キーワード分析による言語情報の定量化方法

山本千雅子　博士（工学）
グラデュウス・マルチリンガルサービス株式会社
〒060-0807 札幌市北区北7条西2丁目 37山京ビル302
TEL 011-717-8770 FAX 011-717-8772
ＵＲＬ：http://www.gradus.net/

キーワード分析による言語情報の定量化方法

1. キーワード分析の定義

対象とする「言語データ」に含まれる頻度の高い「文意」を、その文意を表すときに高頻度で使われるキーワードを用いて検索し、その頻出回数を用いて重要度の定量化を行うことと定義する。キーワードとは、文章の単語リストをより無色で一般的な語彙リストと比較した際に、統計的に特に有意に頻度が多い単語をいう。

2. 行政サービスに用いる目的

1）行政サービスに対し、「いつ、だれが、どこで、どんな」要望やニーズを持っているかについて、既存の市民の声情報を活用する。
2）同じ意見を持つ回答者数を使い、要望・ニーズの大きさを定量化する。

3. 特徴

1）アンケート作成側が気づいていなかった項目が判明し、定量化される。
2）項目を変えた集計ができる。

4. アンケート実施の留意点

1）シーン展開を行えるよう背景情報（回答者の属性、いつ・どこのことか）も入手する
2）質問は、回答者が具体的なシーンや理由を思い浮かべて書けるようにつくる。

例： 1） 質問：「なぜXXXXXXですか？」
回答：「XXXXX」だからです。
2） 質問：「XXXXについてXXXXなことは何ですか？」
回答：「XXXXX」です。

5. アンケート回答データの分析

以下は、エクセルを用いた分析方法である。

5．1 データ入力時の注意点

a） 1人の回答者の回答は、すべて1行に入力する。
b） 各回答者に固有の番号を割りあて、入力する。
c） 数字を入力するときには、単位は別のセルに入力する。
d） 一番上の行に、設問の番号等を入力する。
この行は、後でキーワードの入力にも用いる。
e） 入力時の漢字などの誤変換に注意する。

連絡先：グラデュウス・マルチリンガルサービス株式会社 chigako.y@gradus.net

5．2　キーワード分析

（1）準備作業

1）文の終わりの特定。

文の終わりを明確にする。文末に「。」が無いときには、適切なところに「。」を挿入する。

2）重文と複文の単文化。

日本語の文には、単文、重文、複文がある。分析対象の文が重文や複文のときには、「単文＋単文」の文に書き換える。表1はその例である。

表1　重文と複文の単文化

単文	「道幅が(主語)－狭い(述語)」
重文	「道幅が(主語1)－狭く(述語1)、路面が(主語2)－凍結(述部2)」 ↓ 「道幅が(主語)－狭い(述語)」＋「路面が(主語)－凍結(述部)」
複文	「道幅が(主語)狭く(述語)(全体で主部)、歩きにくい(述部)」 ↓ 「道幅が(主語)－狭い(述語)」＋「歩行者が(主語を加える)－歩きにくい(述部)」

（2）キーワード分析

図1は、キーワード分析のフローである。以下は、その流れを、冬期道路状態の問題点に関する記述式アンケートの分析を事例に説明する。

1）頻度の高いキーワードを抽出する。

頻度の高い「文意」に使われる語彙をキーワードとして、品詞にかかわらず抽出する。

2）キーワードで代表させる「文意」を決める。

単独のキーワード、あるいはキーワードの組み合わせに代表させる「文意」を一意に定める。表2は、冬期道路状態の問題点を記述した回答から抽出された「キーワード」と「文意」の例で、表3は、「キーワードの組み合わせ」と「文意」の例である。

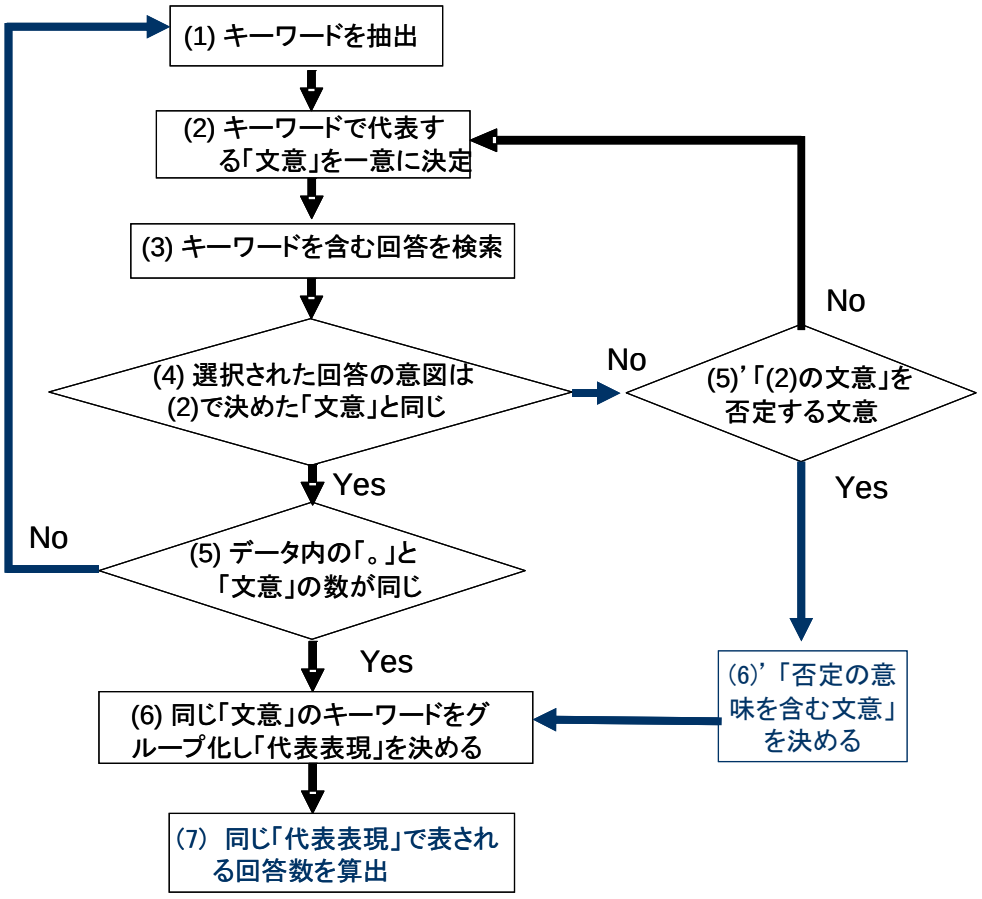

図1　キーワード分析のフロー

表2　抽出されたキーワードと文意

文　意	キーワード				
幅員不足	車線	狭	せま(い)**	幅	雪山
滑りやすい	すべ	滑	氷	凍	アイス
	スリップ	ツルツル	つるつる	圧雪	
溶けた雪	融	溶	とけ(る)**	シャーベット	
路面の雪	雪	除雪*	雪山*	雪道*	排雪*
不陸	悪	がたがた	ガタガタ		
排雪作業	作業				
路上駐車	駐車				
交通量増加	都心	ラッシュ			
冬期道路	雪道	冬道			

*「路上の雪」については、「雪」に加えて*のキーワードが回答に含まれるときには、別の文意となる。「雪」というキーワードは文脈によって「堆雪」、「路上の雪」、「溶けた雪」などがあるので、他のキーワードとの組み合わせを用い、全体の文脈から判断する。
**(　)は活用形により、べつの文字となることもある

表3　キーワードの組み合わせ」と「文意」

文　意	キーワードの組み合わせ
交差点が滑りやすい	「交差点」＋「滑／すべ（る）／凍結」
滑りやすい路面のため渋滞	「渋滞」＋「滑／すべ（る）／凍結」
歩道が狭い	「歩道」＋「狭／せま（い）」
路面が凸凹・轍	「路面（省略されることもあり）」＋「轍／でこぼこ／がたがた／そろばん」

3）キーワードを含む回答を検索する。

エクセルのサーチ関数(「SEARCH」か「SEARCHB)」を使い、キーワードを含む回答を検索する。図2は、サーチ関数による検索結果の例である。サーチ関数は、検索した文字列の位置を数字で返すので、キーワードが含まれる文には「数字」が、含まれない文のときは「#VALUE!」が表示される。

回　答	車線	狭	せま	すべ	凍	アイス	スリップ	圧雪
	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
・車の渋滞。・雪による車線の減少。・道路の凍結。	12	#VALUE!	#VALUE!	#VALUE!	22	#VALUE!	#VALUE!	#VALUE!
・バス遅延(定刻発車より10分) ・雪のため道幅が狭く、右折、直進の2車線道路が1車線となっていた	36	26	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
・冬型自然渋滞・のろのろ運転	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
・冬道の渋滞。車線が狭くなっている(駐車違反)。・交差点がすべっている人が多い。	8	11	#VALUE!	30	#VALUE!	#VALUE!	#VALUE!	#VALUE!
・屋根(車)の雪おろしとアイドリングに時間がかかった ・駐車場が混み合っていた	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
25日はいつも混む(交通量が多い)。雪のため、道幅が狭くなっている。	#VALUE!	27	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!

図2　エクセルのサーチ関数(SEARCH)を用いたキーワードの検索

4）上記2）で決めたキーワードの「文意」と同じ「文意」を持つ回答を特定する。

キーワードを含むことが検索で分かった「アンケート回答」の意図するところと、2)で決めたキーワードが代表する「文意」が等しいかどうかひとつひとつ検討する。

「文意」が異なるときには、新しい「文意」を決め直す(図1の(5)')。「アンケート回答」がキーワードを代表する「文意」を否定する意味のときは、新しく「否定の意味を含めた文意」を定める(図1の(6)')。

5）各データ中の「。」の数量を数える。そのデータに含まれる「文意」の数と一致していないときは、そのデータにはまだ抽出されていない「文意」があるので、キーワードを抽出し、1）からのプロセスを繰り返す。

6）同じ「文意」のキーワードをグループ化し、「代表表現」を決める。

図3は、徒歩による道路利用者が指摘した冬期道路の問題点のキーワード分析の例である。この例で、キーワードは「凍結」、「すべ(り)」、「滑」、「ツルツル」、「氷」で、文意は「滑りやすい歩道」なので、「滑りやすい」を代表表現とした。

AA	AO	AP	AQ	AR	AS	AT
回答	凍結	すべ	滑	ツル	氷	滑りやすい
歩道が確保されておらず、また車道の状態が悪かった為。	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	
雪道のため	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	
でこぼこで歩きにくい。	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	
除雪作業中だったので。	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	
雪、氷のため歩きづらかった。	#VALUE!	#VALUE!	#VALUE!	#VALUE!	3	1
道路の状況が悪かった(雪解けで、なかなか狭い)	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	
雪、氷のため歩きづらかった。	#VALUE!	#VALUE!	#VALUE!	#VALUE!	3	1
雪がとけて歩きづらかった。	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	
凍結のため	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!	1
凍結のため	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!	1
歩道がすべりやすく歩くのに時間がかかった。	#VALUE!	#VALUE!	4	#VALUE!	#VALUE!	1
雪道が滑る為。雪が融けてぐちょぐちょしていた為、歩きにくかった。	#VALUE!	4	#VALUE!	#VALUE!	#VALUE!	1
	7	4	3	1	2	17

図3　冬期道路の問題点(歩行者)

7)同じ「代表表現」で表される回答数を合計する。

図3の一番下の行は、それぞれのキーワードを含む回答の数である。これは、SEARCH 関数が数値を返した箇所数をカウント関数(「COUNT」)で得たものである。代表表現の「滑りやすい」の列は、その左にあるキーワードに対して SEARCH 関数が数値を返しているときに「1」を記入する。同じ文意のキーワードがひとつの「回答」に複数個含まれているときにも「1」と記入する。最下行は列の合計である。この合計の「16」が、全回答(データ数 40 個)の中で「滑りやすい歩道」を問題であると回答したデータ数である。

5．3　重要度

各「代表表現」で表される回答数の全回答数に対する割合(%)が、「代表言語」で表わされる意見の割合である。 図3の例では、全回答数が42なので、(16÷40)×100＝38.1(%)となり、38.1%を重要度とする。表4は、同じ事例のキーワード分析で得た全問題点の重要度である。

表4　キーワード分析で得た重要度

（事例：歩行者にとっての冬期道路の問題点）

問題点	頻度	重要度
路面の積雪	12	30.0
狭い歩行者空間	3	7.5
滑りやすい路面	17	42.5
スラッシュ	7	17.5
歩道が除雪されていない	1	2.5
でこぼこ	1	2.5
水たまり	1	2.5
除雪作業	1	2.5
	43*	

*一つのデータで問題点を複数指摘しているものがあるので合計は、データ総数と一致しない。

4．注意事項

1）あくまで、同じ「文意」の回答数を算出するための「鍵」としてキーワードを用いる。キーワードの言葉としての意味ではなく、回答された文章中の意味と「文意」をリンクさせる。

2）頻度が高い文意が複数のキーワードの組み合わせで表されるときには、キーワードの組み合わせに対して、上記のプロセスを適用する。

3）苦情電話の内容など、アンケート以外のデータを用いるときには、ひとつのデータの「単位」に注意する。単純にキーワードの頻度で重要度を求めると、ひとつのデータに同じ代表表現のキーワードが複数含まれていると、ダブルカウントやトリプルカウントされる可能性が高い。

（以上）

連絡先：グラデュウス・マルチリンガルサービス株式会社　chigako.y@gradus.net