Mining Software Repositories: Étude Exploratoire du Dataset AIDev

Nihed Jamaoui

1 Introduction

Avec l'essor des agents de programmation basés sur l'intelligence artificielle, de nouvelles formes de collaboration émergent dans le développement logiciel. Ces agents sont désormais capables de proposer des modifications de code sous forme de *pull requests* dans des projets open-source hébergés sur GitHub.

L'objectif de ce travail est d'explorer un dataset réel issu de GitHub afin d'analyser le devenir et les caractéristiques des pull requests générées par des agents de programmation IA. Cette étude s'inscrit dans le domaine du *Mining Software Repositories (MSR)* et s'appuie sur le dataset AIDev, proposé dans le cadre du MSR Mining Challenge.

2 Objectifs du Travail

Les objectifs de ce travail sont les suivants :

- Étudier la structure et le contenu du dataset AIDev
- Identifier les pull requests générées par des agents de programmation IA
- Définir des questions de recherche objectives et mesurables
- Analyser les données à l'aide de métriques simples
- Extraire des observations exploratoires sur l'intégration des agents IA dans le développement logiciel

3 Présentation du Dataset AIDev

Le dataset **AIDev** est un dataset public qui regroupe des *pull requests* générées par des agents de programmation IA dans des projets GitHub open-source. Il constitue un dataset à grande échelle permettant l'étude des contributions automatisées dans des projets réels.

Le dataset est disponible sur la plateforme Hugging Face et est principalement stocké au format **Parquet**, un format optimisé pour l'analyse de grands volumes de données.

3.1 Chargement du Dataset

Le dataset a été chargé à l'aide du langage Python et de la bibliothèque `datasets`. La Figure 2 illustre le chargement et la structure générale du dataset.

```
from datasets import load_dataset

dataset = load_dataset("hao-li/AIDev")
print(dataset)
```

```
DatasetDict({
    train: Dataset({
        features: ['id', 'number', 'title', 'user', 'user_id', 'state', 'created_at', 'closed_at', 'merged_at', 'repo_url', 'repo_id', 'html_url', 'body', 'agent'],
        num_rows: 932791
    })
})
```

Figure 1: Chargement du dataset AIDev

3.2 Structure du Dataset

L'analyse des colonnes du dataset met en évidence les informations suivantes :

- informations générales sur les pull requests (identifiant, titre, description)
- état des pull requests (fusionnées ou fermées)
- dates de création et de fermeture
- informations sur les repositories GitHub
- identification de l'agent ayant généré la pull request (agent IA ou humain)

```
dataset["train"].column_names
```

```
['id',
 'number',
 'title',
 'user',
 'user_id',
 'state',
 'created_at',
 'closed_at',
 'merged_at',
 'repo_url',
 'repo_id',
 'html_url',
 'body',
 'agent']
```

Figure 2: Chargement du dataset AIDev

3.3 Subset du Dataset Utilisé

Dans cette étude, nous utilisons un sous-ensemble du dataset AIDev constitué des pull requests pour lesquelles les informations suivantes sont disponibles :

- l'identification de l'agent générateur de la pull request,
- l'état final de la pull request,
- les dates de création et de fermeture.

Les pull requests incomplètes ou ne contenant pas ces informations ont été exclues afin de garantir la cohérence de l'analyse.

4 Conception de l'Étude : Approche GQM

4.1 Goal

Le but de cette étude est d'analyser le devenir et le comportement des pull requests générées par des agents de programmation IA dans des projets GitHub open-source, afin de mieux comprendre leur intégration dans le processus de développement logiciel.

4.2 Questions de Recherche

À partir de cet objectif, nous formulons les questions de recherche suivantes :

- **RQ1 :** Quel est le devenir des pull requests générées par des agents IA en comparaison avec celles générées par des développeurs humains (fusionnées ou fermées sans fusion) ?
- **RQ2 :** Le temps de traitement des pull requests générées par des agents IA diffère-t-il de celui des pull requests humaines ?
- **RQ3 :** Quelle est la contribution relative des agents IA dans l'ensemble des pull requests du dataset AIDev ?

5 Métriques Utilisées

Pour répondre aux questions de recherche, nous utilisons les métriques suivantes :

- **Taux de fusion** : proportion de pull requests fusionnées parmi les pull requests générées par des agents IA et par des développeurs humains (RQ1).
- **Taux de fermeture sans fusion** : proportion de pull requests fermées sans être fusionnées (RQ1).
- **Temps de traitement** : durée entre la date de création et la date de fermeture d'une pull request (RQ2).
- **Proportion de pull requests IA** : nombre de pull requests générées par des agents IA rapporté au nombre total de pull requests (RQ3).

6 Résultats et Discussion

Les analyses réalisées montrent que les agents de programmation IA contribuent de manière significative à la création de pull requests dans les projets étudiés. Des différences sont observées entre les pull requests générées par des agents IA et celles générées par des développeurs humains, notamment en termes de taux de fusion et de temps de traitement. Ces résultats doivent être interprétés comme des observations exploratoires, dépendantes du subset étudié et des métriques utilisées.

7 Conclusion

Ce travail a permis d'appliquer une démarche structurée de *Mining Software Repositories* à un dataset réel issu de GitHub. L'étude exploratoire du dataset AIDev met en évidence son potentiel pour analyser l'intégration des agents de programmation IA dans le développement logiciel. Les résultats obtenus constituent une première étape et ouvrent la voie à des analyses plus approfondies dans de futurs travaux de recherche.