

ソフトウェア開発効率化のための パッチ分類システムの構築

所属 富士通株式会社 名前 味曾野 智礼 メールアドレス misono.tomohiro@fujitsu.com

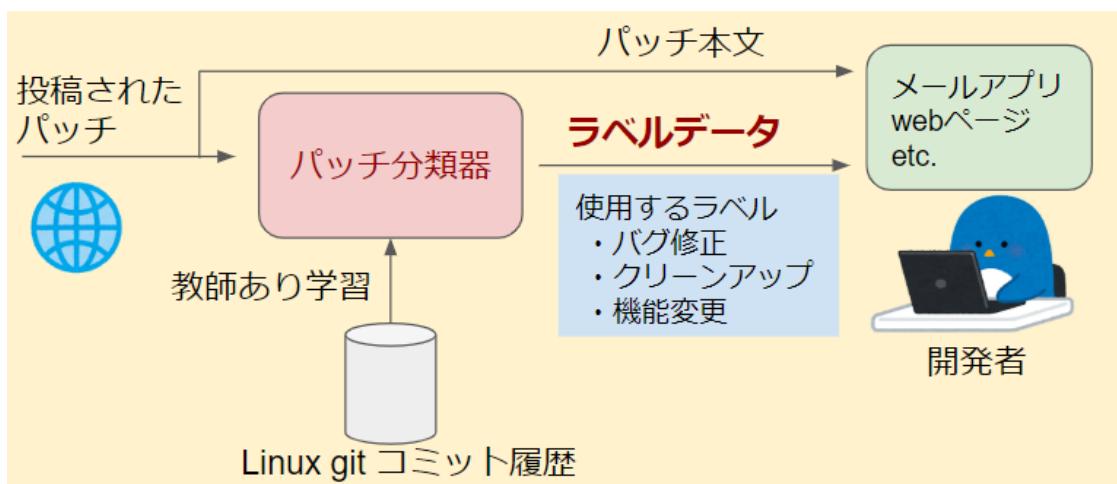
Linux開発作業における問題

- Linuxは従来よりメーリングリスト(ML)で開発が行われており、パッチもメールで投稿される
- 1日に1000件以上届くメールの中から目的のパッチを探すことは開発者にとって大変な作業
- パッチの種類(バグ修正、機能追加など)を自動で付与してくれるシステムがあると便利

手法・ツールの適用による解決

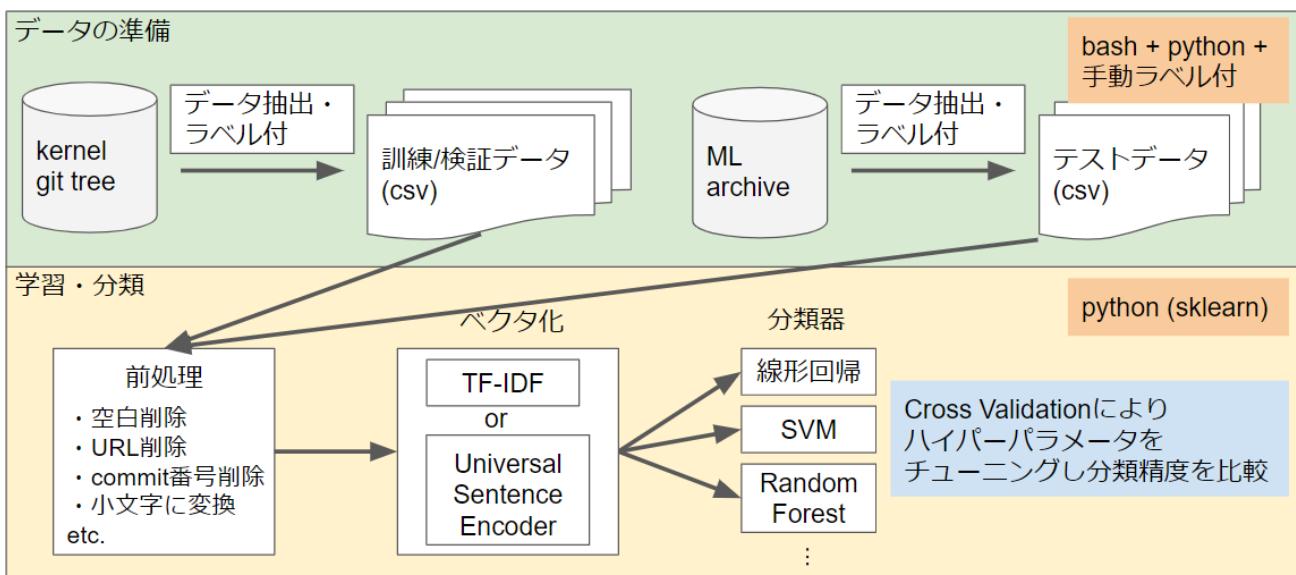
- Linuxでは習慣的にパッチの説明文(gitのコミットログに相当する部分)は長く、筆者の経験的にも多くの場合は(コード部分を見なくても)説明文から大まかなパッチの分類を行うことが可能
- よって本修了制作ではパッチの説明文に注目し、自然言語処理の手法でその文章をベクタ化し、機械学習によるラベル付けの精度を検証した

提案するシステムの概要



- Linuxのコミットログを訓練データとして教師あり学習によるパッチ分類器を作成
- その分類器を用いてMLに投稿されたパッチに対してラベル付けを実施

パッチ分類器の詳細・実験結果



- Linuxコードベース中のファイルシステム関連のコミットから約2000件の訓練データを用意
- 実際にファイルシステム関連MLに投稿されたパッチから100件のテストデータを用意
- 複数のベクタ化・機械学習手法を組み合わせ、学習精度を比較
- 結果として今回はTF-IDFによるベクタ化とSVMを組み合わせた場合が最も良く、約77%の精度で実際のMLに投稿されたパッチを分類することができた

※ 機械学習ライブラリにはscikit-learnを利用