

Mining Email Social Networks

Chirstian Bind, Alex Gourley,
Prem Debanbu, Michael Gertz,
Anand Swaminathan

MSR'06, May 22–23, 2006, Shanghai, China.

背景

- 大規模ソフトウェア開発では開発者間で多くのスケジュール調整とコミュニケーションが行われる
- このコミュニケーションは開発と関係があるかどうか調べたかったが普通の開発では調べることができない
- OSSプロジェクトではネットワーク(Email)を通じて行っているので調べることができる

リサーチクエストション(目的)

RQ1 開発者のネットワーク特性は何か?

RQ2 メーリングリストに多くのメッセージを送る開発者はソースコード変化時においても活発であるか?

RQ3 開発者と非開発者間に異なった行動はあるか?

RQ4 最も活発な開発者はどの状態が最も高いか?

アプローチ

- 電子メールとソフトウェア開発にはどのような相関があるのか？
- Cvsアーカイブのデータを用いて実験を行う
- 期間1999年から2006年の間のデータを使用する

事前準備

- メールアドレスの履歴を追跡したいがメールアカウントを複数使っている人もいる
- そこで似ているアドレス(ファーストネームやフルネームを使っているアドレスなど)を同じメールアドレスとして扱う
- 102611のメールのうち101637を解析することができたが残りの1.3%が解析できなかった

事前準備 例

フルネームがRasmus Lerdorfの人のメールアドレス

rasmus@apache.org,

rasmus@bellglobal.com,

rasmus@lerdorf.ca,

rasmus@lerdorf.com,

rasmus@lerdorf.on.ca,

rasmus@linuxcare.com,

このようなアドレスを同じアドレスとしネットワーク図
を作成する

ネットワーク図(一部)



調査要素

- Out degree 異なった人から受け取った程度
- In degree 異なった人に送信した程度
- Between ness (以下BW)
 - 誰かにメッセージを送信する際に最もそのメッセージを仲介しているかどうかの指標
- BWが高いとそのコミュニティの中心にいることがわかる

実験目的と実験方法

- 開発者と非開発者間の活発度の差を調べる
 - 活発度の中間値をとりt検定を行う
- 開発者の活動要素の関係を調べる
 - 相関表を作って調べる

実験結果 1

	Developer	Non- -Developer	T-value	Significance
Betweenness	0.0114	0.000140	5.07	$p < 0.001$
Out-degree	0.00666	0.000451	8.14	$p < 0.001$
In-degree	0.00794	0.000367	7.54	$p < 0.001$

開発者と非開発者の間にはどの値も大きな差があり
非開発者よりも活発であると理解できる

実験結果2

	changes	srcChanges	docChanges	outdegree
changes	1	0.789	0.932	0.520
srcChanges	0.789	1	0.514	0.712
docChanges	0.932	0.514	1	0.308
outdegree	0.520	0.712	0.308	1
indegree	0.474	0.679	0.263	0.971
betweenness	0.553	0.757	0.327	0.955

	indegree	betweenness	mean	min	max
changes	0.474	0.553	912	0	16289
srcChanges	0.679	0.757	420	0	5741
docChanges	0.263	0.327	492	0	13420
outdegree	0.971	0.955	0.0080	0	0.0396
indegree	1	0.917	0.0067	0	0.0260
betweenness	0.917	1	0.0011	0	0.0965

まとめ

- 開発者と非開発者の活発度には差がある
- 開発者はプログラム変更と文章変更を同じ程度で変更はしていない
- 三つの指標のうちBWの値が最もプログラム変更に影響を与えている