点进行可视化展示, 从而为产业和企业标准布局提供理论依据和实践指导。

1 基于TF-IDF和SGTM模型的标准技术地图构建方法

1.1 总体研究框架

本研究从标准文本信息入手, 通过TF-IDF关键词提取和SGTM方法实现标准关键信息的挖掘和映射, 具体包括标准文本收集、关键词提取、标准-关键词矩阵构建、标准技术地图绘制和分析5个步骤(如图1所示): (1) 依据研究领域选定标准数据库, 构建检索表达式收集技术标准文本。为了提高文本挖掘分析结果的可靠性, 对标准文本数据进行清洗, 剔除相关度较低的文本。由于一项标准通常涉及若干不同的主题, 根据涉及的技术主题不同对标准文本进行分割。(2) 利用TF-IDF算法结合领域专家判断提取关键词。(3) 构造词(关键词)-篇(标准文本)矩阵以表示标准文本信息, 作为下一步标准地图的输入信息。(4) 绘制标准地图。通过SGTM实现多维标准数据在二维地图空间的映射, 利用SGTM的聚类功能对标准地图分布进行有效展示, 通过标准密度信息识别标准热点和空白点。(5) 通过SGTM的逆映射函数来完成标准热点和空白点解释, 系统分

析标准技术发展趋势。

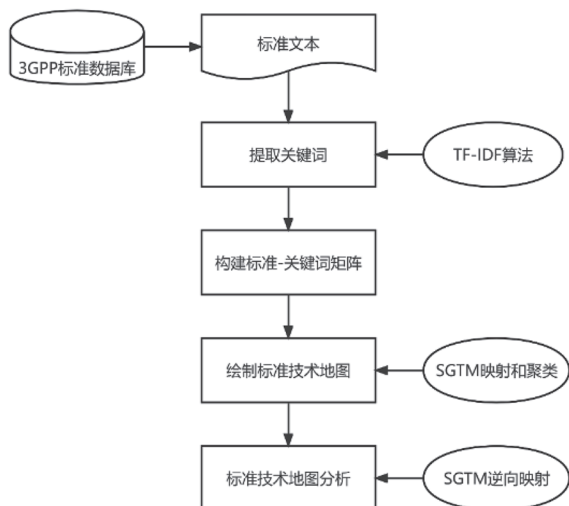


图1 标准技术地图构建思路 and 流程

1.2 主要研究方法

(1) 基于TF-IDF的关键词提取方法

TF-IDF (term frequency-inverse document frequency) 是一种基于概率统计文本的方法,通过分析每个文本的词频-逆文档频率来衡量一个词语的权重,是信息检索与数据挖掘的常用加权技术^[19]。在数据集中,某一字词的重要性随着其在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降,若该字词仅在少数文档中出现,且该词在文档中被频繁使用,则能很好地代表文档的主题。TF-IDF的计算公式为:

$$TF-IDF(t, d) = TF(t, d) \times IDF(t, D)$$

其中 $TF(t, d)$ 是词频 (Term Frequency), 表示词 t 在文档 $d \in D$ 中出现的次数, $IDF(t, D)$ 是逆文档频率 (Inverse Document Frequency), 用于度量词 t 在文档语料库 D 中的稀缺性, 计算公式为:

$$IDF(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|}$$

其中 N 是语料库 D 的文档总数, 分母是包含术语 t 的文档数量, IDF 会降低常用词语的权重, 增加文档集中很少使用词语的权重。借助TF-IDF算法可得到某一领域标准技术初步的关键词集合, 在此基础上结合领域专家判断从而确定关键词列表。

(2) 基于SGTM的标准地图可视化与聚类方法

由于关键词矩阵构成的是高维稀疏空间, 我们需要从众多关键词中求出那些对分类最有效的特征, 从而实现特征空间维数的压缩, 这个过程被称为特征提取, 也称降维。目前的降维方法一般包括线性和非线性两类: 线性降维算法主要有主成分分析 (PCA)、多维标度法 (MDS) 及独立成分分析 (ICA); 非线性降维算法主要有自组织映射网络 (SOM), 生成拓扑映射 (GTM), 自编码神经网络 (ANN) 和线性混合模型 (MLM) 等。其中GTM (Generative Topographic Mapping) 方法作为一种新的降维方法^[20], 具有存在全局损失函数、收敛性确定、参数自动调整等优点^[18], 成为近几年学者们研究的热点。

GTM可以将多维数据空间映射到低维潜在空间, 这个过程通过径向基函数神经网络来实现。图2为二维潜在变量空间与三维数据的映射示意图, 左边的二维潜在变量数据通过非线性函数 $y(x; W)$ 映射至右边三维数据空间上。GTM算法可以依据高维实际数据与反应点之间的依赖关系进行基于潜在空间到数据空间之间的映射变换, 根据文档分散分布的关键词向量, 构建潜在变量分类的数据节点, 生成的每个数据节点都能根据其特征进行映射, 反映原始数据的概率密度, 因而可以绘制包含关键词密度的标准技术分布地图。此外, GTM的逆向映射功能可以实现反向解释专利地图的功能。通过反向绘图功能将地图中的不同位置还原成原始的关键词矢量, 可以对标准地图进行客观而准确的解释, 因此本文选定GTM方法作为标准地图绘制的基本工具。

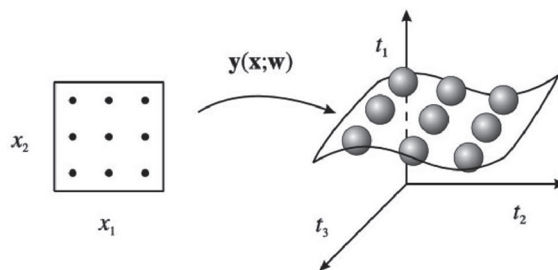


图2 生成式拓扑映射过程示意图

然而, 在一般的GTM映射地图上相邻网格点之间的技术距离是随机分布的, 虽然我们可以通

过逆向映射函数对每一个网格点进行逐一核验,但很难对地图的整体分布结果进行有效的分析和解释。为了克服这一问题,我们借鉴Kaneko的思路将GTM方法与聚类方法进行融合^[21],形成SGTM (Sparse Generative Topographic Mapping)方法,从而同时实现数据可视化和聚类的目的。SGTM的工作原理如下:

在一个 $k \times k$ 的二维地图中,让这两个维度分别为 z_1 和 z_2 ,第 i 个网格点 $z^{(i)}$ 表示为:

$$z^{(i)} = [z_1^{(i)}, z_2^{(i)}]$$

第 j 个数据点的 X 变量表示为:

$$X^{(j)} = [x_1^{(j)}, x_2^{(j)}, \dots, x_m^{(j)}]$$

其中 m 是 X 变量的数量。令 f 表示将 $z^{(i)}$ 转化为 $x(j)$ 的函数,对于所有 i 和 j , $x(j) = f(z^{(i)})$ 。在最初的GTM算法中, $f(z^{(i)})$ 是由 p 个RBFs(径向基函数)的线性组合定义的,但是对于使用神经网络的SGTM模型,RBFs就相当于一个中间层。当第 j 个RBF的中心为 $t^{(j)} = [t_1^{(j)} t_2^{(j)}]$ 时,RBF表示如下:

$$\phi(z^{(i)}) = \exp\left[-\frac{1}{2\sigma^2} \|t^{(j)} - z^{(i)}\|^2\right]$$

其中 σ 是RBF的方差。 $f(z^{(i)})$ 表示如下:

$$f(z^{(i)}) = \phi_1(z^{(i)})w_1 + \phi_2(z^{(i)})w_2 + \dots + \phi_p(z^{(i)})w_p = \Phi(z^{(i)})W$$

其中 w_j 是 m 维的权重向量。从潜在空间 z 中的每个网格点到原始空间 X 的变换具有以 $f(z^{(i)}) = \Phi(z^{(i)})W$ 为中心的正态分布,表示如下:

$$p(X|z^{(i)}, W, \beta) = \left(\frac{\beta}{2\pi}\right)^{m/2} \exp\left[-\frac{\beta}{2} \|\Phi(z^{(i)})W - X\|^2\right]$$

其中 β 是正态分布方差的倒数。于是,给出基于所有网格点的概率分布 $p(X|W, \beta, \pi)$ 如下:

$$p(X|W, \beta, \pi) = \sum_{i=1}^{k^2} \pi^{(i)} p(X|z^{(i)}, W, \beta)$$

其中 $\pi^{(i)}$ 为第 i 次正态分布的权重,因此SGTM中的对数似然函数给出如下:

$$L(W, \beta, \pi) = \sum_{i=1}^n \ln[p(X^{(i)}|W, \beta, \pi)] \\ = \sum_{i=1}^n \ln\left\{\left(\frac{\beta}{2\pi}\right)^{m/2} \sum_{j=1}^{k^2} \pi^{(j)} \exp\left[-\frac{\beta}{2} \|\Phi(z^{(j)})W - X^{(i)}\|^2\right]\right\}$$

与GTM模型中一样, W 的大小受到正则化系数 λ 的限制。在SGTM中,映射的大小、基函数的数量 p 、基函数的方差 σ 和正则化系数 λ 是必须预先设置的超参数。一旦构建了SGTM模型,就可以计算出每个数据点 x_i 在二维地图上每个点 z_j 的存在概率。

为了优化聚类个数,在逐个减少聚类数的同时计算BIC。BIC计算如下:

$$BIC = -2L(W, \beta, \pi) + d \cdot \ln(n)$$

其中 d 是当前聚类的簇数。

可以看出,在原始的GTM算法中每个网格点的权重是恒定的,而在SGTM中修改了图上每个节点对应的正态分布的权重,且每个网格点的权重都是可变的。SGTM为原始GTM算法增加了一些自由度,使其能够解决自组织映射(SOM)遇到的各种问题。此外,SGTM模型通过基于贝叶斯信息准则(BIC)来优化确定适当数量的聚类,使其具有更强的解释力。总的来说,SGTM可以在可视化数据的同时对数据进行聚类,提供了一种更高效的方法来分析数据。

2 V2X标准技术地图实证分析

2.1 研究样本与数据处理

为了验证该方法的有效性,我们拟选取一个关键技术领域进行验证。根据研究目的,研究样本的选择应满足以下两个条件:(1)该领域应为关键或新兴技术领域,标准布局对产业发展存在重要影响;(2)该领域标准技术发展迅速,形成了大量的标准信息,传统的人工手段难以对文本信息进行有效解读和判断。车用无线通信技术(Vehicle to Everything, V2X)是将车辆与一切事物相连接的一代信息通信技术,其中V代表车辆,X代表任何与车交互信息的对象,X主要包含车、人、交通路侧基础设施和网络。V2X技术可以提高驾驶安全性、减少拥堵、提高交通效率、提供车载娱乐信息等,是实现城市交通系统智能化、构建新型智慧城市的核心要素^[7, 22, 23]。世界各地围绕V2X的技术研究和标准制定进行得如火如荼,以期抢占未来发展制高点。因此,本研究以V2X相关的标准为样本进行

研究。

3GPP是当前V2X国际标准制定的主要机构,本文从3GPP获取相关标准信息 and 文档。首先,通过关键词和主题检索的方式获得相关标准文件308项。其次,考虑到每一个标准存在多个不同版本以及草案信息,本文对标准文件进行了去重处理:以标准号对相关信息进行合并,只纳入已经发布的标准文件,同一标准号以最新发布日期为准,最终得到有效标准文件42项。由于一个标准文本可能涉及多个技术主题,本文依据各章节内容对下载的标准文档进行文本分割,共得到技术主题文本497项。

完成文本分割的标准技术主题文档中包含许多与技术无关的词语,需要对其进行预处理:首先,删除如:日期、连接词、网址、数字等符号,进行英文字母大小写的转换;其次,还原词语的词

性,并根据设置的固定词组和同义词进行归一化处理;最后,删除停用词,提高数据的可靠性。接下来使用TF-IDF算法选取TF-IDF值大于0.1的名词词语,结合领域技术专家意见筛选出63个最具解释力的关键词,见表1。

利用上一步确定的63个关键词构建二进制表达的词-篇矩阵。当某标准主题文档包含所确定的关键词时,该矩阵中关键词对应向量元素赋值为1,否则赋为0。得到的词篇矩阵见表2。

2.2 V2X标准地图绘制

在输入“标准文本-关键词”特征矩阵之后,通过定义模型的5个超参数生成标准技术地图,包括网格节点数、高斯基函数个数、每个高斯基函数的方差、EM算法的 λ 值和SGTM模型的训练迭代次数。由于不同的参数设置会极大地影响可视化

表1 标准文本技术关键词

序号	关键词	序号	关键词	序号	关键词
1	PC5-unicast	22	LTE	43	VAE-info
2	server	23	RSU	44	location
3	VAE	24	power	45	V2X services
4	vehicle	25	policy	46	entity
5	MBMS	26	parameter	47	groupcast
6	security	27	communication	48	URI
7	QOS	28	procedure	49	RAT
8	V2X	29	V2V	50	response
9	resource	30	broadcast	51	architecture
10	V2X application server	31	UAV	52	user
11	message	32	user equipment	53	E-utran
12	link	33	control	54	AVP
13	V2X service	34	application	55	consumer
14	V2X communication	35	configuration	56	interface
15	service	36	DRX	57	USD
16	element	37	subscription	58	geographical
17	V2X application	38	sidelink	59	frequency
18	network	39	provision	60	receive
19	V2X message	40	operation	61	change
20	identity	41	notification	62	carrier
21	VAE client	42	LTE-Uu	63	Video

表2 专利和标准文本-关键词特征向量矩阵

Keyword	pc5-unicast	server	VAE	vehicle	Mbms	security	...	Video
Standard 1	0	0	0	0	0	1	...	0
Standard 2	0	0	0	1	0	1	...	0
Standard 3	0	0	0	1	0	0		0
...	...							
Standard 496	1	0	0	0	0	0		1
Standard 497	0	0	0	0	0	0	...	1

结果, 本研究使用灵敏度分析以及EM算法确定参数的最优选择。最终确定使用 10×10 的网络方格作为潜在变量的映射点, 使用36个径向高斯基函数, 每个基函数的方差设置为 2^{-1} , 权重正则化系数为0.001, 训练周期设置为500次, 使用的软件工具为Python。图3显示了不同聚类个数的BIC值, 我们可以看到当聚类个数为5时出现了明显的拐点, 因而将聚类个数设置为5个。生成的结果如图4所示, 标准文本对应的点可能归属于不同的技术领域, 用不同深浅表示; 空白区域即为标准空白点, 表示未有与之对应的标准文本存在。

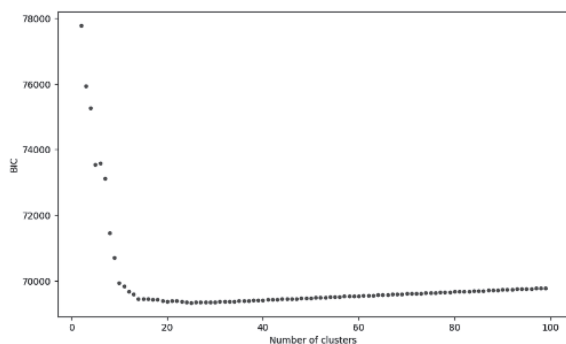


图3 SGTM的BIC值

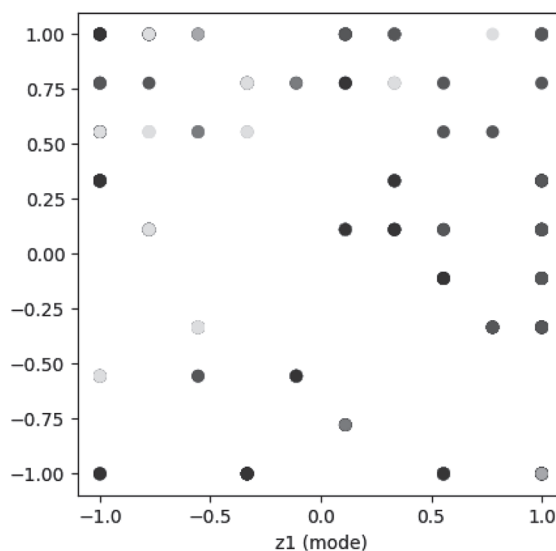


图4 基于SGTM的V2X标准技术地图

为了更为准确地分析标准布局情况, 本文根据映射关系进一步计算了标准地图上每一个位置的标准密度信息, 以气泡图的形式予以呈现, 结果如图5所示。图形的大小代表该位置标准分布的多少, 气泡越大所对应的标准越多, 气泡越小所对应的标准越少, 还有58个空白点代表着截至目前这些位置上仍没有相应标准出现。

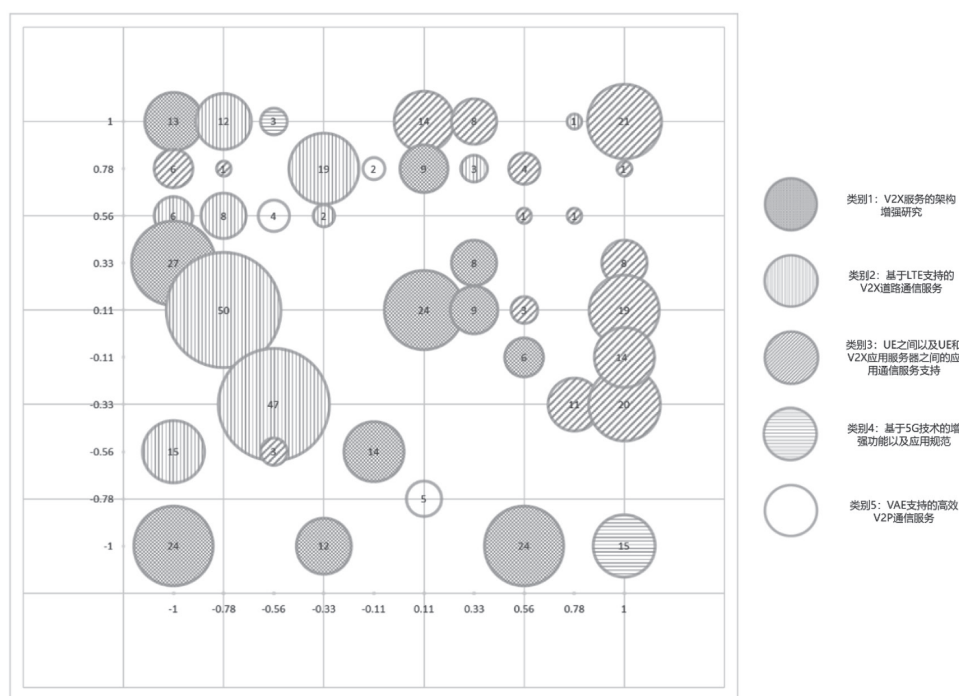


图5 V2X技术标准布局图

2.3 结果分析

在上一节中,我们用SGTM方法绘制出了已有的技术标准热点和标准空白点分布情况。为了深入分析标准布局,我们需要将每个位置所在的矢量空间反向映射到原始数据空间,进而对标准布局的内容进行有效解读。由于算法逆向映射的结果为关键词的向量表示,所以还需要将向量的数值与设置的阈值相比较,参照已有文献的做法本研究设置阈值为0.3,最终得出每一个矢量空间所对特征的关键词组合,在此基础上对标准布局现状进行分析。

(1) V2X领域标准布局分析

V2X是以车辆为主体,依靠与车、人、网、基础设施的通信网络互联实现互联互通、信息共享,进而达到保障交通安全、拓展智能服务等实现智慧交通解决方案的目标。从标准地图发现,目前V2X标准制定大致分布在5类主题,其中类别1、2、3代表的标准文本占有所有文本的94.2%(如图6所示),构成了当前V2X标准的主体。

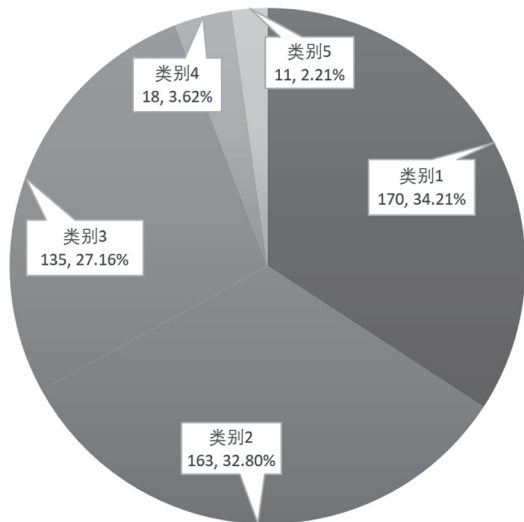


图6 不同类别包含的标准文本数量比重

类别1包含的标准文本主要是面向V2X服务的架构增强研究,该类标准占文本总量的34.21%。这一类标准针对PC5接口定义了全新的帧结构、资源调度和数据重传方式,支持单播、组播和广播3种模式;在Uu口引入了V2X通信切片、边缘计算、QoS预测等特性,从而能够满足车联网低时延、高可靠和大带宽等需求,并且可以在增强移动宽带

(eMBB)和高可靠低时延(URLLC)场景实现ToC向ToB的业务拓展。类别2的标准主要规定了基于LTE支持的V2X道路通信服务。具体又分为安全与非安全两类:与安全相关V2X的应用场景包括各等级的自动驾驶、前向碰撞预警、协同自适应巡航控制、紧急制动、车辆排队、通过基础设施提供道路安全服务等,而非安全相关的V2X场景包括移动高数据速率娱乐、移动热点/办公室/家庭、动态数字地图更新等。因而这一类标准对于V2X的各类服务要求做出了较为全面的规定,在标准文本中占比约为32.8%。类别3包括的标准主要是UE之间以及UE和V2X应用服务器之间的应用通信服务支持,用以解决支持行人感知车辆、基础设施及应用服务器的用户设备之间的通信问题,规定了UE在用于V2X服务时的应用层支持的协议,该类别文本比例在27%左右。上述3类标准搭建了V2X的核心协议规范和技术要求,为V2X的产业化落地搭建了较为坚实的基础。

随着V2X技术的演进、应用场景的丰富和部署的完善,V2X在智能网联汽车和自动驾驶中将发挥更多的协同作用,逐步实现网联协同决策与控制。因而,类别4主要针对的是基于5G技术的增强功能以及在增强场景中的应用规范问题,如:Sidelink增强、Sidelink中继增强、LTE V2X与NR V2X共存等;类别5规定了VAE支持的高效V2P通信服务,VAE层可以通过整合V2P的应用服务要求,并使通信流量模式与PC5 QoS设置和AS层配置保持一致,从而为支持V2P应用提供支持功能。总体而言,这两类标准包含的相关标准文本较少,未来仍有很大的标准制定空间。

(2) V2X潜在标准机会分析

在技术发展过程中,往往存在着不同领域的技术开发进度和标准制定的不平衡性,从技术创新的角度来看标准空白点在一定程度上隐含着标准技术进步的方向。挖掘出这些标准空白信息有助于我们识别技术机会,把握技术创新的方向,为未来的标准技术开发提供一定的参考。我们可以通过逆向映射确定标准空白点代表的关键词组合,结果见表3所示。以空白点1为例,它所对应的

表3 空白点对应关键词向量

Keyword	pc5-unicast	server	VAE	vehicle	Mbms	security	...	Video
空白点1	0	0	0	0	0	1	...	0
空白点2	1	1	0	0	0	0	...	0
空白点3	0	0	0	0	0	0		0
...	...							
空白点57	1	0	1	0	0	0		0
空白点58	0	0	0	0	0	0	...	0

关键词为Network, AVP, V2X, Link, Service, RSU, Entity, V2X Services, LTE, Control, QOS, Message, Resources, Carrier, V2X Application, UEs, Vehicle等, 因而代表的是一个“基于LTE的UE与V2I、V2P、V2V等设备间的网络通信服务”的标准机会, 其他技术空白可以按照类似的原理进行判断。当然, 这些关键词组合可行性如何, 是否需要再进行前瞻性的布局需要领域专家共同参与加以研判。

综合而言, 根据空白点对应关键词可以发现, V2X技术标准网络设备支持、用户应用服务、信号接收、位置信息更新等领域仍存在较多技术空白。目前, 车联网产业正在进入以行业实际应用需求和市场发展为牵引进行小规模部署与应用实践的新阶段, 大量增强型应用场景对基于V2X的车路协同决策提出了更高的要求, 可以瞄准产业需求和标准空白, 加快推动重点V2X产业技术的研发和标准制定。

3 结论与建议

借助 TF-IDF和SGTM方法, 本文提出了一套基于标准文本信息构建标准技术地图的方法。该方法从技术角度分析标准化活动和布局, 通过文本挖掘和可视化工具绘制出某一技术领域的标准技术地图, 为标准化战略、技术研发和产业技术竞争等创新决策提供科学依据。基于V2X技术的实证研究发现: 在V2X领域3GPP已经制定出了较为完备的技术规范, 这些技术规范分布在基于LTE技术的服务要求和传输层支持、与用户设备相关的通信规范和V2X服务网络协议等领域。而基于5G技术的增强功能以及在增强场景中的应用规范问题相关标准虽有出现, 但总体数量较少, 未来还有较大的完

善空间。另外, 标准地图上的58个标准空白点为我们提示了标准开发的潜在方向, 从空白点所对应的关键词可以看出主要包括道路、行人及基础设施等智能化UE的搭载, 基于5G技术系统的V2X增强功能以及应用规范等领域。

上述研究验证了本文所提出的研究框架具有可行性, 同时对V2X标准布局的研究结果为我们提出如下启示: (1) 3GPP的V2X基础协议规范已经较为成熟, 但在UE间通信支持、高精度地图更新、信息传输增强、用户应用服务等领域还有较多空白点。近年来我国在《国家车联网产业标准体系建设指南(智能网联汽车)》的指导下, 协同推进LTE-V2X全协议栈标准制修订, 完成了覆盖总体要求、接入层、网络层、应用层等各个环节的技术标准规范制定。未来可结合国际标准空白点, 选择我国已具备较好基础的重点标准推向国际, 进一步提升国际话语权和影响力。(2) 整体而言, 针对V2X增强业务场景的标准尚在研究阶段, 由于无线通信技术的演进节奏明显快于垂直行业的部署周期, 导致场景需求和技术开发推进力度不一, 一定程度上阻碍了标准的开发。下一步需要汽车、交通、公安和通信等行业通力协作, 从技术开发和市场需求两方面综合考虑V2X的技术成熟度和部署可行性, 加快对增强场景的标准研制。

本文是利用标准文本信息进行深度挖掘和可视化呈现的初步尝试, 一定程度上可以弥补当前研究仅针对浅层结构化信息进行统计分析的不足。未来的研究可以向两个方向进一步探索: (1) 本文提出的标准技术地图构建方法可以根据分析目的进一步拓展和改进, 形成不同侧重点的标准地图。比如: 我们可以在本文基础上构建标准起草人-关键词矩阵, 从竞争主体的视角形成标准竞争地图, 更

为直观地揭示标准竞争态势。(2)当前研究仅从标准文本本身挖掘标准布局信息,未来可以考虑将标准和专利、科学论文等多源信息纳入统一的研究

框架,从而对未来的技术发展趋势和标准布局态势进行更为准确的判断。

参考文献

- [1] SMES E C D I M I E A. An EU Strategy on Standardisation. Setting global standards in support of a resilient, green and digital EU single market [J/OL]. 2022.2, <https://www.europeansources.info/record/an-eu-strategy-on-standardisation-setting-global-standards-in-support-of-a-resilient-green-and-digital-eu-single-market/#:~:text=By%20setting%20global%20standards%2C%20the%20EU%20exports%20its,the%20global%20role%20of%20the%20European%20Standardisation%20System>.
- [2] United States Government National Standards Strategy For Critical And Emerging Technology [J/OL]. 2023.5, <https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf>.
- [3] GB/T200001-2014, 标准化工作指南 第1部分: 标准化和相关活动的通用术语 [S].
- [4] HENKEL J. Licensing standard-essential patents in the IoT - A value chain perspective on the markets for technology [J]. Research Policy, 2022, 51(10).
- [5] 张震,甘克勤.基于Solr的大规模标准文献可视化分析系统[J].计算机系统应用, 2016(3):67-71.
- [6] 王季云,王寒.旅游业标准起草单位合作网络研究——基于社会网络分析方法[J].旅游论坛, 2018, 11(2):1-9.
- [7] 陈山枝. C-V2X车联网赋能智能驾驶和智能交通[J]. 智能网联汽车, 2023(1):36-39.
- [8] 闫坤,苏丹,李文宇. 智能交通中的V2X技术专利分析[J]. 信息通信技术与政策, 2021(7):93-96.
- [9] 汪传雷,胡蕾,刘书美,等. 基于ucinet的物流标准文件的社会网络分析[J]. 物流技术, 2022, 41(5):7-14.
- [10] 杨德相,李剑锋. 基于知识图谱的标准知识管理研究[J]. 标准科学, 2023(4):39-45.
- [11] 李国鹏,李景,甘克勤. 2010~2014年国内标准文献使用情况的数据挖掘分析——以中国标准化研究院服务数据为例[J].标准科学, 2017(12):113-118.
- [12] 于晓明,史胜楠,甘克勤. 基于Tableau的标准起草单位大数据可视化分析[J]. 标准科学, 2019(9):80-84.
- [13] CHOWDHURY S, ALZARRAD A. Applications of Text Mining in the Transportation Infrastructure Sector: A Review [J]. Information (Switzerland), 2023, 14(4).
- [14] WANG X, WANG H, JI H, et al. New Frontiers of Scientific Text Mining: Tasks, Data, and Tools; proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2022, August 14, 2022 - August 18, 2022, Washington, DC, United states, F, 2022 [C]. Association for Computing Machinery.
- [15] 梁薇,马万钟,王文君.基于知识地图的标准知识链接服务研究[J].标准科学, 2015(6):23-26.
- [16] 孙孝静,张雪玫,王美远.团体标准研究知识图谱:基于VOSviewer的可视化研究[J].标准科学, 2019(9):46-50.
- [17] JEONG S, YOON B. A systemic approach to exploring an essential patent linking standard and patent maps: Application of generative topographic mapping (GTM) [J]. EMJ - Engineering Management Journal, 2013, 25(1): 48-57.
- [18] 吴菲菲,陈明,黄鲁成. 基于GTM的3D生物打印专利技术空白点识别[J]. 情报杂志, 2015, 34(3):58-64.
- [19] QIN J, ZHOU Z, TAN Y, et al. A big data text coverless information hiding based on topic distribution and tf-idf [J]. International Journal of Digital Crime and Forensics, 2021, 13(4): 40-56.
- [20] BISHOP CM, SVENSÉN M, WILLIAMS C. GTM: The Generative Topographic Mapping [J]. Neural Computation, 1998, 10(1): 15-34.
- [21] KANEKO H. Sparse Generative Topographic Mapping for Both Data Visualization and Clustering [J]. Journal of Chemical Information and Modeling, 2018, 58(12):2528-2535.
- [22] 陈山枝. 蜂窝车联网(C-V2X)及其赋能智能网联汽车发展的辩思与建议[J]. 电信科学, 2022, 38(7):1-17.
- [23] 陈山枝,葛雨明,时岩. 蜂窝车联网(C-V2X)技术发展、应用及展望[J]. 电信科学, 2022,38(01):1-12.