

楽天市場商品評価データに基づく 店舗間比較

南木望，三道弘明（関西学院大学総合政策学部）

背景・目的

近年、テキストデータの分析手法としてテキストマイニングは注目されている。
しかし、単に頻出語を抽出し、それらの関係を共起ネットワーク等に表現しても
経営戦略に有益な知見を得ることは困難である。

外部変数やコーディングを利用しクロス集計や対応分析を行うことで
統計的な有意差の存在を確認したり、外部変数やコードの視覚的なポジショニング
を行うことが可能となる。

楽天市場の商品レビュー(テキストデータ)
を分析し店舗ごとの特徴を抽出し提示

⇒各店舗の経営戦略に役立てる
(商品開発にフィードバック)

方法

テキストマイニング

⇒テキストデータからコンピュータを使って
何らかの情報を抽出する

【使用データ】

楽天市場の商品評価レビューのうち
最新の2012年12月のデータ
2,181,861レコード/15フィールド

そのうち、季節を問わず購入されると考えられる
「珈琲」、「サプリメント」の商品評価データ

■KH Coder

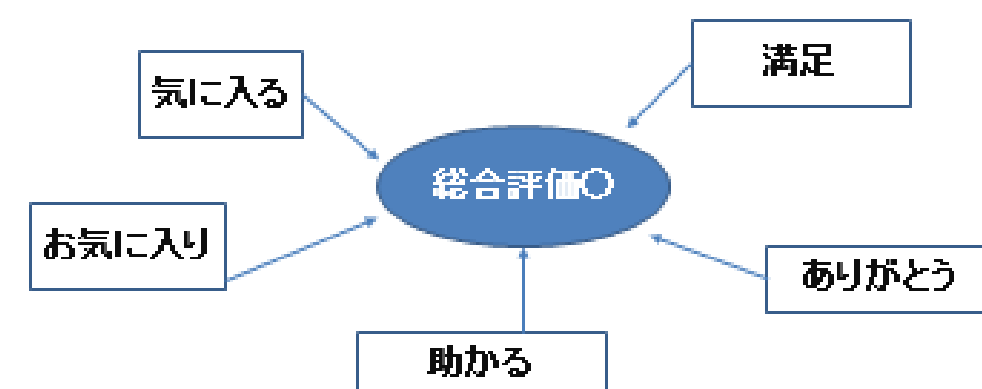
⇒テキストデータを分析するためのフリーソフトウェア

開発者：樋口耕一（立命館大学准教授）

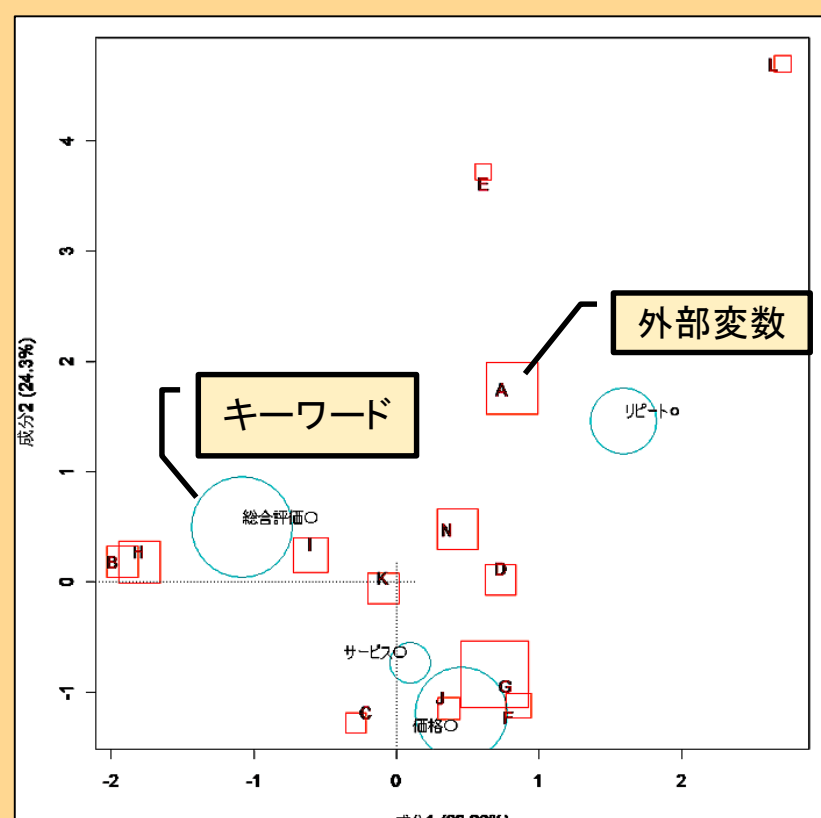
□機能

- ✓形態素解析(茶釜)
- ✓頻出語の抽出
- ✓外部変数の利用
- ✓コーディングルールの適用
- ✓種々の統計的分析(R) 等

複数の単語を一つのキーワードに集約させる



分析結果



分割表による独立性の検定

出現の割合が少ない

小規模	*価格○	*サービス○	*リピート○	*総合評価○	ケース数
A	75 (18.62%)	9 (1.18%)	83 (10.88%)	101 (13.24%)	763
B	38 (14.07%)	10 (3.70%)	6 (2.26%)	88 (39.30%)	270
C	288 (10.09%)	54 (4.47%)	119 (9.77%)	184 (14.33%)	1284
D	74 (11.10%)	20 (4.00%)	14 (2.80%)	184 (37.59%)	490
E	63 (18.79%)	14 (4.76%)	20 (6.90%)	104 (30.89%)	338
F	79 (28.09%)	7 (2.62%)	34 (12.79%)	87 (32.58%)	267
G	67 (14.39%)	15 (3.21%)	47 (10.05%)	77 (16.49%)	467
H	816 (17.39%)	180 (3.39%)	417 (8.03%)	873 (20.61%)	4721
合計	1481 (6.84%)	218 (1.00%)	534 (2.47%)	265 (1.21%)	

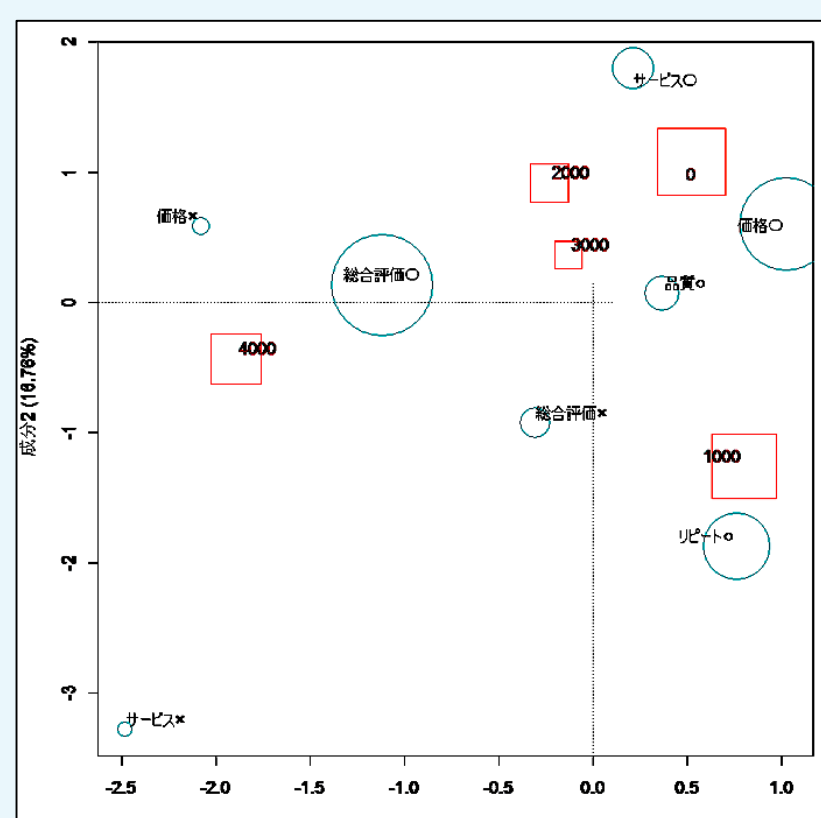
統計的に有意差が認められた

出現の割合が多い

※：有意水準1%で有意
※：有意水準5%で有意

小規模	特徴
A	価格○、サービス○、総合評価○の出現割合が少ない
B	リピートに関する語の出現は少ないが総合評価は高い
G	価格とサービスに関する語の出現が多い
H	総合評価が高い
I	サービスが良い
K	価格が安くリピートしやすい
N	リピート○が比較的多い

対応分析
データの持っている特徴を
二次元平面に表現した図



価格帯

分割表による独立性の検定 (サプリメント)

■価格帯ごとにコーディングルールを適用

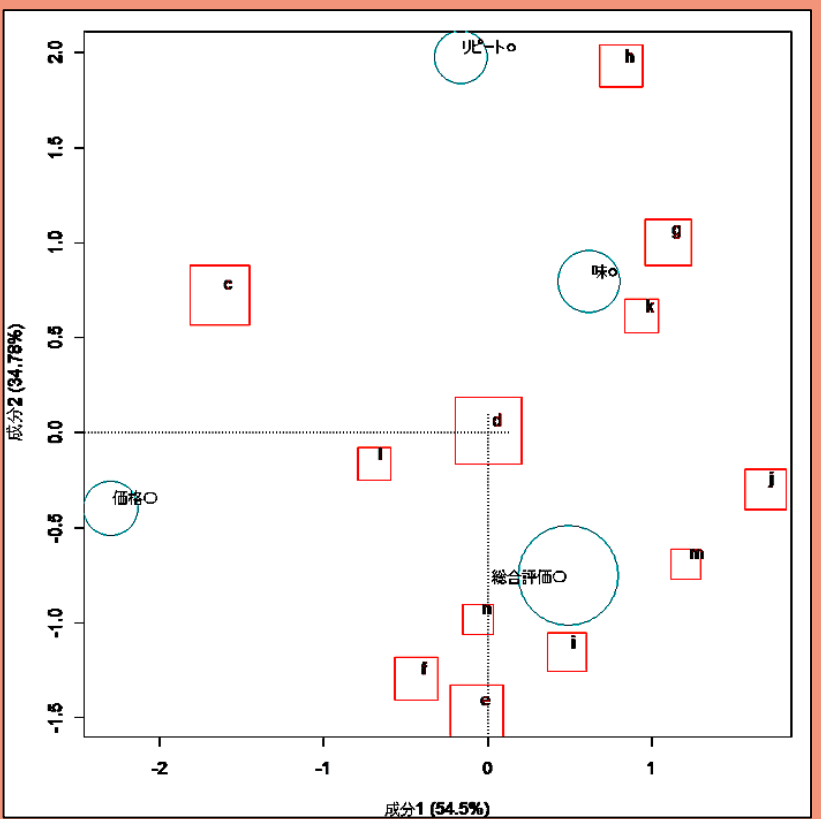
大規模	*価格帯	*サービス○	*リピート○	*総合評価○	ケース数
0	172 (19.07%)	20 (2.22%)	140 (16.42%)	401 (44.46%)	907
1000	756 (25.40%)	76 (2.56%)	724 (24.37%)	1550 (52.51%)	2971
2000	1071 (24.27%)	170 (3.83%)	138 (25.39%)	2270 (51.41%)	4413
3000	174 (19.18%)	41 (4.52%)	272 (30.57%)	477 (52.59%)	907
4000	103 (22.06%)	23 (4.93%)	92 (19.70%)	417 (53.25%)	467
合計	2278 (23.56%)	330 (3.42%)	2315 (23.96%)	4960 (51.33%)	9660

統計的に有意差が認められた
⇒価格帯の間に差

小規模	*サービス○	*リピート○	*総合評価○	ケース数
0	54 (3.39%)	115 (7.14%)	282 (17.50%)	1611
1000	41 (2.89%)	187 (13.39%)	446 (16.99%)	1446
2000	25 (4.82%)	43 (8.29%)	122 (23.51%)	519
3000	15 (5.88%)	28 (10.20%)	57 (22.33%)	255
4000	25 (2.87%)	66 (7.59%)	84 (30.84%)	970
合計	160 (3.38%)	447 (9.83%)	977 (21.97%)	4741

価格帯4000は総合評価が高い

サプリメント

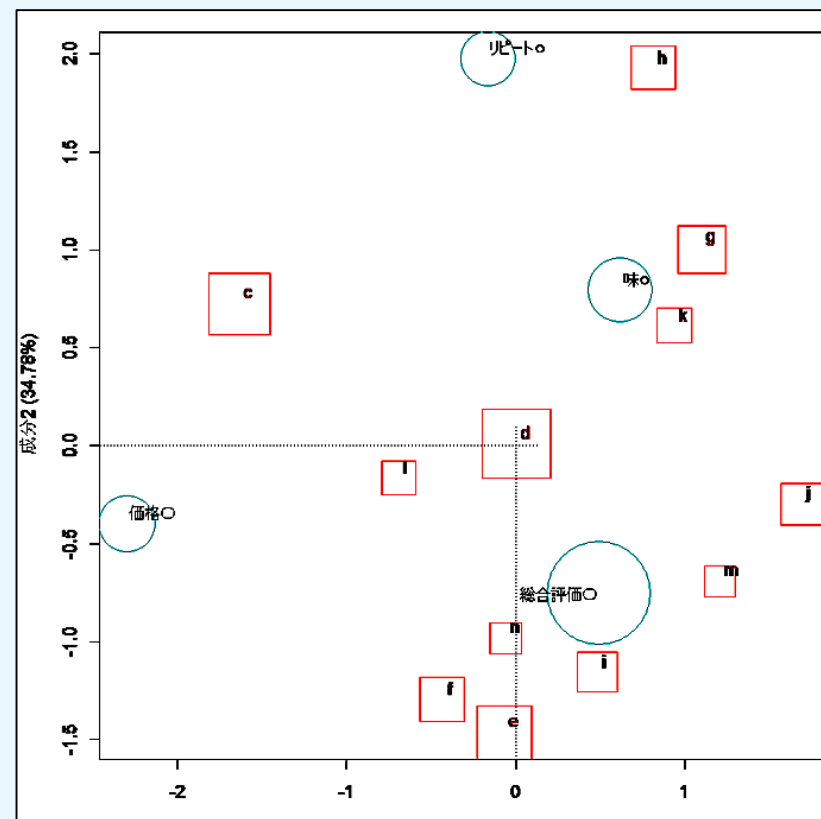


分割表による独立性の検定3 (珈琲小規模)

小規模	*価格○	*サービス○	*リピート○	*総合評価○	ケース数
c	73 (33.18%)	54 (24.55%)	116 (52.73%)	51 (23.18%)	220
d	38 (13.57%)	38 (13.57%)	130 (46.43%)	48 (17.14%)	280
e	29 (16.28%)	15 (8.47%)	102 (57.62%)	27 (15.25%)	177
f	25 (22.12%)	7 (6.19%)	65 (57.52%)	27 (23.89%)	113
g	10 (7.88%)	21 (15.91%)	62 (46.87%)	35 (26.52%)	132
h	11 (9.81%)	28 (25.23%)	56 (50.45%)	25 (21.97%)	111
i	10 (10.87%)	7 (7.61%)	45 (48.91%)	14 (15.32%)	92
k	7 (7.72%)	13 (18.06%)	41 (56.94%)	19 (26.39%)	72
l	17 (28.76%)	15 (22.79%)	46 (79.70%)	12 (18.18%)	66
m	5 (8.93%)	5 (8.93%)	39 (58.82%)	14 (25.00%)	56
n	9 (15.25%)	7 (11.86%)	33 (55.93%)	8 (13.56%)	59
合計	238 (16.07%)	227 (15.33%)	798 (53.88%)	309 (20.80%)	1481

統計的に有意差が認められた

小規模	特徴
c	価格が安くリピートしやすい
d	プラス評価の語のヒット数が少ない
e	価格○が比較的多い
f	リピートに関する語の出現数が少ない
g	価格はいいけど味に関する記述が多い
h	リピートが多く味に関する記述も多い
i	味に関する記述が少ない
k	リピート○、味○が比較的多い
l	価格が手ごろで評価が高い
m	総合評価は高い
n	味に関する記述が少ない



リピートされやすい価格帯→1000-2000
サービス評価が高い価格帯→3000-4000
キーワードによる語の出現数及び評価が低い価格帯→0-1000

分割表による独立性の検定 (珈琲)

■価格帯ごとにコーディングルールを適用

大規模	*価格帯	*サービス○	*リピート○	*総合評価○	ケース数
0	172 (19.07%)	20 (2.22%)	140 (16.42%)	401 (44.46%)	907
1000	756 (25.40%)	76 (2.56%)	724 (24.37%)	1550 (52.51%)	2971
2000	1071 (24.27%)	170 (3.83%)	138 (25.39%)	2270 (51.41%)	4413
3000	174 (19.18%)	41 (4.52%)	272 (30.57%)	477 (52.59%)	907
4000	103 (22.06%)	23 (4.93%)	92 (19.70%)	417 (53.25%)	467
合計	2278 (23.56%)	330 (3.42%)	2315 (23.96%)	4960 (51.33%)	9660

統計的に有意差が認められた

小規模	*サービス○	*リピート○	*総合評価○	ケース数
0	79 (12.76%)	305 (49.27%)		619
1000	85 (20.37%)	284 (54.32%)		488
2000	20 (13.51%)	86 (56.11%)		146
3000	21 (18.03%)	59 (50.82%)		131
4000	19 (19.59%)	59 (60.82%)		97
合計	238 (16.07%)	798 (53.88%)		1481

価格帯4000は総合評価が高い

珈琲

結論・考察

レビューデータにコーディング・ルールを適用，また外部変数も利用することで
全て同じ条件のもと店舗ごとの特徴を抽出することができた

結果を各店舗に提示

商品の開発・経営戦略に生かすことができる