言語データ解析による “売れる”要因発見の方法

学校法人産業能率大学　総合研究所
経営管理研究所　福岡　宣行

●論旨

本稿は、製品企画段階の耐久消費財において、製品の“売れる”要因を発見し、製品企画の意思決定に活用できる情報収集・分析の方法について新しい提案を行うものである。近年、Webサイトの口コミのように言語データは、大量かつ簡易に入手が可能になり、コストをかけずに言語データを用いた解析が行える環境が整っている。本稿では言語データを用い、製品の販売量向上に寄与する製品特性を把握する実用的な方法について検討し体系化した。

分析には、言語データ解析のためのテキストマイニングと統計的手法を用いる。これにより通常のアンケート調査では把握することが難しい、製品のデザインや使い勝手等の顧客が感覚的に評価する製品特性を定量的に把握し、製品企画に役立つ情報として提供できるようにした。また、データ解析はオープンソースでフリーウェアの統計解析ソフトRを用い、コストをかけずに分析ができる環境で行っている。本稿ではこの方法について事例研究を行い、その有効性を検証した。

●はじめに

製品企画やマーケティングの領域で市場調査は、古くから多くの企業で実践され、実績が積まれてきた。市場調査の方法には、製品の仕様や価格など数値で示される定量データを扱う調査方法と、人間の行動や言語など数値で示すことのできない定性データを扱う調査方法がある。特に近年は後者が注目されており、人間の行動や言語の分析から顧客の潜在的なニーズを発見するアプローチに関心が高まっている。

本稿では、定性データの中でも特に言語データの有用性に着目し、製品企画の実務にそれらを活用する方法について研究を行った。言語データの分析は、近年テキストマイニングの領域で研究が進んでいる。テキストマイニングはコンピュータによる大容量のデータ解析によって有益な知識を取り出すデータマイニングの一種であり、言語データを対象とし、大量の文章を単語や文節で区切り、単語の出現頻度や前後関係を解析することで、その文章データが持つ傾向を定量的に把握する。

しかし、言語データを扱う分析には膨大なコストと手間がかかるため、これまで一部の企業でしか使われてこなかった。言語データは、音声の記録や手書きの文書などで保存されることが多く、それらをテキスト形式の電子データに変換する手間がかかることから、導入されているのはコールセンター業務の改善等、大規模な投資に対して効果が得られる特定の分野に限定されていた。しかし、近年耐久消費財においてはWeb上のSNSや口コミサイトから、顧客ニーズ・満足度に関する大量のデータが、簡易かつ低コストで入手することができる。また近年、音声言語認識（speech recognition）技術や光学文字認識（optical character recognition）技術が発達し、これらのフリーウェアソフトも実務で使えるレベルになってきている。さらに、言語データを分析するためのツールもフリーウェアの有効なソフトが公開されている。本稿で使用している統計解析ソフトRはオープンソースの開発環境として世界的に多くの研究者が開発に参加し、専門的な学術論文にも用いられているものである。

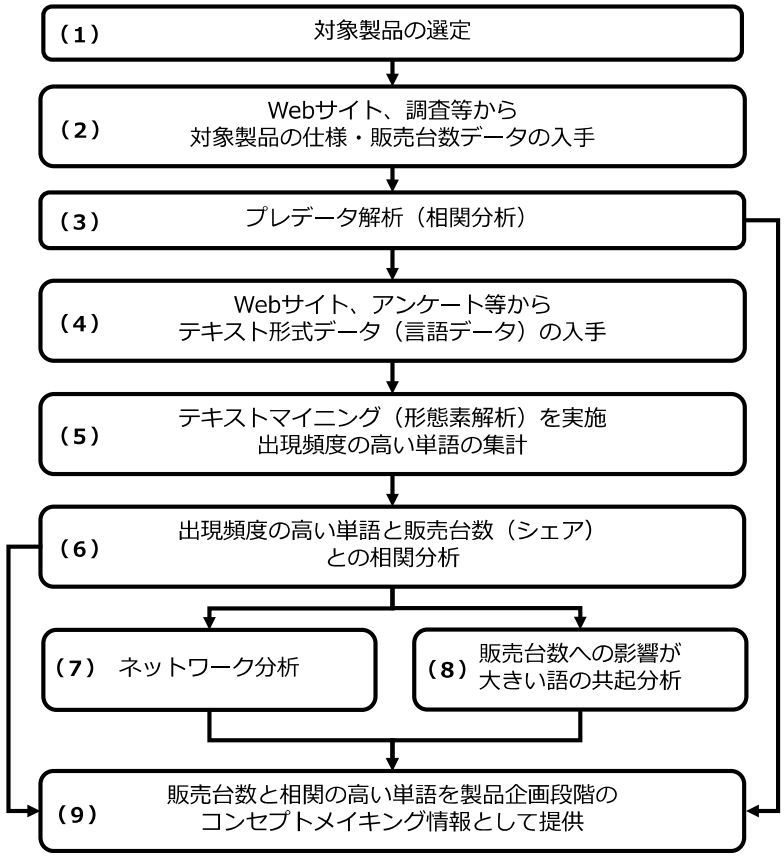
テキストマイニングは「仮説を発見する」ことを目的とした調査・分析手法である。そのため、実務で活用する上で、得られた結果を「どのように解釈し意思決定に活用するか？」というプロセスが重要である。それらのプロセスを導くための方法論はテキストマイニングの実践において必要であるが、明確に体系立てられたものは見当たらない。本稿では、耐久消費財の製品企画の実務を対象として、言語データによる調査・分析を行い、その結果を製品企画の意思決定に活用できる情報として提供するまでの一連の方法を体系化した。

＜キーワード＞
テキストマイニング、相関分析、形態素解析、ネットワーク分析

１．分析の概要

本稿の方法の特長は、テキストマイニングによって得られた出現頻度の高い単語と、製品の売上実績との相関分析を行い、「売れる」製品の要因を客観的に言語情報で得られるようにしたことである。テキストマイニングを実施すると、出現頻度の高い単語が明らかになるが、その結果についての良し悪しを判断する情報は得られない。よって、テキストマイニングの結果を製品企画の意思決定に活用するために、結果を解釈する必要がある。当然、その解釈はマーケッターの力量に左右される。結果の解釈は個人の主観的要素が入りやすいため、製品企画のための情報として用いるには説明力に欠けることもあり、結果として補足的に使われるに留められることも少なくない。
本稿の方法は、これらの結果の解釈を客観的に行えるようにするため、売上実績というパラメーターを取り入れて、主観的要素を極力排除できるようにした。分析の手順を以下の図表－１に示す。

図表－１　分析の手順



まず(1)対象となる製品を選定し、(2)対象製品の代表的な製品特性を表す仕様のデータと、対象となる製品の販売台数に関するデータを入手する。そして言語データ解析を行う前に、このデータでプレデータ解析を行い、製品の持つ特性の全体傾向について把握する。その上でWebサイト等からテキスト形式の言語データを入手し、(5)テキストマイニングを実施して、言語データの中から出現頻度の高い単語を抽出する。そして、(6)出現頻度の高い上位の単語と製品の販売台数との相関分析を行い、販売台数との相関が0.7程度以上確認できる単語をピックアップする。さらに、ネットワーク分析と(8)共起分析によってそれらの単語の前後関係に着目してより明確に意味を抽出する。(3)(7)(8)で得られた結果を、販売量への影響力を示す製品特性の情報として整理し、(9)製品企画段階のコンセプトメイキング活動に提供する。

以下、この分析の流れに沿って事例研究を行った結果を報告する。

２．事例研究

(1) 対象の選定

分析の対象となる製品のカテゴリから代表的な製品を10種類ほど選定する。ここでいうカテゴリとは、その製品を大まかに分ける分類であり、乗用車の例でいうと車型（セダン、ワゴン、ミニバン等）が相当する。

本稿で提案する方法は、特に新製品開発の競争が厳しい市場に有効なため、事例研究では見込生産方式で製品を投入する成熟製品の市場を想定し、デジタルカメラを対象製品とした。デジタルカメラには、コンパクトデジタルカメラ、一眼レフデジタルカメラ、および、その中間に位置する高級コンパクトデジタルカメラの３つのカテゴリが存在する。今回はこのうち、高級コンパクトデジタルカメラを分析対象のカテゴリとし、2014年9月の販売台数（シェア）トップ10の10機種を選定した。

(2) 基本情報の収集

対象として選定した機種については、下記の図表－２のように、その製品を特徴づける代表的な特性を示す仕様と価格、販売台数等の基本情報を収集する。これはまず、大まかな傾向を知るために基本情報によってプレデータ解析を行うもので、そもそも言語データ解析をする意味があるかどうかを確認するというねらいもある。

図表－２　対象製品の基本情報

メーカー	機種	画素数（万）	光学ズーム（倍）	撮影枚数（枚）	最低価格（円）	初値（円）	販売台数シェア（%）
ソニー	Cyber-shot RX100	2,020	3.6	330	34,885	62,860	18.0
キヤノン	PowerShot S120	1,210	5.0	230	26,980	44,926	15.0
富士フイルム	FUJIFILM XQ1	1,200	4.0	240	27,700	40,399	8.6
ニコン	COOLPIX P340	1,219	5.0	220	23,332	40,299	7.1
オリンパス	OLYMPUS STYLUS XZ-2	1,200	4.0	310	22,779	57,899	6.7
キヤノン	PowerShot S200	1,010	5.0	200	16,866	31,399	5.1
ソニー	DSC-RX100M3	2,010	2.9	320	73,012	85,499	4.6
リコーイメージング	MX-1	1,200	4.0	290	27,772	44,259	4.3
富士フイルム	FUJIFILM XF1	1,200	4.0	300	19,800	49,319	3.8
オリンパス	OLYMPUS STYLUS 1	1,200	10.7	410	47,300	62,385	3.5

※最低価格は2014年9月30日時点の価格ランキングサイトにて発売日以後最も安かった価格

(3) プレデータ解析

基本情報により、販売台数（シェア）と代表的な製品特性との相関分析を行う。これにより、販売台数へ影響を与える要因を大まかに確認する。この段階で、販売量との相関が高い製品特性が見つかれば、それらは製品企画の上で売れる要因として重要な位置づけを持つものと考えてよい。しかし、必ずしもこの段階で相関の高い特性が見つかるとは限らない。また高い相関を持つ特性が特に見つからない場合は、この後のテキストマイニングを行う余地はあると考えてよい。

デジタルカメラの事例研究では、製品の特性を示す代表的な仕様として、画素数（万画素）、光学ズーム（倍）、撮影枚数（枚）、最低価格と販売台数（シェア）の情報をWebサイトより入手し、相関分析を行った。

図表－３　基本データの相関分析結果

	画素数（万）	光学ズーム（倍）	撮影枚数（枚）	最低価格（円）	販売台数シェア（%）
画素数（万）	1.000				
光学ズーム（倍）	−0.319	1.000			
撮影枚数（枚）	0.390	0.498	1.000		
最低価格（円）	0.709	0.110	0.560	1.000	
販売台数シェア（%）	0.423	−0.204	−0.132	−0.089	1.000

図表－３の結果は画素数、光学ズーム、撮影枚数、最低価格、販売台数シェアの各相関係数をマト

リクスにして出力したものである。最低価格の列を見ると、画素数との相関が0.70以上あることが確認でき、画素数が大きいほど価格が高くなる傾向があることが確認できる。しかし、販売台数シェアについては相関は最も高いもので画素数の0.42であり、これだけの情報では販売量を説明することは難しいことが分かる。

※データ解析のために、本稿ではデータ解析ソフトRを用いる。Rは筑波大学のWebサイト[1]でダウンロードできる。インストールの方法やRの基本操作については紙面の都合上割愛するが、その方法は本稿の末にRの基本操作方法が載っているWebサイト[2]および書籍を掲載しておくので参考にしていただきたい。またテキストデータ解析には、日本語の形態素解析ツールも必要である。本稿ではフリーウェアの解析ツールMeCabおよび、その実行をRで制御するパッケージRMeCabを用いた。

このように基礎情報だけで販売量を説明できない場合は、言語データ解析を用いることで、主要な製品の仕様だけでは説明できない販売量に影響している要因を発見できる可能性がある。以後にその手順について示す。

(4) 言語データの収集

顧客の声は口コミという形で、インターネット販売を行う企業のサイトや、それらとリンクするランキングサイト[3]で入手することが可能である。言語データ収集の注意点は、どの時点の言語データを対象とするかを考慮して収集することである。例えば、毎月のシェアのランキングが激しく入れ替わる製品カテゴリで、１年以上前の口コミデータを利用しても売上との関係がうまく発見できるとは考えにくい。製品カテゴリの販売量の変動を加味して、妥当だと思われる投稿時期や期間で収集する必要がある。

事例研究では、2014年9月の販売量シェアを用い、その2か月前までの2014年8月1日～9月30日までの高級コンパクトデジタルカメラに関する口コミデータを、製品ランキングサイト[3]から入手した。口コミデータは、製品のデザインや性能、使い勝手等に関する満足度についてユーザーが自由に記載した文章である。

(5) 形態素解析による品詞別頻度の分析

入手した言語データを実行環境（Rおよび解析ツール）で解析する。日本語の意味の最小単位（形態素）に分解する形態素解析を行い、それらの出現頻度をカウントすることで口コミの言語データから、対象製品が持つ特性の意味を抽出する。

例えば、「このカメラは使いやすい」という文章を形態素解析すると、「この（連体詞）」「カメラ（名詞）」「は（助詞）」「使う（動詞）」「やすい（形容詞）」の５つの形態素に分割・変換される。動詞の「使い」が「使う」に変換されているように、ただ文章を分割しているだけでなく、動詞の原形に変換することで、同じ意味としてカウントしている。

事例研究では、入手した10機種の高級デジタルカメラに関する口コミデータを解析し、名詞、動詞、形容詞の３つの品詞に絞って集計した。

※それぞれの機種によって収集できる口コミデータの文章量が異なるため、単純な出現頻度を比較することは望ましくない。よって文書量による影響を調整する正規化を行って重み付けをしている。

(6)出現頻度の高い単語と販売台数（シェア）との相関分析

形態素解析の結果得られた単語を、名詞、動詞、形容詞等の品詞別に、出現頻度の高い順でソートし、その中で上位20程度の単語と販売量シェアとの相関分析を行う。

事例研究では、名詞、動詞、形容詞ごとに実施したところ、それぞれ上位に0.7程度の相関係数を持つ単語が現れた。

図表－４　出現頻度の高い名詞とシェアの相関係数

単語(名詞)	シェア	単語(名詞)	シェア
性	0.572	画質	0.670
X0	-0.425	感	0.700
の	0.462	操作	0.553
カメラ	0.471	X.	-0.098
撮影	0.383	さ	0.212
X1	0.292	購入	0.674
機種	0.660	デザイン	0.573
機能	0.487	携帯	0.554
こと	0.253	X..1	-0.217

図表－５　シェアとの相関が高い単語

名詞	
感	0.7
購入	0.67
画質	0.67
機種	0.65
動詞	
絞る	0.69
できる	-0.62
形容詞	
素晴らしい	0.71
大きい	-0.64
早い	0.63
多い	0.61

図表－４は出現頻度の高い名詞と販売量シェアとの相関分析を行った結果である。この結果の中には名詞でないものも含まれているが、これは形態素解析において品詞の誤判定が起きているためである。（誤判定の回避は、事前に語句を登録する必要がある。）

「感」という語が0.7となったがこれは、「～～感」のように他の語と合わせて使うことで、人間の感覚を表現する形で用いられ、感覚的な要素が売上の向上に関係していると考えられる（「感」は名詞ではないが、このように重要な表現であることから削除せずに残した）。

また、「画質」も売上との相関が高いことが分かるが、(3)のプレデータ解析で、「画素数」は販売台数シェアとの相関が0.4しかなかった。これは画素数だけでは売上との相関は明確でなく、画素数に加えてイメージセンサーの大きさ等で決まる「画質」が売上との明確な相関があり、売上の向上において重要な要素であることを示していると考えられる。名詞、動詞、形容詞のそれぞれの販売量シェアと相関が高かった上位の単語を表－２に示す。

このように出現頻度の高い品詞と販売量シェアの相関分析を行うことで、プレデータ解析では見えなかった売上を向上させる要因が見えてくる。しかし、まだこの情報では製品企画に十分な情報提供ができるとは言えない。例えば、「感」はその前にくる単語によって意味は全く異なってくるし、動詞の「絞る」などの専門的な操作を意味する単語は、ある程度の意味の想定はつくものの、やはり動詞だけでは意味を特定しづらい。このように、出現頻度の高い単語だけを見て意味を解釈できないものも存在する。

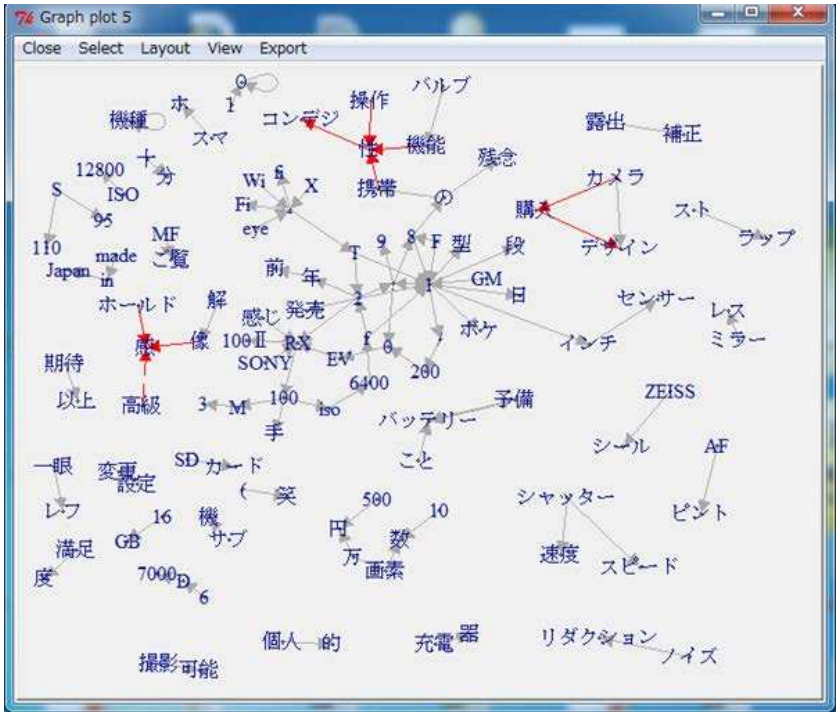
よって、この後さらに、単語が構成している文章の文脈から意味を抽出する分析が必要である。

(7) 単語の前後関係（共起関係）を用いたネットワーク分析

単語の意味を厳密に解釈するには、それらの語の前後関係を確認しなければならない。しかし、この分析は大量の言語データを扱うことが前提であり、全ての文章に目を通すのには限界がある。そこでテキストマイニングには、ある単語の前と後に出現する（共起）単語の出現頻度をカウントすることで、語のつながりの多さから意味を抽出する方法がある。

これらのつながりは通常、連鎖しているため、全体を見やすくするための方法として、ネットワーク分析がある。これもRのパッケージを利用すれば、図表－６のような図を簡単に描くことが可能である。

図表－６　売れている製品群に共通のネットワーク図（名詞）



ネットワーク図の単語から出ている矢印は、文章の中でその語の後に出現している語を指している。全ての関係を表示すると見にくくなるため、ここでは相対的に出現回数が高いものだけを表示する。図表－６の例では３回以上の出現頻度があるものを表示した。ネットワーク分析は通常個々の製品別に実施するが、ここでは販売量シェアの高い製品の特徴を発見したいため、販売量シェアの高い製品を１つのグループにして分析する。

事例研究では、販売量シェアでトップを走るCyber-shot RX100とPowerShot S120（２つ合わせてシェア33%）の２機種を１つのグループとして実施した。このグラフに出てきた語のうち、販売量シェアとの相関が特に高い語との関係を示す矢印は赤で記してある。赤い矢印の周辺を確認すると、「ホールド感」、「高級感」、「操作性」、「携帯性」などが、シェアの高い製品にとって特に重要なキーワードであることが分かった。

このようにネットワーク分析を行うことにより、製品群の全体像をとらえながら、売上との相関の高い語とその周辺の関係を見ることで製品企画において有益な情報を提供することができる。

(8) シェアの影響が高い単語の共起分析

ネットワーク分析は語の共起関係をカウントするため、(6)で抽出した品詞が全て出現するとは言えない。そのため、(6)で明らかになった出現頻度の高い品詞のうち、ネットワーク分析で出てこなかった語についても共起分析を行うことで、新しい発見がある可能性がある。

事例研究でも、ネットワーク分析では販売量シェアとの相関が高い上位８つの単語を見ても、その単語の共起関係は３回以上の出現頻度に絞っているため、一部しか確認することができない。よって、８つの単語の共起関係を分析し、この中で共通して見られた共起頻度の高い語をカウントした。結果

を図表－７に示す。

図表－７　シェアと相関の高い単語のバイグラム（２語の共起関係）

「感」のバイグラム		「購入」のバイグラム		「画質」のバイグラム		「機種」のバイグラム	
[ホールド-感]	21	[購入-する]	25	[画質-良い]	5	[機種-機種]	15
[感-ある]	5	[する-購入]	4	[する-画質]	3	[画質-機種]	2
[高級-感]	4	[カメラ-購入]	3	[思う-画質]	3	[機種-購入]	2
[像-感]	3	[グリップ-購入]	2	[いい-画質]	2	[機種-背景]	2
[クリック-感]	2	[円-購入]	2	[いる-画質]	2	[機種-不明]	2
[安定-感]	2	[機-購入]	2	[しまう-画質]	2	[機種-夜景]	2
[感-エッジ]	2	[機種-購入]	2	[画質-期待]	2	[買う-機種]	2
[感-感じる]	2	[購入-考える]	2	[画質-機種]	2	[EV-機種]	1
[重量-感]	2	[グリップホルダー-購入]	1	[画質-変わる]	2	[2-機種]	1
[グリップ-感]	1	[シルバー-購入]	1	[コンパクト-画質]	1	[シャープ-機種]	1

「絞る」のバイグラム		「素晴らしい」のバイグラム		「早い」のバイグラム		「多い」のバイグラム	
[絞る-優先]	3	[素晴らしい-カメラ]	3	[スピード-早い]	1	[こと-多い]	2
[絞る-シャッター	2	[これ-素晴らしい]	2	[減る-早い]	1	[荷物-多い]	1
[3200-絞る]	1	[画質-素晴らしい]	1	[消耗-早い]	1	[機種-多い]	1
[できる-絞る]	1	[解像度-素晴らしい]	1	[早い-快適]	1	[機会-多い]	1
[絞る-F]	1	[素晴らしい-オプション]	1	[早い-狙う]	1	[機能-多い]	1
[絞る-暗い]	1	[素晴らしい-ジャケット]	1	[早い-注意]	1	[撮影-多い]	1
[絞る-開く]	1	[素晴らしい-完全]	1	[早い-予備]	1	[出番-多い]	1
[調整-絞る]	1	[素晴らしい-機種]	1	[速度-早い]	1	[場合-多い]	1
[最大-絞る]	1	[素晴らしい-景色]	1			[多い-動画]	1
[撮る-絞る]	1	[素晴らしい-動画]	1			[多い-携帯]	1

「感」は、ネットワーク分析によって、「ホールド感」や「高級感」が多いことが分かっていたが、この分析で他にも「高級感」、「重量感」、「クリック感」、「安定感」などの多様な言い回しに利用されていることが分かる。これは定量的には表すことのできない感覚的な満足感が多数あることが、売上の向上に影響を与えていると考えられる。

「購入」は「購入する」という使い方が多く、その前に状況を示す語を伴うことが多い。例えば、「円-購入」はいくらの価格でこの製品を購入したのかを意味するときに使われる。このように購入者が口コミで伝えたくなるような購入の動機となる状況があることを示している。

「画質」については、「画質－良い」が多く、画質の高さによる満足度はやはり売上への影響が大きく、製品の基本的機能として力を入れるべき重要な特性であると考えられる。

「機種」については「購入」と同じく、この機種を購入した状況が表現されている。背景や夜景などとのつながりもあり、さまざまな利用シーンにおける製品の能力について満足している様子が伺える。

「絞る」はこの前後のつながりより、光量の調節に関する操作に関するものであることが分かった。これも定量的には表すことのできない、使い勝手に関する機能の優秀さについての満足が伺える。

「素晴らしい」は満足度の高さを表す単語であるが、それは画質や解像度、それに伴う利用シーンでの満足度において、購入者が期待以上であったことを表現していると考えられる。これは顧客の期待を上回って感動を生むレベルであったことが伺える。

これらの結果から、販売量で圧倒的にシェア差をつけている製品は、上記のように基本的な性能が高いだけでなく、デザインや利用時の多様な感覚的操作や質感に優れ、またそれらが顧客の期待を裏切るようなレベルであることが分かった。さらに顧客の感動がこのような口コミとなって、潜在顧客に伝わり、良い循環を生んでいることも伺える。

よって、売れる高級コンパクトデジタルカメラの製品企画のためには、これらのことを十分に配慮してコンセプトメイクを行うことが重要である。製品企画段階では、これらの分析結果を意思決定者と共有しておくことで、コンセプトの方向を定め、企画を推進するのに役立つと考えられる。

３．結語

本稿では、テキストマイニングを販売量との相関を見ながら実施する新しい方法を提案し、高級コンパクトデジタルカメラの事例研究を通じてその有用性を確認した。

事例研究では紙面の都合上割愛したが、売れる製品以外にも、売れない製品についても分析を試みたところ、分かったことがある。それは、決して売れない製品を購入した顧客は、満足度が低いというわけではないが、感動レベルの満足度を示す記述は、売れる製品ほど多くないということである。

製品コンセプトに込められた企業の“思い”が顧客の期待を超えるときに“売れる”製品となることは多くの実務経験者が肌で感じていることだと思われる。しかし、このことはこれまで目に見えないものとして、マーケッターの「経験」の中で培われてきた。そのため、売れるかどうかは結果論として扱われるケースも少なくなかった。

本稿はそのような企業の“思い”が顧客に伝わり“売れた”製品を分析し、顧客の口コミからその要因を発見することを目指し方法論を構築した。この方法は製品企画で特に難しい定性的な製品の特性を決める際に、新たな仮説を発見し意思決定者を支援する情報として提供できるようにしたものである。この方法が製品企画者の一助となれば幸いである。

＜参考Webサイト＞

［1］筑波大学CRAN　http://cran.md.tsukuba.ac.jp/
［2］統計ソフトRの使い方　https://sites.google.com/site/webtextofr/
［3］BCNランキング　http://bcnranking.jp/　価格.com　http://kakaku.com/

＜参考文献＞

1)石田基広：「Rによるテキストマイニング入門」森北出版株式会社(2008)
2)鈴木努：「ネットワーク分析」共立出版(2009)