トピックモデリングに基づく開発者検索手法の構築

オープンソース・ソフトウェア工学研究室

60190101

福井 克法

背景

- 効率的な開発には適切な人材配置が重要
- 開発者の全ての業務経験を把握することは困難

人材情報の管理が
困難

先行研究と問題点

- **先行研究**
  - ファイル変更履歴に基づいた検索手法[1][2]

- **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対象とできない

[1]Mockus, A. and Herbsleb, J. D.: Expertise Browser: A Quantitative Approach to Identify- ing Expertise, Proceedings of the 24th International Conference on Software Engineering, ICSE ’02, pp. 503–512 (2002).

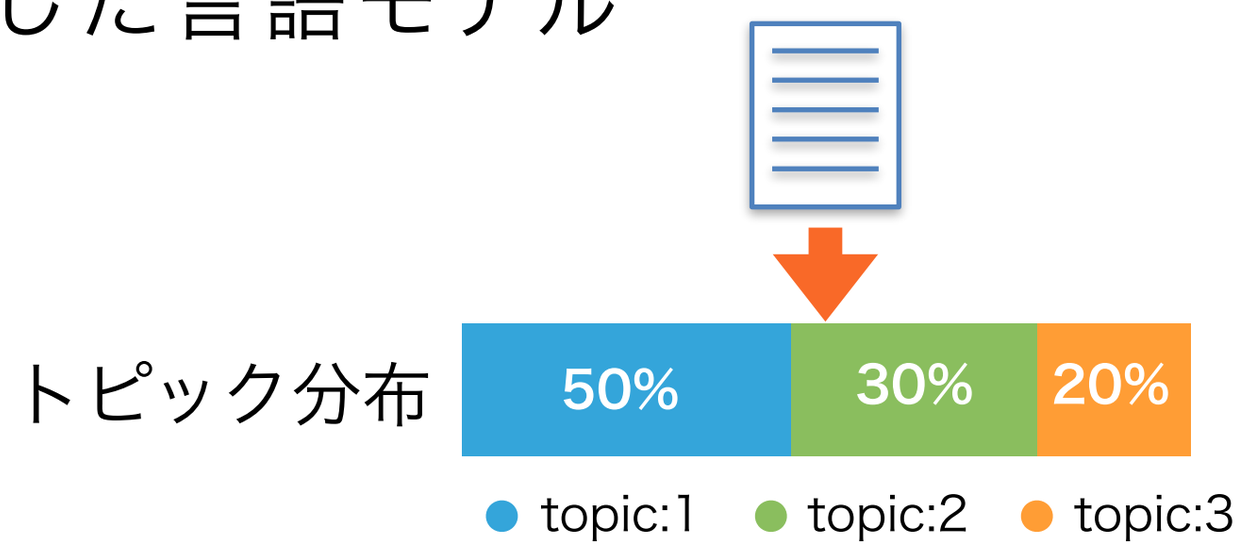
[2]Robbes, R. and R¨othlisberger, D.: Using Devel- oper Interaction Data to Compare Expertise Met-rics, Proceedings of the 10th Working Conference on Mining Software Repositories, MSR ’13, pp. 297– 300 (2013).

目的とアプローチ

- **目的**
  - 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
- **アプローチ**
  1. メーリングリストアーカイブにトピック分析を適用
  2. 結果に基づいて検索単語により開発者を検索

トピック分析

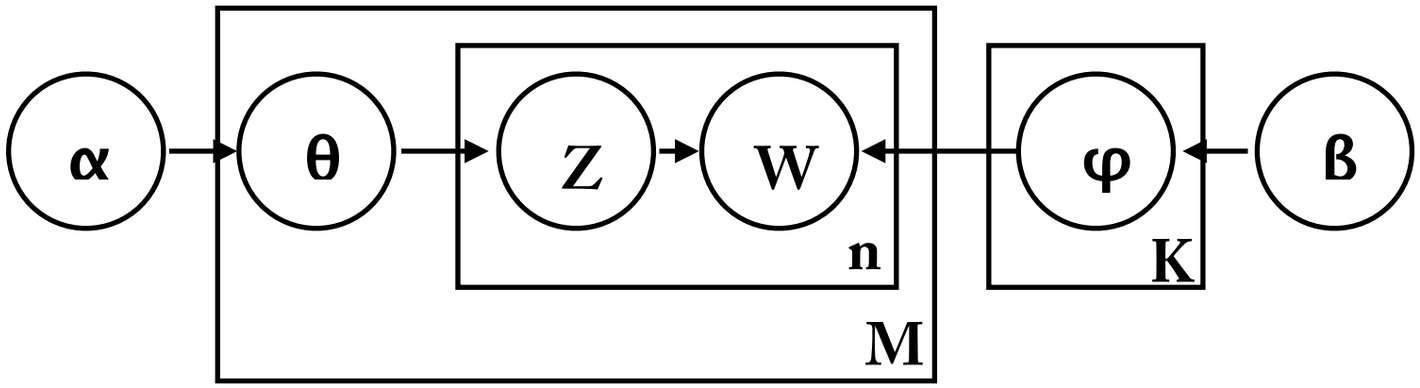
- **トピック分析**
  - 文章に含まれる可能性あるのトピック（話題）を抽出する
- **Latent Dirichlet Allocation(LDA)**
  - 1つの文書が複数のトピックからなることを仮定した言語モデル



  - トピック分布，トピックごとの単語分布を求める

Latent Dirichlet Allocation

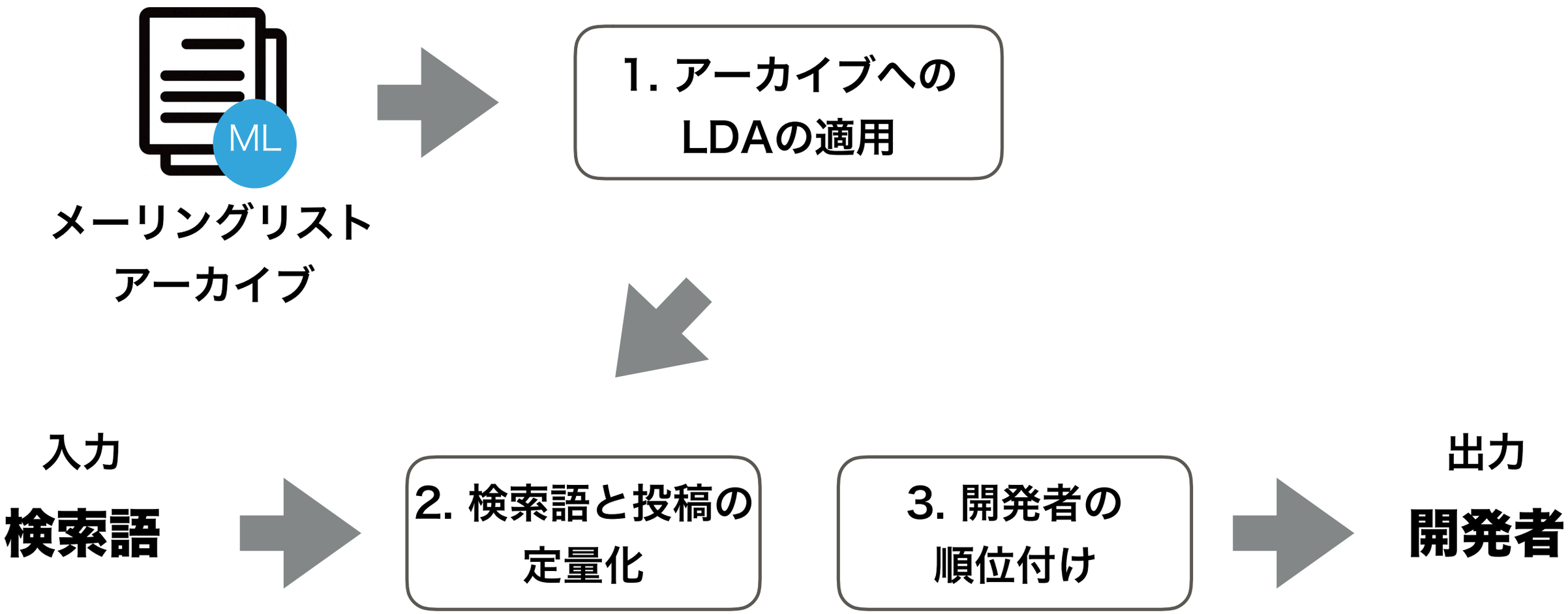
- 文書生成過程



LDAのグラフィカルモデル

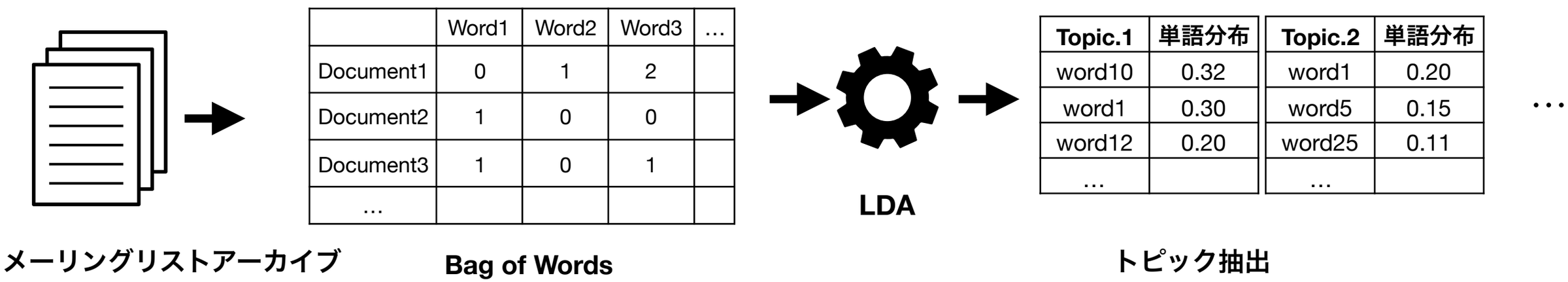
θ:文書におけるトピックの単語分布
φ:トピックにおける単語の多項分布
Z:語Wに割り当てられたトピック
K:トピック数
M:文書数
n:単語数
α:トピック数 θが従うディリクレ分布のハイパーパラメータ
β:単語分布φが従うディレクリ分布のハイパーパラメータ

提案手法の概要

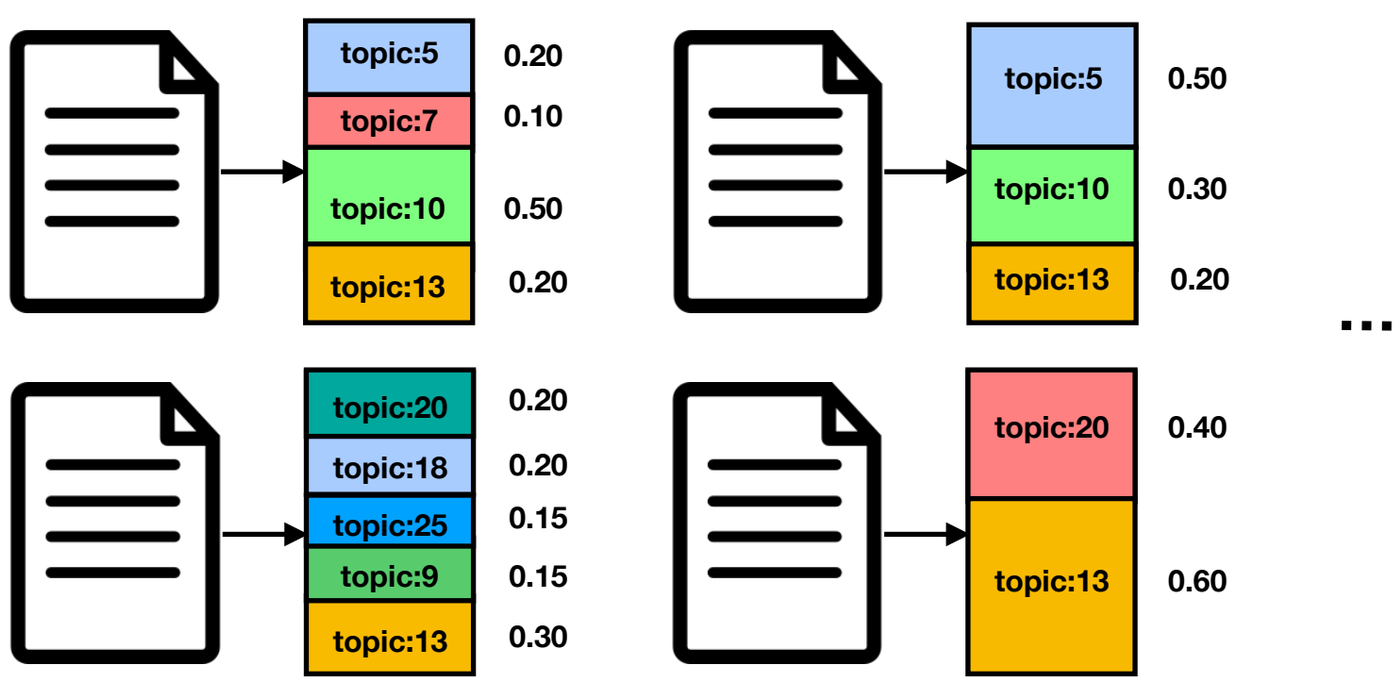

ML
メーリングリスト
アーカイブ
1. アーカイブへの
LDAの適用
入力
検索語
2. 検索語と投稿の
定量化
3. 開発者の
順位付け
出力
開発者

1. 1. メーリングリストアーカイブへのLDAの適用

1. メーリングリストアーカイブ全体のトピックと単語分布の抽出



2. メーリングリストアーカイブへの各投稿のトピック分布の抽出



2. 検索語と各投稿の定量化

- 検索語が出現するトピックを特定

検索語の生起確率の
高いトピックを特定

検索語 →

トピック:1

単語分布	生起確率
Word_A	0.20
検索語	0.13
Word_B	0.10
Word_C	0.08
Word_D	0.05

トピック:3

単語分布	生起確率
検索語	0.09
Word_E	0.08
Word_A	0.06
Word_C	0.02
Word_D	0.01

…

トピック:20

単語分布	生起確率
検索語	0.12
Word_A	0.10
Word_C	0.08
Word_F	0.40
Word_G	0.20

…

トピック:2

単語分布	生起確率
検索語	0.10
Word_A	0.90
Word_B	0.82
Word_E	0.50
Word_C	0.20

トピック:4

単語分布	生起確率
Word_E	0.10
Word_D	0.80
Word_A	0.70
Word_C	0.50
検索語	0.04

…

トピック:21

単語分布	生起確率
Word_B	0.22
検索語	0.10
Word_C	0.81
Word_F	0.70
Word_A	0.50

…

2. 検索語と各投稿の定量化

- 検索語と各投稿，各開発者を定量的に評価する

1.DocumentScoreの算出（検索語と投稿の関係）

検索語の生起確率　投稿のトピック分布

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

2.AuthorScoreの算出（検索語と開発者の関係）

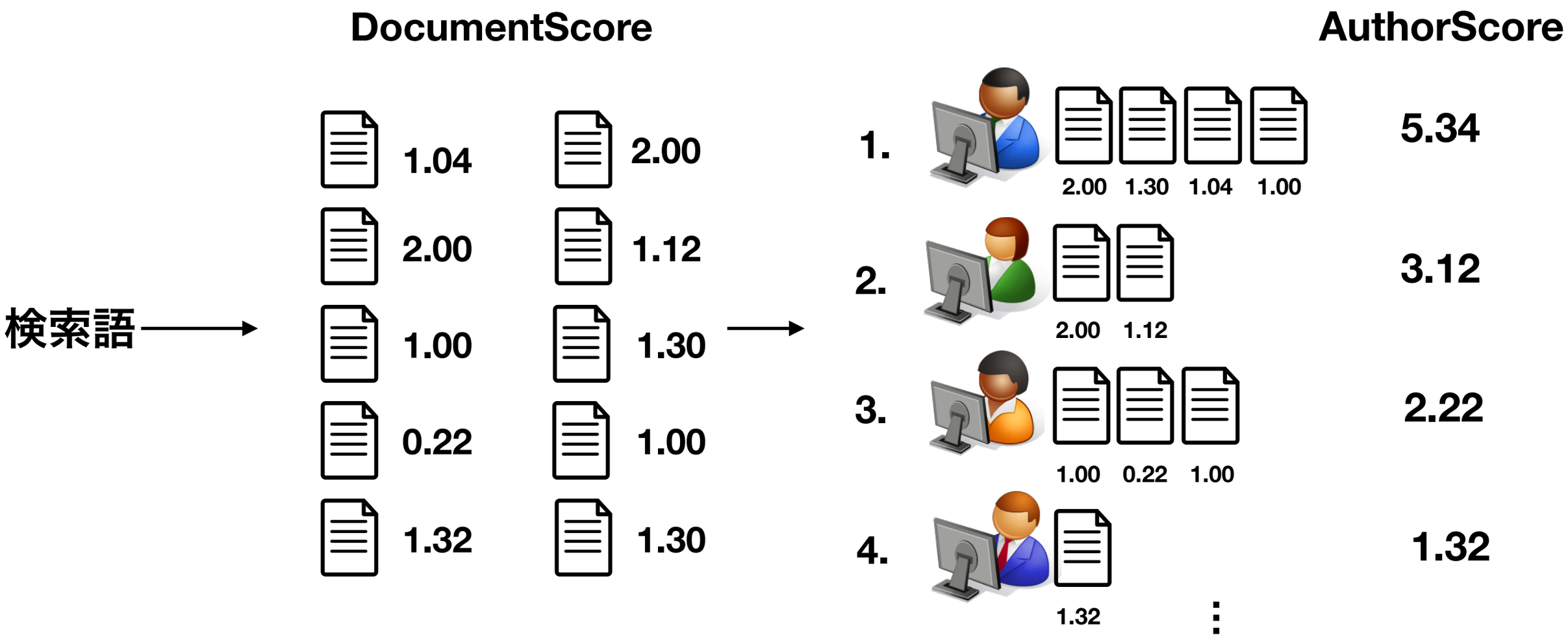
$$\sum_{m=1}^{NumberOfMail} DocumentScore_{w_i,d_m}$$

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

3. 開発者の順位づけ

- 検索語と関係のあると考えられる開発者を順位付ける



検索語と投稿の関係　　検索語と開発者の関係

評価実験

- **目的：有用性の評価**
  - キーワードに関係する開発者を正しく検索できていることを確認する
- **実験方法**
  1. 正解集合を作成するために複数プロジェクトのメーリングリストを合成
  2. 各プロジェクトに関する語を用いて開発者を検索
  3. 開発者の正答率で評価を行う

データセット

- Hadoopのサブプロジェクト
  - 比較的内容が近いと考えられるため

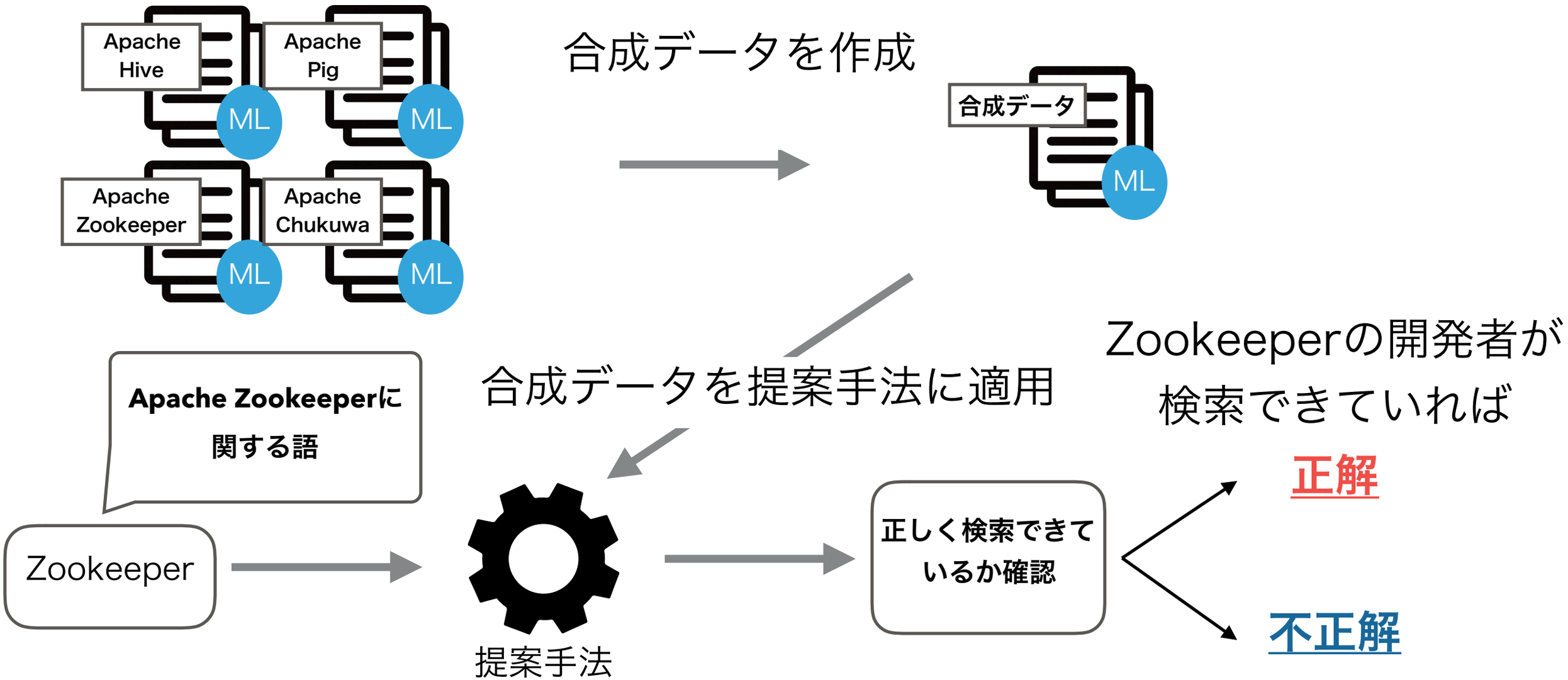
プロジェクト	期間	件数	開発者数
Apache Hive	2008/11/11~2010/10/07	1,881	28
Apache Pig	2007/10/31~2010/09/28	1,662	44
Apache Zookeeper	2007/10/31~2010/09/28	1,506	23
Apache Chukuwa	2009/03/10~2013/10/08	877	25
合計		4,209	118

実験手順

1. 各OSSプロジェクトから検索語を抽出
   - 特徴語(tf-idf値を利用)と頻出語(df値を利用)
2. LDAのトピック数を100, 500, 1000に設定し手法を適用
3. 手法により出力された上位10名の開発者の正答率を算出

実験の概要

- Zookeeperの開発者を検索する例



抽出した特徴語

- クラス名や変数名が抽出できている

	Apache Hive	Apache Pig	Apache Zookeper	Apache Chukuwa
1	hiveopensourc	physicallay	nioservercnxn	adapter
2	srcpart	pig	zookeep	chukwa
3	metaexcept	mapreducelay	nioservercxn	hicc
4	jobconf	cocoon	todd	demux
5	thrift	gld	clientcnxn	collector
6	srcbucket	inactiveaccount	socket	seqfilewrit
7	protect	piggybank	peer	agent
8	numreducetask	eq	antlib	row
9	processnam	acct	hunt	depart
10	libjar	foreach	elect	scienc

抽出した頻出語

- プロジェクト名やソフトウェア開発で頻繁に利用される語が抽出できている

	Apache Hive	Apache Pig	Apache Zookeper	Apache Chukuwa
1	hiveopensourc	pig	zookeep	chukuwa
2	hive	physicallay	hunt	adaptor
3	srcpart	mapreducelay	nioservercnxn	hicc
4	thrift	schema	clientcnxn	demux
5	metaexcept	foreach	todd	collector
6	jobconf	gate	elect	scienc
7	srcbucket	cocoon	socket	depart
8	protect	piggybank	nioservercxn	jiaqi
9	numreducetask	gld	lock	agent
10	metastor	javacc	peer	seqfilewrit

特徴語を用いた結果

- 各検索語で検索を行った際の上位10名での正答率平均

プロジェクト	トピック数		
	100	500	1000
Apache Hive	81%	61%	62%
Apache Pig	91%	100%	99%
Apache Zookeeper	61%	70%	73%
Apache Chukuwa	71%	67%	78%

頻出語を用いた結果

- 各検索語で検索を行った際の上位10名での正答率平均

プロジェクト	トピック数		
	100	500	1000
Apache Hive	78%	65%	66%
Apache Pig	86%	99%	91%
Apache Zookeeper	66%	75%	75%
Apache Chukuwa	71%	69%	82%

正答率の低い検索語

正答率が特に低かった検索語

thrift	**libjar**	**javacc**	**socket**
antlib	**row**	**scienc**	hicc
peer	**numreducetask**	srcbucket	seqfilewrit
jobconf			

- ソフトウェア開発において頻繁に利用される用語
- 複数のプロジェクトで使われる語が多い
  - thrift：Apache Thrift
  - javacc：Java Compiler Compiler　．．．

まとめと今後の課題

- **まとめ**
  - MLアーカイブにトピック分析を適用し検索手法を提案
  - 合成データを用いて評価実験の実施
    - 正答率は6割〜10割
- **今後の課題**
  - 評価実験に合成データを用いている
  - 検索語の抽出の再検討

大学院での予定

- 別アルゴリズムとの比較
  - Word2vecなど
    - Word2Vecにより，比較的意味の近い語が取得できることを確認（Apache httpdメーリングリストの語を学習）

入力語

類似度の高い語

database	network
rdbm	packet
msql	gigabit
access	zero-copi
postgr	latenc
queri	ethernet
informix	sniffer
postgresql	sendfil
postgres95	congest
metabas	socket
mysql	traffic