研究成果報告
1

アンケート調査の自由記述文章を“定量分析”

（一財）和歌山社会経済研究所 主任研究員

長谷川　強

1. はじめに

アンケートにおける自由記述の設問に対する回答は、回答者の「生の声」が反映されており、非常に有用である。一方で、自由がゆえに内容が捉えにくく情報量も多いことから、大まかな傾向（がどれだけあるのかを含めて）をつかむにも、全ての記述に目を通す必要があり、多大な労力を要する。そのため、分析結果の質は分析者の集中力、記憶力などの能力に依存するところが大きく、分析者の主観が入り込む余地も大きい。逆に言えば、それらを克服し、自由記述を少しでも楽に、早く、高精度に、かつ客観的に分析できれば、アンケート活用の場面が広がると考える。

Amazon での本の検索時の「この商品を買っている人は、この本も買っています。」と紹介されるが、これは過去の購入データからユーザー購買行動の類似性、または商品間の共起性（共に起こる商品）を分析し、対象者個人の行動履歴を関連づけることで商品を提示するという手法を用いている。この共起性は、アンケート結果においても分析に有効な指標であり、事実の裏付け、新たな知見の獲得などにつなげられると考える。

近年では ICT 技術の発達により、文章をパソコンで処理する技術が手軽に利用できるようになった。そこで、それらの技術を自由記述等の文章の分析に用い、実用性について検証する。

2. 分析の概要

2.1. 分析の流れ

本稿の目的は、「アンケートにおける自由記述」のような複数の文書（ここでは「文書群」という。）の全体的な傾向を把握・説明するために、文書群として特徴的な単語及び単語間の関係（「共起度」、後述）を抽出することである。その流れを図 2.1.1 に示す。

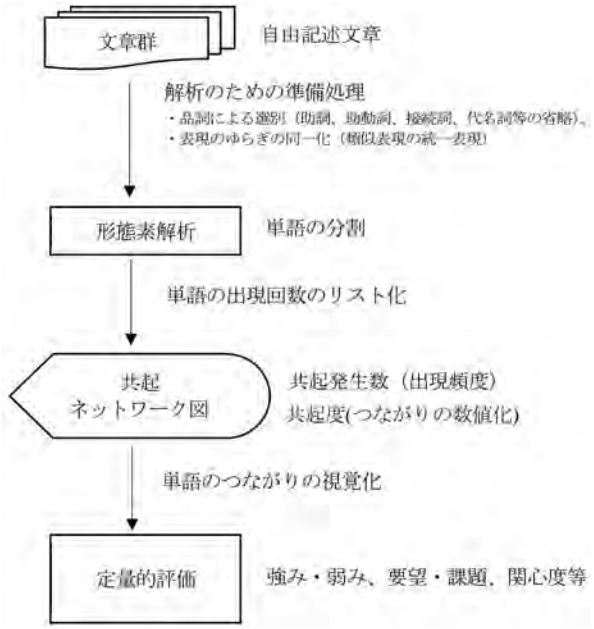

図 2.1.1　自由記述文章の分析の流れ

2.2. 形態素解析

2.2.1. 形態素解析とは

形態素解析とは、普段生活の中で一般的に使っている言葉､つまり｢自然言語｣を形態素(形言葉が意味を持つまとまりの単語の最小単位にまで分割）する技術のことである。

自然言語による文章を分析するには、まず文章を単語レベルに分割する必要がある。英語をはじめとする多くの言語は、単語の区切りにスペース（空白）が用いられるが、日本語の文章は基本的に句読点及び改行以外の明示的な単語の区切りがない。そこで、それらを文字の並びから単語の区切りを推測するという作業が必要になる。図 2.2.1 に短文を形態素解析する例を示す。

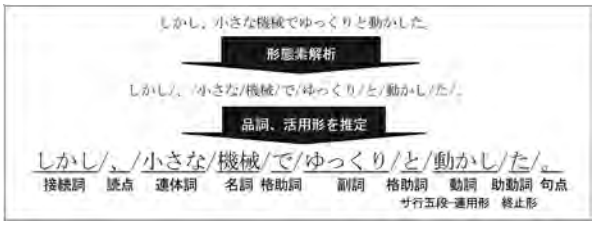

図 2.1.1　自由記述文章の分析の流れ

「しかし､小さな機械でゆっくりと動かした。」という文章を形態素解析すると、「しかし /、/ 小さな / 機械 / で / ゆっくり / と / 動かし / た /。」のように分割される。

こうして形態素解析により最小単位になった単語を、辞書（データベース）などの情報と照らし合わせ、それらの単語の品詞の種類、活用形の種類などを割り出していく。実際の活用事例は 3.3 で述べる。

2.2.2. 形態素解析の歴史

形態素解析という技術は 1970 年代前半、日本語ワードプロセッサに搭載される仮名漢字変換装置の開発において着目された 。その後 70 年代後半に同製品が実用化され､発売に至った。このころから 1980 年代にかけては、文法と制約に基づく「合理的」な解析であった。例えば、名詞→格助詞→動詞（などの活用品詞）語幹→助動詞という並び順が多い、などの文法的な規則と、文節数が最小になるような分割の仕方を採用する（文節数最小法）などの制約を組み合わせて解析を行っていた。

1990 年代はインターネットの商用利用開始とパーソナルコンピューターの性能向上により、情報量と情報処理能力が大幅に増大した。それに伴って、形態素解析に用いられるデータベース、すなわち文章と解析結果がセットになった見本「コーパス」の充実と、これを処理する解析エンジンの高度化及び解析アルゴリズムの多様化がもたらされた。このような情報量の多さを活かした「経験的」な解析は、これまでの「合理的」な解析の限界を克服するものとして主流となった。

2000 年代になると、形態素解析のノウハウが MeCab、Chasen などオープンソースプログラムとして公開され、研究者以外のユーザーが利用できるようになった。それに伴い、検索エンジンの最適化、テキストマイニング、他の自然言語処理といった応用が盛んになった。

2.3. 共起ネットワーク

共起ネットワークは、「どんな言葉が多く出てきていて、どの言葉とどの言葉が一緒に使われていたのか」を探るためのもので、個々のコメント間の単語の共通性を共起度と共起頻度で俯瞰されるネットワーク図により可視化し、特徴を読み解くことが可能となる。

以下にサンプル回答群を示す。

(1) 今度は温泉と食事に行くのを楽しみにしている。
(2) 温泉に入った後の酒は楽しい。
(3) 温泉には行きたいが、暇がない。
(4) 旅行したいなあ。

これらを形態素解析すると、

(1) 今度 / は / **温泉** / と / 食事 / に / 行く / の / を / **楽し** / み / に / し / て / いる /。
(2) **温泉** / に / 入っ / た / 後 / の / **酒** / は / **楽しい** /。
(3) **温泉** / に / は / 行き / たい / が /、/ 暇 / が / ない /。
(4) 旅行 / し / たい / なあ。

ここで、「温泉」、「楽しい」、「酒」という単語に注目し、共起発生数、共起度を求め、共起ネットワーク図を作成する。

2.3.1. 共起の発生

「共起の発生」とは、2 つの単語が 1 文に出現することである。「温泉」「楽しい」「酒」という単語に注目すると、サンプル（1）から（4）においては、「温泉」と「楽しい」が 2 回、「温泉」と「酒」、および「楽しい」と「酒」が 1 回である（表 2.3.1）。

表 2.3.1　形態素の発生数と共起発生数の関係

回答群	温泉	楽しい
(1)	1	1
(2)	1	1
(3)	1	0
(4)	0	0
共起発生数	2	

温泉	酒
1	0
1	1
1	0
0	0
1	

楽しい	酒
1	0
1	1
0	0
0	0
1	

2.3.2. 共起度

文章群中における単語間の共起度はさまざまな捉え方があるが、ここでは Cosine（コサイン）の方法を用いる。これは、2 つの単語の発生数をベクトルに見立て、その「角度の余弦」$\cos\theta$ を「共起度」と定義し、単語間の「接近度」を測定する方法である。

表 2.3.2 に示すように二つの単語ベクトルとして、

「温泉」$a = (1,1,1,0)$、「楽しい」$b = (1,1,0,0)$

の角度の余弦を算定すると、「温泉」と「楽しい」の共起度は 0.816 となる。このように、単語間の共起関係の強弱（数字の範囲は 0 から 1 で、数字が大きいほど強い）を評価される。

表 2.3.2　形態素の発生数と共起度の関係

回答群	温泉	楽しい
(1)	1	1
(2)	1	1
(3)	1	0
(4)	0	0
共起度	0.816	

温泉	酒
1	0
1	1
1	0
0	0
0.577	

楽しい	酒
1	0
1	1
0	0
0	0
0.707	

2.3.3. 共起ネットワーク図

共起ネットワーク図は、単語間のつながりを見える化したもので、出現回数を相対的に円の大きさ（node）示し、単語間は、共起度で強弱を表した線（edge）で結ばれている。共起度が大きい円どうしは、近い「距離」にあり、共通に出現していて共起関係の結果である。

今回の回答群の例を共起ネットワークとして書き起こしたものを図 2.3.1 に示す。ネットワーク図は、全体を定量的に俯瞰したり、一部に焦点を当てて読み取る（クラスター化）ことも有効で、大量の文書から特徴を客観的に見つけ出す可能性がある。

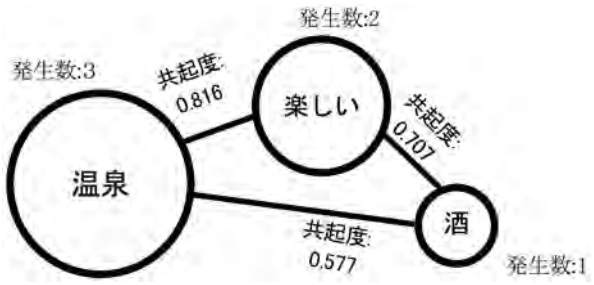

図 2.3.1　サンプル回答群の共起ネットワーク図

3. 文章解析の実践

3.1. アンケートの自由記述

当研究所では、和歌山県内の地方自治体から受託して、訪れた観光客へのアンケートを実施し、その際に自由記述欄を設けた。本稿では、これを実例として分析を試みる。

3.2. 解析のための準備処理

① 品詞による選別

助詞、助動詞、接続詞、代名詞等は分析結果から省略する。出現数は多いものの、文章の分析に有効な情報ではないためである。

例）観光アンケートの自由記述を形態素解析し、すべての品詞を表示した結果は 表 3.2.1 のとおりとなった。文章（群）の意味を見出すには不適当なデータである。

表 3.2.1 すべての品詞を含めた形態素　頻出度上位 10 語

抽出語	品詞	出現回数
た	助動詞	322
ます	助動詞	270
の	助詞	265
て	接続助詞	260
が	助詞	255
に	助詞	251
は	助詞	189
を	助詞	184
する	動詞 B	176
です	助動詞	176

② 表現の揺らぎ

例えば今回の事例では、「もも」「モモ」「桃」は同じものを示すが、実際の文章では回答者によって表現が異なる。また、「組み立て」「組立て」「組立」のような送り仮名のバリエーションも存在する。これらを同一のものとして処理した。

例）「みやげ」「おみやげ」「土産」「土産物」、「卵」「玉子」「たまご」「タマゴ」、「おいしい」「美味しい」

③ コーディング

コーディングとは、「特定の記述がデータ中にあれば、そのデータを 1 つのカテゴリーに分類すること」である 。今回の事例では、「宣伝」「アピール」「PR」という単語がほぼ同義で用いられているため、同一単語として取り扱った。

例）「フルーツ」「果物」、「おいしい」「旨い」

3.3. 形態素解析

194 件ある自由記述の文章を形態素解析によって単語に分解し、その単語を出現数の上位 50 語のリストを表 3.3.1 に示す。

表 3.3.1　形態素解析結果　頻出語上位 50 語

rank	抽出語	品詞	出現回数
1	する	動詞 B	176
2	思う	動詞	82
3	良い	形容詞	73
4	ある	動詞 B	72
5	行く	動詞	51
6	かつらぎ	タグ	41
6	来る	動詞	41
8	山	名詞 C	36
9	高野	タグ	35
9	町	名詞 C	35
11	なる	動詞 B	34
12	もっと	タグ	31
12	橋本	タグ	31
14	とても	副詞 B	30
15	観光	サ変名詞	29
15	宣伝	サ変名詞	29
17	できる	動詞 B	28
17	果物	名詞	28
19	おいしい	形容詞 B	27
19	ない	形容詞 B	27
21	訪れる	動詞	26
22	温泉	名詞	25
22	道の駅	タグ	25

24	今回	副詞可能	22
25	欲しい	形容詞	21
26	知る	動詞	19
27	宿	名詞 C	18
28	市	名詞 C	17
28	場所	名詞	17
28	食べる	動詞	17
28	人	名詞 C	17
28	大阪	地名	17
33	いる	動詞 B	16
33	柿	名詞 C	16
33	買う	動詞	16
33	和歌山	地名	16
37	嬉しい	形容詞	15
37	残念	形容動詞	15
37	宿泊	サ変名詞	15
37	神社	名詞	15
37	大変	形容動詞	15
37	利用	サ変名詞	15
43	初めて	副詞	14
44	わかる	動詞 B	13
44	見る	動詞	13
44	少ない	形容詞	13
44	情報	名詞	13
44	土産	名詞	13
49	たくさん	副詞可能	12
49	食事	サ変名詞	12
49	多い	形容詞	12
49	丹生都比売	タグ	12
49	訪問	サ変名詞	12
49	良い	形容詞（非自立）	12

3.4. 形態素の共起と共起ネットワーク図

文章を形態素解析した結果を元に、形態素ごとに出現した数を一覧表にしたものを表 3.4.1 に示す。この表を用いて、共起する形態素を抽出する。例えば、表 3.4.1 における文章 L11 においては『温泉』と『美味しい』がそれぞれ 1 語ずつ存在するので、「文章 L11 では『温泉』と『美味しい』が共起している」としてカウントする。

これら文章群中における共起の発生確率（Cosine 値）を算定した結果を図 3.4.1 に示す。

表 3.4.1　文章を形態素解析した結果（抜粋）

	単語	温泉	美味しい	訪れる
	品詞	名詞 - 普通名詞 - 一般	形容詞 - 一般	動詞 - 一般
	総数	20	17	16
ID	文章			
L 11	ドライブによく訪れます。路上販売の柿や黒枝豆はすごくおいしく利用させてもらってます。すごく落ち着く場所です。地域の人（住人）も親切でいい人が多いです。ゆの里温泉もすごく大好きです。水のサービスがうれしい。オムレツは食べそこねて残念です。	1	1	1
L 13	京奈和がつながったので、日高地方から行きやすくなりました。美味しいもの（卵、フルーツ、野菜）さがしに、何度でも行きたいと思っています。次は天野へ…の予定です。	0	1	0
L 18	お米、果物、農作物等、身体に良いおいしいものをこれからも沢山つくって下さい。また来たいと思います。ありがとうございました。	0	1	0
L 19	丹生都比売神社に初めて訪れた時、和歌山県にこんな立派な所、「天空の城」があったのかと感動しました。それから 3 度行かせてもらいました。また行きたいと思っています。知り合いにも紹介していきます。	0	0	1

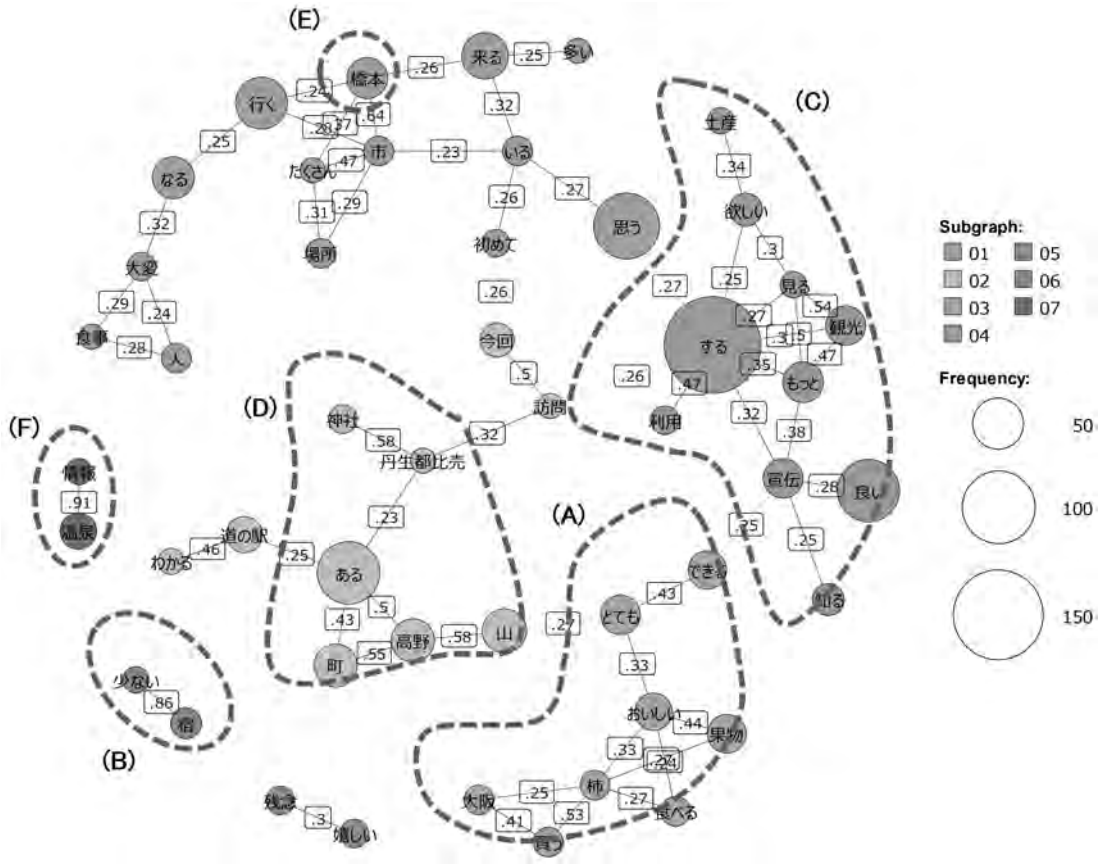

図 3.4.1　自由記述文章　共起ネットワーク図

ネットワーク図のエッジ上に示されている数値は、発生確率（Cosine 値）である。今回の事例では、Cosine 値が 0.25 以上、単語の発生数が 8 以上のものを表示した。Cosine 値または単語の発生数が小さいものまで表示すると見にくくなるため、表示する対象を絞り込んで分析、解釈が可能なレベルにする必要がある。

3.5. 分析結果

図 3.4.1 を用いることによって、当該エリアの評価、強みと要望・課題、弱みを抽出し、出現数（図上では単語を囲む○の大きさ）と併せて訪問者の関心度を定量的、客観的に評価することができる。

これらの「つながり」すなわちクラスターにおいて、この文章（群）に特徴的なもの、及びその解釈として想定できるものを以下に示す。

○強み…図 3.4.1 中（A）

- 「柿」をはじめ「果物」が「おいしい」
- 「大阪」から「柿」を「買い（←買う）」に来た

○弱み…同（B）

- 魅力的な「宿泊施設」の選択肢が「少ない」

○要望・課題…同（C）

- 「もっと」「宣伝」「した（←する）」方が「良い」
- 「土産」になりそうないいものが「欲しい」

○観光地のつながり…同（D）（E）

- 「丹生都比売」「神社」と「高野」「山」のつながり（D）は見えるが、一方でそれらと橋本とのつながり（E）は見えにくい。

○訪問者の関心度…同（F）

- 「温泉」の「情報」がもっと欲しい

4. まとめと今後の課題

以上、文章を形態素単位に分割し、それを再構成することで分析、定量化を行った。その中で明らかになった点を以下に述べる。

- 単語を抽出、定量化することにより、回答内容の把握を高速かつ明確にすることができた。
- 単語の結びつきを視覚化することにより、回答者意見の一定の傾向を見出すことができた。

なお、再構成の方法によりさまざまな分析が可能となるので、今後取り組んでいきたい。以下にその例を示す。

- アソシエーション分析…2つ以上の形態素（単語）で構成される組合せを抽出し、条件確率などの指標を用いて重要なキーワードを探る。
- 数量化Ⅲ類…文章における形態素（単語）の含まれ方を数値・ベクトル化し、数値の近いものの類型化（クラスター化）や、類似する文章が近くになるように並べ替える序列化などを行う。
- 数量化Ⅱ類…文章における形態素（単語）の含まれ方を数値化し、例えば満足度（満足した・しない）などの離散的な回答との関連性を見る。
- 数量化Ⅰ類…文章における形態素（単語）の含まれ方を数値化し、例えば消費額など連続的な数値の回答との関連性を見る。
- 機械学習（AI）への応用

これら文字情報分析を活用し、今後ともクライアントに提供する情報の有効・有用性、迅速性など品質の向上に努めたい。

以　上

1　海野裕也、(2011年10月11日)、形態素解析の過去・現在・未来、slideshare: https://www.slideshare.net/pfi/ss-9805912
2　松本裕治、(2008年)、自然言語処理における制約と選好、コンピュータ ソフトウェア 25巻3号
3　天野真家, 森健一、(2002年11月)、漢字・日本語処理技術の発展：日本語ワードプロセッサの誕生とその歴史、IPSJ Magazine Vol.43 No.11, 1217
4　樋口耕一、(2014年)、社会調査のための計量テキスト分析、ナカニシヤ出版