北航-华为关键软件联合实验室

项目开题报告

（2024年度）

项目名称：基于IDE开发流程与代码仓库的代码

深度理解技术研究

负 责 人： 石琳

编制时间：

目 录

[一、 课题背景 1](#_Toc16076)

[二、 研究目标 1](#_Toc27622)

[三、 技术方案与技术路线 2](#_Toc19302)

[四、 整体计划和关键路标 5](#_Toc32490)

[五、 风险、问题和建议 5](#_Toc6917)

1. 课题背景

大语言模型（LLM）的发展为软件工程智能化提供了广阔的前景。目前，大语言模型已经应用与部署到广泛的软件工程任务当中，辅助软件开发，其中最具代表性的任务就是代码生成。GPT-4等大语言模型已经能够集成到IDE中，在真实开发环境中根据开发人员的编码需求进行自动代码生成、自动代码建议等任务。然而，目前真实场景下基于大语言模型的代码生成技术依然具有一系列局限性。

**首先，大语言模型难以挖掘开发者真实意图。**在开发过程中，大语言模型往往只能根据程序员输入提示与当前代码文件上下文等知识给出代码生成建议，然而从些知识中可能无法获取开发人员真实意图。例如，程序员需要开发一个数据处理模块，由于数据处理算法的源代码细节依赖于具体的数据类型与格式，因此大模型根据自然语言描述以及代码上下文直接生成的算法可能由于不符合真实的数据处理场景而无法应用。这些隐藏的真实意图信息往往无法直接反馈给大模型，此时大模型生成的代码难以贴合开发项目特点。

**其次，大语言模型在仓库级别的代码生成上表现不佳。**代码仓库往往包含错综复杂的结构，不同文件的代码之间也存在依赖关系。因此将仓库代码上下文输入给大模型时，大语言模型可能无法理解代码仓结构与代码之间的依赖关系，同时由于输入给大模型的提示词具有长度限制，大模型也可能无法学习到需要的上下文知识。此时生成的代码可能无法应用到代码仓库中。

1. 研究目标

本项目旨在提出一种基于IDE开发流程与代码仓库的代码深度理解技术，通过开发者意图预测以及代码仓库地图检索加强大模型对IDE开发场景代码的理解，增强IDE集成大模型的代码生成效果。

**基于时序模型的开发者真实意图预测。**IDE 开发者的真实意图可能蕴含在开发者之前的动作中，例如，开发者在开发某个功能之前可能进行查看相关代码文件、搜索相关API等操作，因此可以通过开发者历史动作来预测当前开发者的编码意图。基于IDE开发者动作序列信息，利用时序模型对开发者意图进行预测。

**基于代码地图的代码仓深度理解。**代码仓库结构复杂，代码之间依赖关系不明显，因此需要基于代码仓构建一种数据结构来抽象整个代码仓库的信息，显示地表示代码结构关系与依赖关系。用图结构表示代码仓库，称为代码地图，并通过检索代码地图帮助大模型获取有效上下文信息，加深大模型对代码仓库的理解。

1. 技术方案与技术路线

本项目整体技术方案如图1所示，包括两个部分：（1）基于IDE开发动作序列的意图预测；（2）基于仓库代码地图的关联代码检索。通过开发者在IDE开发历史中的动作序列预测开发者编码意图，并通过提取仓库代码依赖关系与历史更改信息来构建代码地图，根据开发者意图与代码地图来加深大模型对代码的理解，增强代码生成效果。除此之外，本项目将基于IDE开发一个开发者意图理解和代码地图应用的原型系统，实现意图预测以及代码地图检索等功能。两部分技术路线如下。

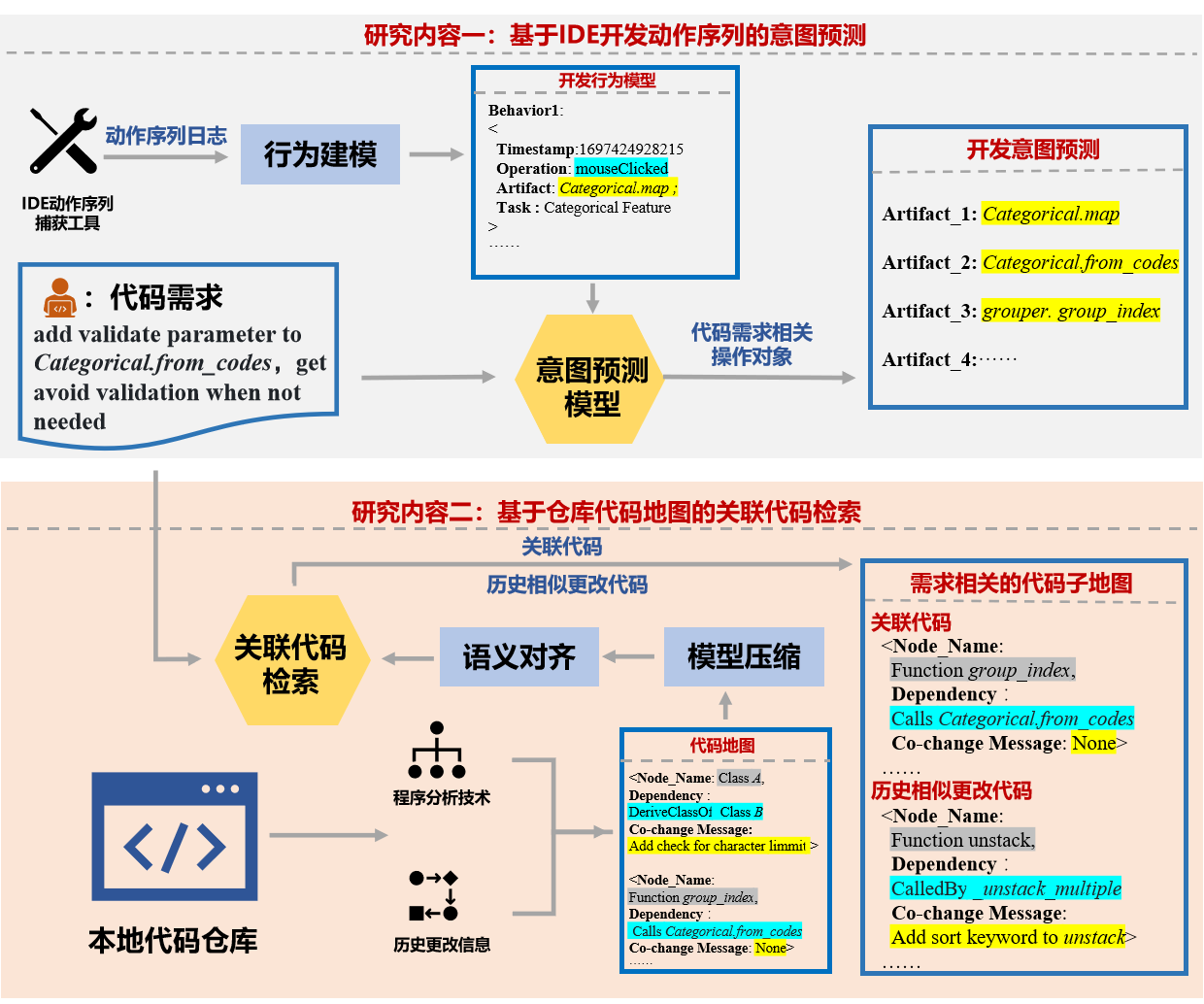


图 1 整体技术方案

**3.1 基于IDE开发动作序列的意图预测**

IDE开发流程中隐式蕴含了开发者的编码意图与开发思想，利用开发者的开发意图信息辅助代码生成会使得生成的代码更加贴近于开发者开发思维以及项目特点。因此本项目拟通过对IDE开发者行为进行建模，利用时序模型对开发者意图进行预测。具体技术路线如下：

**（1）开发者行为建模。**将开发者动作序列表示为：

其中代表开发者行为，包含行为发生时间、行为操作、操作工件以及开发者阶段性任务。操作工件

代表开发者行为操作对象，如搜索过的API、阅读的代码片段、搜索过的编码方法、检查过的功能、使用的代码生成提示等。

**（2）基于IDE的开发者动作数据收集。**为了构建开发者意图预测模型，需要用IDE开发者行为数据对模型进行训练。本项目拟通过IDE开发者行为跟踪插件来获取IDE项目开发场景中的实际开发者动作数据。

**（3）基于时序挖掘的程序员开发意图推荐。**基于当前的代码生成需求，将当下的代码生成行为表示为：

行为发生时间为当前时间点，行为操作为执行代码生成过程，操作工件为代码生成提示。基于包含代码生成需求的代码生成提示以及开发者历史动作序列，利用时序模型对当前代码生成行为进行意图预测：

意图是开发者开发历史中与代码生成需求相关的操作工件。

**3.2 基于仓库代码地图的关联代码检索**

进行仓库级代码生成的关键是充分获取代码仓库的相关信息。由于整个代码仓库的代码数量众多，同时其中有很多信息与代码生成需求无关，因此需要一种需求关联代码检索技术。本项目拟利用代码仓库依赖关系与开发历史变更关系来构建全局代码地图，依据代码生成需求对代码地图中的需求关联代码进行检索。具体技术路线如下：

**（1）基于程序分析技术构建代码仓库依赖关系。**利用仓库代码结构关系、控制流图CFG，数据流图DFG等构建依赖关系，表征仓库代码的结构与依赖关系。

**（2）基于开发历史构建协同演化关系。**依据仓库项目的历史更改，构建仓库代码的变更关系，历史变更代码信息可以为相似的代码生成需求提供参考。

**（3）基于和构建代码地图。**将代码仓库依赖关系与开发历史变更关系合并去重，得到代码地图模型。

（**4）根据需求和代码地图高效检索代码。**使用文本嵌入对代码地图进行压缩，

同时使用语义嵌入模型使需求文本与代码达成语义对齐，即

然后利用检索模型，对代码地图中与代码需求相关联的代码节点进行检索：

其中包括了与需求相关、具有依赖关系的代码以及具有与需求相似的历史更改的代码。

1. 整体计划和关键路标

|  |  |  |  |
| --- | --- | --- | --- |
| 项目阶段 | Q1 | Q2 | Q3 |
| 起止日期 | 2024.9.1-  2024.12.31 | 2024.1.1-  2024.4.31 | 2024.5.1-  2024.8.31 |
| 工作描述 | 完成IDE动作序列获取模块与开发者意图预测模型 | 完成仓库代码地图获取模块与关联代码提取模型 | 完成开发者意图理解和代码地图应用的原型系统开发 |
| 开发成果 | 《基于IDE开发流程和代码仓库的代码深度理解技术报告》中的“基于IDE开发动作序列的意图预测技术”部分 | 《基于IDE开发流程和代码仓库的代码深度理解技术报告》中的“基于仓库代码地图的关联代码获取技术”部分 | 开发者意图理解和代码地图应用的程序代码  开发者意图理解和代码地图应用的原型系统 |
| 验收标准 | 技术报告内容完整包括实验结果,预测精度不低于0.7 | 技术报告内容完整包括实验结果，NDCG@10指标不低于0.65 | 原型系统应实现开发者意图预测和代码地图关联代码获取的功能 |
| 验收方法 | 华为验收小组验收 | 华为验收小组验收 | 华为验收小组验收，包括原型工具功能测试 |

1. 风险、问题和建议

本项目在实际开发过程中可能存在以下难点：

**开发者历史动作数据获取。**为了训练意图预测模型，需要获取IDE 开发者的历史动作数据。目前已有相关IDE插件功能可以实时监测IDE开发者操作，并将动作数据保存在本地日志，然而需要结合真实IDE开发场景。因此为了收集真实场景IDE开发者动作数据，需要华为提供真实工程项目相关数据。

**构建有效实用的时序模型。**意图预测模型的训练数据集需要考虑开发历史动作序列时间点的选取、清除无效动作数据等问题；时序模型要考虑模型参数量大小与意图预测准确率的平衡，保证模型在基于IDE 的应用开发实用性。

**代码地图模型高效检索。**为了保证基于代码地图的检索模型应用可落地，需要考虑模型规模与检索策略优化提高关联代码的检索速度与准确率；同时要面向鸿蒙应用场景，实现自然语言与代码之间的语义对齐，需要联合ArkTS，仓颉编程语言工程师联合开发。