

研究発表

OSSコミュニティの経時的変化を 考慮したコミッター候補者予測

山崎 大輝

背景

- OSSが社会に普及するにつれて、コミュニティへの不具合報告が増加
- 不具合を解決するためにパッチ投稿が増加し、それを検証するコミッターへの負担が増えている

コミッターに関する問題

- コミッター候補はプロジェクト管理者や既存コミッターが推薦[1]
- コミッター候補となる人物は、早期に発見しないとプロジェクトを脱退する恐れ
- 早期にコミッター候補となる人物を見つける必要があるが、推薦のためには長期にわたる活動が求められる

先行研究

- コミッターと一般開発者の活動量を元に
コミッター候補者を予測する^[1]
- OSSコミュニティは時期によって環境が変わると考えられる→データの性質が変わっている可能性が高い
- 時期を限定して予測モデルの説明変数にした方が、より高い精度でコミッター予測できるのではないか？

[1]伊原彰紀, 亀井靖高, 大平雅雄, 松本健一, 鵜林尚靖: OSSプロジェクトにおける開発者の活動量を用いたコミッター候補者予測, 電子情報通信学会論文誌,

目的とアプローチ

- 目的
 - コミッター候補となるべき人物の早期発見
- アプローチ
 - データセット期間を調整することでコミッター候補者予測の精度がどう変化するか調べる

データセット



期間 2001/10 - 2012/10

コミッター 195人

一般開発者 13,297人



期間 1999/07 - 2013/01

コミッター 530人

一般開発者 76,931人



期間 2000/12 - 2014/04

コミッター 363人

一般開発者 8,017人

活動量メトリクス

開発期間

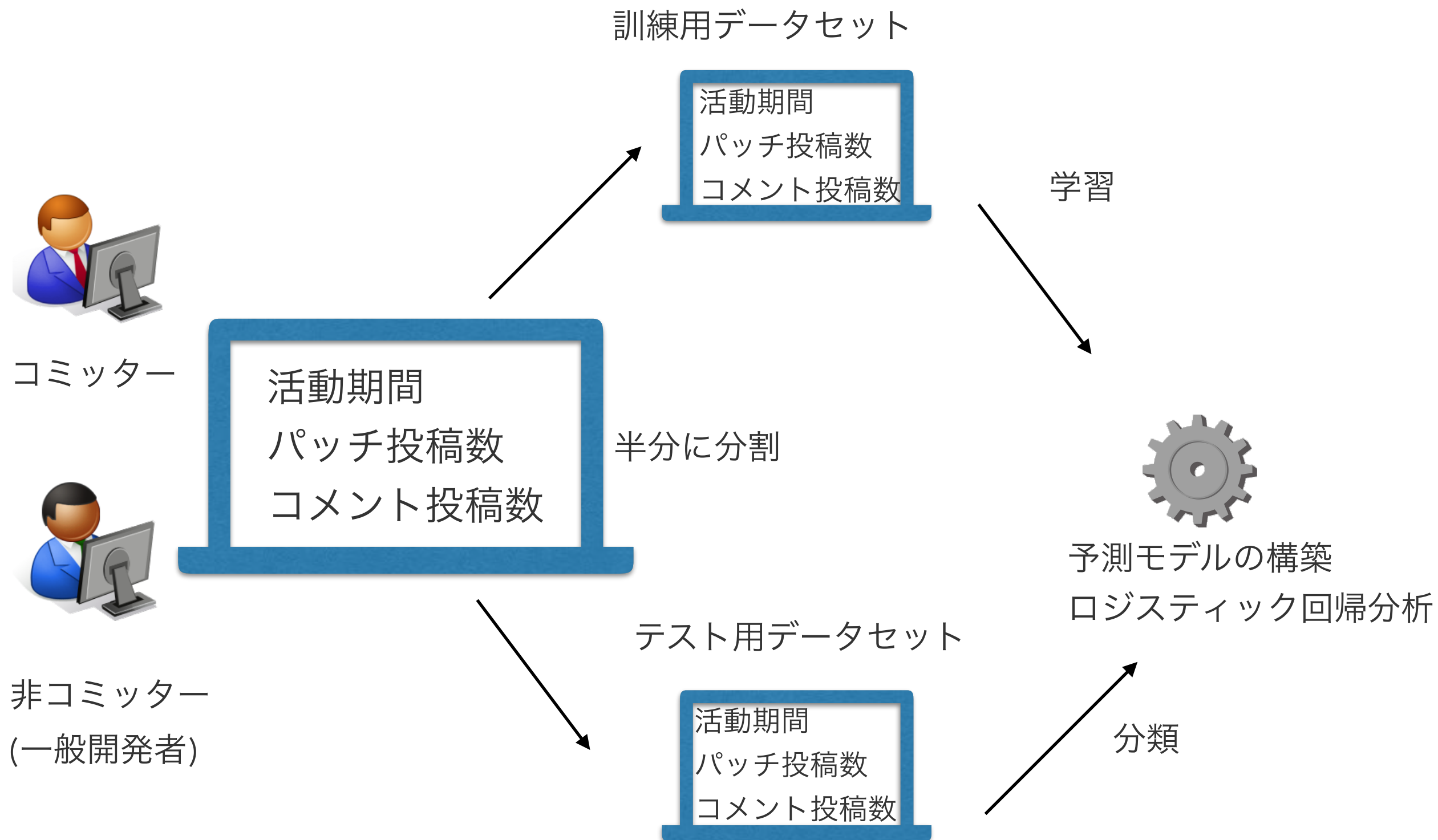
月当たりのコミット数

総コミット数

月当たりのパッチ投稿数

総パッチ数

コミッター候補者予測モデル



データセット

従来のデータセット

活動期間



t

データセット内全ての開発者の
活動量メトリクスを使用

データセット

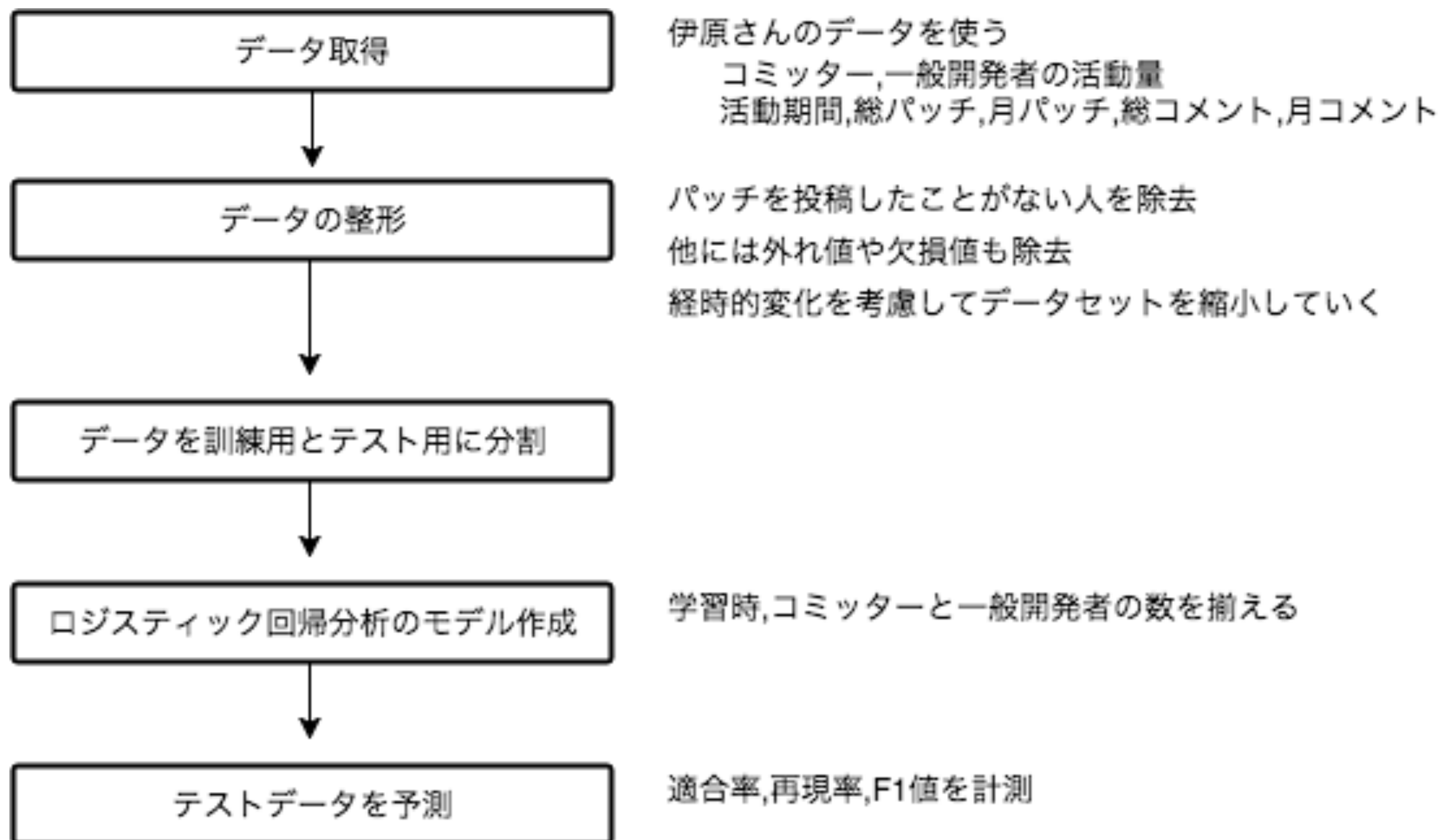
従来のデータセット

本研究のデータセット

コミュニティのプロセス改善など途中で
データの性質が変わっている可能性がある

期間を絞り込んでモデル構築

分析方法



結果

- Eclipse(2001~2012)

	Committer	Developer	Precision	Recall	F1
2001~	139	343	0.605	0.754	0.667
2002~	123	343	0.556	0.721	0.626
2003~	110	343	0.513	0.709	0.594
2004~	93	343	0.440	0.674	0.534
2005~	81	318	0.408	0.675	0.509
2006~	63	260	0.368	0.645	0.472
2007~	52	203	0.361	0.615	0.457

結果

- Webkit(2000~2014)

	Committer	Developer	Precision	Recall	F1
2000~	360	2499	0.440	0.9	0.591
2001~	359	2498	0.438	0.899	0.589
2002~	359	2498	0.438	0.899	0.589
2003~	359	2498	0.438	0.899	0.589
2004~	359	2498	0.438	0.899	0.590
2005~	340	2408	0.438	0.906	0.590
2006~	331	2280	0.450	0.915	0.604
2007~	321	2066	0.465	0.919	0.617
2008~	284	1782	0.463	0.915	0.614
2009~	214	1481	0.440	0.907	0.593
2010~	136	1121	0.407	0.912	0.562
2011~	62	710	0.304	0.871	0.451

結果

- Firefox(1999~2013)

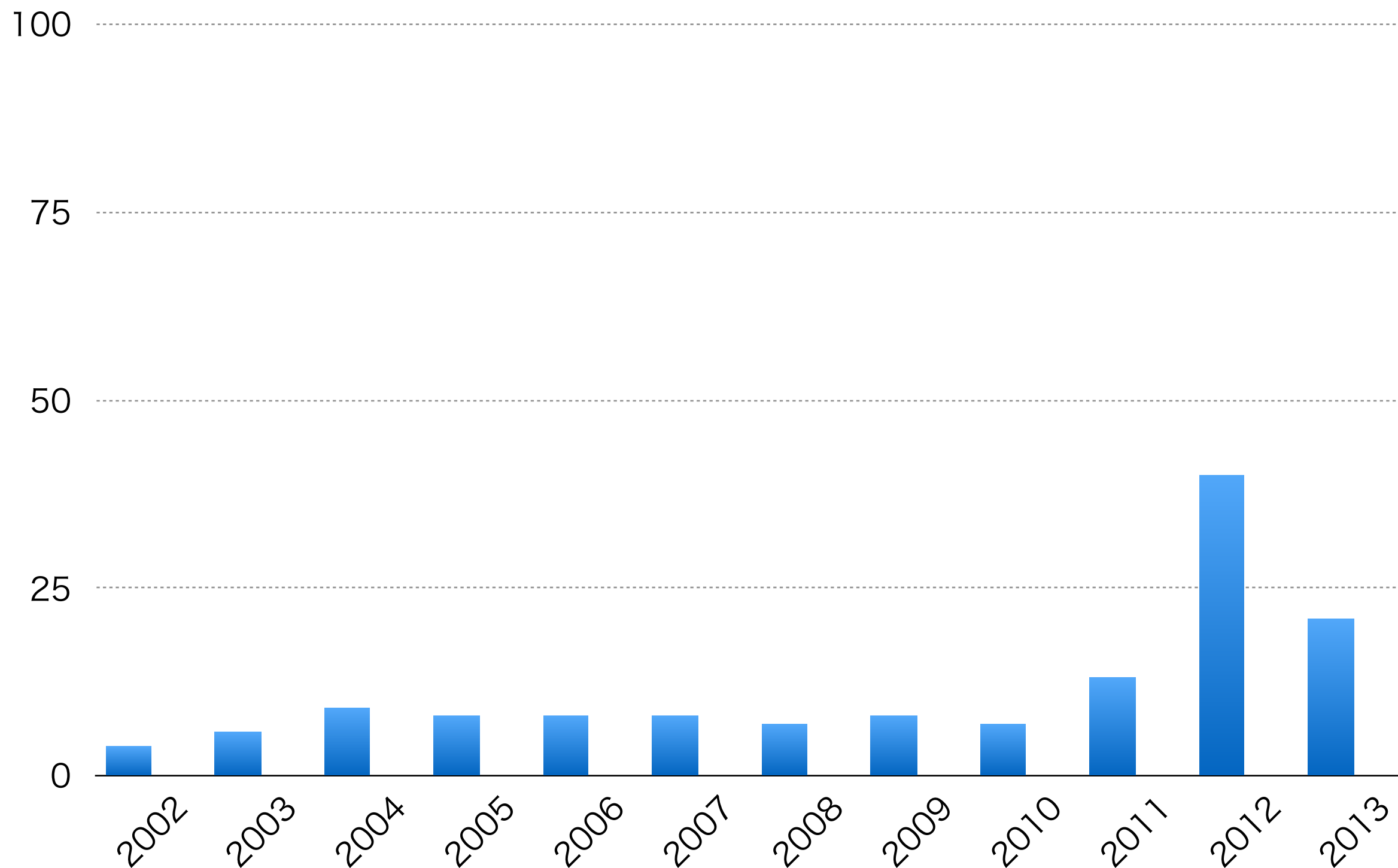
	Committer	Developer	Precision	Recall	F1
2000~	378	730	0.553	0.587	0.571
2001~	373	722	0.556	0.597	0.574
2002~	356	668	0.565	0.590	0.579
2003~	329	572	0.593	0.598	0.595
2004~	283	434	0.629	0.596	0.613
2005~	253	333	0.664	0.595	0.629
2006~	220	268	0.690	0.627	0.656
2007~	202	221	0.717	0.634	0.670
2008~	143	129	0.738	0.676	0.703
2009~	114	80	0.733	0.737	0.735

実験について

- データの区分間違えた可能性が高いので組み直す
 - どうデータを分割するか再考する
- 現在はその期間内に活動していた開発者をデータセットにしているが、その期間内にコミットした人をデータセットにする
- 交差検証のやり方が間違っている気がする
 - 今の実験ではコミッターになる人の予測ではなくコミッターかどうかの判別になっているので

コミッター昇格時期

Eclipse(計139人)



コミッター昇格時期

Firefox(計379人)

