テキストデータを扱う多変量解析について

Kosugitti の研究ノート

2008/02/16

1　はじめに：テキストマイニングとデータ

1.1　データの切り出し方が一番の問題

テキストマイニングで分析されるデータは、話し言葉であったり、書き言葉であったり、単語であったり、長文であったり、と実に様々である。言い換えれば、研究者や回答者にとって、自由度の大きいデータだといえる。社会調査で用いられる調査票や、心理学や福祉の分野で開発されてきた尺度は、一定の手続きを経て用語や項目が選定され、回答方法も決められている。これに対して、テキストデータのような、回答形式に束縛がない自由というのは、研究者にも調査協力者にもメリットがある。たとえば研究者は調査票や尺度を作る手間が少なくてすむし、調査協力者も心理的・身体的負担が少なくてすむ、などである。

しかし、この「自由度が大きい」ことは、データを分析するときは逆に、デメリットとなる。つまり、データが膨大すぎて、どこから手をつけて良いのか、どのように分析すればよいのか、どうすれば意味のある情報と意味のない情報とを区別することができるのか、について、研究者はひとしきり頭を悩ませることになるからだ。

テキストマイニングとは、得られたデータを極力客観的に（分析者の主観がはいらないように－誰が分析しても同じ結果が出るように）扱いつつ、意味のある情報を引き出そうとする試みである。その最大の難点は、どのようにコード化するべきか、そこからどのようにして情報を引き出すか、の二点に集約できる。

例えば、「阪神タイガースに対してどのようなイメージをお持ちですか」という質問項目に、自由記述で回答を求めたとしよう。

ここで得られる回答が、「好きやで」とか「生まれたときからずっと応援しています」とか、「ガンバレ～」といったような、自由なコトバであったとする。これを集計したところで、例えば甲子園球場のお客さんと東京ドームのお客さんを比較して、どちらの方が強く応援しているかといった結論は出てこない。いくつか回答例を選び出したとしても、「甲子園球場から出てきたお客さんには『ガンバレ』という方がいらっしゃいました。東京ドームから出てきたお客さんの回答からは『今年は強いね』という声が聞かれました。」としか言えない。『ガンバレ』と『今年は強いね』を比較して、どちらが強く応援しているとか、愛しているといったことまではわからないのである。さらに、その言葉の選び方が分析者の自由裁量になることも問題である。このやり方では、違う分析者が同じデータを見たら、違う報告書ができあがる可能性があるのだ (どの回答をピックアップするかによって、結果は大きく左右されるだろう)。

では、適当に言葉をピックアップするのではなくて、何か客観的な指標に基づいて分析すればどうだろうか。例えば「一番多かった回答例」という選び方ならどうだろうか？

例えば「甲子園球場から出てきたお客さんで一番多かった回答は『ガンバレ』、東京ドームから出てきたお客さんの回答で一番多かった回答は『今年は強いね』でした」という分析であれば、なるほど先ほどよりは客観的である。なぜなら、別の人が同じデータを見ても、回答の数が一番多かったもの、という基準で選べば、結果が変わるはずがないからだ。もちろん、ここから「地元ファンの方が愛している」という結論は出てこないが、『回答場所によって、回答パターンは明らかに違う』といった結論ぐらいなら導き出せそうだ。

しかし実は、テキストマイニングをするときは、もうひとつ微妙な問題がある。色々な回答、たとえば、「ガンバレ～！」、「ガンバレ～」、「が・ん・ば・れ」、「頑張れ」、「頑張ってください」があったとしよう。このとき、「ガンバレ　5票」なのだろうか。それとも、各1票なのだろうか。

これを明確に規定するルールはない。この点は、あくまでも研究者の判断に任されている。言葉の頻度だけで選ぶのか、回答者の意味を汲み取ってカテゴライズ（分類）するのか。ここはテキストマイニングの、永遠の悩みとでも言うべき点である[*1]。

数量化三類などの多変量解析にかける場合、度数分布表を細かく分解していくことになるのだが、連続値のデータと違って、度数がひとつ増えるか増えないかということが結果に大きく影響する。この意味でも、データの切り出し方が一番の問題になる。

1.2　テキストマイニングで扱うデータの種類

テキストマイニングで扱うデータは、以下のような種類があるだろう。

- ある人から得られた発話、記述物 (日記など) をデータ全体として、そこからキーワードを抽出し、データとする。
- あるテーマについて、複数の人間から言葉が発せられている、会議・討論・グループワークなどから得られたデータを全体とし、そこからキーワードを抽出、データとする。
- 質問紙調査で自由記述してもらい、それをそのままデータとする方法。これには更に2つのパターンがあって、
  - 文章で回答してもらったデータ。例えば「私は○○について、▲▲と考えています」など。
  - 単語で回答してもらったデータ。例えば「きれい」「面白い」など。

  にわけられる。

テキストマイニングでは、特にどのような使用法がよいとか、どのようなデータに向いている、といった性質は特にない。ただし、うまく聞き出さないといい回答が得られないし、とりとめもない言葉が頻出して、十分まとまったクロス集計表が作れない (ほとんどが度数1の表になってしまう、など) ということになると、後々の分析が難しくなるということは覚えておいて欲しい。

1.3　テキストマイニングの流れ

テキストマイニングは以下のような流れで進められる。

1. テキストデータの収集
2. テキストデータを語の単位に分割

[*1] 言語学の分野では、「オランウータン」「オラウータン」などの誤表記とみられるものも、とりあえず区別してあつかうようである。

3. 不要な語を削除し、データを見やすくする
4. 度数分布表の作成
5. 多変量解析的アプローチ (数量化など)

具体例で見てみよう。次の文は、とあるラーメン屋さんに対する評である。

> 味わいは、これまたトンコツで美味しかったのだが、ややマイルドすぎる感があった。

まず、得られたデータが普通の文になっている場合、これを分かち書きにする必要がある。

> 味わい は 、これ また トンコツ で 美味しい た の だが、やや マイルド すぎる 感 が ある た 。

そこからキーワードを抜き出す。すなわち、句読点、記号、助詞や指示代名詞を省き、名詞、動詞、形容詞が残るように分割するのである。

> 味わい　トンコツ　マイルド

次に、データ間の関連性を見いだすべく、度数分布表を作成する。このときは、データ間の関連を見ても良いし、調査協力者間の関連や他の質的変数 (属性など) を絡めたクロス集計表でもよい。

そして最後に、このクロス集計表から線形関係を見いだす多変量解析的アプローチを採る。これについて、以下で詳しく論じる。

2　質的データの多変量解析

2.1　対応する分析手法

質的データの分析に有効な林の数量化は、質的なデータに適切な数値を与える方法である。数値の与え方については、目的によって以下の四種類に分かれる (表.1 参照)。

表 1　林の数量化

手法	外的基準	データ	目的	関連する手法
数量化一類	量的変数	質的変数	外的基準の予測	重回帰分析
数量化二類	質的変数	質的変数	外的基準の判別	判別分析
数量化三類	なし	質的変数	変数間の関係の要約と記述	正準相関分析
数量化四類	なし	対象間の類似度	対象間の関係の要約と記述	多次元尺度法

テキストマイニングで用いられるのは数量化三類に関連する手法で、「質的変数を、外的基準に合わせるのではなく、その内部構造を明らかにするべく分析し、変数間の関係を要約・記述する」分析法である。このようにややこしい表現をしたのは、この分析手法が、日本では林の数量化、フランスではコレスポンデンス・アナリシス、オランダでは等質性分析[*2]、カナダでは双対尺度法、というそれぞれ異なる名称で発展してきているからである。それぞれに、それぞれの哲学が背景にあり、微妙に異なるのだが、ねらいとしているところは

[*2] SPSS にはこの等質性分析がパッケージングされている

同じなので、以下では双対尺度法をもとに考えていく。

ちなみに、双対尺度法は主成分分析法、正準相関分析とも概念的に近い分析方法である。この違いを表.2 にまとめてみた。

表 2　双対尺度法と主成分分析法、正準相関分析

分析方法	もとになる行列	分解方法	得られる重み
主成分分析法	n × n	固有値分解	項目に対する重み
双対尺度法	n × m	特異値分解	行と列に対する重み
正準相関分析	n × n と m × m の行列	固有値分解	二つの項目群それぞれについての重み

主成分分析と双対尺度法は、もとになる行列が違うので、分解方法や重みが変わってきてはいるが、基本的に考え方はその内部構造を知ろうとしているという意味で同じである。しかし、表.1 にあるように、双対尺度法の量的変数バージョンに対応する分析としては、正準相関分析になる。

正準相関分析とは、例えば二つの尺度間の相関係数を最大にするように、各項目に対する重みをつけるものである。すなわち、二つの変数群、X と Y があり、それぞれに重み付けをして合成変量を作れば $U = a_1X_1 + a_2X_2$ と $V = b_1Y_1 + b_2Y_2 + b_3Y_3$ のようになろう。このとき、U と V の相関 $r_{U,V}$ が最大になるように、a_1, a_2, b_1, b_2, b_3 を求めようというものである。つまり、n 次の正方行列と m 次の正方行列の二つを用いて、共通次元の固有値をひとつ抽出し、それに対する各項目への重み付けを算出するという方法である。

2.2　双対尺度法

2.2.1　行と列の双方向に重みをつける

質的なデータを分析する前に、演算操作が簡単な量的変数の場合を復習しておこう。量的変数の場合、その内部構造を明らかにする方法として、主成分分析や因子分析があげられよう。主成分分析は、$X_1, X_2, X_3, \ldots$ という量的変数に重み $w_1, w_2, w_3, \ldots$ をつけることで得られる合成変数 Y を作る方法である (式 1)。この合成変数 Y の分散を最大にするように重みを決定するのが主成分分析であった。

$$Y = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_nX_n \tag{1}$$

$$\text{ただし、} w_1^2 + w_2^2 + w_3^2 + \ldots + w_n^2 = 1.0$$

合成得点の分散の最大値は固有値に等しく、このとき作られる合成変数を主成分、重みを負荷量と呼ぶ。この場合、得られた量的データが少なくとも間隔尺度以上の水準である必要がある。

さてしかし、主成分分析だけではわからないことがある。例えば血圧と年齢のデータがあって、クロス集計表を作ったとする (表.3)。

表 3　血圧と年齢

血圧	20-34 歳	35-49 歳	50-65 歳
高い	0	0	4
普通	1	4	1
低い	3	1	1

このような線形関係が明らかな場合は、主成分分析をすることに意味があるだろう。しかし、表.4 のような場合、一件線形性が見られないこれら二変数間に、関連がないといえるだろうか？

表 4　血圧と頭痛

血圧	ない	たまに	時々
高い	0	0	4
普通	3	3	0
低い	0	0	5

このような場合は、「血圧が高い人か低い人は、頭痛がある」という傾向が見て取れる。しかし、二つの項目の相関係数としては $r = -0.06$ であり、主成分分析では意味のある主成分を抽出しにくい。

これはなぜか、というと主成分分析は線形性の制限があるからである。すなわち、「1. 血圧高い」、「2. 血圧普通」、「3. 血圧低い」という順番が保存されていなければならない。例えばこれが、表.5 のようであれば、綺麗な線形関係になっているし、主成分分析もその威力を発揮したであろう。

表 5　並び替えられた「血圧と頭痛」

血圧	ない	たまに	時々
高い	0	0	4
低い	0	0	5
普通	3	3	0

双対尺度法は、このように**行・列の双方を並び替えることによって、非線形関係において線形性が最大になるものを見つける方法**であると言えよう。

3　他の手法との違い

さてここで、数量化三類とコレスポンデンス分析、および双対尺度法とよばれる一群の手法が、それぞれどのように異なっているのか明らかにしておく。

3.1　数量化三類とコレスポンデンス分析

数量化三類とコレスポンデンス分析は、分析のもとになる分割表が異なっている。例えば、「はい」と「いいえ」で回答が得られる項目が三つあったとしよう。このとき、数量化三類が分析対象にする行列は、「はい」を 1、「いいえ」を 0 とした分割表.6 のようになっている。

これに対して、コレスポンデンス分析（および双対尺度法）が分析対象とする行列は、表.7 のようになっている。

要するに、この分析するもととなる行列の違いがあるだけで、それを特異値分解して行・列の両方に重みを与える点では同じ分析手法である。

表 6　数量化が分析する行列

被験者	項目 1	項目 2	項目 3	合計
S1	1	0	1	2
S2	0	0	1	1
S3	1	1	0	2
S4	1	0	1	2
S5	1	1	1	3
	4	2	4	

表 7　コレスポンデンス分析、および双対尺度法が分析する行列

被験者	項目 1(Yes)	項目 1(No)	項目 2(Yes)	項目 2(No)	項目 3(Yes)	項目 3(Yes)	合計
S1	1	0	0	1	1	0	3
S2	0	1	0	1	1	0	3
S3	1	0	1	0	0	1	3
S4	1	0	0	1	1	0	3
S5	1	0	1	0	1	0	3
	4	1	2	3	4	1	

3.2　コレスポンデンス分析と双対尺度法

コレスポンデンス分析と双対尺度法の違いは、もう少し説明がややこしい。これらの違いについて考える前に、双対の関係（双対性）について知る必要がある。双対の関係とは、式 (2) のようなものである。

$$\begin{cases} Y_i &= \dfrac{1}{\rho}\dfrac{\sum f_{ij}X_j}{f_{i.}} \\ X_j &= \dfrac{1}{\rho}\dfrac{\sum f_{ij}Y_i}{f_{.j}} \end{cases} \tag{2}$$

さて、この双対の関係が意味するのは、Y のベクトルが張る空間と X のベクトルが張る空間は等しくなく、一方が他方に射影されるためには図.1 のように特異値をもちいて変換しなければならない[*3]。

$$\rho Y_i = \frac{\sum f_{ij}X_j}{f_{i.}}, \rho X_j = \frac{\sum f_{ij}Y_i}{f_{.j}} \tag{3}$$

このように列および行の項目を、それぞれ行ベクトルの作る空間か、列ベクトルの作る空間か、どちらかに射影することを選択させるのが双対尺度法であるが、コレスポンデンス分析はこれをひとつの空間に射影する。ρY_i と ρX_j は、ノルムを整えてあるので分散が同じである。そこで、この ρY_i と ρ_j をひとつの空間に射影しようじゃないか、というのがコレスポンデンス分析のやり方なのである。これは、数学的に正しくないが、アウトプットが見やすいというメリットがある同じ空間にプロットできているわけではない点に注意が必要である。

[*3] 被験者 Y1 が、X1,X2,X3 という 3 つの項目を選択した場合、射影される重みは X が作る三角形の重心に来る。

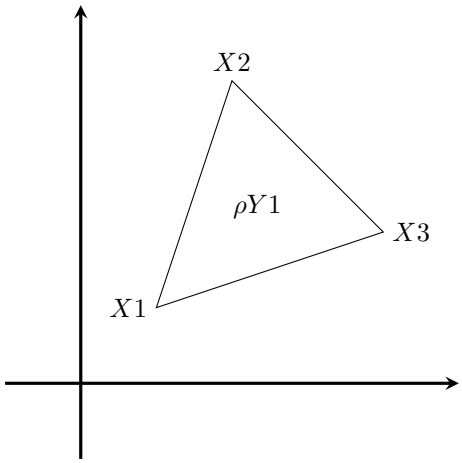

図 1　X の空間に対する Y の射影

4　引用文献

芝　祐順　1979　因子分析法　第二版　東京大学出版会

渡部　洋（編）　1988　心理・教育のための多変量解析法入門　基礎編　福村出版

繁桝算男・柳井晴夫・森　敏昭（編）　1999　Q＆Aで知る統計データ解析 DOs and DON'Ts　サイエンス社