

概要

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。この変遷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、この方法で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 (例えば、人名では「池原悟 (先輩) →村上仁一 (後輩)」のような先輩後輩関係の対、分野名では「統計 (ルーツ) →統計的機械翻訳 (派生概念)」のような派性関係の対) を自動的に抽出する方法を提案した。その結果、変遷の対を抽出することができた。

また、今回の手法では重み付けを行うことでルーツと考えられる対を抽出した。本研究では重み付けを行わない場合 ($a=0$, または $a=1$ の場合) との比較を行った。その結果、重み付けを行う方法では人名では $0.64\sim 0.66$, 分野名では 0.65 の正解率であり、重み付けを行わない方法では、人名では $0.45\sim 0.50$, 分野名では 0.60 の正解率であった。これにより、重み付けをする方法の方が正解率が高く、重み付けをする方法の有効性を確認した。

川中ら [1][2] もソーシャルブックマークを用いて概念の変遷の抽出を行っている。本研究では川中らの方法と比較実験を行った。その結果、人名では 0.25 , 分野名では 0.61 の正解率であった。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した。

目次

| | | |
|-------|--------------------------|----|
| 第1章 | はじめに | 1 |
| 第2章 | 関連研究 | 2 |
| 第3章 | 提案手法 | 3 |
| 3.1 | 人名の変遷情報の推定方法 | 3 |
| 3.2 | 分野名の変遷情報の推定方法 | 7 |
| 3.3 | 重み付け手法 | 8 |
| 3.4 | 出力例 | 11 |
| 3.4.1 | 人名の出力例 | 11 |
| 3.4.2 | 分野名の出力例 | 12 |
| 第4章 | 抽出実験 | 13 |
| 第5章 | 評価実験 | 16 |
| 5.1 | 評価基準 | 16 |
| 5.2 | 正解率算出方法 | 16 |
| 5.3 | 提案手法の評価実験結果 | 18 |
| 第6章 | 先行研究との比較実験 | 20 |
| 6.1 | 先行研究の手法 | 20 |
| 6.2 | 先行研究の手法を用いた人名の変遷情報の推定方法 | 20 |
| 6.3 | 先行研究の手法を用いた分野名の変遷情報の推定方法 | 20 |
| 6.4 | 提案手法と先行研究の手法との違い | 21 |
| 6.5 | 先行研究の手法を用いた実験結果 | 22 |
| 6.6 | 先行研究の手法を用いた実験の評価実験 | 24 |

| | |
|--|-----------|
| 第7章 考察 | 25 |
| 7.1 誤り分析 | 25 |
| 7.1.1 実験で使用したデータ10件の中にそもそも正解が入っていない場合 | 25 |
| 7.1.2 候補が多数ある場合 | 27 |
| 7.1.3 正解が候補にない場合 | 28 |
| 7.1.4 正解が2番目以降に出力されてしまう場合 | 29 |
| 7.2 性能を考慮した利用方法 | 30 |
| 7.3 性能を改善するための方策 | 30 |
| 7.4 a の値(0.1~0.9)によって正解率が異なる原因 | 31 |
| 7.5 先行研究と大きく差がついている原因 | 32 |
| 7.6 先行研究の手法 y について | 33 |
| 第8章 おわりに | 34 |

表 目 次

| | | |
|-----|---------------------------|----|
| 4.1 | 実験で抽出された人名の変遷例 | 13 |
| 4.2 | 実験で抽出された分野名の変遷例 | 14 |
| 5.1 | 提案手法の人名の実験結果 | 18 |
| 5.2 | 提案手法の分野名の実験結果 | 18 |
| 6.1 | 実験で抽出された人名の変遷例 | 22 |
| 6.2 | 実験で抽出された分野名の変遷例 | 23 |
| 6.3 | 先行研究の人名の実験結果 | 24 |
| 6.4 | 先行研究の分野名の実験結果 | 24 |

目次

| | | |
|------|---------------------------------------|----|
| 3.1 | 著者名データの例 | 3 |
| 3.2 | 「松吉俊」を人名 A として抽出した例 | 4 |
| 3.3 | 「松吉俊」を人名 A として抽出した B_i の例 | 5 |
| 3.4 | 人名の変遷情報の推定方法 | 6 |
| 3.5 | 論文データのタイトルの例 | 7 |
| 3.6 | タイトルより名詞連続を抽出した例 | 8 |
| 3.7 | 不要な語の例 | 8 |
| 3.8 | 重み付けの例 | 9 |
| 3.9 | 「松吉俊」を人名 A とした場合の出力例 | 11 |
| 3.10 | 「重要文抽出」を分野名 A とした場合の出力例 | 12 |
| 5.1 | 正解率算出方法の例 | 17 |
| 5.2 | $a = 0, a = 1$ と提案手法との比較 | 19 |
| 7.1 | 「関根聡」の著者名データの例 | 25 |
| 7.2 | 「河原達也」の出力例 | 26 |
| 7.3 | 「徳久雅人」の出力例 | 27 |
| 7.4 | 「山本英子」の出力例 | 28 |
| 7.5 | 「FrancisBond」の出力例 | 29 |
| 7.6 | 「今村賢治」の出力例 | 31 |
| 7.7 | 「丸山岳彦」の提案手法での実験結果 | 32 |
| 7.8 | 「丸山岳彦」の先行研究の手法での実験結果 | 33 |

第1章 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。この変遷とは本研究では、人名の場合「誰が誰を教示している」というような先輩後輩関係、また分野名の場合は「どの分野からどの分野が発生した」というような派性関係を示す。

この変遷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、この方法で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(例えば、人名では「池原悟(先輩)→村上仁一(後輩)」のような先輩後輩関係の対、分野名では「統計(ルーツ)→統計的機械翻訳(派生概念)」のような派性関係の対)を自動的に抽出する方法を提案する。

川中ら [1][2] 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相互情報量に基づく方法により解析し、概念の派性関係の対(例えば「SNS(ルーツ)→mixi(派生概念)」のような対)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。

第2章では、本研究の関連研究を述べ、第3章で提案手法の説明を行う。第4章では、提案手法を用いた抽出実験を行った結果を示し、第5章では提案手法の評価実験を行い、結果を示す。また、第6章では先行研究の比較実験について述べる。先行研究の手法と評価実験の結果はこの第6章で示す。第7章では、本研究全体に対しての考察を述べる。

第2章 関連研究

関連研究としては以下のものがある。

川中ら [1][2] 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相互情報量に基づく方法により解析し、概念の派性関係の対 (例えば「SNS(ルーツ) → mixi(派生概念)」のような対) を自動的に抽出した。

松尾ら [3] 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。

Referral Web[Kautz97][4] は、ある人物から対象人物への繋がりを Web 上の情報から順次発見していくものである。ある人物の名前と共起する名前を抽出し、更にその名前から次の名前を抽出するという方法を用いている。

原田ら [5] はある単語で検索した Web ページ集合から固有表現抽出により人物名を抽出し、独自に定義した共起度を使用し、共起関係から人物の関係を表すネットワークを抽出する方法を提案している。

村田ら [6] は検索エンジンで検索された件数を使用して Web ページ間の関係を発見する手法を提案している。

Adar ら [7] 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、時間の情報から解析するモデルを提案した。

丹羽ら [8] 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し、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。

第3章 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.

図 3.1 は著者名データの例である. 「出現年, 著者名, 著者名, 著者名, ...」という記述となっている.

——— 著者名データ ———

1995, 荒木哲郎, 池原悟, 横川秀人
1995, 荒木哲郎, 池原悟, 四方啓智
1995, 白井諭, 横尾昭男, 木村淳子, 小見佳恵
...
2001, 関根聡
2001, 下畑光夫, 隅田英一郎
2001, 村上裕, 神光太郎, 兵藤安昭, 池田尚志
...

図 3.1: 著者名データの例

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 (下の例を参照).

図 3.2 は「松吉俊」を人名 A として抽出した例である。「松吉俊」が出現する著者名データの中から、出現年の早いものから 10 件を取り出している.

— 著者名データ —

2005, 松吉俊, 佐藤理史, 宇津呂武仁
2005, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一
2005, 松吉俊, 佐藤理史, 宇津呂武仁
2006, 松吉俊, 佐藤理史, 宇津呂武仁
2006, 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一
2006, 高木俊宏, 注連隆夫, 土屋雅稔, 内元清貴, 松吉俊, 宇津呂武仁, 佐藤理史
2007, 土屋雅稔, 注連隆夫, 松吉俊, 宇津呂武仁, 佐藤理史, 中川聖一
2007, 注連隆夫, 土屋雅稔, 松吉俊, 宇津呂武仁, 佐藤理史
2007, 松吉俊, 佐藤理史
2007, 杉山賛太, 松吉俊, 佐藤理史

図 3.2: 「松吉俊」を人名 A として抽出した例

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．

図3.3は「松吉俊」を人名Aとして抽出した B_i の例である．「松吉俊」と共起した人名を抽出している．

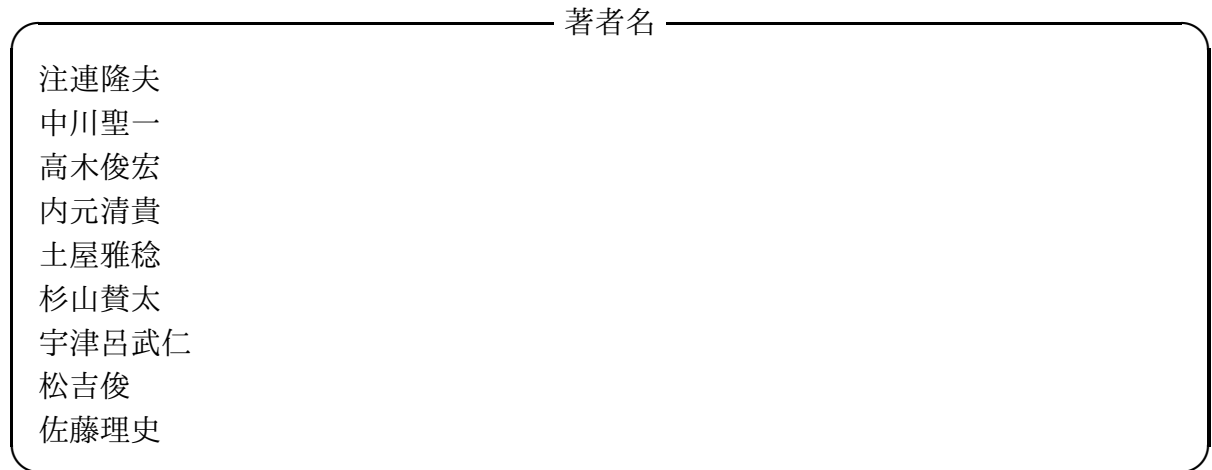


図 3.3: 「松吉俊」を人名Aとして抽出した B_i の例

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3.4に図式化したものを載せる。

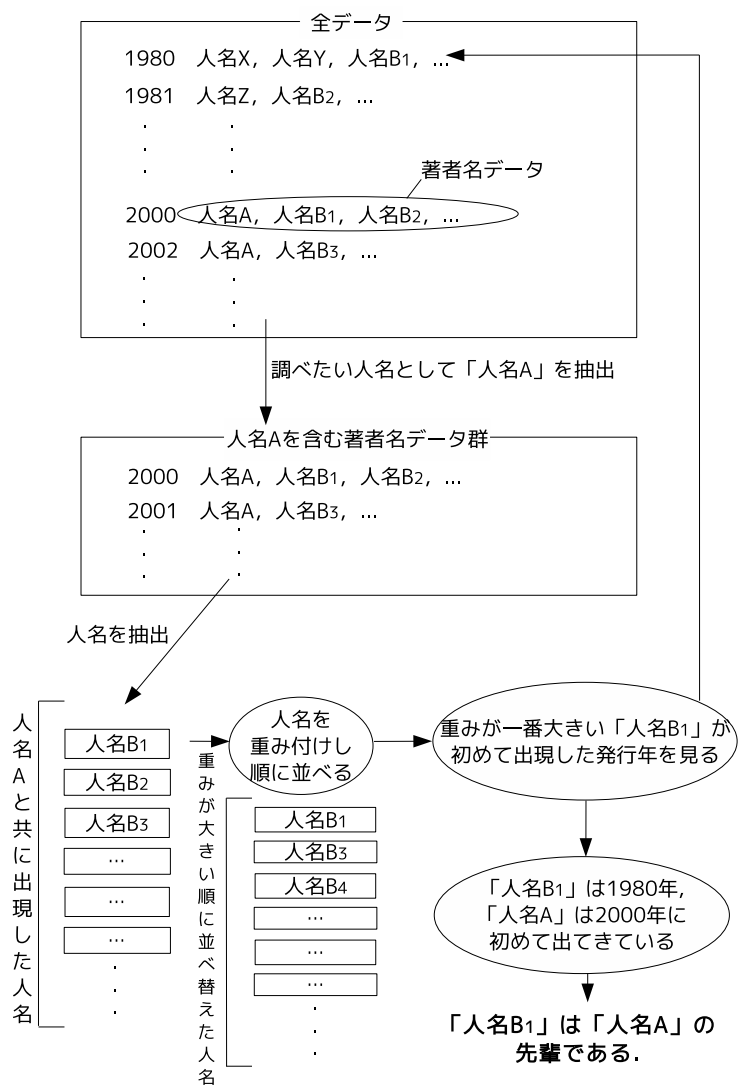


図 3.4: 人名の変遷情報の推定方法

3.2 分野名の変遷情報の推定方法

手順1 「言選」[9]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．その中から指定した名詞連続を抽出し分野名Aとする．

以下，[手順2]から人名の変遷情報の推定方法と同様．

例を用いて説明する．図3.5は論文データのタイトルの例である．

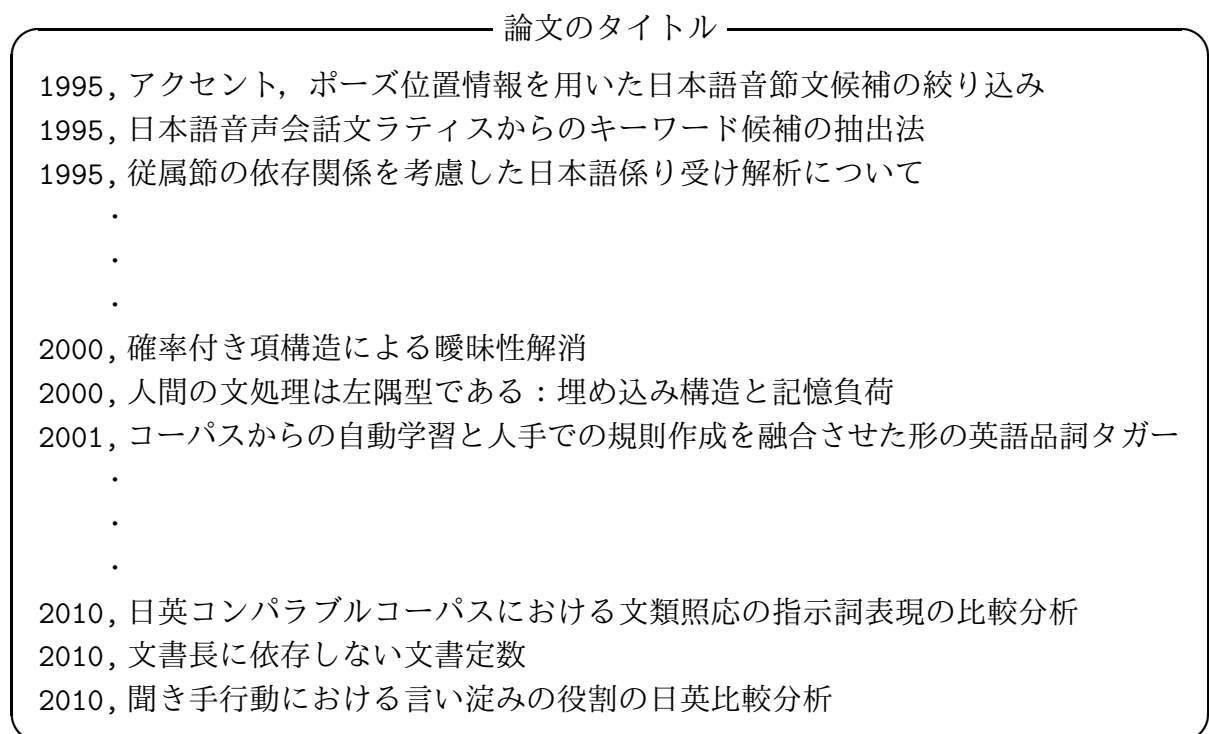


図 3.5: 論文データのタイトルの例

図3.5のようなデータより，「言選」を使用して名詞連続を抽出し，図3.6のようなデータを作成する．

図3.6のようなデータから，不要な語を取り除く．図3.7は削除した不要な語の例である．

—— タイトルより名詞連続を抽出したデータ ——

1995, アクセント, ポーズ位置, 情報, 日本語音節文候補, 絞り込み, ...
1995, 日本語音声会話文, ラティス, キーワード, 候補, 抽出, ...
1995, 従属節, 依存関係,, 日本語係り受け解析, 解析, ...
.
.
.

図 3.6: タイトルより名詞連続を抽出した例

—— 不要語 ——

語 文 法 解 抽出 解析 情報 日本語 表 言語 システム 手法
検索 表現 利用 モデル 文書 意味 評価 処理 分類 翻訳 構造
テキスト 学習 単語
.
.
.

図 3.7: 不要な語の例

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^{j_k} ($j_k = k$ 番目の論文の出現年-初出現年, ただし $j_k = 0$ の場合は重み 1) 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.

例を図 3.8 に示す.

重み付け手法

★人名Aを調べたい人名とし，人名Aの初出現年
が1990年であった場合

| 発行年 | 著者名データ | 重み |
|-------|--------------------|-------|
| 1990年 | 人名A, 人名B, 人名C | 1 |
| 1991年 | 人名A, 人名B, 人名D | a |
| 1992年 | 人名A, 人名C, 人名E, 人名F | a × a |

★ここで，a = 0.5とした場合

- ・ 人名B, Cに 重み1
- ・ 人名B, Dに 重み0.5
- ・ 人名C, E, Fに 重み $0.5 \times 0.5 = 0.25$

↓ 重みを人名ごとに加算

人名Bの重み = $1 + 0.5 = 1.5$
 人名Cの重み = $1 + 0.25 = 1.25$
 人名Dの重み = 0.5
 人名Eの重み = 0.25
 人名Fの重み = 0.25

一番重みの大きい人名Bを
人名Aのルーツ(先輩)と考える

図 3.8: 重み付けの例

図 3.8 のようなデータがあるとする．このように人名 A を調べたい人名とし，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，
- 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．この場合では 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．

例えば，重み a を 0.5 とした場合を考える．この場合，1990 年に人名 A と共起した人名 B, C に重み 1，1991 年に人名 A と共起した人名 B, D に重み 0.5，1992 年に人名 A

と共起した人名 C, E, F に重み 0.5×0.5 , すなわち重み 0.25 を与える. ここで, 重みを人名ごとに加算する. 図 3.2 より, 人名 B の 1.5 が一番重み大きい. よって, 人名 B を人名 A のルーツと考えることができる.

$$score(x) = \sum_{k=1}^{10} f_k \cdot g_k(x) \quad (3.1)$$

$$f_k = \begin{cases} 1 & (j_k = 0) \\ a^{j_k} & (j_k > 0) \end{cases} \quad (3.2)$$

ただし $g_k(x)$ は, k 番目の論文が, x という人名を持つ場合 1, そうでない場合 0 を取る. 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.

3.4 出力例

3.4.1 人名の出力例

図 3.9 は人名の出力例である。

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。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。

出力

```
人名：松吉俊 (2005)
a = 0
3 --- 宇津呂武仁 (1995)
3 --- 佐藤理史 (1995)
1 --- 中川聖一 (1995)
1 --- 土屋雅稔 (2003)
0 --- 内元清貴 (1997)
-----
a = 0.1
3.34 --- 佐藤理史 (1995)
3.32 --- 宇津呂武仁 (1995)
1.22 --- 土屋雅稔 (2003)
1.11 --- 中川聖一 (1995)
0.1 --- 内元清貴 (1997)
-----
...
```

図 3.9: 「松吉俊」を人名 A とした場合の出力例

3.4.2 分野名の出力例

図 3.10 は分野名の出力例である。

出力例の、最初の「分野名：」の部分に表示された分野名が分野名 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 --- 分野名」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分野名を分野名 B(先輩)とする。この分野名の例で説明すると、「重要文抽出」は分野名 A で、分野名 B にあたる分野名は「抽出システム」である。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。

出力

```
分野名：文節解析 (1999)
a = 0
1 --- 解析システム (1996)
0 --- コーパス (1995)
0 --- 規模コーパス (1996)
-----
a = 0.1
1.00101 --- 解析システム (1996)
0.001 --- コーパス (1995)
0.001 --- 規模コーパス (1996)
-----
a = 0.2
1.00832 --- 解析システム (1996)
0.008 --- コーパス (1995)
0.008 --- 規模コーパス (1996)
-----
...
```

図 3.10: 「重要文抽出」を分野名 A とした場合の出力例

第4章 抽出実験

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，人名と分野名の変遷の対を抽出した。

抽出した人名と分野名の例を 20 件ずつ表 4.1，表 4.2 に示す．この例は， $a = 0.5$ のものである。

表 4.1: 実験で抽出された人名の変遷例

| 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 |
| 河原達也 | 駒谷和範 |
| 石川徹也 | 藤井敦 |
| 難波英嗣 | 奥村学 |
| 松吉俊 | 佐藤理史 |
| 竹内孔一 | 影浦峽 |
| 橋本力 | 奥村学 |
| 松本忠博 | 伊佐治和哉 |
| 土屋雅稔 | 佐藤理史 |
| 高村大也 | 奥村学 |
| 乾孝司 | 乾健太郎 |
| 酒井浩之 | 増山繁 |
| 風間淳一 | 辻井潤一 |
| 浅原正幸 | 松本裕治 |
| 松尾義博 | 白井諭 |

実験により抽出した人名 277 件中の 20 件を表 4.1 に示している．全データより抽出された人名すべてで実験を行ったわけではなく，ルーツとなる人名が出現するデータを確

保するため、人名 A として、以下の二つの条件を両方とも満足したものをを用いて、ルーツである人名を特定した。

1. 現時点で使用できるデータ中で一番古い年のものの3年(一番古いデータが1995年であれば、1998年)以上後に出現した人名であること。
2. 全データ中に3回以上出現のあったもの(それ以下の出現回数ではデータ不足とみなし、使用していない)であること。

この2つの条件を満たした人名は555件であり、この半数の277件を人名 A として利用し、ルーツを特定する実験を行った。

表の第1列目に示す人名 A(後輩)がルーツを特定したい人名であり、第2列に示す人名 B(先輩)が提案手法を用いて特定したルーツであり、人名 A の先輩と予測される人名である。

表 4.2: 実験で抽出された分野名の変遷例

| 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 |
| 質問応答システム | 数量表現 |
| 自動構築 | コーパス |
| マイニング | コーパス |
| 表現抽出 | 抽出システム |
| 文抽出 | 抽出システム |
| 識別 | 自動獲得 |
| 重要文抽出 | 抽出システム |
| 固有表現抽出 | 表現抽出 |
| 文書要約 | G D A |
| テキスト分類 | S V M |

実験により抽出した人名 94 件中の 20 件を表 4.2 に示している。全データより抽出された分野名すべてで実験を行ったわけではなく、人名の実験の場合と同様の条件を持つ分野名は 187 件であり、この半数の 94 件のみを使用して、分野名 A として、ルーツを特定する実験を行った。

表の第 1 列目に示す分野名 A がルーツを特定したい分野名であり、第 2 列に示す分野名 B(ルーツ) が提案手法を用いて特定したルーツであり、分野名 A のルーツと予測される分野名である。

第5章 評価実験

人名 A として 40 件，分野名 A として 31 件を使用し，ルーツ特定の評価実験を行う。出力のルーツが正解であるかの評価は，言語処理学会に精通している人物が行う。

評価をする際には以下のデータを利用した。

- 提案手法で結果として出力される人名 B(または分野名 B)
- 6.1 章で述べる先行研究の手法で結果として出力される人名 B(または分野名 B)
- それぞれの手法で出力される人名 B(または分野名 B) の候補すべて (ただし、人名 A (または分野名 A) より前に出現していない人名 B (または分野名 B) は除く。)

このデータをすべて混ぜ合わせてランダムな順番に表示し，その中より評価者が正解となる人名または分野名を選ぶ。その正解の情報により，評価を行う。また変遷情報の推定のために手法が使用した 10 件の論文データ (論文の著者名とタイトルの情報を含むもの) も参考のため，評価者に見せ，評価者はこの情報も含めて正解の判定を行う。

5.1 評価基準

評価の基準を以下のように定める。

| | |
|-----|-------------------------------------|
| 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。 |

この基準を満たした人名 (または分野名) を正解とする。

5.2 正解率算出方法

5.1 節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．

図 5.1 の例を用いて説明する．

```
                               出力
人名：酒井浩之（2000）
a = 0
1 --- 大竹清敬（1997）
1 --- 増山繁（1995）
0 --- 梅村祥之（1998）
-----
a = 0.1
1.112212 --- 増山繁（1995）
1.1 --- 大竹清敬（1997）
0.00011 --- 梅村祥之（1998）
-----
a = 0.2
1.259648 --- 増山繁（1995）
1.2 --- 大竹清敬（1997）
0.00192 --- 梅村祥之（1998）
-----
.
```

図 5.1: 正解率算出方法の例

この例では、「増山繁」が正解である． $a = 0.1$ ， 0.2 の場合は一番目に正解を表示しているため，得点 1 をつける．しかし， $a = 0$ の場合，正解と同じ得点をもつものが他に 1 個存在しているため，得点は $1/2$ 点となる．パラメータ (a の値) ごとに得点を算出し，全体の件数で割る．

また，人名 A (分野 A) のルーツである人名 (分野名) のどれもが人名 A (分野 A) と同時に出現している論文が 2 個以上ない場合，その人名 A (分野 A) は，ルーツを特定するにはデータが不足していると考え，実験には用いていない．先に記述した，評価実験に用いる 40 件の人名 A と 32 件の分野名 A にはこのようなデータは含めていない．

より詳細に説明すれば，4章で述べた実験に用いるデータ (人名 277 件，分野 94 件) のうち，ルーツが人名 A (分野 A) と同時に出現する論文が 2 個以上あったものが，人名で

40 件で、分野で 31 件であった。この 40 件の人名と 31 件の分野を評価実験に用いた。

5.3 提案手法の評価実験結果

評価実験結果を以下の表 5.1, 表 5.2 に示す。 a を 0 から 1 の 0.1 刻みとした結果を示している。正解率は 5.2 節の方法で算出する。

表 5.1: 提案手法の人名の実験結果

| a | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|-----|------|------|------|------|------|------|------|------|------|------|------|
| 正解率 | 0.50 | 0.65 | 0.65 | 0.65 | 0.64 | 0.64 | 0.66 | 0.66 | 0.61 | 0.56 | 0.45 |

表 5.2: 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 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|-----|------|------|------|------|------|------|------|------|------|------|------|
| 正解率 | 0.60 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.60 |

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、 $a = 0$ の場合と $a = 1$ の場合を説明する。この 2 つの場合は他の場合と異なっている。 $a = 0$ は最初に共起したもののみを考慮した場合であり、重み付けを行っていない。 $a = 1.0$ 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、これも重みを考慮していない。

例を用いて説明する。図 5.2 に例を示す。

a=0,a=1と提案手法との比較

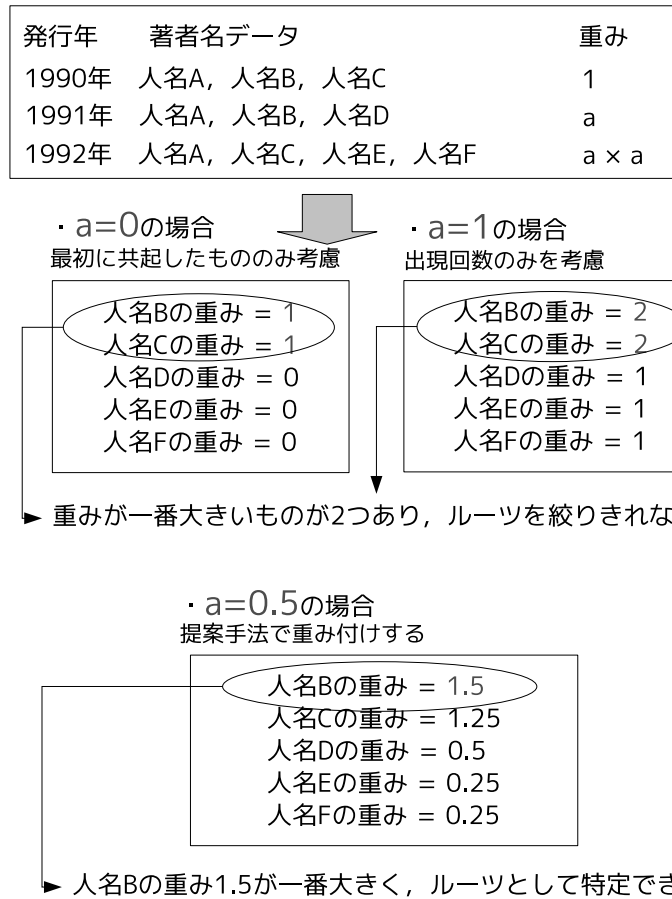


図 5.2: a = 0, a = 1 と提案手法との比較

$a = 0$, または $a = 1.0$ の場合では、重みが一番大きいもの2つありルーツを絞りきれない。しかし、提案手法により重み付けをした場合、ルーツとして特定することが可能である。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。

第6章 先行研究との比較実験

6.1 先行研究の手法

川中ら [1] の先行研究では，ソーシャルブックマークサービス (SBM) におけるタグ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．

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．

6.2 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 の初出現年の年から y 年分のすべての著者名データから，人名 A が出現している著者名データ (多くとも 10 件まで) を取り出し，そこから人名 B の候補を抽出する．

手順 3 人名 A の初出現年の年から y 年分のすべての著者名データを用い，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．

6.3 先行研究の手法を用いた分野名の変遷情報の推定方法

手順 1 提案手法の分野の変遷情報の推定と同様．

手順 2 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．

手順 3 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．

手順4 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.

なお, 6.2節の手順2で用いる「人名Aの初出現年の年から y 年分」の y の値を1から10に変化させ, 違いを見る.

6.4 提案手法と先行研究の手法との違い

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.

$$AEMI(a, b) = MI(a, b) + MI(\bar{a}, \bar{b}) - MI(a, \bar{b}) - MI(\bar{a}, b) \quad (6.1)$$

$$MI(a, b) = P(a, b) \log \frac{P(a, b)}{P(a)P(b)} \quad (6.2)$$

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, $P(\bar{a})$ 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.

6.5 先行研究の手法を用いた実験結果

先行研究の手法を用いた実験を行った。抽出例を表 6.1, 表 6.2 に示す。この例は, $y = 5$ のものである。

表 6.1: 実験で抽出された人名の変遷例

| 人名 A(後輩) | 人名 B(先輩) |
|----------|----------|
| 村上仁一 | 池原悟 |
| 徳久雅人 | 村上仁一 |
| 馬青 | 内元清貴 |
| 宮尾祐介 | 鳥澤健太郎 |
| 関根聡 | 山田篤 |
| 丸山岳彦 | 熊野正 |
| 黒田航 | 仲本康一郎 |
| 河原達也 | 駒谷和範 |
| 石川徹也 | 藤井敦 |
| 松吉俊 | 土屋雅稔 |
| 竹内孔一 | 影浦峽 |
| 橋本力 | 奥村学 |
| 松本忠博 | 伊佐治和哉 |
| 土屋雅稔 | 中川聖一 |
| 高村大也 | 高橋和子 |
| 乾孝司 | 福島俊一 |
| 酒井浩之 | 増山繁 |
| 風間淳一 | 牧野貴樹 |
| 浅原正幸 | 松本裕治 |
| 山本英子 | 神崎享子 |

実験により, 抽出した人名 277 件中の 20 件である。なお, 全データより抽出された人名すべてで実験を行ったわけではなく, 提案手法の人名の実験の場合と同様の条件を持つ人名 555 件の半数の 277 件のみを使用して, 人名 A としてルーツを特定する実験を行った。

左に示す人名 A(後輩) がルーツを特定したい人名であり, 右に示す人名 B(先輩) が提案手法を用いて特定したルーツであり, 先行研究の手法において人名 A の先輩と予測される人名である。

表 6.2: 実験で抽出された分野名の変遷例

| 分野名 A | 分野名 B(ルーツ) |
|----------|------------|
| 自動評価 | 文字単位 |
| 統計的機械翻訳 | 訳語推定手法 |
| サンプリング | コーパス |
| タグ付きコーパス | 自動作成 |
| 音声対話システム | 対話システム |
| 語義曖昧性解消 | 曖昧性解消 |
| 翻訳自動評価 | 文字単位 |
| 情報分析 | テキスト分類 |
| 言語横断情報検索 | L S I |
| 質問応答システム | 数量表現 |
| 自動構築 | 省略解析 |
| 表現抽出 | 抽出システム |
| 文抽出 | 抽出システム |
| 重要文抽出 | 抽出システム |
| 固有表現抽出 | 表現抽出 |
| 文書要約 | G D A |
| 文節解析 | 規模コーパス |
| 単語分割 | HMM |
| テキスト分類 | WWW情報探索支援 |
| 辞書構築 | 格解析 |

実験により、抽出した分野名 94 件中の 20 件である。なお、全データより抽出された分野名すべてで実験を行ったわけではなく、提案手法の人名の実験の場合と同様の条件を持つ分野名 187 件の半数の 94 件のみを使用して、分野名 A としてルーツを特定する実験を行った。

左に示す分野名 A がルーツを特定したい分野名であり、右に示す分野名 B(ルーツ) が提案手法を用いて特定したルーツであり、先行研究の手法において分野名 A のルーツと予測される分野名である。

6.6 先行研究の手法を用いた実験の評価実験

評価方法は5節，評価基準は5.1節，正解率算出方法は5.2節に示しているものを利用する。

結果を表6.3，表6.4に示す。

表 6.3: 先行研究の人名の実験結果

| y | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----|------|------|------|------|------|------|------|------|------|------|
| 正解率 | 0.25 | 0.25 | 0.20 | 0.18 | 0.15 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 |

表 6.4: 先行研究の分野名の実験結果

| y | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----|------|------|------|------|------|------|------|------|------|------|
| 正解率 | 0.61 | 0.48 | 0.35 | 0.29 | 0.26 | 0.23 | 0.26 | 0.26 | 0.26 | 0.26 |

この実験ではどちらも $y = 1$ 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0.64~0.66，分野名は0.65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0.61の正解率)よりも提案手法の方が正解率が高かった。

第7章 考察

7.1 誤り分析

本研究の正解率は、良い結果のパラメータ(a)で人名では0.64~0.66, 分野名では0.65である。重み付けをしていない場合や先行研究の手法より正解率が高いと言えども、一般的には正解率が高いとは言い難い。ここで誤りの例を示し、そこから改善の策を講じ、より高い正解率を目指していく。

7.1.1 実験で使用したデータ 10 件の中にそもそも正解が入っていない場合

実験で使用した 10 件のデータの中にそもそも正解が入っていない場合がある。図 7.1 に例を示す。

——— 著者名データ ———

人名：関根聡
1999, 関根聡, 江里口善生
2000, 内元清貴, 関根聡, 井佐原均
2000, 関根聡
2001, 関根聡
2001, 野畑周, 関根聡, 村田真樹, 内山将夫, 内元清貴, 井佐原均
2002, 内元清貴, 関根聡, 井佐原均
2002, 野畑周, 関根聡, 井佐原均, R a l p h G r i s h m a n
2002, 野畑周, 関根聡
2003, 内元清貴, 野畑周, 山田篤, 関根聡, 井佐原均
2003, 関根聡

図 7.1: 「関根聡」の著者名データの例

この「関根聡」の正解のルーツとなる人名は、この10件の論文には現れていない。先輩であっても共著となっていない場合はこの場合に当てはまる。

また、本来は指導的立場は逆であるが、言語処理学会での発表という意味では先輩にあたる人は出力されているが、本来的な意味で先輩にあたる人は出力されていない例もある。図7.2に例を示す。

```
                                出力
人名：河原達也 (2000)
parameter:0
1 --- 駒谷和範 (1998)
0 --- 井佐原均 (1995)
0 --- 内元清貴 (1997)
0 --- 隅田英一郎 (1995)
-----
parameter:0.1
1.111101 --- 駒谷和範 (1998)
1e-05 --- 井佐原均 (1995)
1e-05 --- 内元清貴 (1997)
1e-05 --- 隅田英一郎 (1995)
-----
parameter:0.2
1.249664 --- 駒谷和範 (1998)
0.00032 --- 井佐原均 (1995)
0.00032 --- 内元清貴 (1997)
0.00032 --- 隅田英一郎 (1995)
-----
.
.
.
```

図 7.2: 「河原達也」の出力例

この例では、「河原達也」に対する出力のルーツは「駒谷和範」となっているが、言語処理学会での発表という意味では「駒谷和範」が先輩であるが、本来は「河原達也」の方が「駒谷和範」の先輩にあたる人物である。

分野名の場合も実験で使用した10件のデータの中にそもそも正解が入っていない場合がある。これは「言選」で抽出できていない単語があることが考えられる。また、不用語として間違っって省いているものがあることも考えられる。

7.1.2 候補が多数ある場合

実験で出力された人名(または分野名)の候補が多数ある場合がある。図7.3に例を示す。

```
                               出力
人名：徳久雅人 (2003)
parameter:0
3 --- 池原悟 (1995)
3 --- 村上仁一 (1998)
0 --- 藤本敬史 (2002)
0 --- 吉田真司 (2002)
-----
parameter:0.1
3.7 --- 池原悟 (1995)
3.7 --- 村上仁一 (1998)
0.1 --- 藤本敬史 (2002)
0.1 --- 吉田真司 (2002)
-----
parameter:0.2
4.4 --- 池原悟 (1995)
4.4 --- 村上仁一 (1998)
0.2 --- 藤本敬史 (2002)
0.2 --- 吉田真司 (2002)
-----
.
.
.
```

図 7.3: 「徳久雅人」の出力例

「徳久雅人」の正解は、「池原悟」である。この例では「池原悟」と「村上仁一」が $a=0.1$ や 0.2 のときですらまったく同一の重みが付けられている。これは、「池原悟」と「村上

仁一」が、実験で使用した著者名データの中では全て「徳久雅人」と共起しているからである。これは正解と同じ出力がある場合となるため、得点 1/2 となり正解率が下がってしまう。

7.1.3 正解が候補にない場合

正解が、実験で出力される候補にない場合がある。この場合の例を図 7.4 に示す。

```
                                出力
人名：山本英子 (1999)
parameter:0
1 --- 山本幹雄 (1996)
0 --- 村田真樹 (1995)
0 --- 井佐原均 (1995)
0 --- 神崎享子 (1998)
0 --- 馬青 (1998)
0 --- 乾裕子 (1997)
0 --- 木田敦子 (1998)
-----
parameter:0.1
1 --- 山本幹雄 (1996)
0.300023 --- 井佐原均 (1995)
0.200022 --- 神崎享子 (1998)
0.10002 --- 木田敦子 (1998)
0.100001 --- 馬青 (1998)
0.1 --- 村田真樹 (1995)
0.1 --- 乾裕子 (1997)
-----
.
```

図 7.4: 「山本英子」の出力例

「山本英子」の正解は、「梅村恭司」である。この例では、「梅村恭司」がまったく出ていない。この実験で用いた 10 件の著者名データでは「梅村恭司」が出現していたため、

正解がもたらなかったわけではない。これは、「梅村恭司」が先輩ではあるが、「山本英子」と同じ年に初出現したためであると考えられる。

7.1.4 正解が2番目以降に出力されてしまう場合

正解とされる人名または分野名が2番目以降に出力されてしまう場合がある。この場合の例を図7.5に示す。

```
                                     出力
人名 : FrancisBond(2005)
parameter:0
2 --- 藤田早苗 (2003)
2 --- 田中貴秋 (1998)
1 --- 中岩浩巳 (1996)
0 --- 井佐原均 (1995)
0 --- 橋本力 (2001)
0 --- 内元清貴 (1997)
0 --- 神崎享子 (1998)
-----
parameter:0.1
2.201 --- 藤田早苗 (2003)
2.2 --- 田中貴秋 (1998)
1 --- 中岩浩巳 (1996)
0.101 --- 橋本力 (2001)
0.002 --- 内元清貴 (1997)
0.001 --- 井佐原均 (1995)
0.001 --- 神崎享子 (1998)
-----
.
.
.
```

図 7.5: 「FrancisBond」の出力例

この例の「FrancisBond」の正解は「中岩浩巳」であるが、3番目に出力され不正解となっている。この場合は「中岩浩巳」が一番目の著者名データでのみ共起しているからであると考えられる。

7.2 性能を考慮した利用方法

ユーザの視点から見て、1番目に出たもののみを出力するよりも、候補を全て見せ、先輩である可能性が高いものから並べるといった方が有用性があると思われる。これは、ユーザがルーツの可能性の高いものから順に見ることができるため、参考になる情報が増え、情報を選ぶことができる。また、同じ理由で実験で使用した10件の著者名データも見せることも有効である。

上記のように、複数の候補のうちどれかが正解ならば、役立つという場合を考えれば、第5章で示した評価方法とは別のものが考えられる。例えば、現在は1番最初に出たもののみを正解としているが、上位いくつかを正解とする、またはMMRで評価をすれば正解率は上がると思われる。MRRとは、 n 番目に正解を出力した場合に得点 $1/n$ をつける方法である。例えば、1番目に正解を出力した場合に得点1を、2番目に正解を出力した場合に得点 $1/2$ を、3番目に正解を出力した場合に得点 $1/3$ を... というように得点を付与していく方法である。

7.3 性能を改善するための方策

分野名の実験の場合は「言選」だけでなく、形態素解析などを用いて候補を増やす、または不用語を見直すことが有効ではないかと考えられる。

7.4 a の値 (0.1~ 0.9) によって正解率が異なる原因

人名の実験では、 $a=0.1\sim 0.9$ の間で正解率が異なる場合がある。例を図 7.6 に示し、説明する。

```
                                     出力
人名：今村賢治 (2001)
.
.
.
parameter:0.8
1.8 --- 柏岡秀紀 (1995)
1.8 --- 鷹尾和享 (2000)
1.70496 --- 隅田英一郎 (1995)
1 --- 下畑光夫 (1998)
0.64 --- 松本裕治 (1995)
-----
parameter:0.9
2.64708 --- 隅田英一郎 (1995)
1.9 --- 柏岡秀紀 (1995)
1.9 --- 鷹尾和享 (2000)
1 --- 下畑光夫 (1998)
0.81 --- 松本裕治 (1995)
-----
.
.
.
```

図 7.6: 「今村賢治」の出力例

この例の「今村賢治」の正解は「柏岡秀紀」である。 $a=0.8$ の場合は正解を出力しているが、 $a=0.9$ の場合は「隅田英一郎」が一番最初に出力されており、不正解となっている。この原因は、出現年を見ると早くに共起しているのは「柏岡秀紀」であるが、出現回数を見ると「隅田英一郎」の方が多いためである。これにより、重み a が 1 に近づくほど、出現回数の方を重視してしまう傾向にあることが原因である。

同じことが 0 付近の重み a にも言える。重み a が 0 に近づくほど、出現回数よりも出

現年が早いものの方を重視してしまう傾向にあるということが言える。

この2つ事柄と実験の結果より、やはり出現年と出現回数のどちらも考慮した方が有効であることがわかる。

7.5 先行研究と大きく差がついている原因

先行研究と提案手法の人名の実験で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。

例を図7.7, 図7.8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」が下の順位に表示され不正解となっている。これは、「柏岡秀紀」が有力な研究者であり、他の人名と多く共起するため、「柏岡秀紀」に対する先行手法のAEMI値が下がってしまうことが原因である。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 $a = 0.5$ のもの、先行研究の方では $y = 1$ のもの)。

| 実験結果 | |
|--------|------------------|
| 人名 | : 丸山岳彦 (2001) |
| a | = 0.5 |
| 2.25 | --- 柏岡秀紀 (1995) |
| 1.25 | --- 熊野正 (1996) |
| 0.625 | --- 田中英輝 (1996) |
| 0.375 | --- 内元清貴 (1997) |
| 0.25 | --- 井佐原均 (1995) |
| 0.125 | --- 稲垣康善 (1995) |
| 0.125 | --- 松原茂樹 (1995) |
| 0.0625 | --- 柏野和佳子 (1999) |
| 0.0625 | --- 山口昌也 (1996) |

図 7.7: 「丸山岳彦」の提案手法での実験結果

人名：丸山岳彦 (2001)
y=1
4.94875915448601 --- 熊野正 (1996)
4.76940036975427 --- 柏岡秀紀 (1995)

図 7.8: 「丸山岳彦」の先行研究の手法での実験結果

7.6 先行研究の手法 y について

先行研究の手法を用いた実験では、 $y=1$ 、つまりルーツを特定したい人名の初出現年だけの著者名データを使用する方法が一番よい結果となった。これは、単に候補が少なくなるからであると考えられる。結果を見てもわかるように、 y を増やせば徐々に正解率が減っていく。特に人名の場合は 7.5 節より他の人名と共起しやすいと数値が下がるため、候補が増えれば増えるほど、正解とされるデータが 2 番目以降に出力されてしまう。やはり候補が増えれば増えるほど、正解とされるデータが 2 番目以降に出力されてしまいやすい。よって、候補数が少ない $y=1$ が一番良い結果となった。

第8章 おわりに

本論文では研究情報の派生概念 (例えば, 人名では「池原悟 (先輩) →村上仁一 (後輩)」のような先輩後輩関係の対, 分野名では「統計 (ルーツ) →統計的機械翻訳 (派生概念)」のような派性関係の対) を自動的に抽出する新しい方法を提案した. その方法は, 以下の特徴に基づいている.

- ルーツを特定したい人名 A(または分野 A) と, 人名 A の出現年になるべく近い時期に共起している人名 B(または分野 B) ほど人名 A のルーツとなりやすい.
- ルーツを特定したい人名 A(または分野 A) と, 共起している回数が多い人名 B(または分野 B) ほど人名 A のルーツとなりやすい.

具体的には, 早い時期に人名 A と共起している人名の重みを高くしつつ, 人名 A と共起している回数分その重みを加算してその合計の値が最も大きい人名を人名 A のルーツとする方法をとった. この提案手法は, 先に述べた 2 つの特徴をどちらも考慮できている方法となっている.

付与する重み a を 0~1 で 0.1 刻みに変化させ実験を行ったところ, 提案手法は最高で人名で 0.64~0.66, 分野名で 0.65 の正解率を出した.

また, 重み付けをしないものと同様のものとなる $a=0$, または $a=1$ の場合を考える. $a=0$ の場合は初出現時で共起したもののみを考慮する方法であり, また $a=1$ の場合は出現回数は考慮するが, 出現時期は考慮していない方法である. この 2 つの場合は, 人名で 0.45~0.50, 分野名で 0.60 の正解率であり, 重み付けをした提案手法の方が正解率が高いことがわかり, 提案手法の有効性を確認できた.

さらに, 相互情報量に基づく先行研究との比較実験を行った. 先行研究の手法での正解率は最大で人名で 0.25, 分野名で 0.61 の正解率であり, 相互情報量を用いる方法よりも提案手法の性能が高いことを確認した.

謝辞

本研究を進めるに当たり，終始に渡り研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科計算機工学講座Cの村田真樹教授に心から御礼申し上げます。また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授，徳久雅人講師に心から御礼申し上げます。加えて，種々の御助言を龍谷大学理工学部数理情報学科の馬青教授に頂きました。ここに深く感謝いたします。その他様々な場面で御助言を頂いた計算機工学講座C研究室の皆様に感謝の意を表します。

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] 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時系列的な依存関係の解析”, 情報処理学会研究報告, ICS-152, pp.25-32, 2008.
- [3] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満: “Web上の情報から人間関係ネットワークの抽出”, 人工知能学会論文誌, vol.20, No.1, pp.46-56, 2005.
- [4] Kanz, H., Selman, B., and Shah, M.: “The Hidden Web”, AI magazine, Vol.18, No.2, pp.27-35, 1997.
- [5] 原田晶紀, 佐藤進也, 風間一洋: “Web上のキーパーソンの発見と関係の可視化”, 情報処理学会研究報告, DBS-130/FI-71, pp.17-24, 2003.
- [6] 村田剛志: “参照の共起性に基づく Web コミュニティの発見”, 人工知能学会誌, Vol.16, No.3, pp.316-323, 2001.
- [7] Adar, E. Adamic, L. A. : “Traking Information Epidemics in Blogspace”, In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp.207-214, 2005.
- [8] 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”, 人工知能学会研究会, SIG-SWO-A602-07, pp.1-8, 2006.
- [9] 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>