

研究発表

オープンソースソフトウェア工学研究室

福井 克法

背景

- 効率的なソフトウェア開発には適切な人材配置が重要
- 組織の規模がある程度大きくなると開発者すべての業務経験を人為的に管理することは困難



先行研究

- ファイル変更履歴に基づいた検索手法
 - 構成管理システムに記録される活動を利用し、ファイルごとに関係のある開発者を検索する手法[1]~[5]
- 実際の開発組織で利用する上での問題点
 - 要件定義や基本・詳細設計にのみにしか関わっていない開発者を検索することができない

[1]Expertise Recommender: A Flexible Recommendation System and Architecture, CSCW, 2000, pages231-240

[2]Expertise browser: a quantitative approach to identifying expertise, ICSE, 2002, pages503-512

[3]Recommending emergent teams, MSR. 2007,

[4]Mining usage expertise from version archives, MSR, 2008,pages121-124

[5]Using Developer Interaction Data to Compare Expertise Metrics, MSR, 2013, pages297-300

目的とアプローチ

- 目的

- ソフトウェア開発の全工程を対象として開発者を広く検索できる手法を構築し効率的な人材配置を支援する

- アプローチ

- コミュニケーションデータ（メーリングリスト）を利用して検索手法の構築を行う
 - ・ コミュニケーションツールはソフトウェア開発において全工程の開発者が利用している

直近の進捗

- 1.サーベイ論文の調査
- 2.先行研究・関連研究の調査
- 3.今後の方針（提案手法）の設定

サーベイ論文の調査

- 「A Survey on Expert Recommendation in Community Question Answering」
 - 2018年のサーベイ論文
 - 開発者検索手法のサーベイ論文（QAサイトがメイン，構成管理システム，電子メールも少し）
- ▶ 開発者検索に用いる手法のデータ，アルゴリズム，評価法をメインにサーベイ

既存研究のアルゴリズム

Table 3. A Comparison of Inputs and Outputs of Representative Expert Recommendation Methods

Category	Representative method	I0	I1	I2	I3	I4	I5	I6	I7	I8	I9	IA	IB	IC	O1	O2
Simple methods	Votes			✓								✓				✓
	Best answer proportion			✓							✓					✓
	Consine similarity based on TF-IDF [19]	✓		✓			✓									✓
Language models	QLL [23–25]	✓		✓			✓									✓
	Category-sensitive QLL [27;27]	✓		✓			✓	✓								✓
	Expertise-aware QLL [6;28]	✓		✓			✓				✓					✓
Topic models	PLSA [15;34], LDA [36], STM [37], UQA [3]	✓		✓			✓									✓
	DRM [35]	✓	✓	✓			✓									✓
	TagLDA [39]	✓		✓					✓							✓
Network-based methods	Indegree [41], PageRank [47;48], HITS [46;53]													✓		✓
	z-score [41], Expertise-aware methods [14;41;45;55;56]			✓							✓			✓		✓
	Reputation-aware methods [60;61]												✓	✓		✓
	Category-sensitive methods [57;58]	✓						✓						✓		✓
	Graph embedding method [63]	✓	✓	✓			✓			✓	✓			✓		✓
Classification methods	SVM [65;66], C4.5 [65], RF [48], GBDT [70]	✓		✓	✓		✓		✓		✓	✓	✓	✓	✓	
	LTR [72]	✓		✓	✓		✓		✓		✓	✓	✓	✓		✓
Expertise probability model	Bernoulli MAP model [77]			✓							✓					✓
Collaborative Filtering	MF [79]	✓		✓	✓		✓									✓
	Tag-based PMF [81]	✓		✓	✓				✓		✓					✓
Hybrid methods	QLL+LDA [30], Topical PageRank [83], TEL [87], LDA+TF [90]	✓		✓			✓									✓
	Topical PageRank+Expertise [89]	✓		✓			✓				✓					✓
	QLL+LDA+userActivity+Indegree [30]	✓		✓		✓	✓				✓			✓		✓
	Indegree+Clustering [55]	✓		✓			✓				✓			✓	✓	

最近の手法はネットワークベース+ α が多い

対象がE-mailの先行研究

タイトル	会議	年度	概要	アルゴリズム	データ
Expertise Identification using Email Communications	CIKM	2003	・ E-mailデータを用いて組織内のエキスパートを探し出す ・ 検索キーワードを含む投稿でネットワーク手法を適用	HITS（ネットワーク手法）	E-mail
Graph-Based Ranking Algorithms for E-mail Expertise Analysis	DMKD	2003	・ E-mailデータを用いて組織内のエキスパートを探し出す ・ ネットワークアルゴリズムの比較（Pagerankが良い）	Pagerank, Power, HITS, Affinity	E-mail
Discovering Important Nodes through Graph Entropy The Case of Enron Email Database	KDD	2005	・ E-mailデータを用いて組織内の重要な役職の人物を探し出す ・ カウントベースとネットワークベースで結果を比較	中心性？	E-mail

対象がE-mailの先行研究

- 特徴

- ネットワークベースの手法が一般的
- 探したい人に関連するE-mailを集める部分がキーワード検索
 - ・ たぶん論文が古いから
- 単純に投稿数を数えるよりネットワークのほうが精度が良い

- 問題

- キーワード検索の部分が大変になる（キーワードを含んでいない投稿は検索できない）
- 全期間のエキスパートが出てくる（プロジェクトの重鎮ほど出てきやすくなる？）
 - ・ 時間で重み付けとかできそう？

関連研究

タイトル	会議	年度	概要	アルゴリズム	データ
Topical Authority Identification in Community Question Answering	ジャーナル	2012	・ 質問に対して適切な回答者を推薦する ・ 質問とユーザーが過去に回答した質問との類似度を算出したのち、ネットワーク手法を適用	LDA + Network 手法	Yahoo!Answer
Knowledge Graph Based Question Routing for Community Question Answering	ジャーナル	2017	調査中	Knowledge Graph Embedding	StackOverflow
Text-Based Question Routing for Question Answering Communities via Deep Learning	SAC	2017	調査中	DeepLearning	StackOverflow

これらの研究は、 質問に対して適切な開発者を推薦する手法

SE分野の先行研究

タイトル	会議	年度	概要	アルゴリズム	データ
Expertise browser: a quantitative approach to identifying expertise	ICSE	2002	・ 特定のファイルやメソッドに精通している開発者を検索する手法 ・ ファイルの作成履歴と変更履歴をカウント	カウントベース (ファイルの変更回数をカウント)	構成管理システム (ファイルの変更履歴)
Mining usage expertise from version archives	MSR	2008	・ メソッドごとにエクスポートを探し出す ・ メソッドの呼び出し回数をカウント	カウントベース (メソッドの呼び出し回数をカウント)	構成管理システム (ファイルの変更履歴)
Using Developer Interaction Data to Compare Expertise Metrics	MSR	2013	・ 時間を考慮した手法 ・ 最近に変更した開発者のほうがスコアが高くなる重み付け	カウントベース (ファイルの変更回数をカウント)	構成管理システム (ファイルの変更履歴)
Matching GitHub developer profiles to job advertisements	MSR	2015	・ ジョブサイトの広告の求人情報を開発者ごとに最適化する手法 ・ 求人情報と開発者のプロフィールの類似度でマッチング	NER, NED (テキストベースの手法)	Githubデータ (ReadMeファイル)

SE分野の先行研究の問題点

- ソースコードを変更する開発者しか対象にできない点
 - 要件定義や基本・詳細設計にのみにしか関わっていない開発者を検索することができない
- ファイルの変更・利用履歴に着目している点（カウントベースの手法）
 - ネットワークベースの手法の妥当性は示されている[6]
 - ネットワークベースの手法は、カウントベースの手法よりも優れているといった報告がある[7]

[6]Socio-technical developer networks: should we trust our measurements?, ICSE, 2011, Pages281-290

[7]Classifying developers into core and peripheral: an empirical study on count and network metrics, ICSE, 2017, Pages164-174

提案手法の方針

E-mailを対象とした既存手法の問題点の指摘，改善手法の提案

1. キーワード検索の部分の改善

- ネットワークベースの手法のE-mailを集める（キーワード検索）の部分を自動化

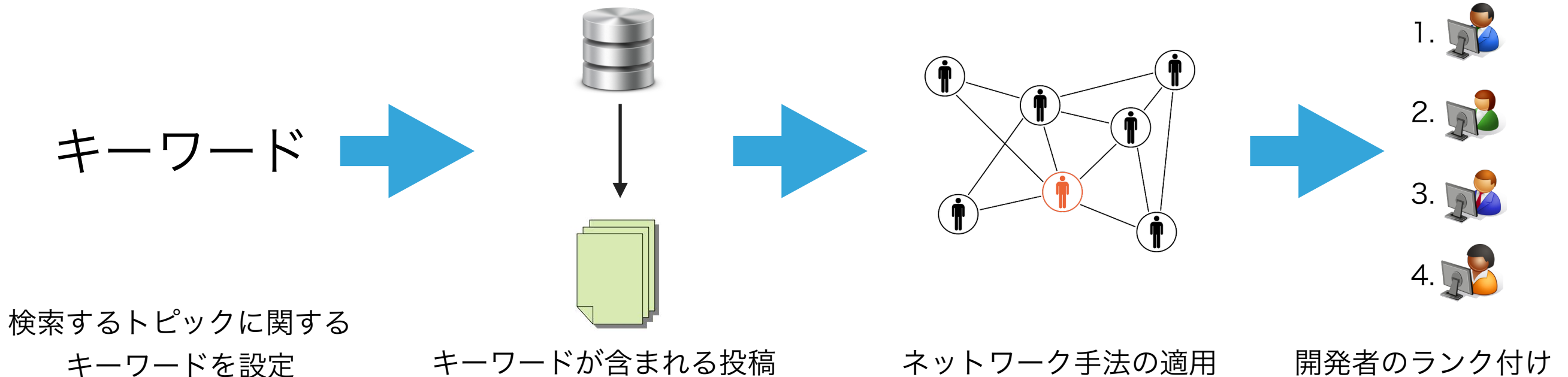
2. 提案手法に時間の重み付けを付け加えるか検討

- より最近キーワードを用いた人の方がキーワードに詳しい

3. 最新のネットワーク手法の調査（もしあれば比較）

1.キーワード検索の部分を改善

- ネットワークベースの既存手法

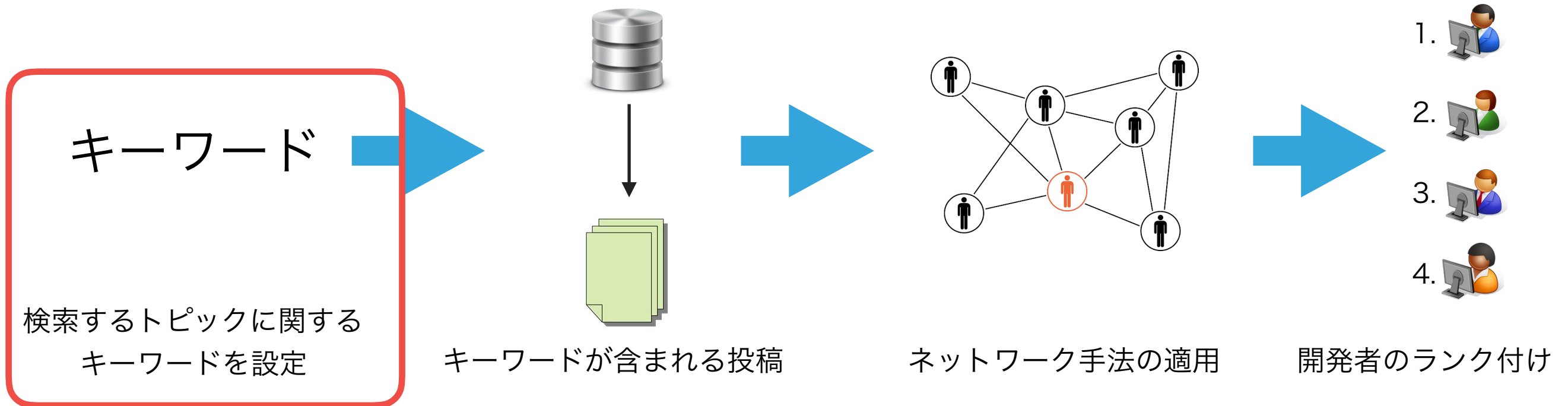


- 既存手法の問題点

- キーワードを設定する必要がある
 - 1つのキーワードでは不十分な可能性がある
 - 複数のキーワードを設定するのは手間

1.キーワード検索の部分を改善

- ネットワークベースの既存手法

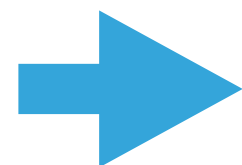


- 改善案

例：

java

キーワードの設定

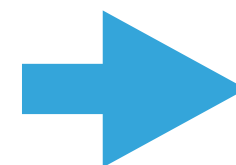


java



jdk
j-unit
maven
jsp

Term Expansion
(Javaに関係のある用語をあつめる)



java maven
jdk jsp
j-unit

キーワードとして設定

1.キーワード検索の部分を改善

- **Term Expansion**を行うアルゴリズムの候補
 - トピックモデルを用いて拡張
 - LDAを利用して共起関係にある語を取得
 - トピック分布を利用して文書を取得
 - 単語の分散表現を利用して拡張
 - Word2Vec, Fastext, Gloveなどを利用して類似している語を取得
- ▶ どのアルゴリズムでクエリを拡張するのが一番良いかの実験を行う

2. 提案手法に時間の重み付けを付け加えるか検討

- 時間によって重み付けを行っている論文[5][8]を参考に
 - 最近, 検索キーワードを使った開発者に高いスコアを付ける
 - ・ プロジェクトに参加したての開発者も検索対象に含めることができる
 - 重みをつける期間を指定できていいかも..
 - ・ 最近, 検索キーワードを使っている人は, 他の開発者も知ってる人の可能性がある

[5]Using Developer Interaction Data to Compare Expertise Metrics, MSR, 2013, pages297-300

[6]Uncovering the Dynamics of Crowdlearning and the Value of Knowledge, WSDM, 2017,

直近の予定

1. ストックしている論文の調査（残り6本程度）
2. ネットワークベースのベースライン手法の実装（この方針でいくなら）
3. 最新のネットワークベースのアルゴリズムの調査