ニーズに関する多量の言語データの分析支援

A method of supporting the process of analyzing large size userneeds textdata.

経営システム工学専攻　篠崎 亮

１．研究目的

新製品開発においては、消費者のニーズを正確かつ迅速に把握し、そこから新製品を企画していくことが重要である。

ニーズの把握は、ニーズの構造を明らかにする段階と明らかにした構造に基づいて各項目の重視度を調査する２段階に分けられる。

ニーズの構造を明らかにする手法の１つに「要求品質展開表」がある。これは集められたニーズを読み解き、表にまとめる手法である。この段階で扱われるデータはアンケート調査や行動観察などで集められた言語データである。この処理については形態素解析（文をそれ以上細かくできない意味のある単語（形態素）に分解するもの）やニューラルネットワークなど解析に使えそうな手法もいくつかあるが、そのままでは数値化ができないものが多く、機械的に解析するには困難な面がある。現在では解析者が１つずつ読んで解析している所も珍しくない。

言語データを人手で読み解くには時間も手間もかかり、解析できる数に限りがある。反面、今日ではインターネットの普及によって多すぎる程のデータが集まってくるため、集まったデータの中の一部しか解析出来ない状況に陥っている。

本研究では、要求品質展開表作成の際の言語データの処理に着目し、人手による手間を、形態素解析、ニューラルネットワークなどを用いて支援し、効率化する方法を考えることを目的とする。

２．言語データの人手による分析とその支援方法

２．１　人手による要求品質展開表の作成フロー

人手で分析を行う場合、まず、顧客の声（原始データ）を「要求品質」に変換するという作業が行われる。要求品質とは、原始データを簡潔な品質表現に直したものである。これをその類似性に着目して展開表の形にまとめたものが要求品質展開表である。要求品質展開表作成の流れについては様々な文献で紹介されているが、それらをまとめてフローチャートにしたものを図１に示す。

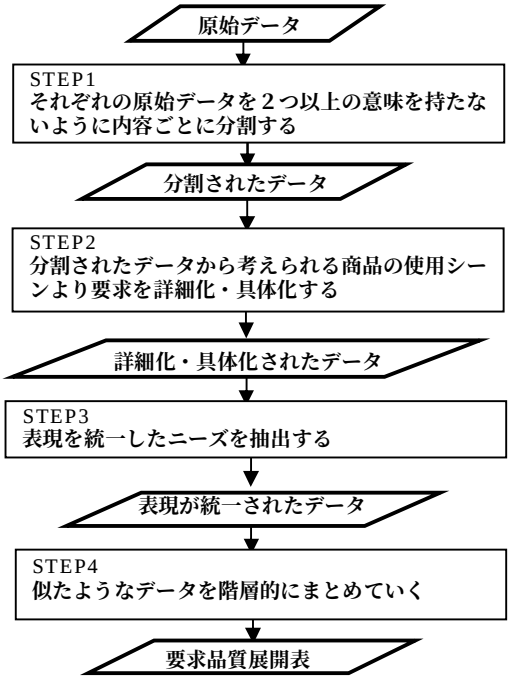

図１　品質展開表作成のフローチャート

２．２　支援方法の概要

図１を、計算機を用いて支援する方法には様々な形が考えられるが、本研究では表１に示す支援方法を考えた。

２．３　STEP１の支援方法

STEP１の支援方法としては、データの区切り文字「＊」を決め、区切り文字のある部分を Excel の機能を用いて区切っていくことを考えた。

表１　図１を計算機を用いて支援する方法とネックとなる部分

STEP	支援方法	支援の際ネックとなる部分
1	区切り文字を定め、内容の分岐点として文を分ける	・どのような区切り文字を使用するか ・定めた区切り文字で分割出来なかったところ、間違って分割してしまったところをどのようにするか
2	STEP1で出力されたデータと予め用意した使用シーンをニューラルネットワークを用いて付き合わせ、それぞれのデータを詳細化できる使用シーンを判別する。	・使用シーンの判別の際に用いる教師信号の作成 ・どのようにニューラルネットワークに与えるパラメータを決めるか
3		
4	STEP2・3 で出力されたデータでクラスター分析を均等な重みでスタートし、途中で止めて得られたクラスターの郡内分散に応じて重みを変えて先を続ける。	・使用する変数の選び方 ・どこで重みを変えるか

２．４　STEP２・３の支援方法

STEP２・３では形態素解析とニューラルネットワーク(階層型)を用いる。支援方法のアルゴリズムを以下に記す。

１．予め調査しておいた利用シーンのアンケートデータから典型的シーン（１、２、・・l）を作る。

２．コンビニエンスストアに対するニーズ調査[2]の元データ（$X_1, X_2, \ldots, X_n$）と典型的シーン（１、２、・・l）を付き合わせ、当てはまる場合は１、当てはまらない場合は０とする。また、得られたデータをYjiとする（$1 \le i \le n, 1 \le j \le l$）。

表２　シーンのつき合わせ例

	1	2	・・・	l
X_1	0	1	・・・	0
X_2	0	0	・・・	1
：	：	：	：	：
X_n	0	0	・・・	0

３．元データ X_i を形態素解析して、結果V_{ki} を得る（$1 \le k \le m$）。

表３　形態素解析結果例

	V_{1i}	V_{2i}	・・・	V_{mi}
X_1	0	1	・・・	0
X_2	0	0	・・・	1
：	：	：	：	：
X_n	0	0	・・・	0

４．$V_{k\bullet}$（$V_{k1}, V_{k2}, \ldots, V_{mn}$）に$Y_{j\bullet}$を加えたデータ$W_{pi}$を作る（$1 \le p \le q$）。

表４　シーンの付加例

	V_{1i}	V_{2i}	・・・	V_{mi}	$Y_{j\bullet}$
X_1	0	1	・・・	0	Y_{j1}
X_2	0	0	・・・	1	Y_{j2}
：	：	：	：	：	：
X_n	0	0	・・・	0	Y_{jn}

５．　W_{pi}を教師信号として、ニューラルネットワークのモデルを作る。

Input　：　$W_{1i}, W_{2i}, \ldots, W_{mi} (i = 1,2,\ldots,n)$

Output：　Y_{ji}

６．まだ使用していない元データ X_i（$n+1 \le i \le N$）を形態素解析し、結果V'_{ki} を得る。

７．５．のモデルに $V'_{\bullet i}$ を入力し、Y'_{ji} を求める。

８．４．～７．を　$j = 1, \ldots, l$ について繰り返す。

９．$Y'_{j1} = Y'_{j2} = \ldots = Y'_{jn} = 0$　となるデータ をはじき出す。（はじき出されたものは人手によって解析する。）

２．５　STEP４の支援方法

STEP４ではクラスター分析を用いる。ただし、均等な重みでスタートし、途中で止めて得られたクラスターの郡内分散に応じて重みを変えて先を続けるという２段階での適用を考えた。これは人手による方法（ＫＪ法）を真似たものである。

１．コンビニエンスストアに対するニーズ調査[2]の元データ X_i を形態素解析して、結果V_{ki} を得る（$1 \le i \le n, 1 \le k \le m$）

表５　形態素解析結果例

	V_{1i}	V_{2i}	・・・	V_{mi}
X_1	0	1	・・・	0
X_2	0	0	・・・	1
：	：	：	：	：
X_n	0	0	・・・	0

２．V_{ki} を用いて階層的クラスター分析を行う。この際、ウォード法（各ステップで形成される任意の２つのクラスターのグループ内平方和を最小にする距離定義）、ユークリッド距離を用いて解析を行う。

ユークリッド距離の定義式：

$$d(X_a, X_b) = \sqrt{\sum_{k=1}^{m} (V_{ka} - V_{kb})^2} \quad \cdots (1)$$

３．２．で得られた結果を同一クラスターごとにまとめる。

４．クラスター l （$l = 1,2,\ldots,j$）の変数 k の自由度 ϕ_{kl}、偏差平方和 S_{kl} を求める。

５．変数 k について分散のプーリングを行う。

$$V_k = \frac{\sum_{l=1}^{j} S_{kl}}{\sum_{l=1}^{j} \phi_{kl}} \quad \cdots\cdots (2)$$

６．５．で得た分散V_k から重みω_k を求める。

$$\omega_k = \sqrt{\frac{1}{V_k}} \quad \cdots\cdots (3)$$

７．データに重みω_k をかけ、結果V'_{ki} を得る。

$$V'_{ki} = V_{ki} \times \omega_k \quad \cdots\cdots (4)$$

８．V'_{ki} を用いて階層的クラスター分析を行う。この際、重み付きユークリッド距離を用いて解析を行う。

重み付きユークリッド距離の定義式：

$$d(X_a, X_b) = \sqrt{\sum_{k=1}^{m} \omega_k (V'_{ka} - V'_{kb})^2} \quad \cdots (5)$$

３．STEP１におけるパラメータ最適条件の探索

２．で述べた支援方法におけるパラメータの最適条件について、コンビニエンスストアに対するニーズを調査したデータ[2]を用いて検討を行った。この調査ではインターネットを用いてデータを集めている。

３．１　STEP１のパラメータと条件の候補

STEP1 で定めるべきパラメータは「何を区切り文字とするか」である。ここでは候補として表６に示すものを考えた。

表６　STEP１のパラメータについて候補とした条件

パラメータ	候補とした条件	
使用する区切り文字	1	句点とピリオド
	2	句点とピリオド＋“読点”
	3	句点とピリオド＋“・”
	4	句点とピリオド＋“や”＋“と”
	5	句点とピリオド＋“？”＋“，”

３．２　STEP１の最適条件の評価方法

最適なパラメータの条件を決めるための評価方法としては以下の２つを用いた。

（１）分割することが適切と考えられる箇所のうち、解析によって正しく分割出来た箇所の割合

（２）分割することが適切でないところで分割してしまった箇所の数

３．３　STEP１の評価結果

人が原始データを読んで判断した場合、分割することが適切と考えられるデータの総数は１３２件である。また、データ中に分割したい箇所が複数あるものもあり、それらの総数は１７２箇所である。

区切り文字を句点とピリオドにした場合の結果を表７に記す。

表７　区切り文字「句点とピリオド」の場合の解析結果

分割したくないところで分割した箇所			5	
分割したいデータ	分割できた箇所	94	172	合計
	分割出来なかった箇所	78		

これらの検討より、以下のことが分かった。

（１）分割したくないところで分割してしまった箇所は５箇所であった。この５箇所では分割することで意味が分からなくなるということはないので分割しても大きな問題はない。

（２）分割したい所の約 50%がうまく分割できた。

（３）区切り文字に読点（、）や「・」「？」などの文字を加えると、分割したいところについて分割できたものの割合は増加するが、分割したくないものが分割される数が急激に増える。

以上の結果を総合し、ここでは「句点とピリオドを区切り文字に用いる」という条件を最適条件とした。

４．STEP2・3 におけるパラメータ最適条件の探索

４．１　STEP2・3 のパラメータと条件の候補

STEP２・３で定めるべきパラメータは、「使用する変数」と「中間層のユニット数」の２つである。各パラメータの候補となる条件については、表９に示すものを考えた。

表９　STEP2・3 のパラメータについて候補とした条件

パラメータ	候補とした条件	
使用する変数	1	名詞＋動詞（以下、変数 a）
	2	名詞＋動詞＋形容詞（以下、変数 b）
ユニット数	1	４個
	2	８個

４．２　STEP２・３の最適条件の評価方法

最適なパラメータの条件を決めるための評価方法としては、「ニューラルネットワークによって得られた結果のうち、人手によって得られた結果と同様の結果となったデータの割合（正答率）」を用いた。この正答率を「当てはまらない」「当てはまる」「全体」の３つで比較した。一例を表１０に記す。

表１０　評価の際に用いるクロス表例

		クラスター分類					
		1	・・・	7	合計	分割	比率
人手グループ分類	A	2	・・・	5	42	7	19.05
	：	：	・・・	：	：	：	：
	G		・・・	2	4	3	50.00
	合計	12	・・・	21	141		
	分割	5	・・・	5		35	平均比率
	比率	50.00	・・・	33.33			39.61

４．３　STEP２・３の評価結果

シーンについては、消費者から別途集めたデータをもとに、コンビニエンスストアの典型的な利用シーンとして、「他の店が閉まっているとき」「立ち読みするとき」などの 18 シーンを作成した。結果の一部を図２に示す。

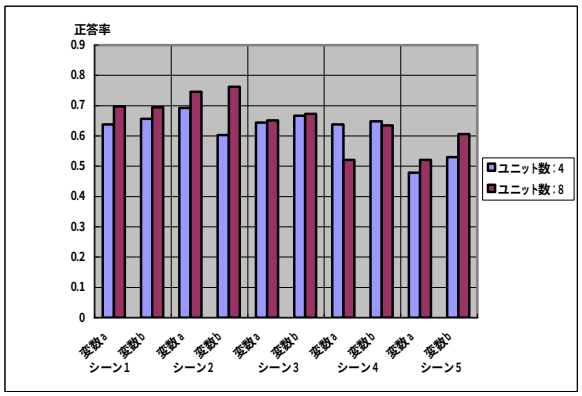

図２　候補となる条件による比較

これより、以下のことが分かった。

（１）全体的に変数ｂ（名詞・動詞・形容詞ごとの数）を用いた方が正答率が高い。

（２）全体的に「ユニット数：８」の方が正答率が高い。この傾向は変数ｂの場合に顕著である。

（３）シーンによっては、最適な条件が異なる場合がある。

以上の結果を総合し、ここでは「変数ｂ、ユニット数８」を最適条件とした。

５．STEP４におけるパラメータ最適条件の探索

５．１　STEP４のパラメータと条件の候補

STEP４で定めるべきパラメータは「使用する変数」と「重みを決める段階」、「シーンの有無」である。「重みを決める段階」とは２段階でクラスター分析を提供する際、１段階目をどこで止めるかということである。各パラメータの候補となる条件としては表１１に示すものを考えた。

表１１　STEP４のパラメータについて候補とした条件

パラメータ	候補とした条件	
使用する変数	1	名詞＋動詞（以下、変数 a）
	2	名詞＋動詞＋形容詞（以下、変数 b）
重みを決める段階	1	（データ総数の）８０％
	2	（データ総数の）５０％
	3	（データ総数の）２０％
	4	重み付けなし
シーンの有無	1	シーンを付け加えない
	2	シーンを付け加える

５．２　STEP４の最適条件の評価方法

最適なパラメータの条件を決めるための評価方法としては、人手による分類とクラスター分析による分類の二元表を作成した上で、以下の２つの指標を求めた。（１）は１に近ければ良く、（２）は１００％に近ければ良い。

（１）分類の数の一致度：人手による分類の数÷二元表の０でないセルの数

（２）最大セルの比率：行・列ごとの最大セルの値 ÷ 当該の行・列の合計

５．３　STEP４の評価結果

結果を表１２に記す。

表１２　一致度と比率

			80%	50%	20%	なし
一致度	シーン無し	変数a	0.206	0.2	0.206	0.2
		変数b	0.2	0.212	0.194	0.194
	シーン有り	変数a	0.206	0.212	0.2	0.219
		変数b	0.212	0.212	0.212	0.206
比率	シーン無し	変数a	39.89	39.61	40.37	37.38
		変数b	38.42	38.24	37.61	39.77
	シーン有り	変数a	44.82	47.3	45.18	45.41
		変数b	45.6	47.06	43.05	44.51

これより、以下のことが分かった。

（１）シーン有りと無しを比較すると、シーン有りの方がよい。

（２）分類数の一致度を見ると、シーン有りの場合についてはやや変数ｂの方が良いが、大きな差はない。

（３）最大セルの比率を見ると、シーン有りの場合については変数 a の方が良い。

（４）重みを決める段階をみると、分類数の一致度、最大セルの比率とも、５０％が良い。

以上の結果を総合し、ここでは「シーン有り、重みを決めるのは５０％を併合した段階」という条件を最適条件とした。変数に関しては、最適条件のシーンありの５０％から「変数 a」を最適条件とした。

６．結論と今後の課題

本研究で考えた多量の言語データの分析支援における各 STEP のパラメータの最適条件を表１３に示す。

表１３　各 STEP の最適条件

STEP	最適条件		精度
STEP１	1	区切り点を「句点」と「ピリオド」	精度 ５５％
STEP２・３	1	変数は「名詞＋動詞＋形容詞」	正答率 ６０％
	2	ユニット数は８個	
STEP４	1	変数は「名詞＋動詞」	一致度 ２１％
	2	重みを決める段階を５０％	
	3	シーンを加える	比率 ４７％

各 STEP の精度はまだ低く、改良の余地がある。

STEP1 については、記号だけで分割するには限界であるので、単語の内容や品詞を区別し、これを区切り点にすることで精度が上がると考えられる。また現在では単語同士の係り方を解析できるソフトも開発されており、これを取り入れることで改良につながると考えられる。

また、STEP2・3 および STEP4 については、使用する変数を品詞によって設定するのではなく、単語の意味を見て解析者自らが厳選することで改良につながると考えられる。

さらに、アンケートのとり方を考えることでデータの質（誤字・脱字、文法間違い、抽象的な記述など）をさらに上げることも改良につながると考えられる。

参考文献

[1]大藤正,小野道照 他：品質展開表（１）

[2]正木喜雄：インターネットを用いた潜在ニーズの調査における静止画・動画の活用（2007,中央大学大学院修士論文）