自然言語処理　114－1
（1996. 7. 18）

補完類似度を用いた固有名詞のグルーピングの試み

松本兼一† 梅村恭司‡

† matumoto@avenue.tutics.tut.ac.jp
‡ umemura@tutics.tut.ac.jp
豊橋技術科学大学 情報工学系 梅村研究室
〒441 豊橋市天伯町雲雀ヶ丘1-1

あらまし　現在、様々な分野で様々な情報が散在している。自分が知らない分野について何か調べようとする時、何に注目して良いかわからない。そこで、注目すべき固有名詞群を探すために統計処理を行うことにした。本研究では、新聞記事の中から固有名詞の出現回数を計数し、その出現分布を元に固有名詞間の関係を見い出した。そこで関係を求めるために、文字認識などで用いられる補完類似度を用いた。その結果より、出現分布に偏りがある人々について、出現分布が似ている人々のグラフを求めた。このグラフの部分グラフとして妥当な固有名詞のグループが存在することを確認した。

キーワード　補完類似度、固有名詞、情報検索

A Trial of Grouping Proper Nouns using Complementary Similarity Measure

Kenichi Matsumoto† and Kyoji Umemura‡

† matumoto@avenue.tutics.tut.ac.jp
‡ umemura@tutics.tut.ac.jp
Umemura Laboratory, Department of Information and Computer Sciences,
Toyohashi University of Technology
1-1, Tempaku, Toyohashi, Aichi 441, Japan

Abstract

Currently, various information is dispersing in various fields. When we examine unfamiliar fields, we waver where to start with. We have performed statistical calculation in order to find appropriate nouns. As a result of this research, we obtain graph of people that resemble distribution. We evaluated that this subgraph have appropriate group of proper nouns.

key words　Complementary Similarity Measure,Proper Noun,Information Retrieval

1 はじめに

現在、様々な分野で様々な情報が存在している。中には良く知っている分野もあれば、あまり良く知らない分野もある。良く知っている分野について何か調べようとする時は、自分の知識の中に注目すべき固有名詞があるだろう。しかし、あまり良く知らない分野について何か調べようとする時には、どんな固有名詞に注目したら良いのかわからない。自分の知らない分野、例えば、医学について何か知りたい時には、注目する価値がある固有名詞、同じ症状からなる病名や病名とその治療法などの固有名詞のグループを得られると有用であると考えられる。予め注目すべき固有名詞群がある場合とない場合では、ある場合の方が便利であると考えらる。そこで、注目すべき固有名詞群と固有名詞間の関係を統計処理で得る方法を試みた。

2 注目すべき固有名詞を発見し固有名詞をグルーピングする方法

我々は、新聞の記事から固有名詞の出現数を計数し、その出現数を1カ月毎に集計し、出現分布のヒストグラムを生成した。その中から、後続の処理のため出現数の多いものと1カ月にある程度出現する固有名詞を選び出した。さらに、固有名詞の出現分布の類似度を比較し、その類似度の高いものを固有名詞間の関係が深いと定義し、固有名詞間の関係を求めた。得られた結果はグラフで示した。そして、得られた結果と実際の関係との検証を行なった。

方法を説明する前に本研究で用いた実験データを挙げる。情報源として用いたデータは、新聞の記事4年分である[1]。固有名詞の候補としては、有名人の人名を用いた。では、処理の流れを順を追って説明する。

2.1 人名の出現数を計数する方法

まず、最初に行なった新聞記事から人名の出現回数を計数する方法を説明する。人名の候補を先頭から1つずつ取り出し、その人名が新聞記事に出現したら、その記事が書かれた日の日付を取り出す。本研究では、日付毎で分けられているデータを1カ月毎に集計した。その結果、出現分布に特徴ができた。

2.2 人名を選出する方法

本研究の目的は、検索に効果がある固有名詞群を発見することであったので、計数された人名の中から、57カ月で100回以上出現する人名と、1カ月で15回以上出現する人名を選び出した。ここで人名の数が39個になった。本来なら全員を行なうべきであるが、固有名詞間の関係が求まっているかどうかを評価できないので足きりを行なった。

2.3 出現分布の類似度を比較する方法

先に示した手順で選出された人名について、出現分布を大きさについて正規化し類似度の比較を行なった。出現分布は2次元で時間軸と出現数の軸である。ここで正規化を行なわない場合には、出現分布は似ているが、一方は出現数が多く、他方は少ない場合など類似度が低くなってしまう。本研究で類似度を比較するために用いた関数については次節で詳しく述べる。選出された人名についてすべての可能性のある組合せについて比較を行なった。39個の人名について比較を行なったので741組の組合せができる。図1は出現分布の類似度が最も高いと考えられる2つのヒストグラムである。

2.4 得られた結果と実際の関係との検証方法

得られた結果の関係が実際の関係に現れているか検証する。ここでは、機械に自動的に検証を行なわせることができないので、人の手で検証した。

3 補完類似度の適用

この補完類似度とは、文字認識などで用いられる類似度である[2][3]。本研究では、この補完類似度を元に類似度を比較する関数を定義し用いた。では、本研究で用いた関数を説明する。この方法は、2値画像特徴と呼ぶ正規化2値画像(n= N×M画素)そのものを特徴ベクトルに用いて個別パターンを認識する。入力の2値画像特徴はn次元の2値特徴ベクトル $\vec{F} = (f_1, f_2, ..., f_i, ..., f_n)(f_i =$ 0または1)で表現し、比較対象のパターンも2値比較対象パターン $\vec{T} = (t_1, t_2, ..., t_i, ..., T_n)(t_i =$ 0または1)で表現する。Tを比較対象パターンの1の数とすると、$\vec{F}$ の $\vec{T}$ に対する用いた関数 S_c は、次式で定義される。

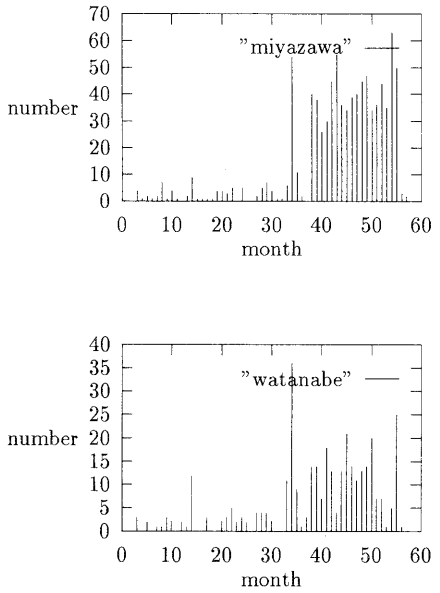

図 1: 出現分布の類似度が最も高かった 2 つのグラフ

$$S_c(\vec{F},\vec{T}) = a{\cdot}d - b{\cdot}c \qquad (1)$$

ここで、

$$a = \sum_{i=1}^{n} f_i{\cdot}t_i, b = \sum_{i=1}^{n}(1-f_i){\cdot}t_i,$$

$$c = \sum_{i=1}^{n} f_i{\cdot}(1-t_i), \qquad (2)$$

$$d = \sum_{i=1}^{n}(1-f_i){\cdot}(1-t_i)$$

$$a+b+c+d = n \qquad (3)$$

$\vec{T}$ がかすれたパターンを入力した場合には $b = 0$、$\vec{T}$ が汚れたパターンを入力した場合には $c = 0$ となるため、かすれたパターンでも汚れたパターンでも $b{\cdot}c$ の項は 0 となる。そのため、類似度 S_c は比較対象パターンのかすれと汚れの両方に高い類似度を維持できるので、雑音にロバストな認識ができる。

では、実際のデータに基づいて具体的な例を図 1を用いて示す。まず次元数 n は n=57 × 100=5700 画素とした。次に正規化だが、宮沢喜一のヒストグラムの最大値は、54 カ月目の 63 回で、正規化を行なうとすべての値について 100/63≅1.6 を掛ける。渡辺美智雄のヒストグラムの最大値は、34 カ月目の 36 回で、正規化を行なうとすべての値について 100/36≅2.8 を掛ける。2 つのヒストグラムを正規化したら一方を $\vec{F}$、他方を $\vec{T}$、今回は $\vec{F}$ を宮沢喜一、$\vec{T}$ を渡辺美智雄とする。そこで、(1),(2) 式に数値を当てはめる。まず、(2) 式の a については宮沢喜一・渡辺美智雄の両方が 1 の時で 821、b については宮沢喜一は 0 で渡辺美智雄は 1 で 138、c については宮沢喜一は 1 で渡辺美智雄は 0 で 629、d については宮沢喜一・渡辺美智雄の両方が 0 の時で 4112 である。(1) 式より $S_c(\vec{F},\vec{T}) = 821{*}4112 - 138{*}629 = 3289150$ となる。ここで示したものは最も類似度が高いものであるが、最も類似度が低いものは、江副浩正と宮沢喜一の関係で類似度の値は −265100 である。

4 実験結果

この実験で得られた結果を類似度が高いペアから 10 組と 20 組と 40 組を図 2,3,4 にグラフで示す。今回は、10 組と 20 組と 40 組の結果を示す。固有名詞間を結んでいる線は何らかの関係がある、すなわち出現分布の類似度が高いと考えられる固有名詞の間を結んでいる。固有名詞が並んでいる順番は、左上から回数が多い順に時計回りで並んでいる。この様に、選び出された固有名詞とその固有名詞間の関係がグラフで見ることができる。

5 得られた結果と事実関係との検証

ここでは、先に示した結果のグラフを用いて、そのグラフの関係が実際の関係に現れているかを検証する。結局選び出した人名は有名なものなので、人名の間には何らかの関係があると考えられるが、関係が大きいと思われるものを示す。では図 2で、なるべく大きな完全グラフとなっている関係について、実際の関係を示す。宮沢喜一・渡辺美智雄・加藤紘一は、第 1 次宮沢内閣の閣僚である。宮沢喜一・加藤紘一・田辺誠は、PKO 法案設立時の中心人物である。金丸信・小沢一郎は、東京佐川急便事件に引き続く竹下派会長問題の中心人物である。宮沢喜一・梶山静六は、自民党総裁・自民党幹事長という党の要職についている。宮沢喜一・河野洋平は、第 2 次宮沢内閣の閣僚である。しかし、渡辺美智雄も第 2 次宮沢内閣の閣僚であるが、この関係は図 4にも現れていない。結局、類似度が高い方から 76 番目に登場しているが、この理由として、河野洋平が宮沢内閣閣僚の後の、自民党総裁として新聞記事に出現する

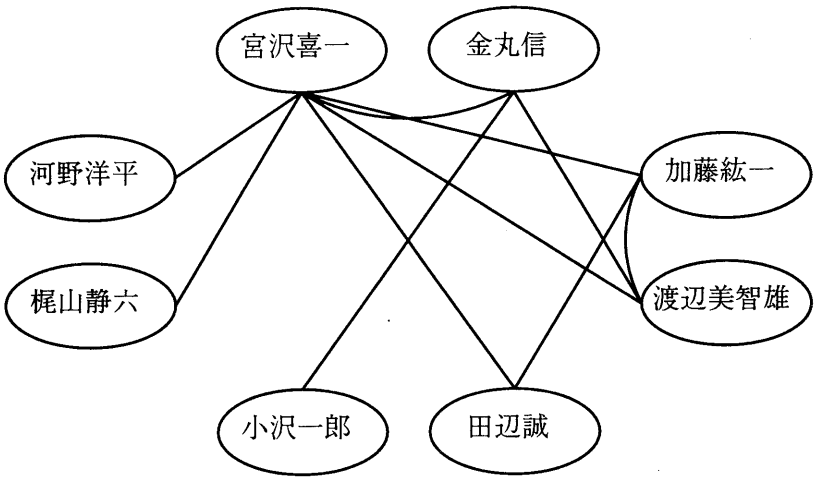

図 2: 類似度が高い方から 10 組選んだグラフ

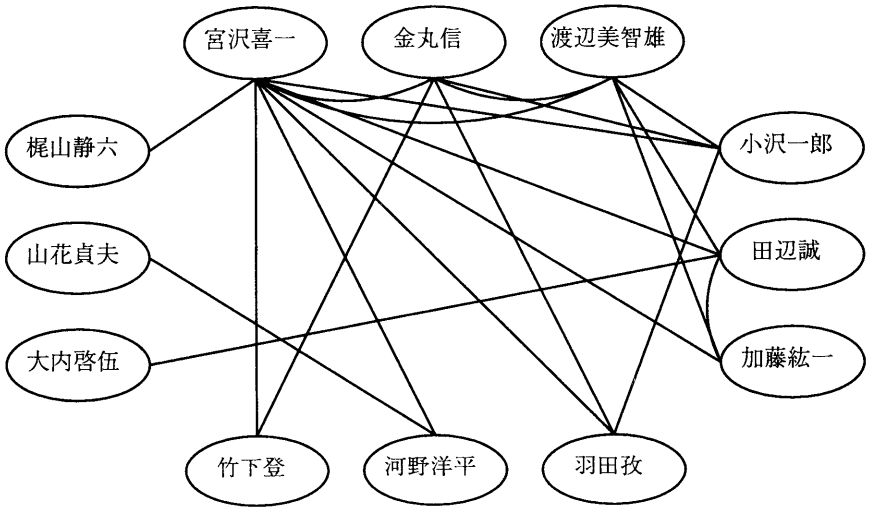

図 3: 類似度が高い方から 20 組選んだグラフ

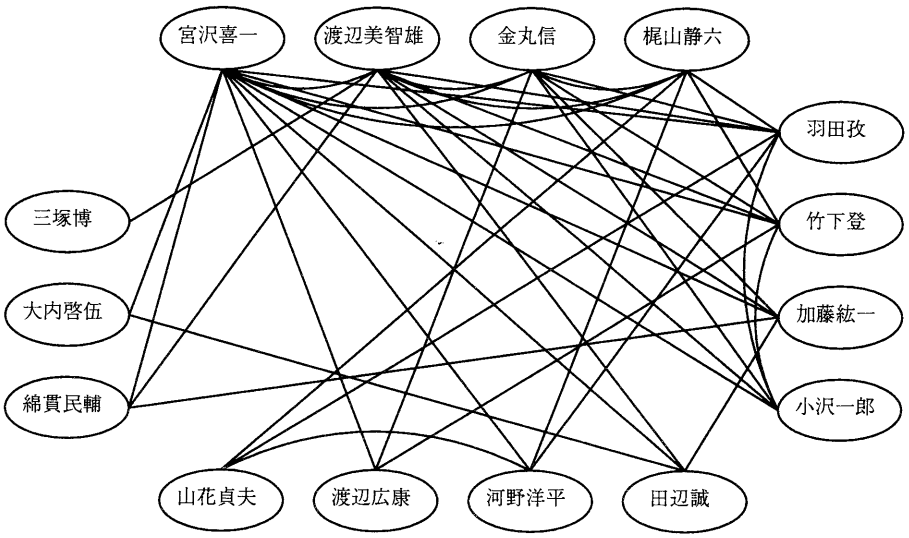

図 4: 類似度が高い方から 40 組選んだグラフ

方が多いからだと考えられる。

この様に、類似度が高い方から 10 組の場合については、得られた結果の関係のほとんどが実際の関係に現れている。

次に図 4について簡単な検証を行なう。ここで、渡辺広康に注目する。この人は、元東京佐川急便社長で東京佐川急便事件の中心人物である。東京佐川急便事件に関わった人物は、金丸信、竹下登で、渡辺広康・金丸信・竹下登との関係は図 4で現れている。しかし、渡辺広康は宮沢喜一とは全く関係がない。それなのに渡辺広康と宮沢喜一の間に関係が現れている。この理由は、東京佐川急便事件が宮沢内閣の時期に起こり、かつ宮沢喜一の出現数が多かったからだと考えられる。

検証を行なった結果、得られた結果が完全ではないことがわかった。

6 問題点

本研究での問題点は、4 つある。まず 1 つ目の問題点だが、前節の検証の項で少し触れている。出現分布の類似度を比較し、分布が似ていれば固有名詞間に関係があるとしてしまっていることである。同時期に全く別の出来事でそれぞれが記事に出現する場合もある。この問題を解決することは難しいと考えている。しかし、別の見方をすれば、同時期に起きた出来事に出現する人名の間で関係が得られることも検索の参考になると考えられる。

2 つ目の問題点は、得られた結果、すなわち出現分布の類似度が高い方からどこまでが有効な範囲かということである。この問題を解決するためには、情報源と固有名詞の範囲を拡大し、さらに実験を重ねれば、ある程度は有効な範囲を求めることができると考えている。

3 つ目の問題点は、得られた結果の一部が実際の関係に現れていることは検証し確かめたが、実際の関係の全てが結果に現れていることは検証を行なっていないことである。この問題については検証を行なえば済むことだが、前節で述べた通り、有名政治家の間には何らかの関係があると考えた方がいいので、検証は難しいと考えられる。

4 つ目の問題点は、出現分布の類似度を比較する方法でしか、固有名詞間の関係を得ていないことである。他の方法、相互情報量 [4] や、相関関数を用いる方法等を試みなければならない。

7 まとめ

我々は、新聞記事の中から検索に効果のある固有名詞群を発見することを試みた。次に補間類似度の高さより得られるグラフを生成した。その部分グラフと事実とを比較した。実験の範囲では妥当なグループが求まっているのがわかった。まだ実験は1つのデータに適用しただけなので、他のデータにも応用し、検証を行ないたい。

8 関連する研究

関連する研究としてシソーラスを用いたキーワードの抽出[5]と類似度行列の固有ベクトルを用いたクラスタ抽出法[6]等がある。どちらもキーワードと成り得る名詞、固有名詞のグループを抽出する点は同じであるが、本研究では、単に固有名詞の出現分布の類似度に注目し統計処理だけでキーワードと成り得る固有名詞のグループを抽出しているが、[5]では、シソーラスを用いて関連している名詞群を抽出しテキストのキーワードとしている。同じく[6]では、類似度行列の固有ベクトルを用いたクラスタ抽出法で互いに類似した対象の集合を抽出している。

謝辞

本研究に際し「CD-毎日新聞91-94年版」の使用の許可を下さった株式会社毎日新聞社に深く感謝します。

参考文献

[1] 日本アソシエーツ(株): CD-毎日新聞91, 92, 93, 94年版。

[2] 澤木美奈子、萩田紀博: "補完類似度に基づく新聞見出し文字の領域抽出と認識" 信学技報 TECHNICAL REPORT OF IEICE. PRU95-106,pp.19-24,1995.

[3] 澤木美奈子、萩田紀博: "補完類似度による劣化印刷文字認識" 信学技報 TECHNICAL REPORT OF IEICE. PRU95-14,pp.101-108,1995.

[4] 中川聖一: 『情報理論の基礎と応用』近代科学社(1992年4月10日)

[5] 鈴木斎、増山繁、内藤昭三: "語の意味分類の出現傾向を考慮したキーワード抽出の試み" 自然言語処理研究会 NL98-10,pp.73-80,1993.

[6] 津田宏治、仙田修司、美濃導彦、池田克夫: "共起行列の固有ベクトルを用いる単語クラスタリング法" 自然言語処理研究会 NL 103-6,pp.41-48,1994.

[7] 塚本明利: 分布特徴を反映したポテンシャル関数によるパターン識別" 信学技報 TECHNICAL REPORT OF IEICE. PRU95-216,pp.1-8,1996.

[8] 湯浅夏樹、上田徹、外川文雄: "大量のデータから自動抽出した名詞間共起関係による文書の自動分類" 自然言語処理研究会 NL 98-10,pp.81-88,1993.

[9] 原正巳、中島浩之、木谷強: "単語共起と語の部分一致を利用したキーワード抽出法の検討" 自然言語処理研究会 NL106-1,pp.1-6,1995.

[10] 松川智義、中村順一、長尾真: "共起関係の注目したDM分解と確立的推定による単語のクラスタリング" 自然言語処理研究会 NL 72-8,pp.1-8,1989.

[11] 崔進、小松英二、安原宏: "EDR電子化辞書を用いた単語類似度計算法" 自然言語処理研究会 NL93-1,pp.1-6,1993.

[12] 平岡冠二、松本裕治: "コーパスからの動詞の格フレーム獲得と名詞のクラスタリング" 自然言語処理研究会 NL104-11,pp.79-86,1994.

[13] 佐々木一朗、増山繁、内藤昭三: "語彙的結束性に着目した文章抄録法の提案" 自然言語処理研究会 NL98-9,pp.65-72,1993.

[14] 白倉悟子、梅村恭司、小川貴英: "新聞記事における事件特定のための単語群の抽出" 自然言語処理研究報告 NL113-17,1996.