多変量解析による文章の類似性解析

田中　佑一

2008 年 2 月 14 日

1　発表内容

複数の著者によって書かれた複数の著作物を，文章の統計解析の手法で識別する。

本研究では，読点の打ち方に着目する先行研究（金，1993）を元に，より少ない変数に着目しながら著者の文章の傾向を捉えることを目的とする。

また，シミュレーションによる仮想著者から多くのデータを収集し，解析法の信頼度を検証することを目的とする。

題材：セルジオ越後，金子達仁，戸塚啓のコラム

分析法：クラスター分析

2　クラスター分析とは

クラスター分析とは，特定の計算手順により，多くの観測対象について ‘‘ 似ているもの ’’ を集めて分類する数学的手法である。

ある対象と対象が似ているということは何らかの意味における ‘‘ 距離 ’’ が小さいことである。

クラスター分析を行なうには，対象間の距離とクラスター間の距離を決める必要がある。

対象間の距離の例：ユークリッド距離，標準ユークリッド距離，都市ブロック距離，相関係数など

クラスター間距離の例：最近距離，最遠距離など

3　文章の類似性解析

3.1　使用する文章

・セルジオ越後（サッカー評論家）

日刊スポーツのブログ「ちゃんとサッカーしなさい」より，2005 年 3 月から 2007 年 8 月コラム

・金子達仁（スポーツライター）

スポニチのサッカーコラムサイトより，

2006 年 8 月から 2007 年 8 月のコラム

・戸塚啓（スポーツライター）

スポニチのサッカーコラムサイトより，

2005 年 12 月から 2007 年 8 月のコラム

必要統計量を得るまでの手順は次のようにまとめられる。

文章の電子テキストを取得する。

形態素解析を行なう（JUMAN）。

欲しい統計データの取得（自作プログラム）。

3.2　特定文字に読点を打つ割合

[本解析をする前に行なっていること]

・注目した文字に読点を打つ割合が落ち着くに足る文量を収集しているかの確認。

・注目した文字に前回読点を打ったかどうかと，次回読点を打つかどうかが，独立試行であると考えられることの調査。

・それぞれのコラムを複数のグループ化するため，副助詞「は」が 200 個出現するごとに，1 つの作品として区分した。

解析方法：特定文字「は」「が」「と」「に」「で」に読点を打つ割合に着目して解析を行なう。

この解析における**特定文字「X」**とは，形態素解析をした結果，X が単独で分割されたものと，分割された文字の最後が X であるもの（○○…○ X）を示す。

著者識別の指標として次の値を定める。

いるべきクラスターではないクラスターに属した作品の数を**ペナルティ**と定める。また，

正答率＝ $\frac{\text{全グループ数}-\textbf{ペナルティ}}{\text{全グループ数}} \times 100$ と定める。

特定文字を単独で使用した解析を行なう。

特定文字「は」を必ず含む 16 通りについて解析

を行なう。

特定文字の著者識別能力

特定文字の著者識別能力は,

特定文字「は」＞「が」＞「と」≫「に」≈「で」

特定文字の組み合わせとしては,

2 変数「は」「に」・3 変数「は」「が」「に」が著者識別能力最大である。

3.3 シミュレーションを用いた解析

実験目的

3 人の現存する文章には限りがある。3 人の著者が特定文字に読点を打つかどうかをシミュレーションする。そして，どの文字に着目すべきかを解析する。

シミュレーションの仕方

・特定文字に読点が打たれなかったら 0，打たれたら 1 を入力すると考える。プログラムで 01 数列を作る。1 の出現率を実測データをもとに設定して，乱数を発生させる。(1 の出現した回数）/（0 と 1 の個数）× 100 を求める。

・3 人分（1 人 10 作品× 3 人=30 作品）のシミュレーション結果をクラスター分析にかける。

・特定文字「は」を必ず含む計 16 通りの変数の組み合わせを試し，それぞれペナルティを取得する。

・シミュレーション→クラスター分析→ペナルティ取得，という作業を 50 回繰り返し（生成した総作品数 30 × 50=1500)，どの変数を使用するとペナルティが低くなるのかを調べる。

50 回の解析の結果，特に識別能力が高かったのは変数（は，が，と)，変数（は，が，で，と）の組み合わせである。

ペナルティ数の分布

変数（は，が，で，と)，変数（は，が，と）に着目したヒストグラムが描ける。シミュレーション→ペナルティ取得という試行は独立であり，読点の打つ割合は固定されているので同分布である。

中心極限定理より，ペナルティの分布は試行回数→∞で正規分布に近づく。

実験結果

・変数（は，が，で，と）

ペナルティ：平均 7．14　標準偏差 3．00

正答率：平均 76．2 %　標準偏差 10．0

平均ペナルティが最も低い変数の組み合わせとなった。

・変数（は，が，と）

ペナルティ：平均 7．58　標準偏差 2．7

正答率：平均 74．7 %　標準偏差 9．0

標準偏差が最も低かったことから，正答率のゆらぎが小さく，その意味で識別能力が安定しているといえる。

信頼度について

変数（は，が，で，と）を使用してクラスター分析を試みた場合，ペナルティ 2．2〜12．1，正答率 59．7 %〜92．7 %

変数（は，が，と）を使用してクラスター分析を試みた場合，ペナルティ 3．1〜12．0，正答率 59．9 %〜89．5 %

の範囲の実験結果が 90．0 %の確率で得られることが分かる。

参考文献

[1] 水野欽司「多変量データ解析講義」朝倉書店，1996.

[2] 村上征勝「真贋の科学」朝倉書店，1994.

[3] 金明哲，村上征勝，永田昌明，大津起夫，山西健司「言語と心理の統計」岩波書店，2003.

[4] 金明哲，樺島忠夫，村上征勝（1993)「読点と書き手の個性」計量国語学 18，382–391.

[5] 金明哲（1994)「読点の打ち方と文章の分類」計量国語学 19，317–330.

[6] 金明哲（1997)「助詞の分布に基づいた日記の書き手の識別」計量国語学 20，357–367.

[7] 金明哲（2002)「助詞の分布における書き手の特徴に関する計量分析」社会情報 11，15–23.

[8] 吉岡亮衛（1999)「新書の数量的分析」人文学と情報処理 20，51–56.