

Mining Email Social Networks

Christian Bird, Alex Gourley, Prem Devanbu,
Michael Gertz, Anand Swaminathan

発表者：辻 裕真

C&C活動

Communication & Co-ordination

コミュニケーションと調整

大規模なソフトウェア開発では開発者同士のC&C活動が非常に重要

C&C活動

商用ソフトウェアなどの開発でのC&C活動は通常公開されないが,OSSプロジェクトでは公開されている(通常メーリングリスト)

OSSプロジェクトのメーリングリストのアーカイブからC&C活動と実際の開発活動の関係を調査

OSSのML

OSSのMLは公開されており誰でもメッセージを投稿できる

MLには開発者,バグ報告者,パッチ投稿者,一般ユーザが参加しており非常にアクティブ

Apache:2004年 4996件 2005年 2340件

GCC:2004年 19173件 2005年 15082件

ApacheのML

メッセージの約73%には返信があった

最も活発な開発者はプロジェクトの存続期間中に4486件ものメッセージを投稿

最も少ない開発者で10件

ApacheのML

1件のメールに254人からの返信

1人の開発者が281人の異なる開発者に返信

しかし,MLに参加している大多数はメッセージの送信や返信もしない

C&Cと開発のアクティビティ

- ・メーリングリストで多くのメッセージを送信する開発者は、ソースコードの変更に積極的であるか
- ・開発者は、ソーシャルネットワークで非開発者とは異なる役割を果たしているか
- ・最もアクティブな開発者は、開発者の中で最高のステータスを持っているか

MLの分析

個人が複数のメールアドレスを使用
別の人物として分析すると混乱が生じる

エイリアス(メールアドレスに別名を付ける機能)は開発者の名前と関係ないものであることもある

MLの分析

CVS(バージョン管理システム)アカウントは一元管理されているので,メールアドレスとCVSアカウントを紐づけて分析を行う

MLの分析

まず,メールアドレスのヘッダを抽出して
ID,メールアドレス,名前のリストを作成

全てのIDのペアの類似性を調査し,十分に
類似しているものは同じクラスターに配置す
る

ApacheのMLの分析

2544個のIDが見つかり,クラスタリングアルゴリズムを用いて1581個のクラスターに分類

これらのクラスターにはエラーが含まれるので手動で後処理を行う

ApacheのMLの分析

手動での後処理のために意図的にクラスター類似性しきい値を非常に低く設定する

多くのクラスターから類似する2つを統合するよりも、少ないクラスターから手動で異なるものを分割するほうが簡単であるから

クラスタリングアルゴリズム

1. 名前の正規化

句読点, 接尾辞(jr等), 一般的な用語を削除し,
姓と名に分割

2. 名前の類似性

フルネームと姓と名の間の類似性をレーベンシュタイン距離に基づいてスコアを算出

クラスタリングアルゴリズム

3. 名前とメールアドレスの類似性

名前とメールアドレスが一致する場合 スコアを高くする(メールアドレスに姓と名が含まれている場合のみ)

一部のイニシャルとその他全体が一致する場合も高いスコアになる

例) Erin Bird は erinb と ebird に一致

クラスタリングアルゴリズム

4.電子メールの類似性

2つのメールアドレス(@以前)のレーベンシュタイン距離からスコアを算出

5.2～4のスコアの最大値をそれぞれのIDの類似性とし,クラスタを統合

クラスタリングアルゴリズム

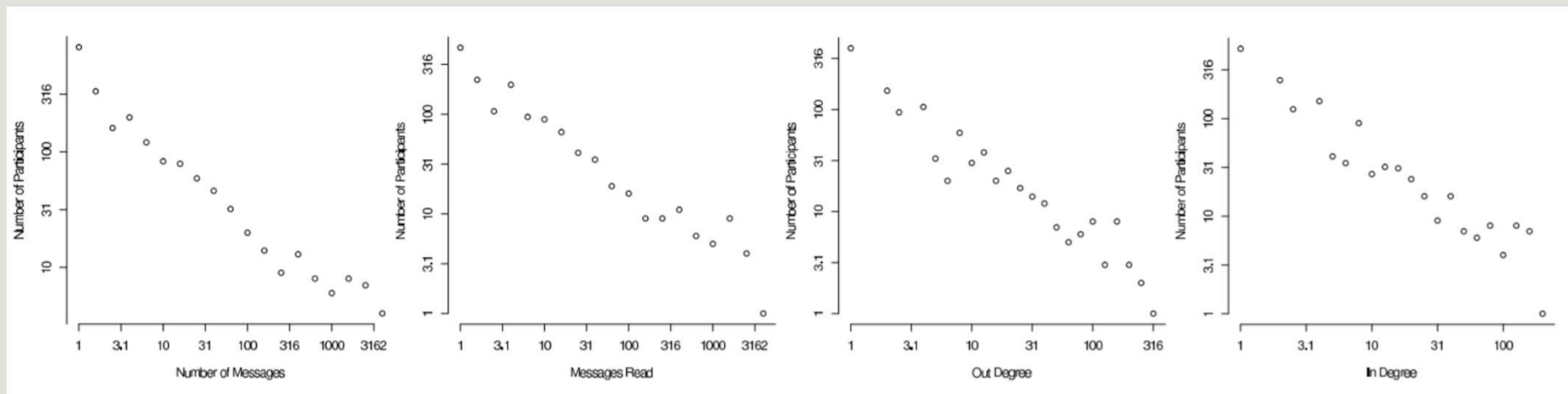
完成したクラスターを手作業で分割し,リストが完成

同様の手法でクラスターとCVSアカウントを紐づけ

データ抽出

ApacheのMLの10,2611件のメッセージから
ヘッダーの形式が正しくないものを除き
10,1637件のメッセージを解析

SMALL WORLD



対数目盛であることに注意

少数の人が送信と返信の大部分を占める

C&Cの活動と開発活動

- ・メーリングリストで多くのメッセージを送信する開発者は,ソースコードの変更に積極的であるか

CVSアーカイブからデータを収集し,行われた変更の数を調べる

C&Cの活動と開発活動

開発期間中にコミットしていた73人について調査

個人によって送信されたメッセージの数とソース変更の数の間のスピアマンの順位相関を計算すると約 0.80

強い相関があるので開発活動を多く行くとC&C活動も増える

開発者の相対的ステータス

73人の開発者のアクティビティを調査

	changes	srcChanges	docChanges	outdegree	indegree	betweenness	mean	min	max
changes	1	0.789	0.932	0.520	0.474	0.553	912	0	16289
srcChanges	0.789	1	0.514	0.712	0.679	0.757	420	0	5741
docChanges	0.932	0.514	1	0.308	0.263	0.327	492	0	13420
outdegree	0.520	0.712	0.308	1	0.971	0.955	0.0080	0	0.0396
indegree	0.474	0.679	0.263	0.971	1	0.917	0.0067	0	0.0260
betweenness	0.553	0.757	0.327	0.955	0.917	1	0.0011	0	0.0965

左から合計変更数,ソースコード変更数,ドキュメント変更数

開発者の相対的ステータス

メーリングリストでのアクティビティのレベルは、ソースコードの変更アクティビティと強く相関しており、ドキュメントの変更アクティビティとの相関性は低い

結論

少数の開発者がMLのメッセージの大半を占めている

MLのアクティビティのレベルとソースコードの変更のアクティビティには強い相関があったが、ドキュメントの変更のアクティビティとの相関は強くはなかった