检索增强的网络流量预测方法

Retrieval-Augmented Network Traffic Prediction Method

常远/CHANG Yuan[1]，吴春鹏/WU Chunpeng[2]，
王峰/WANG Feng[1]
（1. 中国电信研究院，中国 北京102209；
2. 中国电力科学研究院，中国 北京100192）
(1. Research Institute of China Telecom, Beijing 102209, China;
2. China Electric Power Research Institute, Beijing 100192, China)

DOI：10.12142/ZTETJ.202505005
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20250926.1430.006
网络出版日期：2025-09-26
收稿日期：2025-07-25

摘要：网络流量预测是保障网络服务质量的关键技术，现有时序模型难以融合文本描述的变更事件信息。提出一种融合时序大模型与语言大模型的协同预测框架，实现变更事件驱动的网络流量动态预测。针对变更事件稀疏性及专业语义理解难题，设计基于检索增强生成（RAG）的变更影响知识库，通过检索历史相似变更的流量影响特征，构建可解释的上下文提示。模型采用双阶段架构：首先使用时序大模型生成基础流量预测，继而由语言大模型结合检索的变更案例及当前变更描述，对预测结果进行语义推理修正。实验表明，在真实网络运维数据集上，模型在变更事件场景下的预测误差相较于仅通过时序预测的方法有明显下降。

关键词：流量预测；检索增强生成；时序预测

Abstract: Network traffic prediction is a critical technology for ensuring network service quality, yet existing time series models struggle to incorporate textual descriptions of change events. This paper proposes a collaborative prediction framework that integrates large-scale time series models and large language models to achieve change-driven dynamic network traffic forecasting. To address the sparsity of change events and challenges in professional semantic understanding, we design a retrieval-augmented generation (RAG)-based change impact knowledge base. This retrieves traffic impact characteristics from historically similar changes to construct interpretable contextual prompts. The model adopts a two-stage architecture: First, a large time series model generates baseline traffic predictions; subsequently, a large language model performs semantic reasoning-based refinement of these predictions by incorporating both the retrieved change cases and the current change description. Experiments on real-world network operation datasets demonstrate that our framework significantly reduces prediction errors in change event scenarios compared to time series-only approaches.

Keywords: network traffic prediction; retrieval-augmented generation; time series prediction

引用格式：常远, 吴春鹏, 王峰. 检索增强的网络流量预测方法 [J]. 中兴通讯技术, 2025, 31(5): 25-29. DOI: 10.12142/ZTETJ.202505005
Citation: CHANG Y, WU C P, WANG F. Retrieval-augmented network traffic prediction method [J]. ZTE technology journal, 2025, 31(5): 25-29. DOI: 10.12142/ZTETJ.202505005

随着网络规模的不断扩大以及业务需求的日益复杂，网络流量预测作为保障网络服务质量、优化资源调度以及提升运维效率的关键技术，正受到越来越多的关注。准确预测网络流量的变化趋势，不仅有助于提前识别潜在的拥塞风险，还可以为网络容量规划、故障恢复以及安全防护提供决策支持。然而，传统的流量预测方法[1-4]主要依赖于历史流量数据，通常采用统计模型或深度学习模型对时间序列进行建模，以捕捉流量的周期性、趋势性和突发性特征。尽管这些方法在常规场景下表现良好，但在面对网络中发生的计划性变更事件（如端口关闭、服务迁移、带宽调整等）时，往往难以准确反映变更对流量模式的潜在影响。

近年来，随着大语言模型[5-8]在自然语言处理领域的广泛应用，研究者开始尝试将语言模型的能力引入时间序列预测任务中，以处理与文本信息相关的预测问题。例如，在经济预测、天气预报等领域，已有工作尝试将新闻事件、政策文本等非结构化信息与时间序列预测模型相结合，提升预测结果的语义解释能力和准确性。然而，将这一思路应用于网络运维场景仍面临诸多挑战。一方面，网络环境中的变更事件具有显著的稀疏性，即在长时间的流量数据积累中，真正发生且对流量产生显著影响的变更操作相对较少。这使得基于监督学习的方法难以获得足够的训练样本，从而限制了对语言模型进行微调或端到端训练的有效性。另一方面，网络

基金项目：国家电网有限公司总部管理科技项目（5700-202358842A-4-3-WL）

运维领域具有高度的专业性，涉及复杂的协议、拓扑结构及服务依赖关系。通用语言模型在缺乏领域知识的情况下，往往难以准确理解变更描述的语义，并据此推理其对网络流量的具体影响。

为应对上述挑战，本文提出了一种融合时序大模型与语言大模型的协同预测框架，旨在实现变更事件驱动的网络流量动态预测。该方法的核心思想在于将变更事件的语义信息与流量预测任务进行有机结合，通过构建基于检索增强生成的变更影响知识库，解决历史变更数据稀疏和语义理解受限的问题。具体而言，本文首先将历史变更记录及其对应的流量变化特征组织为结构化的知识条目，并构建高效的检索机制。当面对新的变更描述时，系统将通过语义相似度匹配，从知识库中检索出若干历史相似的变更案例，并将其描述及影响特征作为上下文提示输入给语言大模型。随后，基于历史流量数据，使用时序大模型生成基础预测结果。最终，语言大模型将结合检索到的上下文信息以及当前变更描述，对基础预测结果进行语义推理与修正，从而输出考虑了变更影响的流量预测值。

本文提出的双阶段预测架构不仅有效融合了时序建模与语言理解的优势，还在不依赖大规模标注数据的前提下，提升了模型对网络变更事件的响应能力与泛化性能。通过引入可解释的上下文提示机制，模型在提升预测精度的同时，也为运维人员提供了可追溯的决策依据，增强了系统的透明度与可信度。实验结果表明，在真实网络运维数据集上，该方法相较于仅依赖时序模型的预测方法，在包含变更事件的预测场景中具有显著的误差降低效果，验证了其在复杂网络环境下的实用性与有效性。

1 双阶段协同预测框架

1.1 架构概述

如图1所示，本文提出的网络流量预测方法基于一种融合时序建模与语言理解能力的双阶段协同预测框架，旨在实现对计划性网络变更事件影响的动态预测。该方法的核心设计目标是解决传统时序预测模型难以处理文本描述性变更信息的问题，同时克服变更事件稀疏、难以通过端到端方式训练语言模型的挑战。系统整体架构由3个关键模块构成：变更影响知识库、时序预测模型以及语言大模型推理模块。其中，变更影响知识库作为外部知识源，负责存储与组织历史变更事件及其对应的流量变化特征，并通过检索增强生成机制为语言模型提供可解释的上下文支持；时序预测模型负责基于历史流量数据生成基础预测结果，捕捉流量的时间演化规律；语言大模型则承担语义理解与推理修正功能，结合当前变更描述与检索到的历史相似案例，对基础预测结果进行语义驱动的修正，从而输出考虑了变更影响的流量预测值。

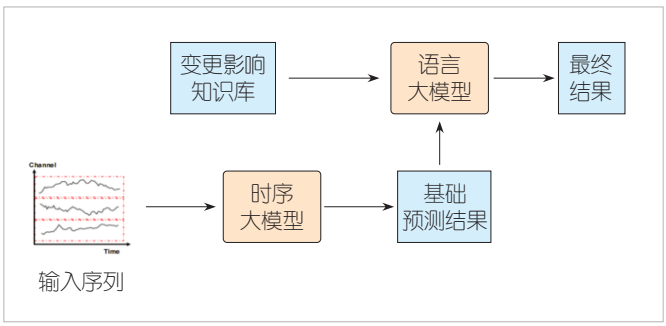

图1 检索增强的网络流量预测架构

系统运行流程分为两个主要阶段：第一阶段为基于时序模型的基础预测阶段，输入为当前时刻之前一段时间内的历史流量数据，输出为不考虑任何变更操作影响的未来流量预测结果；第二阶段为语言大模型主导的修正阶段，该阶段不仅接收来自第一阶段的预测结果，还接收当前计划变更的自然语言描述，并通过检索增强生成（RAG）机制从变更影响知识库中检索出若干历史相似变更事件，将其描述与对应的流量影响模式作为上下文提示注入至语言模型中，以辅助其理解当前变更的潜在影响。语言大模型在此基础上对基础预测结果进行语义推理与调整，最终输出融合了变更语义信息的预测值。这种双阶段架构不仅有效分离了时序建模与语义推理的功能边界，也避免了直接对语言模型进行大规模微调的需求，从而提升了方法的实用性与泛化能力。

1.2 变更影响知识库的构建

为有效解决网络变更事件语义信息难以建模、历史变更样本稀疏以及大语言模型对网络运维领域理解能力受限的问题，本文构建了一个结构化、语义增强的变更影响知识库。该知识库旨在为语言大模型提供可检索、可解释的历史变更上下文信息，从而在不依赖大规模标注数据与模型微调的前提下，增强其对变更事件语义的理解与推理能力。知识库的构建主要包括3个关键环节：数据收集与预处理、知识条目组织以及检索机制设计，它们分别从数据来源、知识表示与信息检索3个维度构建支持系统。总体的构建流程如图2所示。

在数据收集与预处理阶段，知识库的数据源主要来自网络运维日志系统中记录的历史变更事件及其对应的网络流量数据。变更事件通常以自然语言文本的形式记录，包含变更时间、操作类型（如端口关闭、服务迁移、配置更新等），涉及设备或服务、变更原因、变更可能造成的影响等信息。

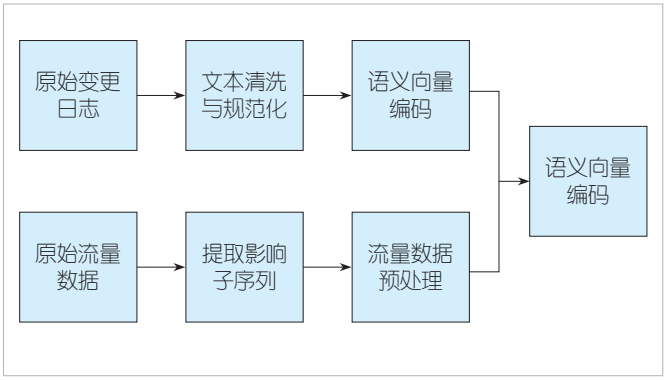

图2 变更影响知识库的构建流程

为了保证知识库中变更条目的质量与可解释性，需对原始变更文本进行清洗与规范化处理。具体而言，首先去除冗余信息与非结构化表达，保留与网络行为直接相关的操作描述；其次，对变更描述中的专业术语进行标准化，例如将“关闭端口443”与“停用HTTPS服务”统一为一致表述，以提升后续语义匹配的准确性。与此同时，针对每条变更记录，提取其发生前后一段时间内的网络流量数据，用于刻画该变更对网络状态量的潜在影响。流量数据通过平滑滤波等手段进行预处理，以消除噪声干扰，保留具有代表性的流量变化模式。

在完成数据预处理之后，下一步是将变更描述与流量序列信息组织为结构化的知识条目，形成可供后续检索与推理使用的知识单元。每个知识条目由两个核心部分构成：变更语义描述与流量影响序列。其中，变更语义描述是对原始变更文本进行语义编码后的向量化表示，本文采用通用文本嵌入模型（BGE模型[9]）对其进行嵌入编码，以保留其语义信息；流量影响序列则直接记录了该变更发生前后一段时间内的原始流量时间序列数据，以数值序列形式保存。通过上述结构化组织方式，每个知识条目不仅保留了变更事件的语义信息，也完整刻画了其对网络流量的实际影响过程，从而为后续的语义检索与上下文提示生成提供了基础支撑。为确保RAG知识库提供准确的修正依据，本系统还引入了人工审核环节，仅收录对流量波动影响显著且语义描述完整的变更事件。

在知识条目的组织完成后，需构建高效的检索机制以保障系统响应效率与预测准确性。本文采用基于语义相似度的检索策略，利用向量空间中的相似性度量方法，从知识库中快速定位与当前变更描述最相似的历史变更条目。具体而言，首先将输入的当前变更描述通过相同的语义编码模型转化为向量表示，然后在知识库中计算其与所有条目中变更语义描述向量的余弦相似度，并依据相似度排序选取Top-K个最相似的历史变更案例作为上下文提示。检索到的Top-K变更条目将与其对应的流量影响序列一同作为上下文信息注入语言大模型，辅助其理解当前变更可能带来的网络流量变化趋势。

1.3 基于时序大模型的基础预测

整个预测过程的第一阶段为基于时序大模型的基础预测。系统接收当前时刻之前一段时间内的历史流量数据作为输入，目标是生成一个不考虑任何变更操作影响的基础流量预测序列。该阶段的核心任务是捕捉网络流量的固有时序演化规律，包括周期性、趋势性、突发性等特征。为实现对复杂流量模式的高效建模，本文选用Moirai作为基础预测模型[10]。Moirai是一种基于Transformer架构的通用时序大模型，具有强大的多变量建模能力与长序列预测性能。该模型无需手工设计特征，能够直接接受原始时间序列作为输入，并通过自注意力机制自动学习变量间的复杂依赖关系。此外，Moirai支持零样本预测，在未见过的目标变量上也能保持良好的泛化能力，这使其特别适用于网络流量预测中可能出现的多维度、多设备、多指标的复杂场景。为进一步提升该模型与网络流量任务的适配性，提高预测的准确率，本文以开源的Moirai模型作为基座模型，利用1年周期的历史流量数据对模型进行微调。在本系统中，Moirai的输入为历史流量序列，输出为未来一段时间内的预测值。模型的输入为历史流量序列：

$$\boldsymbol{X} = [x_{t-T+1}, \cdots, x_t] \in \mathbb{R}^T \quad (1),$$

其中，x_i表示i时刻的流量值，T表示输入序列的长度，例如以5 min粒度采集1 d的序列长度为$T = 288$。模型的输出则为未来一段时间的流量序列：

$$\boldsymbol{Y} = [y_{t+1}, \cdots, y_{t+M}] \in \mathbb{R}^M \quad (2)。$$

1.4 基于语言大模型的语义修正

该阶段的目标是将用户输入的计划性变更描述有效地融入预测结果中，从而生成考虑了变更影响的最终预测流量序列。由于变更事件通常以自然语言形式描述（例如“将于明日10:00关闭服务器A的端口443”），传统的时序模型无法直接理解此类信息，因此需要借助语言大模型来完成语义理解与推理任务。在本阶段，系统首先将当前变更描述输入至预训练语言大模型中，获取其语义表示；随后，通过前文所述的RAG机制，从变更影响知识库中检索出若干历史相似变更事件，并将其变更描述与对应的流量影响序列作为上下文信息注入语言模型。这种上下文提示机制不仅为语言模型

提供了可解释的推理依据，也有效缓解了变更事件稀疏、语言模型缺乏领域知识所带来的语义理解偏差问题。

语言大模型在接收到基础预测结果、当前变更描述以及检索到的历史上下文信息后，通过设计的提示模板引导其对基础预测进行语义驱动的修正。具体而言，提示模板会明确指示模型：基于当前变更的语义描述，并参考历史相似变更的流量影响模式，对基础预测结果进行调整，输出修正后的未来流量预测序列。模型在生成过程中不仅考虑当前变更的直接语义含义，还结合检索到的历史案例的流量变化趋势，进行类比推理与语义泛化，从而实现对变更影响的动态建模。值得注意的是，这一阶段并不依赖于对语言大模型的微调，而是完全基于其基础能力与上下文学习机制，从而提升了方法的部署灵活性与泛化能力。

2 实验验证

2.1 数据集

为验证本文所提出方法在面向计划性变更事件的网络流量预测任务中的有效性，实验聚焦于云池出口链路的网络流量监测数据及对应的变更操作记录。该数据集涵盖了连续8个月的流量观测序列与运维变更日志，时间跨度覆盖了典型的业务高峰期与低谷期，具有较强的代表性与现实意义。其中，网络流量数据以5 min为粒度进行采样，记录了云池出口的总流速（单位为bit/s），经过预处理后形成连续、对齐的时间序列数据。该流量序列整体呈现出较强的周期性特征，如每日早晚的访问高峰、周末与工作日之间的流量差异等，同时也受到突发事件（如服务发布、故障恢复、网络攻击等）的影响，表现出一定的非线性与不确定性。

在流量数据之外，数据集还包含同期记录的网络变更日志，每条变更记录均包含变更发生时间、变更标题、变更描述以及变更影响等字段。其中，变更时间精确到分钟级别，可与流量序列进行时间对齐，变更标题为简要概括变更内容的短语，如“某个区出口切流变更”等，变更描述则以自然语言形式详细说明了变更的背景、具体操作内容等，变更影响字段由运维人员记录了该变更可能对网络服务状态、流量分布等方面造成的实际影响情况。需要指出的是，尽管变更事件在整个时间跨度中相对稀疏，但其对网络流量的影响往往具有显著性和可观察性，因此构成了本文方法中变更影响知识库的核心数据来源。

为了保证数据质量与建模效果，本文在数据预处理阶段进行了多项标准化处理。首先，对流量数据进行了缺失值插补与异常值检测，采用滑动窗口中位数滤波与线性插值相结合的方法填补缺失点。其次，对变更描述文本进行了清洗与规范化，去除冗余信息与非结构化表达，统一术语表述，并对部分描述不完整或语义模糊的变更记录进行了人工补充与标注，以提升后续语义检索与语言模型理解的准确性。最后，将流量数据与变更记录按时间戳进行对齐，构建出每条变更事件对应的变更前后流量时间窗口，用于构建变更影响知识库中的知识条目。

2.2 实验对比

为验证本文所提出方法在考虑计划性变更事件影响下的网络流量预测能力，实验选取了预测时间段内存在明确变更事件的历史样本作为测试集（测试集中的变更事件未参与知识库的构建），重点考察本文方法相较于仅依赖时序大模型的预测方法在变更影响建模方面的优势。具体而言，测试样本均来自数据集中变更发生前后的时间窗口，确保每条测试样本均包含一次计划性变更操作及其对网络流量产生的可观测影响。实验对比的两个方法分别为：1）仅使用Moirai时序大模型的预测方法，即不引入任何变更语义信息，仅基于历史流量数据进行未来流量预测；2）本文提出的双阶段协同预测方法，在Moirai预测基础上，结合变更描述文本与检索增强生成机制，通过语言大模型对预测结果进行语义修正。为直观说明两种方法在变更影响建模上的差异，图3展

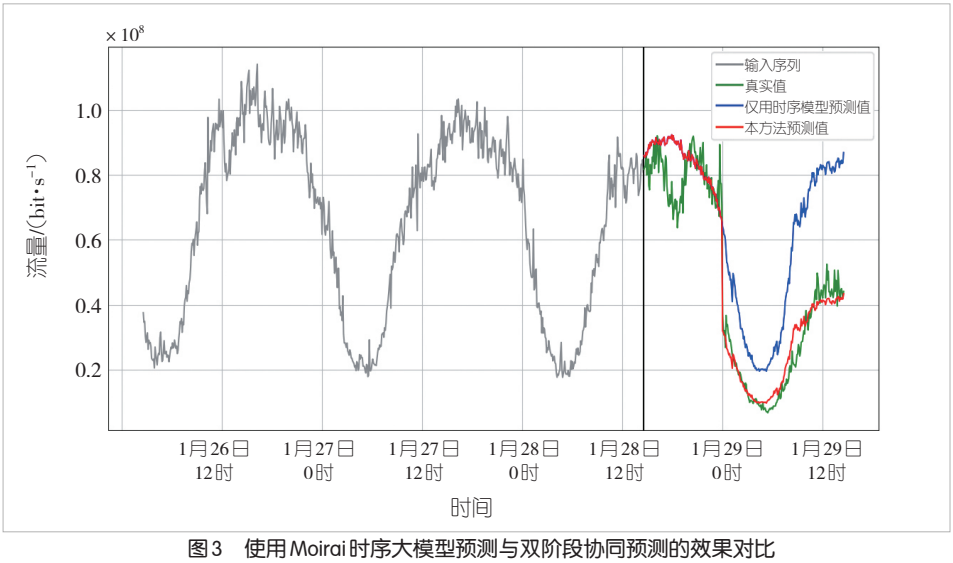

图3 使用Moirai时序大模型预测与双阶段协同预测的效果对比

示了一组测试样本中真实流量曲线、Moirai基础预测曲线与本文方法预测曲线的对比情况。

从图2可以看出，Moirai模型虽然能够较好地捕捉流量的整体趋势与周期性变化，但在面对计划性变更事件所引发的流量突增或突降时，其预测结果仍存在明显偏差。例如，在某次计划性服务迁移操作发生后，真实流量在短时间内出现了显著下降，而Moirai的预测结果则延续了历史趋势，未能准确反映该变更所带来的影响。相比之下，本文方法在引入变更描述与历史相似案例的基础上，能够有效识别出变更事件可能引发的流量变化模式，并对基础预测结果进行合理修正，使其更贴近真实流量变化趋势。

为进一步量化评估两种方法在变更事件影响预测中的性能差异，本文采用均方误差（MSE）和平均绝对误差（MAE）两项指标对测试集上的预测结果进行评价。表1展示了两种方法在包含变更事件的测试样本上的性能对比。

从表1可以看出，本文所提方法在两项评价指标上均显著优于仅使用Moirai的基础预测方法，表明本文方法在预测误差方面具有明显优势。这些结果充分说明，本文所提出的双阶段预测框架能够有效融合变更事件的语义信息，并结合历史相似案例进行上下文推理，从而显著提升在变更驱动场景下的流量预测精度。

表1 使用Moirai时序大模型预测与双阶段协同预测的量化对比

测试方法	MSE	MAE
仅使用时序预测模型	0.057 0	0.176 6
本文双阶段预测框架	0.007 4	0.061 1

MAE：均方误差　MSE：平均绝对误差

3 结束语

本文的研究不仅为网络流量预测提供了一种新的建模思路，也为时序预测与语言理解的跨模态融合提供了可借鉴的范式。未来的工作将进一步探索该方法在多类型网络场景中的适应性，并尝试引入更多上下文信息（如网络拓扑、服务依赖关系等），以提升预测模型的语义表达能力与推理深度。

参考文献

[1] AOUEDI O, LE V A, PIAMRAT K, et al. Deep learning on network traffic prediction: recent advances, analysis, and future directions [J]. ACM computing surveys, 2025, 57(6): 1–37. DOI: 10.1145/3703447
[2] LIM B, ZOHREN S. Time–series forecasting with deep learning: a survey [J]. Philosophical transactions of the royal society of London series A, 2021, 379(2194): 20200209. DOI: 10.1098/rsta.2020.0209
[3] MASINI R P, MEDEIROS M C, MENDES E F. Machine learning advances for time series forecasting [J]. Journal of economic surveys, 2023, 37(1): 76–111. DOI: 10.1111/joes.12429
[4] BENIDIS K, RANGAPURAM S S, FLUNKERT V, et al. Deep learning for time series forecasting: tutorial and literature survey [J]. ACM computing surveys, 2023, 55(6): 1–36. DOI: 10.1145/3533382
[5] ACHIAM J, ADLER S, AGARWAL S, et al. GPT–4 technical report [EB/OL]. (2023–03–15) [2025–08–16]. https://arxiv. org/abs/2303.08774
[6] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models [J]. ACM transactions on intelligent systems and technology, 2025, 16(5): 1–72. DOI: 10.1145/3744746
[7] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [EB/OL]. (2023–03–31) [2025–08–16]. https://arxiv. org/abs/2303.18223
[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre–training of deep bidirectional transformers for language understanding [EB/OL]. (2018–10–11)[2025–08–16]. https://arxiv.org/abs/1810.04805
[9] XIAO S T, LIU Z, ZHANG P T, et al. C–pack: packed resources for general Chinese embeddings [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2024: 641–649. DOI: 10.1145/3626772.3657878
[10] WOO G, LIU C, KUMAR A, et al. Unified training of universal time series forecasting transformers [EB/OL]. (2024–02–04) [2025–08–16]. https://arxiv.org/abs/2402.02592

作　者　简　介

常远，中国电信研究院大数据与人工智能研究所工程师；主要研究领域为大模型、智能体技术等；发表论文10余篇。

吴春鹏，中国电力科学研究院人工智能所副主任；主要研究领域为机器学习、边缘计算、生物启发视觉技术；发表论文50余篇。

王峰，中国电信研究院大数据与人工智能研究所副所长；主要研究领域为云计算、人工智能技术等；发表论文20余篇。