用語類似度のクラスター分析

2014SS101 劉正賢
指導教員：小藤俊幸

1　はじめに

わたしたちは日常的にグループ分けを行っている. 色や数字, 性別, 年代, 国籍などの様々な条件をつけて似たものどうしを分ける. こうした感覚的なグループ分けでなく, 数字のデータに基づいて数学的に似たものどうしを分けるには ”似ている ”ということを定義しておく必要がある. 最もシンプルなのは ”距離が近い = 似ている ”と定義することである. データににによっては ”似ている = 同じような反応をする ”と考えることも可能であり, ”似ている = 相対係数が大きい ”と定義してもよい. 本研究では言語の数理的研究に焦点を当て, 単位語水準での用語類似度を使ったクラスター分析を行う. 文学作品それぞれを各々の表現領域と考え, その上での語彙の似ている度合いを量的に表にして解析を行う.

2　クラスター分析

クラスター分析ではデータのことを個体と呼び, 個体と個体が集まってクラスタを構成する. このクラスタ間の距離を求める方法には 2 つのケースが考えられる. クラスタの成分が 1 個だけからなる場合と,2 個以上からなる場合である. 前者の場合は ”個体と個体の距離 = クラスタ間の距離 ”とすればよいが, 後者の場合は求め方の種類が複数存在する. 主なものとしては

- 最短距離法
- 最長距離法
- 群平均法
- メディアン法
- 重心法
- ウォード法

があり, 本研究では群平均法を使用する. また, 距離の値が大きいほど類似性を高いとする類似度と, 値が小さいほど類似性を高いとする非類似度があるが, ここでは後者を使用し次の手順で解析を行う.

A：1つずつを構成単位とする n 個の個体から始める.
B：クラスター間の非類似度行列から, 最も類似性の高い2つの個体を合わせて 1 つのクラスターを作る.
C：クラスターが 2 つになれば終了するが, そうでなければ D に進む.
D：B で作ったクラスターと他のクラスター及び個体との非類似度を計算して、非類似行列を更新して B に戻る.

2.1　群平均法

クラスタ A とクラスタ B の各々の個体すべての組合せについて距離を求め,

その距離の平均値 = 2 つのクラスタ A,B 間の距離

と定義する.

3　データについて

クラスター分析を行って近代短歌における歌人の個々人の用語情況から根岸派と明星派に分ける.2 つの流派はそれぞれ全盛時代に差があり, より正確な比較のために時期を揃える必要があるので明治 33 年から 41 年までと統一し, この 9 年間で 1000 首以上の作品を発表した両派の有力歌人を対象にする. 歌人の個人的な語彙を論ずるには, 一人当たり 100 首程度の作品を材料にする必要があるからである。対象歌人は,

根岸派 8 名：正岡子規, 香取秀真, 赤木格堂, 伊藤左千夫, 長塚節, 蕨真, 三井甲之, 斎藤茂吉

明星派 9 名：与謝野鉄幹, 鳳晶子, 茅野蕭蕭, 平野万里, 平出露花, 相馬御風, 石川啄木, 吉井勇, 北原白秋

であり, 作品は各歌人の作品の中から 100 首をランダムに選んだ. この 17 名を自派他派を問わず 136 対にした表を使用し, 直接手計算でクラスター分析を行っていく.

	白秋	勇	晶子	露花	啄木	鉄幹	御風	万里	蕭蕭	茂吉	左千夫	子規	秀真	格堂	真	甲之	節
白秋		.35	.32	.34	.36	.28	.27	.32	.29	.26	.25	.25	.24	.23	.25	.23	.24
勇	.35		.30	.34	.36	.34	.30	.32	.27	.26	.26	.26	.27	.24	.26	.23	.24
晶子	.32	.30		.36	.34	.46	.41	.36	.35	.25	.25	.30	.27	.26	.24	.25	.23
露花	.34	.34	.36		.40	.40	.35	.36	.35	.27	.30	.27	.26	.25	.25	.30	.22
啄木	.36	.36	.34	.40		.35	.41	.34	.32	.25	.29	.28	.27	.26	.27	.25	.24
鉄幹	.28	.34	.46	.40	.35		.41	.35	.36	.28	.28	.30	.27	.26	.25	.23	.24
御風	.27	.30	.41	.35	.41	.41		.40	.46	.25	.31	.26	.30	.26	.24	.25	.23
万里	.32	.32	.36	.36	.34	.35	.40		.43	.29	.30	.30	.29	.29	.28	.27	.24
蕭蕭	.29	.27	.35	.35	.32	.36	.46	.43		.28	.29	.30	.30	.30	.28	.26	.25
茂吉	.26	.26	.25	.27	.25	.28	.25	.29	.28		.31	.35	.31	.31	.32	.32	.26
左千夫	.25	.26	.25	.30	.29	.28	.31	.30	.29	.31		.35	.34	.31	.33	.31	.31
子規	.25	.26	.30	.27	.28	.30	.26	.30	.30	.35	.35		.32	.35	.32	.33	.31
秀真	.24	.27	.27	.26	.27	.27	.30	.29	.30	.31	.34	.32		.35	.33	.30	.31
格堂	.23	.24	.26	.25	.26	.26	.26	.29	.30	.31	.31	.35	.35		.31	.32	.30
真	.25	.26	.24	.25	.27	.25	.24	.28	.28	.32	.33	.32	.33	.31		.28	.31
甲之	.23	.23	.25	.30	.25	.23	.25	.27	.26	.32	.31	.33	.30	.32	.28		.28
節	.24	.24	.23	.22	.24	.24	.23	.24	.25	.26	.31	.31	.31	.30	.31	.28	

図 1　136 対の表

4　群平均法での解析

図 1 の表から最小値を探すと, 平出露花と長塚節の距離 0.22 である. この 2 人の作者をクラスタとして, 他 15 人との距離を計算していく. 群平均法を使用するので, 対象となる作者とクラスター間の距離の平均を出せば良い.

4.1　結果

群平均法を用いたクラスター分析の結果, 最終的な 2 つのグループ A,B は次のようになった.

A：北原白秋, 石川啄木, 相馬御風, 伊藤左千夫, 正岡子規, 与謝野鉄幹, 斎藤茂吉, 茅野蕭蕭, 香取秀真, 三井甲之
B：吉井勇, 鳳晶子, 平出露花, 赤木各堂, 平野万里, 蕨真, 長塚節

デンドログラム (樹形図) に描くとこうなる.

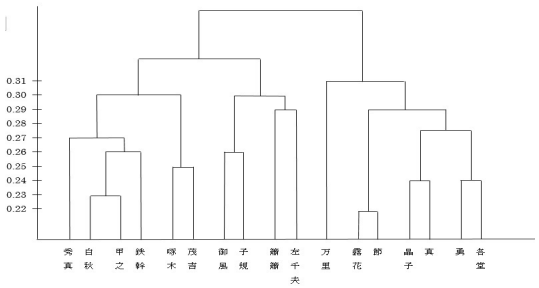

図 2　デンドログラム

4.2　最短距離法での解析

クラスタ A とクラスタ B の各々の個体すべての組合せについて距離を求め, 次のように定義する.

その距離の最小値 ＝ 2 つのクラスタ A,B 間の距離

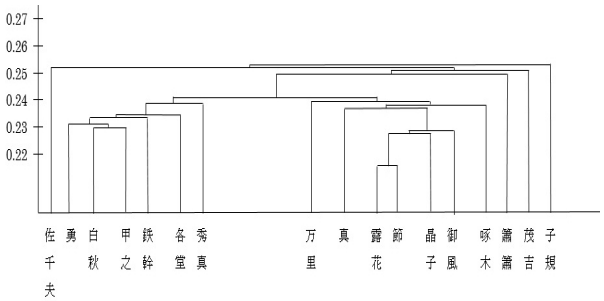

図 3　デンドログラム

4.3　最長距離法での解析

クラスタ A とクラスタ B の各々の個体すべての組合せについて距離を求め, 次のように定義する.

その距離の最大値 ＝ 2 つのクラスタ A,B 間の距離

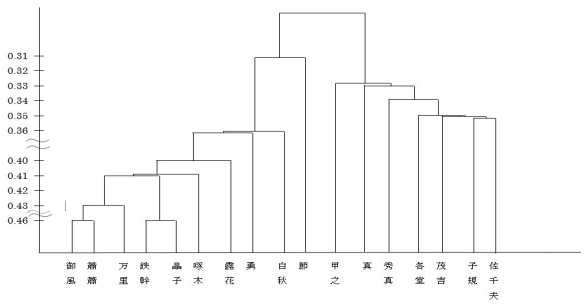

図 4　デンドログラム

5　考察

群平均法, 最短距離法, 最長距離法の 3 つの方法で同じデータを解析した結果, どの結果もあらかじめ分かれていた根岸派と明星派と異なる分かれ方になった. その要因の 1 つに, 解析方法によって図 5(文献 [4]) のように結果が変わったという可能性が考えられる. またデータにおけるグループ分けには各歌人の思想や精神, 活発に活動していた時期などが基準になっていたが, 本研究においては歌人の作品における頻出単語の共通点を基準に数値的な計算でグループ分けを行ったからであると考えられる. また,17 人各々の歌人の 1000 首以上の作品の中からランダムに詩を選んで用語類似度を導きだしたことが結果を大きく変えた要因であると考えられる.

図 5　各計算方による結果の違い

6　おわりに

今回の研究では根岸派と明星派という 2 つのクラスタに分けるという前提があったが, クラスタ分析における最適なクラスタの個数という基準は存在しない. そういった意味では, 研究者の趣向次第では更に細かいクラスタに分けることも可能である. また, クラスター分析を行う 6 つの方法を紹介したが, 図 5(文献 [4]) からわかるように同じデータでもどの方法を使って計算するかで結果が変わってくる. データを観ただけでは最適な方法を見つけることができないため, 最適なクラスタに分けるためにどの方法で計算するかその都度考える必要があると思われる.

参考文献

[1] 石村貞夫, 石村光資郎：『入門はじめての多変量解析』. 東京図書，東京，2007.

[2] H.C. ロムスブルグ（西田英郎, 佐藤嗣二 共訳）：『実例クラスター分析』. 内田老鶴圃，東京，1992.

[3] 水谷静夫：『数理言語学』. 培風館，東京，1982.

[4] クラスター分析の手法 ：[online] www.albert2005.co.jp/knowledge/data_mining/cluster/hierarchical_clustering ＜参照 2017-8-20 ＞