# 研究者および研究分野の変遷の自動推定

堀さな子 *1　村田真樹 *1　徳久雅人 *1　馬青 *2
*1 鳥取大学 工学部 知能情報工学科
*2 龍谷大学 理工学部 数理情報学科
{s072048,murata,tokuhisa}@ike.tottori-u.ac.jp
qma@math.ryukoku.ac.jp

## 1　はじめに

研究者にとって，研究者や研究分野の変遷を知ることは必要不可欠である．これを知るためには一般的に，Web や検索エンジンを使用して情報を得る方法があげられるが，これは網羅的に収集するのが困難であり，かつ多大な労力を要する．そこで，本研究では研究者や研究分野の変遷を自動的に抽出する方法を提案する．

川中ら [1] は，類似する問題点の改善として，ソーシャルブックマークにおける概念を記述するタグを解析することで，概念の派性関係 (概念の変遷情報) を自動的に抽出した．

この川中らの方法も変遷情報の抽出に利用できる．本研究では川中らの方法と比較実験を行い，提案手法の方が川中らの方法より有効であることを確認する．

本研究の主な貢献としては変遷情報の抽出手法として先行研究より性能の高い手法を構築したこと，具体的に提案手法を利用して，自然言語処理研究者にとって有用な情報である自然言語処理分野の研究者および研究分野の変遷情報を抽出したことである．

## 2　関連研究

先に示した川中らの研究以外に関連研究として以下のものがある．

松尾ら [2] は Web 上の情報を用いて共起の強さから人物の関係性の強さを推定し，かつ「共著関係」や「同研究室関係」などの社会的関係性を判別し，その情報が示された人間関係ネットワークを作成した．

Adar ら [3] はブログ上での情報の流れについて，テキストの類似度，リンク，時間情報を元に解析するモデルを提案した．

丹羽ら [4] はソーシャルブックマークにおけるユーザベースの共起度とドキュメントベースの共起度を比較することで，Synonym と呼ばれる同じ意味で用いられる語を共起度の高い精度で発見する手法を提案した．

## 3　提案手法

論文の著者として，ある人名 A が出現した最初の時期に同時に共起し (それもなるべく最初の方で多く共起すると良い)，人名 A より初出現年が早い人名 B は，人名 A のルーツ (先輩) である可能性が高いと思われる．分野名においても同様のことが言える．この仮説に基づき，本実験の手法を示す．

### 3.1　人名の変遷情報の推定方法

**手順 1**　論文から著者名データ (本論文では著者名と共著の人名を合わせたものを著者名データとする) を抽出し，その中から指定した人名を抽出し人名 A とする．

**手順 2**　人名 A を含む著者名データを取り出し，その中より (最初の時期によく共起した情報を取り出したいため) 出現年の早いものから 10 件の著者名データを取り出す．

**手順 3**　その 10 件の著者名データから共起している人名すべてを取り出し，人名 $B_i$($i$ は整数．$B_i$ は共起している人名の異なり数だけ設定する．) とする．3.3 節の方法で重みを付け，出現した論文の分だけ人名 $B_i$ ごとにその重みを加算する．

**手順 4**　初出現年が人名 A の初出現年よりも早く，重みが最も大きい人名 (人名 B) を人名 A のルーツとする．

図 1 に図式化したものを載せる．

### 3.2　分野名の変遷情報の推定方法

**手順 1**　「言選」を使用し，論文データのタイトル (またはアブストラクトも含めてもよい．ただし本研究ではタイトルのみを利用する．) から名詞連続を取り出し，不要な語を人手で省く．

以下，[手順 2] から人名の変遷情報の推定方法と同様．

### 3.3　重み付け手法

本研究では最初の時期に共起するものほど重要と考え大きな重みを付け，また共起回数が多いほど重要とも考え出現した回数だけ重みを加算するという手法を取る．[手順 3] で人名 A を含む論文の著者名データ 10 件の，年毎に出てきた著者名データに含まれる人名すべてに重み $a^{i-1}$($i$ = 出現年-初出現年，$a < 1$) をつける．

例えば，初出現年が 1990 年の場合，1990 年に出てきた論文の著者名データに出現する人名すべてに重み 1，

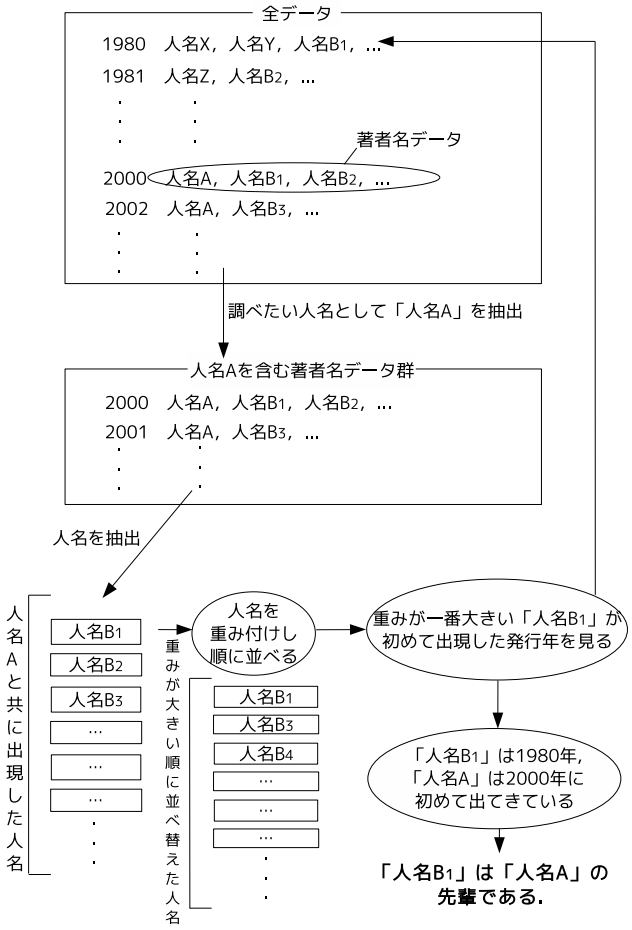


図 1: 人名の変遷情報の推定方法

1991 年に出てきた論文の著者名データに出現する人名すべてに重み $1 \times a$，1992 年に出てきた論文の著者名データに出現する人名すべてに重み $1 \times a \times a$，... を与える．このように年毎に重みを付与する．この例では，1990 年に出てきた論文の著者名データに出現する人名が一番重要と考え，一番大きい重みをつける．その重みを人名ごとに加算し，重みが一番大きいものをその人名の先輩と判断する．

$x$ という人名の重みを加算した $score(x)$ を数式化すると以下のようになる．

$$score(x) = \sum_{i=1}^{10} a^{i-1} g_i(x) \qquad (1)$$

$g_i(x)$ : $x$ という人名がその年の論文に出現していれば 1，そうでなければ 0 をとる．分野名の変遷推定でも同様の方法を取る．なお，$a$ の値は 0 から 1 に変化させ，違いを見る．

# 4 実験結果

言語処理学会年次大会の論文 1995 年から 2010 年の 3,139 件のデータを使用し，人名と分野名のルーツとなるものを抽出した．出力例を以下に示す．

例：「松吉俊」の実験結果

```
人名：松吉俊 (2005)
a = 0
3 --- 宇津呂武仁 (1995)
3 --- 佐藤理史 (1995)
-----
a = 0.1
3.34 --- 佐藤理史 (1995)
3.32 --- 宇津呂武仁 (1995)
-----
a = 0.2
3.76 --- 佐藤理史 (1995)
3.68 --- 宇津呂武仁 (1995)
-----
.
.
```

例：「自動構築」の実験結果

```
分野名：自動構築 (1999)
a = 0
1 --- コーパス (1995)
0 --- シソーラス (1995)
-----
a = 0.1
1.00001 --- コーパス (1995)
0.10001 --- シソーラス (1995)
-----
a = 0.2
1.00032 --- コーパス (1995)
0.20032 --- シソーラス (1995)
-----
.
.
```

出力例の，最初の「人名：」の部分に表示された人名が人名 A にあたる．$a$ を 0 から 1 に変化させたものを順に表示し，「重み — 人名」で重みの大きいものから順に表示させている．一番重みの大きいもの，すなわち一番最初に表示されている人名を人名 B(先輩) とする．この人名の例で説明すると，「松吉俊」は人名 A で，$a = 0$ での人名 B にあたる人名は「佐藤理史」である．また，この例の正解は「佐藤理史」であるがパラメータ 0 では解が「佐藤理史」にしぼりきれない．しかし，$a = 0.1$

または 0.2 では解を「佐藤理史」にしぼることができる．なお，人名と同時に表記される括弧の中身は，その人名が初出現した年号である．分野名の例も同様である．

また，抽出したものの例を表 1，表 2 に示す．この例は，$a = 0.5$ のものである．

表 1: 人名の変遷

| 人名 A(後輩) | 人名 B(先輩) |
|---|---|
| 村上仁一 | 池原悟 |
| 馬青 | 井佐原均 |
| 宮尾祐介 | 辻井潤一 |
| 関根聡 | 井佐原均 |
| 丸山岳彦 | 柏岡秀紀 |
| 黒田航 | 井佐原均 |
| 難波英嗣 | 奥村学 |
| 松吉俊 | 佐藤理史 |
| 竹内孔一 | 影浦峡 |
| 橋本力 | 奥村学 |

表 2: 分野名の変遷

| 分野名 A | 分野名 B(ルーツ) |
|---|---|
| 自動評価 | 機械翻訳 |
| 統計的機械翻訳 | 統計 |
| サンプリング | コーパス |
| タグ付きコーパス | コーパス |
| 音声対話システム | 音声対話 |
| 語義曖昧性解消 | 曖昧性解消 |
| 翻訳自動評価 | 機械翻訳 |
| 情報分析 | 分析 |
| 言語横断情報検索 | 情報検索 |
| 論文要約 | 情報抽出 |

# 5 評価実験

人名 A として 44 件，分野名 A として 32 件を使用し正解率を算出する．評価は言語処理学会に精通している人物が行う．結果として出力される人名 B(または分野名 B) の部分はランダムに表示して評価を行う．評価の基準を次に示す．

| | |
|---|---|
| 人名 | 言語処理学会に初めて発表した当時の指導的立場の人． |
| 分野名 | 言語処理学会においてルーツである分野名の 1 つとして考えられるもの． |

この評価基準に適するものを正解とする．システムの出力の 1 番目に正解を持つ場合に得点 1 を付け，合計を出し，すべての件数で割る．なお，正解と同じ点数を持つものが n 個存在する場合，得点 1/n 点をつける．

また，人名 B(分野名 B) が出現した論文が 2 個以上ない場合，データ不足とし評価対象に入れていない．実験結果を以下の表 3，表 4 に示す．

表 3: 提案手法の人名の実験結果

| $a$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|---|
| 正解率 | 0.45 | 0.59 | 0.59 | 0.59 | 0.58 | 0.58 |
| $a$ | — | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 正解率 | — | 0.60 | 0.60 | 0.56 | 0.51 | 0.41 |

表 4: 提案手法の分野名の実験結果

| $a$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|---|
| 正解率 | 0.40 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 |
| $a$ | — | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 正解率 | — | 0.48 | 0.48 | 0.48 | 0.48 | 0.45 |

結果は，総合すると人名は $a$ として 0.1 から 0.7，分野名は $a$ として 0.1 から 0.9 が比較的良い正解率を出すことがわかった．また，$a = 0$ は最初に共起したもののみ考慮した場合であり，これより重み付けを行って複数の出現を考慮したものの方が正解率が高い．$a = 1.0$ は複数の出現は考慮するが重みづけを行わず出現回数のみを考慮した場合であり，これも他と比べて正解率が低い．これより，提案手法のように重み付けをし複数の出現を考慮した方が性能が高いことがわかった．

# 6 先行研究との比較実験

## 6.1 先行研究の手法

川中らの先行研究では，ソーシャルブックマークサービス (SBM) を解析することで研究を行っている．SBM とは Web 上のブックマーク管理，共有サービスのことである．先行研究では，Web 上の様々なドキュメントについてユーザが付与したタグを用いている．

今回本研究では，論文のデータを用いて，提案手法と先行研究の手法の比較実験を行う．

先行研究の手法を用いた実験の手法を示す．

### 6.1.1 先行研究の手法を用いた人名の変遷情報の推定方法

**手順 1** 提案手法と同様．

**手順 2** 人名 A を含む著者名データを取り出し，その中より出現年の早いものから m 件の著者名データを取り出す．

**手順 3** その m 件の著者名データから共起している人名すべてを人名 B の候補として取り出し，相互情報量に基づく方法で共起度を測り，順に並べる．

**手順 4** 提案手法と同様．

先行研究の手法を用いた分野名の変遷情報の推定方法

については，[手順 1] は提案手法の方法と同様であり，[手順 2] からは先行研究の手法を用いた人名の変遷情報の推定方法と同様．

川中らの先行研究は，上記の通り相互情報量に基づく方法で共起性の高いものを取り，かつ初出現時期が先のものをルーツとする手法を使用していることが本研究と異なっている．川中らの手法では，共起度の指標として AEMI(Augmented Expected Mutual Information) を用いている．AEMI は確率を考慮した精細な共起度を測るための指標であり，次のように示される．

$$AEMI(a,b) = MI(a,b) + MI(\overline{a},\overline{b}) - MI(a,\overline{b}) - MI(\overline{a},b) \quad (2)$$

$$MI(a,b) = P(a,b)log\frac{P(a,b)}{P(a)P(b)} \quad (3)$$

この場合，$P(a)$ は人名 A が出現する確率であり，$P(a,b)$ は人名 A と人名 B が共起する確率である．更に，$P(\overline{a})$ は人名 A が出現しない確率を表す．MI は共起率を評価するための一つの指標であり，AEMI は MI を組み合わせることで，スケールを考慮した確率的な共起度の高さを測ることができる．この式に従い，共起度を求めて一番大きいものをその人名 (または分野名) のルーツとする．

### 6.2 先行研究の手法を用いた実験

先行研究の手法を用いた実験を行った．結果を表 5，表 6 に示す．

なお，6.1.1 節の手順 2 で用いる「最初の著者名 m 件」の $m$ の値を 1 から 10 に変化させ，違いを見る．

表 5: 先行研究の人名の実験結果

| $m$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 正解率 | 0.25 | 0.20 | 0.11 | 0.05 | 0.04 |
| $m$ | 6 | 7 | 8 | 9 | 10 |
| 正解率 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |

表 6: 先行研究の分野名の実験結果

| $m$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 正解率 | 0.39 | 0.30 | 0.20 | 0.18 | 0.18 |
| $m$ | 6 | 7 | 8 | 9 | 10 |
| 正解率 | 0.16 | 0.18 | 0.14 | 0.14 | 0.14 |

この実験ではどちらも $m = 1$ が一番正解率が高かった．結果としては，提案手法は人名で 0.4〜0.6，分野名で 0.4〜0.48 の正解率であり，先行研究の手法 (最大で人名で 0.25，分野名で 0.39 の正解率) よりも提案手法の方が正解率が高かった．

先行研究と結果が大きく差がついている原因は，本来頻度が高く正解であっても他の人名 (または分野名) ともよく共起しているものは AEMI 値が下がってしまうためである．

例をこの節の最後に示す．この例の場合，「柏岡秀紀」が正解であり，提案手法の実験結果は正解を出しているが，先行研究の手法を使った実験結果は「柏岡秀紀」が下に表示され不正解となっている．なお，それぞれの方法で一番精度の良い結果をのせる (提案手法は $a = 0.6$ のもの，先行研究の方では $m = 1$ のもの)．

例：「丸山岳彦」の提案手法での実験結果

```
人名：丸山岳彦（2001）
a = 0.6
2.752 --- 柏岡秀紀（1995）
1.36 --- 熊野正（1996）
0.936 --- 田中英輝（1996）
0.576 --- 内元清貴（1997）
```

例：「丸山岳彦」の先行研究の手法での実験結果

```
人名：丸山岳彦（2001）
m = 1
7.96218332399891 --- 熊野正（1996）
7.89217454680154 --- 柏岡秀紀（1995）
```

## 7 おわりに

本稿では研究情報の関係概念を自動的に抽出する新しい方法を提案した．その結果，ルーツとなる人名または分野名を得ることができた．また，提案手法では重み付けを行ったが，重み付けをした提案手法の方 (人名で 0.58〜0.60，分野名で 0.48 の正解率) が重み付けをしない方法 (人名で 0.41〜0.45，分野名で 0.4〜0.45 の正解率) より性能が高いことも確認した．さらに，先行研究との比較実験を行い，提案手法が先行研究の手法 (最大で人名で 0.25，分野名で 0.39 の正解率) よりも性能が高いことを確認した．

## 参考文献


[1] 川中翔，佐藤周行: “ソーシャルブックマークにおけるタグの派性関係の解析”，第 1 回データ工学と情報マネジメントに関するフォーラム，pp.1-8，2009.

[2] 松尾豊，友部博教，橋田浩一，中島秀之，石塚満: “Web 上の情報からの人間関係ネットワークの抽出”，人工知能学会，pp.46-56，2005.

[3] Adar，E. Adamic，L. A. : “Traking Information Epidemics in Blogspace”，In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Interlligence，pp.207-214，2005.

[4] 丹羽智史，土肥拓生，本位田真一: “Folksonomy の 3 部グラフ構造を利用したタグクラスタリング”，合同エージェントワークショップ&シンポジウム 2006(JAWS2006)，2006.

[5] 専門用語自動抽出サービス「言選 web」，http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html