『学習院大学　経済論集』第52巻　第１号（2015年４月）

Web 上の口コミ情報からの潜在的事前期待の発見
―@コスメにおけるマスカラの評判分析―

白田由香利[1)]，橋本　隆子[2)]

概要

本稿では，経営者のための，初期費用がかからず，対費用効果の高い口コミ分析手法を提案する。提案手法の特徴は，評判の全容の俯瞰，及び潜在的事前期待の発見，ホリスティックな視点からの有益コメントの発見である。その適応事例として@コスメのマスカラの口コミ分析を行った。そして潜在的事前期待として，「まつげへのダメージ回避」，「まつげへのトリートメント効果」を，また，ホリスティックな有益情報として「他のベース化粧品との組合せ」などが発見できた。また，この手法と LSA による分析結果の比較を行ない，分析の本質的な部分は度数分布であり，本提案手法でも把握可能なことを示した。経営者の本当に欲している情報とは，顧客の潜在的事前期待であり，画期的な商品機能改善や，新商品の開発に結び付く未知の機能，サービスであろう。本稿で我々が提案した分析手法は，簡単な方法ではあるが，こうした経営者のニーズに応えている。

１．始めに

ビッグデータをキーワードに，WEB 上のソーシャル・メディアなどの膨大な書き込みから顧客の意見，ニーズなどを分析することが広く行われるようになった。経営，マーケティングの分野でもテキストマイニングによる顧客のニーズ調査が広く行われている。マーケティングにおいてこうした分析を行う目的は，顧客の商品（財）やサービスに対する事前期待と実績評価を知見として得ることである。顧客には，商品などに対する期待がある。その期待に実際の商品などが応えているか，その評価を分析すること。これがマーケティング分野でのテキストマイニングの第1の目的である。そして，第2の目的として，顧客の中に在る潜在的な期待を探るというものがある。これは顧客自身も気付いていない，あるいはそのニーズが潜在的であり顧客自身が表現するところまで到達していない。しかし専門知識のある経営者が書き込みを読

1）　学習院大学経済学部
Gakushuin University, Faculty of Economics
yukari.shirota@gakushuin.ac.jp
2）　千葉商科大学商経学部
Chiba University of Commerce, Faculty of Commerce and Economics

むと，行間から次世代の商品への期待値を察知できる，というケースである。

口コミのテキストマイニングは経営者にとって非常に有益な分析結果をもたらすことは明白であるが，実践していない経営者も多い。その原因のひとつは，「データマイニングは大規模コンピュータシステムをもつ専門業者に依頼しなくては不可能」と思い込み，始めから諦めて実施しないでいる人が多くいるのではないだろうか。本稿の目的は，経営者のための，初期費用がかからず，対費用効果の高い口コミ分析手法を示すことである。事例として@コスメのマスカラの口コミ分析結果を示す。そして，この手法とLSAによる分析結果の比較を行ない，分析の本質的な部分は，本提案手法でも把握可能なことを示す。

第2節では，我々が提案する，簡便な口コミ分析の手法を説明する。第3節では，その手法によるマスカラ口コミ分析結果を示す。第4節では，マーケティング分野でのテキストマイニングに，関する一般的な考察を行う。最終節はまとめである。

2．提案する分析手法

商品の口コミ分析は日進月歩しており，コストをかければ，非常に精緻な分析ができる。しかし，それでは，口コミ分析に馴染みのない経営者にとって敷居が高く実施は困難である。本節では，コストが少なく手軽に行える割に，貴重な情報が発見できる，という対費用効果の高い口コミ分析手法を提案する。本格的なテキストマイニングを始める前の準備としてもよい。

その手法では，事前期待という概念を重視する。顧客には，事前に期待している何らかの期待値があって，それに対して実績評価が大きいと大いに満足し，リピート客化していく。反対に，事前期待値が高いのに実績評価が低いと顧客はがっかりして店から離れる［1，2］。以下の口コミ分析では，特に，潜在的な事前期待を発見することを目的とする（以降のステップC参照）。

テキストの内容把握のために，最も重要な品詞は名詞である。その分野に精通している分析者であれば，抽出された重要名詞だけでも，文章内容がある程度予想可能である。よって，以下の手法では，抽出する品詞は名詞のみとし，名詞のバイグラムを作る。それを複合語あるいはコンセプトと呼ぶことにする。そして出現頻度の高い複合語（コンセプト）を用いて分析を行う。

分析は，AからDまでの以下の4ステップから構成される。以下では，商品ごとの口コミをまとめて1ドキュメントとした。1商品が1ドキュメントを構成する。

A.　その商品の評判の全容俯瞰

商品分析では，始めに全容を俯瞰することが重要である。そのために，全体像をグラフ表示する。

（1）単語のみ抽出し，前後でつながっている2語（バイグラム）を抽出する。出現頻度の高いバイグラムのみを抽出し，それらをグラフ表示する。3-グラムとの比較も必要である。

（2）グラフから，意味的に重要な商品属性と思われる複合語（コンセプト）を抽出する。

（3）そのコンセプトを含む，元の文章を読み，評判の全容を概括する。

B. 語の出現度数分布

語の出現度数分布を調べる。TFIDF などの値を使い，他の商品の口コミになく，その商品にのみ出現している複合語（コンセプト）を発見する。それが，該商品の特徴であるが，特徴には利点と問題点の両方を含むので，本文を読んで調べる。

文章から重要な語をキーワードとして抽出する際には TFIDF がよく用いられる［3］。語 w の TFIDF の定義は，$\mathrm{TFIDF(w)} = \mathrm{TF(w)} \times (1 + \log_2 \frac{D}{DF(w)})$ である。

TF(w)は，1ドキュメント内での語 w の出現回数（局所的出現頻度）である。後ろの項が IDF である。D はドキュメントの総数，DF(w)は語 w を含むドキュメントの数である。ここでは対数の底は2とした。本式は以下のように変形したほうが，式の意味が理解しやすい。

$$\mathrm{TFIDF(w)} = \mathrm{TF(w)} \times (1 + \log_2 D - \log_2 DF(w))$$

IDF において，語 w に依存している部分は DF(w)のみである。上式の意味するところは，多くのドキュメントに含まれる語のほうが，TFIDF の値は小さくなり，語の重要度は下がる，である。少数の文書だけに出現する語の場合は，TFIDF に代わって RIDF を用いるほうがより正確に重要度の指標となる。

C. 潜在的事前期待及び改善希望事項の発見

顧客も気づいていない期待や改善希望点を発見する。ステップ B で発見した，他商品にない重要語の文章を読んでいくが，「xxx だともっといいのに」「これが xxx だったら」というような表現を見つける。これは正規表現で典型的なパターンを抽出する。

D. ホリスティックな視点からの有益コメントの発見

他の商品と組合せで効果が倍増する事例や，利用者の生活全体をホリスティックに見たときに発見される効果などを見つける。

提案する手法では，主成分分析，対応分析などの多変量解析は使っていない。また，LSA や LDA などの機械学習なども使わない。LSA，LDA に関しては考察で後述する。

3．マスカラの口コミ分析

本節では，前節で提案した手法に基づき，実際にテキストマイニングを行う。分析対象は，化粧品の口コミサイト「@コスメ」に投稿されたマスカラの口コミである[1)]。データ採取は2012年である。その当時のマスカラの主力銘柄である3商品についての口コミである。以下では，具体的な銘柄はふせ，商品 #1（2010年12月から2011年8月，260件），商品 #2（2011年11月から2012年7月, 229件），商品 #3（2011年6月から2011年9月, 200件）と呼ぶこととする。上記，カッコ内は書き込みされた期間と採取した口コミ件数を示す。

実験環境を説明する。日本語形態素解析は RMeCab を使っている。R のライブラリとして，

1） @コスメ：http://www.cosme.net/

RMeCab 及び igraph[2]（グラフレイアウト機能）を用いた。これらのツールはフリーソフトウェアである。R のプログラムは，R によるテキストマイニングの教科書を参考にした［4，5］。R に関するテキストマイングのテキストは多数あり，テキストマイニングを初めて行うユーザーはこれらの教科書を参考にしてまずは真似てみることが重要と考える［3，6］。テキストエディターは正規表現での検索機能やパターンマッチング機能が必要であるため，サクラエディタ[3]を使った。またグラフに関しては，R でもグラフは描けるが，分析後のプレゼン資料の作成などとの関係を考えて，EXCEL を使った。TFIDF の計算及びその正規化については，RMeCab ライブラリではなく，EXCEL で行った。EXCEL のマクロは使わなくても計算可能である。

以下では，提案する手法の4ステップに沿って分析結果を説明する。

3.1　ステップ A

図1に，商品 #1の口コミのテキストファイル（ドキュメント）から抽出した名詞のバイグラムのグラフ図を示す。グラフから，この商品の評判に関する重要複合語（コンセプト）が以下のように抽出できた：

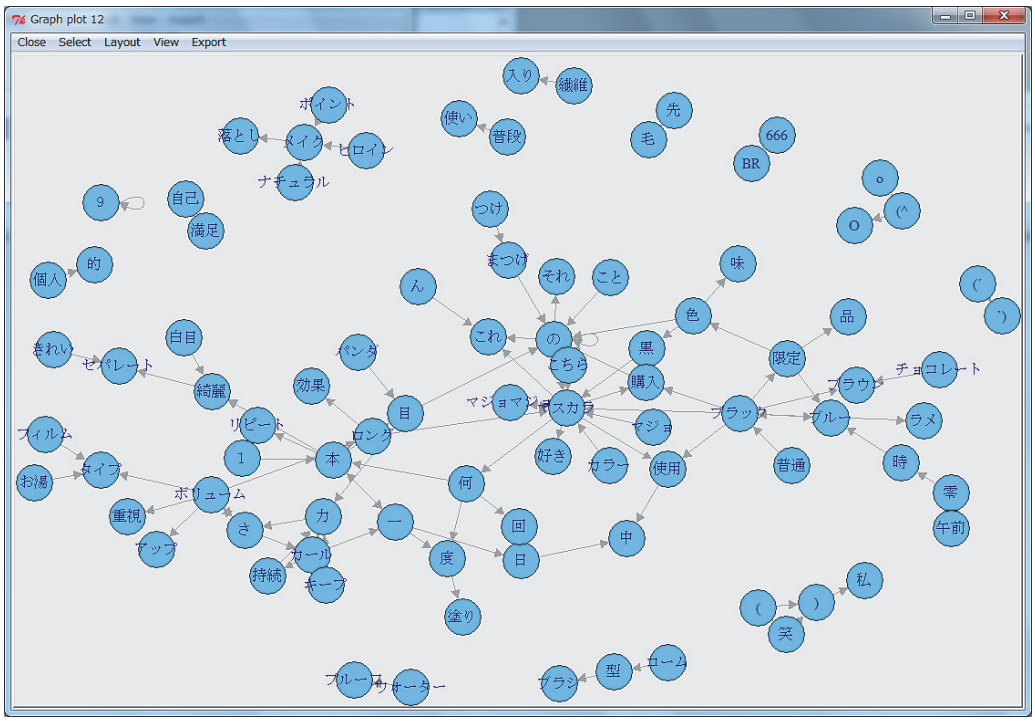

図 1　マスカラ商品 #1について，名詞のみのバイグラムをグラフ表示したようす。
Figure 1　The graph of important noun bigrams about the eyelash liner product #1.

2）　igraph：http://igraph.org

3）　http://sakura-editor.sourceforge.net/ から入手可能。

メイク落とし，ナチュラルメイク，ヒロインメイク，メイクポイント，自己満足，白目綺麗，ロング効果，パンダ目，ボリューム重視，ボリュームアップ，お湯タイプ，フィルムタイプ，カール持続力，カールキープ，ウォータープルーフ，普段使い，繊維入り，毛先，ブラック限定品，ブルーラメ，チョコレートブラウン，毛先

上記のうち下線を引いた複合語は，著者らが特に重要な属性と判断したものである。マスカラにおいては，ロング効果，ボリューム力，カール持続力が重要である。また，マスカラが落ちて目の周りが真っ黒になることを，「パンダ目」あるいは，単に「パンダ」と言う。マスカラでは，それを回避することが重要とされる。

3.2 ステップ B

次に TFIDF を分析する。IDF の式中の項 $\frac{DF(w)}{D}$ は，全体のファイルのうち，何％がその語を含んでいるか，含有ファイル数の比率を示している。その値が大きいほど，全体的に出現することを意味し，語の重要度が下がる。

図2に商品 #1の重要複合語に対して，この項の値を計算したものを示した。全体のドキュメント数 D を3にした場合は，取り得る値は m/3という分数をとる（m＝0, 1, 2, 3）ので4通りで

	3つの DF(w)/D	2つの DF(w)/D
ブルーブラック	0.3	0.0
限定色	0.7	0.5
ロング効果	1.0	1.0
カールキープ	1.0	1.0
ブラック購入	1.0	1.0
限定ブルー	0.3	0.0
一本	1.0	1.0
コーム型	0.7	0.5
零時	0.3	0.0
キープ力	1.0	1.0
パンダ目	1.0	1.0
午前零	0.3	0.0
時ブルー	0.3	0.0
カール力	1.0	1.0
ブラック使用	1.0	1.0
個人的	1.0	1.0
色味	0.3	0.0
カール持続	1.0	1.0

図２　該当複合語が含まれるドキュメント数を，ドキュメント総数で割った値。D=3と D=2の場合の比較。

Figure 2　The values of the number of documents which include the concept divided by the total numbers of documents (D). The comparison of the cases between D=3 and D=2.

ある。図2で，D＝3の際の項 $\frac{DF(w)}{D}$ を見ると，0.3（正確には，1/3。図2では，小数点1ケタまでしか示していない）のときが，その商品だけに出現している複合語であることが分かる。例えば，“ブルーブラック”，“限定ブルー”などがそれに該当する。この項の値だけ見ても，その商品の特徴がある程度推測できる。

図2に D＝2の場合の，項 $\frac{DF(w)}{D}$ の値も示した。取り得る値は，0，0.5，1 の3通りである。この値を見ることで，2商品の比較で，片方の商品にのみ出現する語（0 OR 1），両方の商品に出現する語（0.5）の分類ができる。2つの商品の比較の際，この値を見ることは価値がある。

次に，3種の商品に対する，「正規化された TFIDF」を計算し，それを使って，商品の特徴比較を行う。正規化は，ひとつのドキュメント（ひとつの商品）に関して，各語の TFIDF の2乗の合計が1になるように，比率を決める。

商品比較は，あるひとつの商品の複合語の正規化 TFIDF の値のランキングで，3商品の複合語の正規化 TFIDF を並べた。図3は，商品 #1をベースにした比較である。「限定ブルー」「零時」「午前零」「時ブルー」「色味」など，色に関する語が，商品 #1のみに出現していることが分かる。

これらの複合語を手掛かりに本文を参照すると，この商品は，ブルーブラックという色を限定品として発売し，それが評判になっていることが分かった。他の特徴語は，他の商品にも共通するものが多いが，この限定品に関する語は，この商品だけに出現している。マーケティング分野において，季節限定品，ご当地限定品などの限定品は人気が高い傾向がある。

図4は商品 #2の TFIDF をベースに語を並べたものであるが，商品 #2にのみ出現する語として「お湯」を含むものが数回出ている。商品 #3では，「ゼブラ柄」「コームタイプダマ」などが，

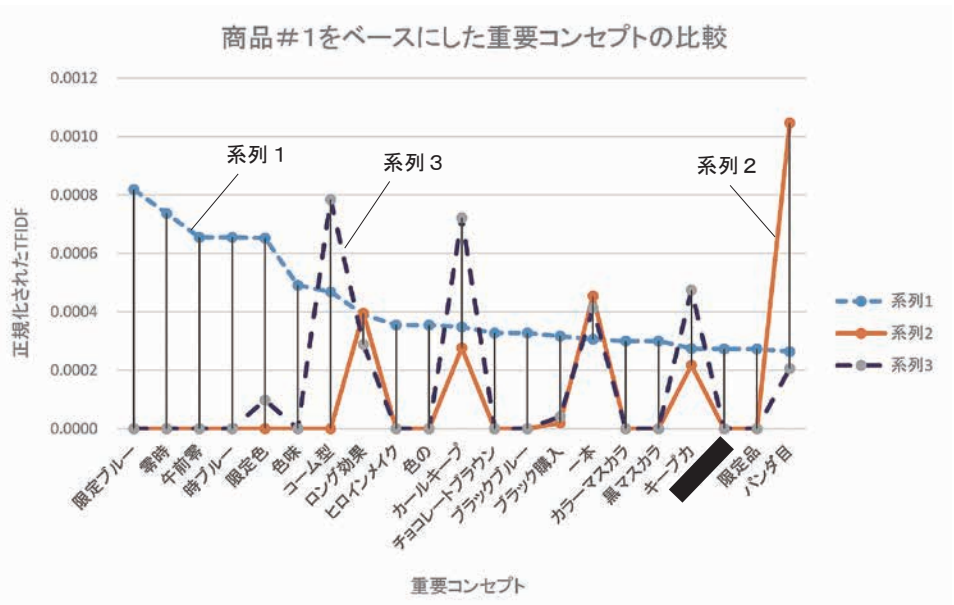

図３　マスカラ #1をベースにした重要コンセプトの比較

Figure 3 Comparison of important concepts of three products sorted by the ranking of the eyelash liner product #1.

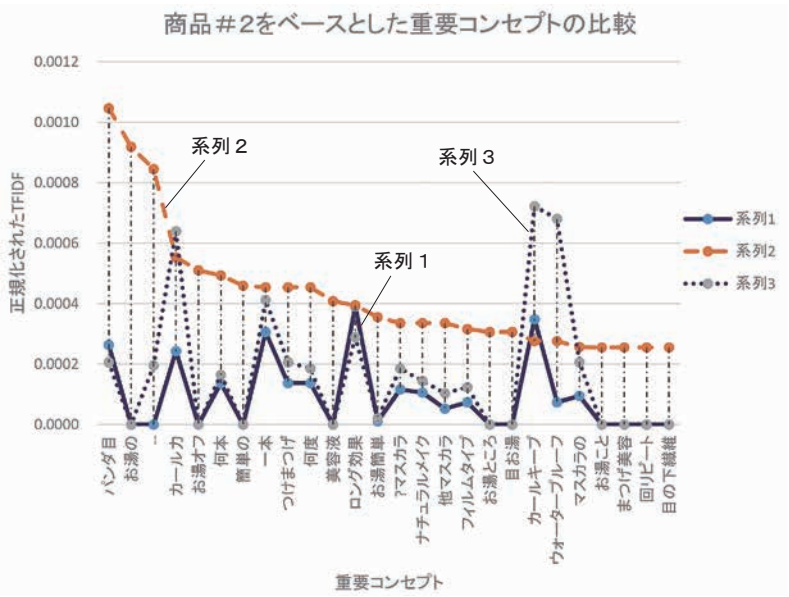

図４　マスカラ #2をベースにした重要コンセプトの比較
Figure 4　Comparison of important concepts of three products sorted by the ranking of the eyelash liner product #2.

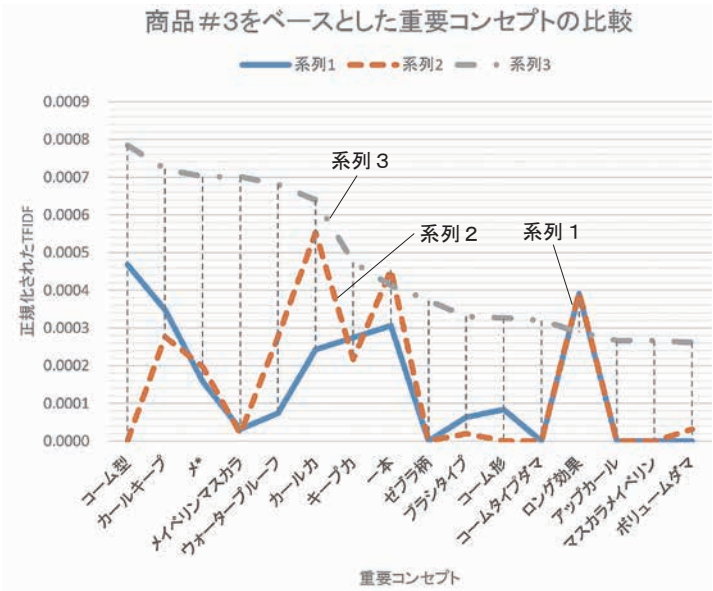

図５　マスカラ #3をベースにした重要コンセプトの比較
Figure 5　Comparison of important concepts of three products sorted by the ranking of the eyelash liner product #3.

この商品だけに出現している。ダマとは，固まりで，マスカラがダマになることは，困ったことで避けたい事態である。ステップ B で発見された，こうした，その商品だけに出現するコンセプトを使い，ステップ C や D を行う。

3.3 ステップ C

潜在的事前期待の発見例を示す。ステップ B において，商品 #2だけに出現するコンセプトとして，「お湯タイプ」があった。「お湯タイプ」という語はどのような文脈で使われているかを調べてみた。本文を読むと，以下のような事前期待が発見された。

> *お湯で落ちるタイプのほうがまつげに良いと思っていたけど以下の理由でこちらのほうが良いことに気付きました。こちらは朝1回のビューラーでカールが一日もつのに対しお湯で落ちるタイプは雨の日など湿気の重みでカールが落ちてビューラーを何回もしなくてはならない。メイクオフするとき，お湯でおちるタイプはまつげをつまないと完全にマスカラがとれない。*

上記の口コミから，「お湯」と「タイプ」は，「お湯で落ちるタイプ」のような語が離れた表現であることが分かる。マスカラにおいて，お湯で落ちるタイプはカールのキープ力が弱いことは既に広く知られているが，「お湯でおちるタイプはまつげをつまないと完全にマスカラがおちない→まつ毛をつまむことはまつ毛によくない」という意見は予期せぬ発見と言えよう。さらには，マスカラへの事前期待として，「まつげへのダメージ回避」，「まつげのトリートメント」という新しい事前期待が考えられる。これは価値ある潜在的事前期待の発見と言えるであろう。

次に改善点の発見例を示す。口コミからの改善要求文を見つけるため，「なければ」「れれば」で検索し，以下の文章を結果として得た。「→」以降のコメントは筆者が追記したものである。

- *わたしの使い方が間違っているのか，まばたきすると乾燥してないマスカラがついてしまいます。しかもなかなか落ちないんですよね。これさえなければリピートするんだけどな。* →これは「まばたきするとマスカラが付着してしまう」という苦情。
- *リムーバー買えって話なんですけどね w 落とすことをめんどくさいとさえ思わなければとてもいい商品だと思います。→「落とすためにリムーバーによる洗浄が必要で面倒」という苦情。*
- *パラパラとマスカラが落ちてきて目に入りゴロゴロして痛くて，クチコミを思い出しました。“目に入ると痛い”と書いてあったので，てっきりしみる物だと思っていたのです。パラパラさえなければ最高の1品だと思います！* →「マスカラのゴミが落ちて目に入って痛い」という苦情。
- *結構な頻度でちょっと繊維が落ちてしまいます（@_@）まあでもこの繊維があの素敵まつげを作り出してくれてるのでそこは目をつむってリピし続けています（=ﾟ-ﾟ）（=。_。）効果はそのままかそれ以上になって，いつかこの“繊維落ち”が改善されればいいなあ（-∀-;）* →表現は「パラパラ」「繊維落ち」と異なる表現を使っているが，「マスカラの繊維が落ちる」という苦情で前項と同じ内容。
- *コームブラシが使いにくい…。自前睫毛が濃く短いせいか，マスカラ液に粘りがあるせいか，ドバッとついてしまい，コーム部分でとかしても繊維が絡まってさらにヒドイ仕上がりに…。ブラシタイプのマスカラコームで根元からとかすと，塗ったマスカラがワサッと*

取れ，自睫毛も数本抜けました…。慣れれば綺麗に塗れるかも知れませんが，ヘタレな私は一回で心が折れました →コームブラシの問題指摘。

●*コームの使い勝手は，いままでブラシ状のものを使っていたせいか最初は難しいなと思っていたのですが，慣れればビシッと付いてくれるので病みつきになります。*→コームブラシの問題指摘。

上記のような，コームブラシへの改善要望，マスカラのダマが落ちることへの改善要望などが発見できた。

3.4 ステップ D

ひとつの商品に限定しない広い視野での使い方への提言（ホリスティックな提言）は，発見するヒューリスティックが特定しにくいので，一般的に言って難しい。しかし，TFIDF で他商品に無い語を見つけ，それを手掛かりに元の文章を読む，というアプローチが考えられる。以下に例を示す。

例：私は XXXX（固有化粧品名）のベース（メーク用化粧品）をプラスしているので，（マスカラの）長さもアップします。［かっこ内は筆者の補足部分である］

上記例は，相性の良いベース（下地作りのための）化粧品の商品目を提案している。こうした，化粧品の合わせ技情報の発見は経営者にとって非常に有益な助言と言える。合わせ技の発見手法としては，正規表現によるパターンマッチングで，商品の固有名詞，「プラス」などの語で検索をするなどの方策が考えられる。

4．考察

テキストマイニングによる商品口コミ分析の一般に関して，考察する。

＜TFIDF だけでも基本的傾向は把握可能＞

提案手法では，TFIDF を使って，複合語の出現頻度の度数分布を見ている。昨今は，LSA（Latent Semantic Analysis）［7］や LDA（Latent Dirichlet Allocation）モデル［8］などの複雑なモデルの分析手法が簡単に使える。しかし，そうしたモデルを使う前に，こうした基本的な TF 及び IDF などで，各商品の度数分布の特徴を把握することが可能であると考える。その上で，LDA などを使った精緻な分類を行うのがよいのではないだろうか。

始めから LDA を用いても，商品口コミに対する俯瞰ができていないと，トピック数を選択するとき，不適切なトピック数にしてしまう可能性があるからである。商品ごとの特徴が掴めていれば，例えば，トピックとして「パンダ目回避」「ボリューム力」「カール持続性」などが商品の重要コンセプトなので，トピック数は3＋アルファが適当である，というような決定ができると考える。

＜LSA との比較＞

以下では，LSA 方式で同じマスカラ口コミ文書を分析した結果を示す。LSA は，語 - ドキュメントの度数行列（以下，行列 A とする）を作り，それを SVD（Singular Value Decompositon）する。重要コンセプトは全商品に対して頻度の高い複合語10種を以下のように選んだ。

ブルーブラック
パンダ目
カールキープ
カール力
一本
ロング効果
キープ力
ウォータープルーフ
コーム型
つけまつげ

文書は，前節の分析で使ったものと同じであるので，D=3である。よって，行列 A は10行3列である。

行列 A は，$\mathbf{A} = \mathbf{U\Sigma}\ \ \boldsymbol{V}^{\mathrm{T}}$のように分解でき，$\mathbf{U\Sigma}$は，複合語の固有ベクトルとなる。複合語の作る3つの固有ベクトルの成分を見てみる（図6参照）。

固有ベクトル1は，「ブルーブラック」の比率が非常に大きい。「パンダ目」「カールキープ」などの共通する属性もあるが，この商品 #1の限定品という特性を反映した内容であることが分かる。

固有ベクトル2は，「ブルーブラック」の比率が1に比べて，マイナスからプラスへ反対方向に振れている。多次元空間で，この2つの固有ベクトルをイメージすると，「ブルーブラック」の軸に関して，反対方向を向いている。固有ベクトル2は，「ブルーブラック」以外の複合語に関しては，固有ベクトル1と方向が同じで似ている。

固有ベクトル3は，「ブルーブラック」に関しては，プラスマイナスの値は中庸であり，固有ベクトル1と2に比較して，「カールキープ」「ウォータープルーフ」「コーム型」の複合語の軸に関して反対方向に振れていることが分かる。このような3つの概念が，主成分に相当する概念として抽出された。

共分散行列 Σ の対角をなす固有値の値は，{159.48, 79.36, 42.20} であった。商品 #1のブルーカラーの限定品の影響が強い固有ベクトル1の固有値が，他の2つの固有ベクトルに比較して159.48と大きいことが分かる。この3商品の比較では，この限定品の影響が大きく，通常のマスカラ口コミの特徴語（パンダ目，カール力など）と違う結果をもたらしていることが分かる。

以上 LSA による分析を示したが，TFIDF だけでも基本的傾向は把握可能と言える。今回の3商品のトピック比較で重要なことは以下であった：（A）商品 #1が，限定カラー商品と言う要因により，通常度数分布から大きくずれた分布を構成していたこと，（B）通常分布では，「カール力」，「ロング効果」，「ウォータープルーフ」，「パンダ目」などが重要要素である。（C）商品 #2にのみ出現する語として「お湯」を含むものが数回出ている。（D）商品 #3では，「ゼブラ柄」「コームタイプダマ」などが，この商品だけに出現している。上記（C）と（D）については，LSA よりも，TFIDF のほうが容易に発見できた。

また，潜在的事前期待を見つけようとする場合は，どうしてもオリジナル文書を読む必要があるので, TFIDF による提案手法でも十分と言えよう。LDA に関しても，同様のことが言える。少なくとも，今回のデータに関しては，TFIDF の度数分布を見るだけでも，LSA に匹敵する

複合語	固有ベクトル1	固有ベクトル2	固有ベクトル3
ブルーブラック	-119.705	47.9498	-3.55113
パンダ目	-39.6182	-33.7729	-28.7017
カールキープ	-43.6236	-18.9589	15.7336
カール力	-36.9031	-29.9777	3.67456
一本	-38.229	-17.5158	-1.31972
ロング効果	-43.2682	-8.77117	-3.99162
キープ力	-33.137	-12.1093	9.01673
ウォータープルーフ	-18.9615	-27.3963	14.9634
コーム型	-32.4086	-4.31432	17.0608
つけまつげ	-20.7043	-17.3288	-8.30911

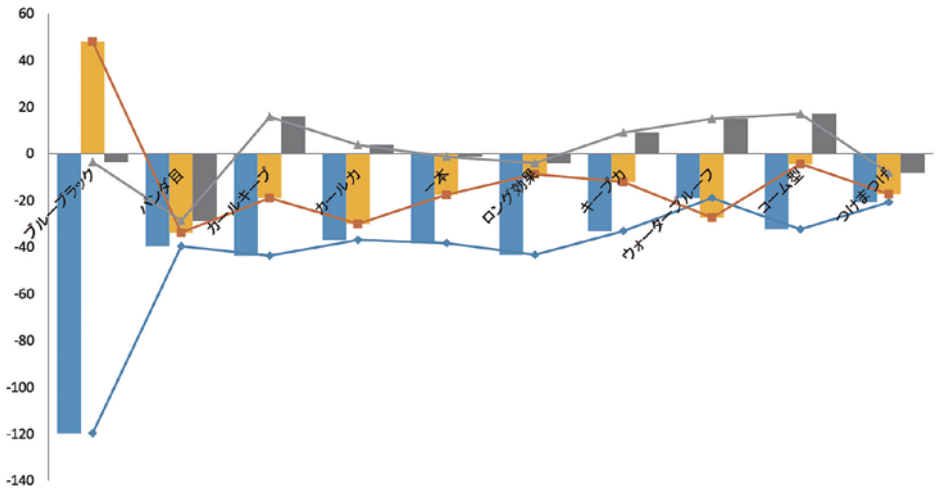

図６　LSA によって同じ口コミ文書を分析した結果の複合語の3つの固有ベクトル。下のグラフの3系列は，各固有ベクトルの複合語の比率を示す。

Figure 6　Three eigenvectors corresponding to the three concepts obtained by the LSA on the same words of mouth messages. The graph under the table shows the individual concept ratios on each eigenvector.

ような知見は得られた，と言えよう。我々の主張は，TFIDF だけで十分でよい，というのではない。LSA や LDA を使う前に，TFIDF などで全体の傾向を掴んでおくと適切に LSA や LDA を使うことができるようになる，ことを主張したい。

＜LDA モデルの適応＞

LDA モデルを使った商品比較を行うアプローチについて言及しておく。例えば，10種類の商品群全体に対して，トピック数5の LDA モデルを決定するとしよう。トピック数は適宜適切な数に変える必要があるが，以下ではトピック数5が適切であると判断されたと仮定する。

まず，10種類の商品群全体に対する LDA モデルを決定する。ここでトピックとトピックご

との語の比率が定まる。そのモデルに対して，商品ごとの個別の文書を入力し，5つのトピックに対する比率を求める。その比率の差異から，商品比較を行う。

この時，トピックの意味解釈はできていることが前提である。繰り返すが，LDA などの複雑な分類器などを用いて分類を行う前に，商品の概要を項 $\frac{DF(w)}{D}$ や IDF で把握しておくことが必要であると考える。

＜予想しなかった事柄の発見＞

テキストマイニングの意義について考察する。膨大なデータがあってそこから統計的に平均や分散に相当するような知見が得られても，その分析結果は現状把握にしかすぎないことが多い。換言すると，母集団の傾向が推測できても，多くの場合予測していたことの確認にすぎない。例えば，「予想した通り，売り上げ数が落ちている」などである。

経営者の本当に欲している情報とは，顧客の潜在的ニーズであり，画期的な商品機能改善や，新商品の開発に結び付く未知の機能，サービスであろう。これは広い意味では，想定外の驚くべき知識の発見をしたい，という要望である。

この例としては，我々が発見した事例を示す。2007年鳥インフルエンザの流行によりデジタル一眼レフカメラの売れ行きが鈍化した［9］。これは，鳥インフルのため，日本全体がゴールデンウィークなどに遠足や運動会，海外旅行などを控えたからである。これはビッグデータという膨大な量を活用して，商品売り上げへの鳥インフルの負の影響を発見した事例である。ここでは全商品に対して無差別的に，鳥インフルの患者数と商品売り上げの相関を探すというアプローチを取っているが，これは費用がかかる手法である。初めてテキストマイニングを行おうという経営者には適さない手法である。

想定外の驚くべき知識発見のための，費用（手間も含む）があまりかからず対費用効果が高く，実践が容易な手法は，有益な少数の書き込みからの予想外の発見・分析であろう。1つの書き込みでも，有益な書き込み情報であれば，それによって得られるメリットは大きいからである。本稿での提案手法は，こうした要望に合致する方法と言える。

＜専門知識をもつ人がやらなくては意味がない＞

ビッグデータ分析においても，データそのものへの専門知識がなければ，質の高い分析結果は得られない。データ分析はその分野の専門知識をもつものが主導を取るべきである。震災後の気仙沼で行われた産総研の絆プロジェクトの事例［10］においても，研究者のシーズから考えるのではなく，現場のニーズを探求することの重要さが述べられている。データマイニングを産業活性化に結び付けるためには，現場の経営者を中心とした分析が行われるべきであると考える。

＜データ書き込み者のレベル＞

今回使ったデータ，@コスメへの書き込み者は，マスカラのヘビーユーザーが多いと推定できる。一般にマーケティングのテキストマイニングの場合，その標本が，ヘビーユーザーの書き込みであるか，初心者あるいは平均的ユーザーのものか，で結果が大きく影響してくる。どちらを標本にするかは，どちらのユーザーの期待購買力が大きいかに依存する。

口コミ解析の将来の課題として，書き込み者のレベル推定，を我々は提案したい。WEB 口

コミを使わないで評判を分析する場合，テスターを雇うか，テストを依頼することで，テスティングしてもらうということが従来から行われていた。その場合でも，こうしたテスターのその商品に対する慣れのレベルは重要であった。分析の信頼性向上のためにはデータ書き込み者のレベルによる区分が必要である。

5．まとめ

本稿では，経営者のための，初期費用がかからず，対費用効果の高い口コミ分析手法を提案した。提案手法を簡単に述べると，(1) 名詞バイグラムのグラフを描き，評判の全容を俯瞰する，(2) 他の商品にない特徴を，TFIDF を用いて発見する，(3)「これが xxx だったら」のような表現から，潜在的事前期待及び改善希望事項を発見する，(4) ホリスティックなコメントを発見する，である。

その適応事例として@コスメのマスカラの口コミ分析を行った。そしてマスカラの潜在的事前期待として，「まつげへのダメージ回避」，「まつげへのトリートメント効果」を，また，ホリスティックな有益情報として「他のベース化粧品との組合せ」などが発見できた。化粧品では商品の評価に「パラパラ」「ベトベト」「うるうる」などのオノマトペが頻出する。評価関係語として，こうしたオノマトペに注目することも重要であろう。また商品のコンセプトを明確に打ち出すため新しいオノマトペを考案するという方策も考えられる。

口コミの語の出現度数分布は，商品の評判を表すものであり，重要な分析対象である。供給者の行う商品の改善とは，口コミの度数分布を思ったように変えることである，と言えよう。提案手法は，語の頻出度数分布を重視する簡便なものである。本稿ではその手法と LSA による分析結果の比較を行ない，分析の本質的な部分は，本提案手法でも把握可能なことを示した。主張は，LDA や LSA のような複雑なトピック抽出手法の前に，十分な全体の傾向把握が必要であり，それは TIIDF などでも行える，ということである。単純な手法のほうが，人間が解釈しやすいという利点もある。LDA や LSA のようなモデルでなくてはできないトピックの表現能力にこだわる前に，扱うデータへの基本的理解が必須であろう。

経営者の本当に欲している情報とは，顧客の潜在的事前期待であり，画期的な商品機能改善や，新商品の開発に結び付く未知の機能，サービスであろう。もちろん，問題や課題には気付いているのだが，技術的ないしはコストが掛かり過ぎて，改善にふみきれない，ということはある。しかし，そのような場合でも，顧客の直接の意見を聞ける WEB 上の口コミ情報は有益である。本稿で我々が提案した分析手法は，簡単な方法ではあるが，こうした経営者のニーズに応えている。経営者にとって対費用効果の高いテキストマイニング法と言える。我々は今後とも，マーケティングの分野でのテキストマイニングを行っていきたい。

参考文献

1. 諏訪良武，北城恪太郎，*顧客はサービスを買っている―顧客満足向上の鍵を握る事前期待のマネジメント*，2009：ダイヤモンド社。
2. 諏訪良武，*サービスサイエンス実践のヒント*，人工知能学会誌，2007. **22**(6)：pp.771-780。
3. 村松真宏，三浦麻子，*人文・社会科学のためのテキストマイニング*，2009：誠信書房。
4. 石田基広，*R によるテキストマイニング入門*，2008：森北出版。

5. 金明哲，*テキストデータの統計科学入門*，2009：岩波書店。
6. Bird, S., et al., *入門自然言語処理*，2010：オライリー・ジャパン。
7. Evangelopoulos, N. and L. Visinescu, *Text-Mining the Voice of the People*. Communications of the ACM, 2012. **55**(2): pp.62-69.
8. Blei, D. M., A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*. Journal of Machine Learning Research, 2003. **3**: pp.993-1022.
9. 橋本隆子，久保山哲二，白田由香利，*ソーシャルメディアを対象としたマーケティング解析―時事問題をきっかけとした想定外の消費行動抽出―*，学習院大学経済論集，2012，47(4)：pp.263-280.
10. 本村陽一，*現場参加型サービス工学―気仙沼～絆～プロジェクトでの気づき―*，情報処理，2014. **55**(2)：pp.161-166。