

トピックモデリングに基づく開発者検索手法

5/14 研究発表

福井 克法

研究背景

- 効率的なソフトウェア開発を行うためには適切な人材配置が重要であるため、各開発者がこれまで携わった業務に関する経験を把握しておくことが望ましい
- しかし、開発組織が大きくなると経験を把握することは難しい
- ソフトウェア開発は、技術の発展が早いためさらに難しくする

関連研究と本研究の目的

- 関連研究

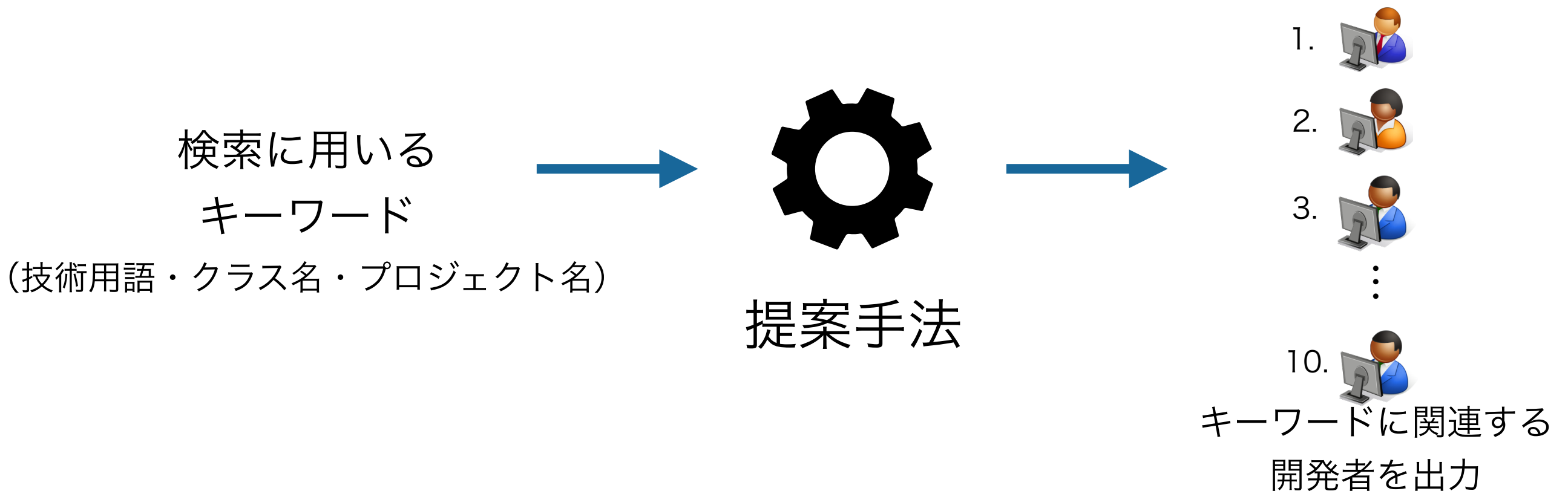
- ソースコードの変更履歴に着目したモジュール単位での適任開発者を推薦する手法が提案されている[1][2]
 - ▶ 特定のモジュールに精通した開発者を検索する

- 本研究の目的

- 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する
 - ▶ 全工程の開発者が共通で使用すると考えられるコミュニケーションツール（メールングリスト）のデータを用いることで手法を構築する

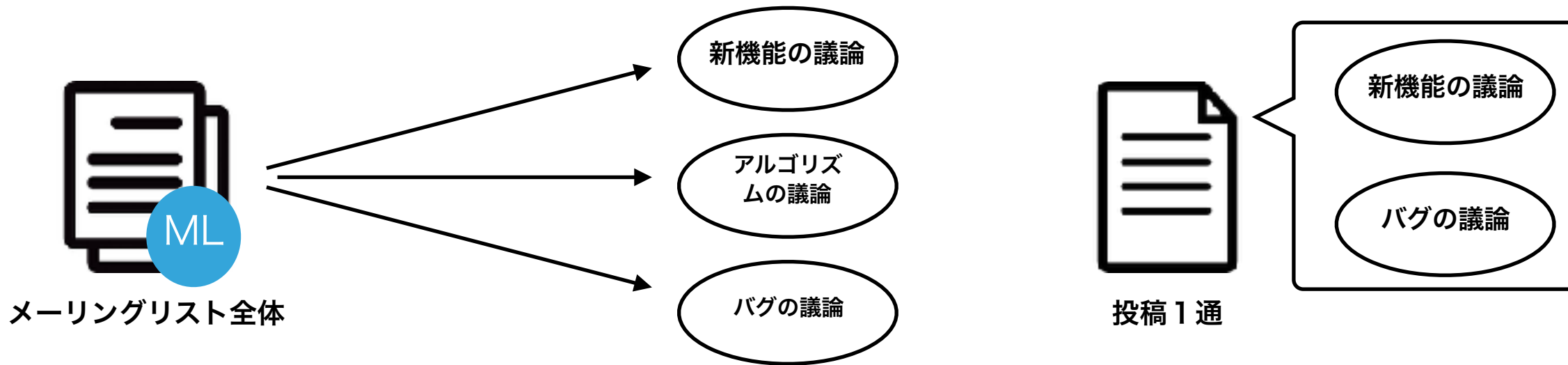
提案手法の概要

- メーリングリストに含まれる用語を利用して開発者を検索する
 - ▶ 各開発者が使用した用語を，開発者の開発に関する経験とする

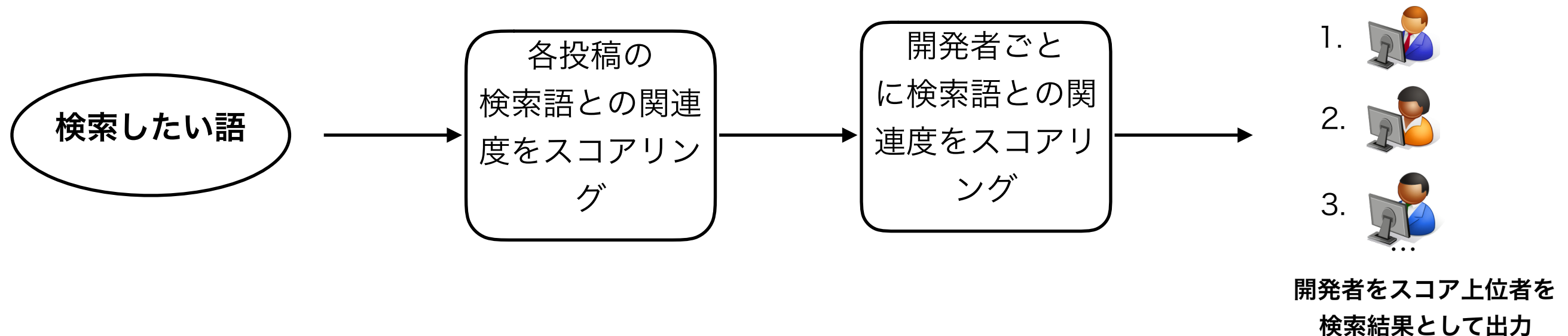


提案手法の構築の流れ

1. メーリングリスト全体に含まれる内容・各投稿の内容を特定

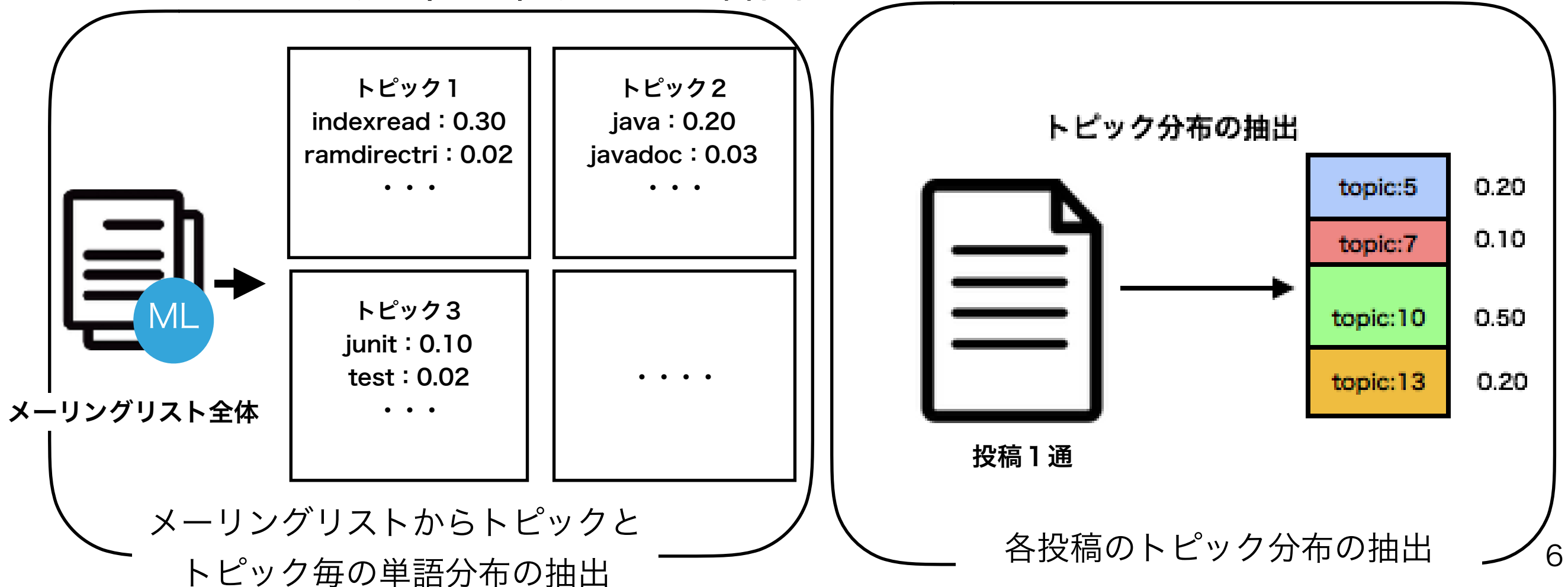


2. 技術用語や特定のプロジェクトに関する用語を入力として、開発者ごとに送信したメールと検索語との関係をスコアリング



1.メーリングリストと各投稿の内容の特定

- メーリングリストを目視で確認して内容の把握を行うことは困難
 - ▶ トピックモデルの潜在的ディリクレ分布（LDA）を用いる
- メーリングリストに含まれるトピック，各投稿の内容をトピック分布を利用して抽出



2.開発者のスコアリング

- 検索語との関連度をトピックの単語分布と各投稿のトピック分布を利用して算出する

1.検索語 w_i に対する各投稿 d_j のスコア(DocumentScore)

$$\sum_{k=1}^5 \theta_{t_k, w_i} \times \phi_{t_k, d_j}$$

検索語の生起確率

投稿のトピック分布

2.検索語に対する各開発者のスコア(AuthorScore)

$$\sum_{m=1}^{NumberOfMail} DocumentScore_{w_i, d_m}$$

θ_{t_k, w_i} : トピック t_k における検索語 w_i の生起確率

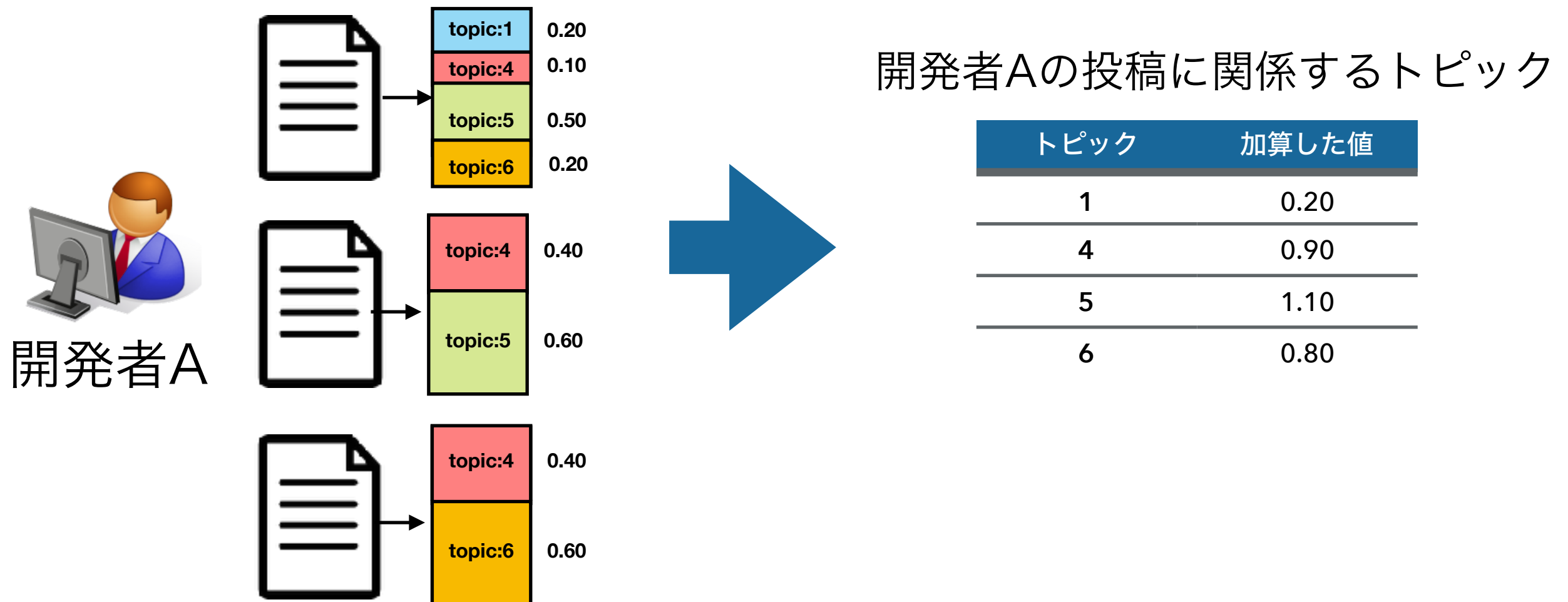
ϕ_{t_k, d_j} : 投稿 d_j におけるトピック t_k のトピック分布の比率

複数の検索語への対応

1. 各開発者ごとに各開発者のメーリングリストへの各投稿のトピック分布を加算する
2. 検索語の生起するトピックの特定と生起確率の加算
3. 開発者と検索語の関係性を測る

各開発者の投稿のトピック分布

- 各開発者の各投稿のトピック分布の合計値を算出
 - 各開発者がどのトピックに関わりがあるかの把握



検索語の各トピックでの生起確率の加算

- 検索語が高い確率で生起するトピックは、検索語に関係のあるトピックであると仮定する

[検索語1, 検索語2]

トピック	生起確率
1	0.25
4	0.10
6	0.05

トピック	生起確率
1	0.10
3	0.10
6	0.05

各検索語のトピックを特定



トピック	加算した値
1	0.35
3	0.10
4	0.10
6	0.10

検索語の各トピックでの生起確率を加算

検索語1, 検索語2が生起するトピックの値が大きくなる₁₀

各開発者と検索語との関係性

- 検索語の生起確率の高いトピックに関係のある投稿を多くしている開発者が、検索したい開発者と考えられる

開発者Aの投稿に関係のあるトピック

トピック	トピック分布の合計
1	0.20
4	0.90
5	1.10
6	0.80

検索語が生起するトピック

トピック	生起確率の合計
1	0.35
3	0.10
4	0.10
6	0.10

$$\text{Score}[\text{dev_A}] = 0.20 \times 0.35 + 0.90 \times 0.10 + 0.80 \times 0.10$$

トピック1 トピック4 トピック6

今後の予定

- 評価実験の方法を考える
- トピック数の決定方法を考える