基于文献计量分析和 LDA 主题模型的自动驾驶研究概述

秦禹
对外经济贸易大学统计学院　北京市　100102

摘　要：随着汽车行业的发展，自动驾驶作为一个新兴技术方向，逐渐进入大众视野。本文旨在通过文献计量分析和 LDA 主题模型分析，结合 Science Direct 上自动驾驶相关的文献，来描述自动驾驶的研究现状和重点发展方向，并阐述了研究领域的一些见解，识别的主题能够反映当前最受关注的自动驾驶潜在机会和挑战。

关键词：自动驾驶　LDA 模型　文献计量分析　文本挖掘

An Overview of Autonomous Driving Research based on Bibliometric Analysis and LDA Topic Model

Qin Yu

Abstract: With the development of the automotive industry, autonomous driving, as an emerging technology direction, has gradually entered the public's field of vision. This paper aims to describe the research status and key development directions of autonomous driving through bibliometric analysis and LDA topic model analysis, combined with the literature related to autonomous driving on Science Direct, and expounds some insights in the research field. The identified topics can reflect the potential opportunities and challenges for autonomous driving, which is very popular nowadays.

Key words: autonomous driving, LDA model, bibliometric analysis, text mining

1　引言

汽车行业的进步促进了自动驾驶技术的不断发展，但由于产业链、基础设施、法规政策和道路场景等差异，造成了自动驾驶不同的发展路线和研发方向。当前对自动驾驶的关注多集中在技术和应用层面，而对于客观量化分析自动驾驶发展方向的研究不是很多。通过对自动驾驶相关文献采取量化分析和主题模型提取等技术分析手段，可以对全面理解自动驾驶的发展现状和行业热点，把握当前发展趋势和机遇，并辅助制定相关的政策和发展战略有很大帮助。

2　自动驾驶概述

自动驾驶是在传统汽车硬件架构的基础上，结合通信技术、网络技术和人工智能等技术来赋予车辆自动行驶的能力，并在安全性、便利性、高效性和低碳性方面有更大优势。随着场景复杂性和车辆智能性的不断提升，自动驾驶的成熟度又被分为不同的等级。汽车行业对自动驾驶的分级，主要参考了美国汽车工程师协会(SAE)提出的J3016自动驾驶等级标准。在这份标准中，基于车辆控制权分配和环境感知程度，自动驾驶被划分为六个等级[1]，分别对应着无自动驾驶、独立辅助驾驶、协同辅助驾驶、条件约束的高级辅助驾驶、场景约束的无人驾驶和无约束无人驾驶这六个阶段。

3　自动驾驶技术架构和发展

一套完整的自动驾驶系统架构，包含环境感知、精确定位、路径决策规划和控制执行。环境感知通过多维传感器采集环境和车辆行驶状态信息，来构建车辆和环境认知模型；精确定位通过高精地图，实时定位车辆的位置和速度等状态；路径决策规划建立了包含障碍区域和自由区域的环境地图，通过路径搜索最优算法快速生成最佳行驶路线；控制执行将智能控制和传统控制相结合，实现在复杂环境中自动行驶。

文献计量学是定量分析文献的有效工具，广泛用于评估多领域的发展现状和研究趋势。随着大数据、5G 和人工智能的不断发展，自动驾驶的重要性不断提升，大量相关的学术文献被发表。通过对文献的量化分析，来了解自动驾驶现状和发展方向是很重要的。主题模型是文本挖掘的有效工具，常用于研究特定领域的主题和热点。LDA（Latent Dirichlet Allocation）是最受欢迎的主题模型之一，在很多领域都有应用[2]。本文结合了文献计量分析和 LDA 模型，从统计分析和文本挖掘的角度，阐述了自动驾驶技术的现状和发展趋势。

4　分析方法

4.1　数据来源

本文数据来自于 Science Direct 网站，检索了从 2002 年到 2021 年包含 Autopilot 关键词同期收录的 4569 篇文献，处理后的数据包含文献标题、刊物、发表年份、作者、关键词、摘要、国家和单位等信息。

4.2 主题识别

4.2.1 主题模型

LDA 模型由 Blei 等人于 2003 年提出[3]，主要用于推测文档的主题分布。它基于三层贝叶斯网络，以概率分布形式给出文档主题，实现主题聚类和主题演变分析等目的。在本文中，提取文献标题、摘要和关键词到模型语料库中。通过 Python Gensim 实现 LDA 模型，超参数设定为：alpha=2，eta=0.9，iterations=1000，passes=20，选取前 15 位关键词结果来解释对应主题。

4.2.2 确定主题数量

主题数会直接影响模型的识别效果，但不同主题领域存在复杂性和多样性，学术界目前并没有统一的方法。基于知识经验并不断优化来主观确定主题数，也有很多成功的实际应用。Blei 提出通过困惑度来确定主题数的方法[3]；Griffiths 提出使用 log− 边际似然函数的方法确定主题数[4]；Teh 提出基于狄利克雷过程的分层 HDP 方法，通过计算 LDA 和 HDP 的困惑度判断主题数[5]；Guan 提出构建主题困惑度 − 方差比值的指标确定主题数[6]；Wang 通过构建困惑度和平均相似度主题交叉曲线来确定主题数[7]。

除困惑度外，还可通过主题一致性来确定主题数。一致性衡量主题词语之间的相对距离，相比于困惑度，一致性得到的结果通常也更接近主观判断。Roder 在主题一致性度量空间的研究中，系统介绍了架构原理，并提出多种度量方法[8]。主题一致性架构为：

$$C=S\times M\times P\times \sum$$

C 代表一致性度量聚合值，S 代表文本语料切分片段集合，M 代表通过计算词频构成的配置空间确认度量，P 代表词频估计方法集合，$\sum$ 代表计算聚合标量值的方法集合。

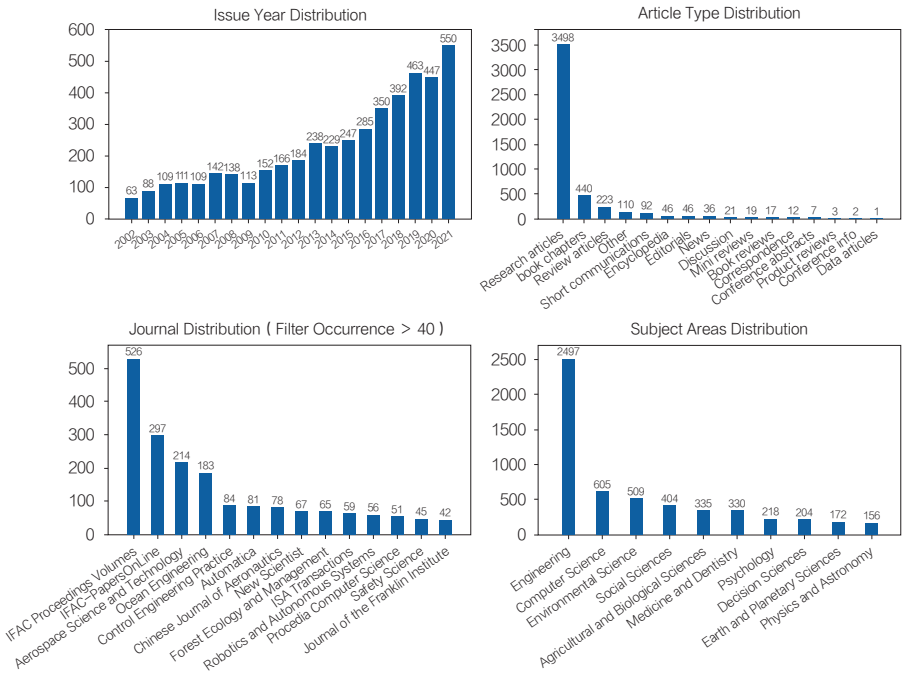
Roder 在此基础上提出一种全新的一致性度量组合方式（Cv），该方法能涵盖现有所有的一致性度量标准，并构建包含间接余弦测量和布尔式滑动窗口的新的度量标准，相比于其他方法，该方法有最好的表现[8]。因此，本文也采用了基于 Cv 的一致性度量方法。

5 结语

5.1 文献计量分析

文献计量分析展示了文献发表时间、类型、期刊和主题领域分布的统计结果（见图 1）。

图1 文献计量分析



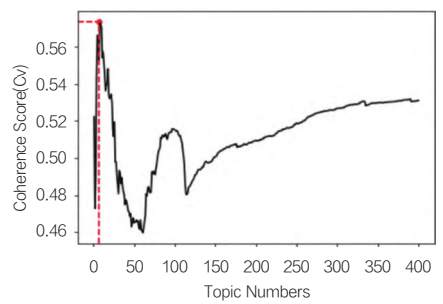
自动驾驶的文献数量在近 20 年内持续增长，近 10 年上升趋势尤为显著。超过 76% 的文献类型为研究类，超过 18% 的文献发表在国际自动控制协会（IFAC）期刊，55% 以上的自动驾驶都分布在工程主题，其次为计算机科学、环境科学、社会科学和农业生物科学主题。

5.2 主题模型分析

5.2.1 主题数量确定

本文根据 LDA 主题模型一致性度量 Cv 来确定最佳主题数。通过一致性与主题数之间曲线（见图 2），当主题数为 7 时，一致性分数达到最大 0.5735，此时模型得到了最佳训练，并保持了足够的主题高频词语义相似性。

图2 主题一致性度量分数



5.2.2 主题结果可视化

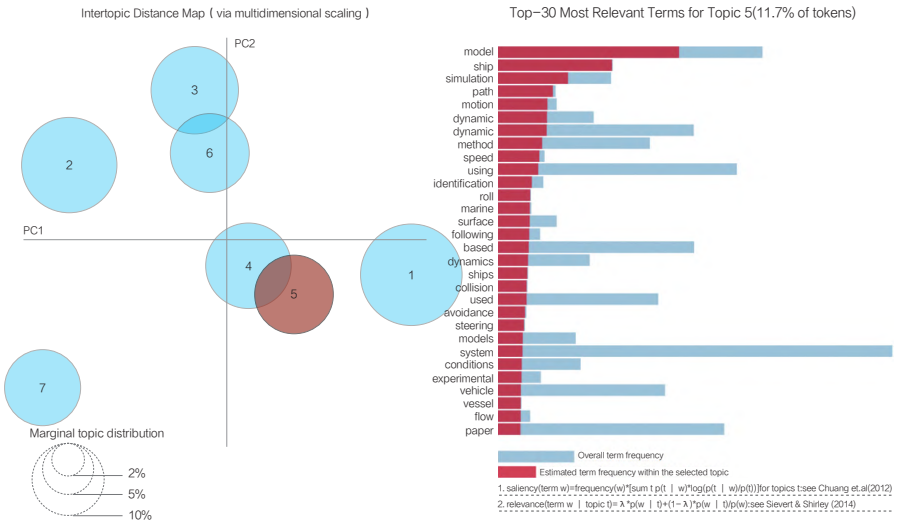
本文通过 pyLDAvis 实现主题结果的可视化，它针对 LDA 模型提供了可以分析主题差异性和高度关联关键词的全局视角框架。每个圆圈代表独立的主题，圆圈大小和重叠面积反映了主题模型的鲁棒性，圆圈中心被映射到两个维度 PC1 和 PC2，圆圈距离由主题降维矩阵计算得出。结果显示有两对主题之间有一定重叠，其他主题均保持了相对的独立性（见图 3）。

5.2.3 主题结果展示

选择概率最高的前 15 个词来解释对应的主题，过滤不相关主题后，最终确定了六个主题：①基于模型的探测，②人体安全，③自适应控制，④神经网络系统，⑤动态路径识别，⑥自主实时导航。结合自动驾驶的架构体系，这些主题可被分为三类。第一类为路径决策规划，包含主题④⑤；第二类为控制执行，包含主题②③；第三类为环境感知，包含主题①⑥。

作为自动驾驶的核心，路径决策规划中基于图搜索、最优化、随机采样或曲线拟合的算法选择，基于规则或强化学习的行为决策，基于故障预警和预留机制的异常处理等，都在不同程度上推动着自动驾驶的不断创新。然而社会层面的因素也不可忽略，比如最近某车型在车祸前一秒退出自动驾驶系统的控制策略，引发了巨大的社会争议和对自动驾驶的顾虑。相关行业标准体系的制定，不仅会推动融合创新生态体系的建立，也对自动驾驶的快速发展提供显著支持[9]。

图3　主题模型可视化结果



自适应控制要实现对环境参数的变化有适应能力的控制策略，基于车辆协同控制及行驶优化技术、多目标优化理论、车辆自主运动决策与高精度跟踪控制等技术，都在推动着此主题的发展。

环境感知主要包含障碍感知和车辆路径感知，过程中有大量数据处理工作。真实场景中收集的各类传感器数据，需要人工标注才能使用，而一些基于模型探测的预训练和测试要在仿真环境实现。自主探测与导航技术的进一步发展，会显著改善计算效率和数据成本。

6　结语

为了全面了解自动驾驶技术的发展，本文基于 Science Direct 网站上过去 20 年发表的自动驾驶相关文献，开展了文献计量分析和主题分析。文献发表数的增长反映了自动驾驶行业的快速发展，行业分析结果则反映了自动驾驶的应用和研发投入分布情况，前五位是工程、计算机科学、环境科学、社会科学和农业生物科学。

通过 LDA 模型完成了自动驾驶主题识别（见表 1），这六个主题能够反映当前的主流研究方向和热点。本文为理解自动驾驶提供了一个宽泛的视角，并分析了发展现状和热点。一方面能够帮助准确把握研究趋势，抓住机遇；另一方面，也能够为科技政策和战略发展提供参考支持。

表1　模型主题识别结果

潜在主题	前 15 位高频词
Model Based Detection 基于模型的探测	data，using，uav，based，estimation，aerial，results，method，used，model，flight，unmanned，landing，detection，accuracy
Human Safety 人体安全	human，study，safety，driving，automation，research，factors，risk，auto mated，chapter，use，de，new，process，technology
Adaptive Control 自适应控制	control，controller，guidance，nonlinear，proposed，design，system，adaptive，systems，paper，law，problem，tracking，robust，linear
Neural Network System 神经网络系统	flight，system，aircraft，design，control，network，neural，energy，fault，systems，power，chapter，learning，networks，analysis
Dynamic Path Detection 动态路径识别	model，ship，simulation，path，motion，dynamic，results，method，speed，using，identification，roll，marine，surface，following
Autonomous Real-time Navigation 自主实时导航	systems，vehicles，autonomous，system，vehicle，paper，unmanned，software，applications，aerial，development，control，research，uavs，uav

参考文献：

[1]J3016_202104，Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles [S].

[2]Jelodar，H.，Wang，Y.，Yuan，C. et al. Latent Dirichlet allocation（LDA）and topic modeling：models，applications，a survey [J]. Multimedia Tools and Applications，2019，（78）：15169-15211.

[3]Blei，D.M.，Ng，A.Y.，& Jordan，M.I. Latent Dirichlet allocation [J]. The Journal of Machine Learning Research，2003，（3），993-1022.

[4]Griffiths T L，Steyvers M. Finding Scientific Topics [C]. Proceedings of the National Academy of Sciences，2004，（S1），5228-5235.

[5]Teh Y，Jordan M，Beal M，et al. Hierarchical Dirichlet Processes [J]. Journal of the American Statistical Association，2007，101（476）：1566-1581.

[6]Guan Peng，Wang Yuefen. Research on the Method of Determining the Optimum Topic Number of LDA Topic Model in Scientific and Technical Information Analysis [J]. New Technology of Library and Information Service，2016（9）：42-49.

[7]Wang Q，Yang K，Zhang Z，Wang Z，Li C，Li L，Tian J，Ye Y，Wang S，Jiang K. Characterization of Global Research Trends and Prospects on Single-Cell Sequencing Technology：Bibliometric Analysis [J]. Journal of Medical Internet Research，2021，23（8），e25789.

[8]RoDer，M.；Both，A.；Hinneburg，A. Exploring the Space of Topic Coherence Measures [C]. Proceedings of the Eighth AMC International Conference on Web Search and Data Mining，2015，（2），399-408.

[9] 工信部：车联网标准体系指南发布 自动驾驶发展迎来新机遇 [J]. 电子元器件与信息技术 . 2018，（05），63-65.

作者简介

秦禹：（1991—），男，满族，内蒙古呼伦贝尔，硕士研究生。研究方向：统计学与大数据。