ブログテキストマイニングによる消費者ニーズの探索的調査手順の再考： KIP の各ステップに残された課題を中心に

Reexamination of the Exploratory Research Procedure of Customers' needs by the Blog Text-mining: On unsolved problems at each step of KIP

加藤 淳一 [1]

Junichi KATO[1]

1 久留米大学 商学部
[1]Department of Commerce, Kurume University

Abstract: The aim of this paper is to arrange the unsolved problems that were pointed out in the development process of KIP or in the experience studies. In this study, the following seven problems are pointed out. The first point is the improvement of dictionary of the morphological analysis mainly on the company and brand names. The second problem is reexamination of data cleaning methods. The third is reconsideration of the methods of characteristic keywords selection including other methods except cosine measure. The fourth point is reexamination of criteria for selection of the characteristic keywords. The fifth problem is the number of author clusters in author analysis. The sixth is reconsideration of the target in nested analysis. The seventh point is the problem in conjunction with the interpretation of the axes of principal component analysis. In future research, we should deal with these problems.

1. 導入

本研究の目的は、市場創造メカニズムの解明に向けて提案された KIP の手順開発および経験研究の蓄積過程で明らかとなった課題を整理することである。KIP は、既に複数の経験研究（例えば、文献[1]など）で使われている。その過程で、開発段階で想定していたよりも多くの課題が指摘されている。だが、それらの課題は個別の研究の今後の課題として指摘されたのであり、このままでは正面から取り組まれることなく散逸してしまいかねない。

ここで、改めて KIP に対して示されている、あるいは著者が自覚している課題を整理する。それにより、今後 KIP の改変に取り組んでいく契機とする。もちろん著者自身もその課題に取り組んでいきたいし、他の研究者が取り組むに当たってもこうした整理があることにより取り組みやすくなる。このように考えられる。

よって、本研究は KIP のこれまでの研究についての知識を与件とする。もちろん、これまでの研究において、何をなしてきたのかを整理することも重要である。今後も折を見て、これらの整理をしていく必要はある。だが既に文献[2]や文献[3]において、このテーマは一応扱ってきた。

これと同様にこれから何をなすのかも重要である。本研究は、KIP の解決されるべき課題に焦点を当てる。とりわけ、KIP の各ステップの課題が検討される。

2. KIP の課題

本研究は、KIP の解決されるべき課題を 7 点に整理する。あらかじめ列挙しておけば、次のようになる。

1. 特に商品名と企業名を中心とした、形態素解析の辞書の充実
2. データクリーニング法の再考
3. コサイン類似度以外の方法を含めた、特性キーワード選択方法の再考
4. 特性キーワード選択基準の再考
5. オーサー分析でのオーサークラスター数の再考
6. 入れ子のクラスタリングの対象の再考
7. 主成分分析の軸の解釈に関連した課題

これらの課題について、現在想定できる解決方法を交えながら整理していく。なお、列挙する順序に重要性などの特段の意図はないものの、原則として KIP の実行手順の順序で説明する。

2.1　形態素解析の辞書の充実

KIP は、ステップ 1（データ収集）において、収集したブログ記事を形態素解析により名詞のみを抽出し、その名詞の使用頻度データを分析の基本としている。この形態素解析では、あらかじめ準備された辞書に登録された単語（名詞）が利用される。よって、この辞書に登録されていない名詞（例えば、新しい商品名や企業名）は、名詞として適切に認識されずにデータとして取り扱われないこともあり得る。

このように考えてくれば、より広範に最新の名詞（商品名や企業名）を含めて辞書に登録するようにしなければならない。これまでのところ、広く知られている辞書の充実方法は次のようである。形態素解析に使用している mecab の辞書(mecab-ipadic-utf8)に Wikipedia 日本語と英語を辞書に入れる。あるいは、他の辞書から単語を抽出して追加する。さらに近年になって、mecab-ipadic-neologd[i]も開発されている。これらが広く知られている。

だが、一般に多様な名詞が登録されるだけでなく、商品名や企業名がより多く含まれることがマーケティングの分析として好ましい。そこで、(例えば、楽天市場[ii]などから）商品名リストを、(例えば、四季報などから）上場企業のリストを取得できると、よりマーケティングの分析にとって有益だと思われる。加えて、一度新しい辞書を入れたということではなく、コマンドで日々辞書をアップデートできると、常に最新の商品名や企業名が反映されてより良い分析を期待できる。

このように、今後の研究において形態素解析の辞書の充実が必要である。とりわけ、マーケティングの分析に関連の深い名詞を充実させる工夫が必要である。

2.2 データクリーニング

KIP は、ステップ 1（データ収集）において、ブログ記事のデータとしての適切さについて評価していない。一般的な調査に倣った表現をするならば、データクリーニングをしていない。よって、そのブログ記事が本当に分析対象とすべきブログ記事か、それとも分析データとしては不適切な記事(例えば、広告記事なのか）を考慮せずに分析している。以下で、適切な記事をクリーンエントリ、不適切な記事をスパムエントリと呼ぶことにする。

すると、これまではデータクリーニングをしていないが、より望ましい手順はデータクリーニングによりスパムエントリを排除してから分析をすべきである。こうした考えで、以前に文献[4]と文献[5]においてデータクリーニングを試みた。ここでは文献[4]に依拠して、そのときのデータクリーニング方法を説明する。その方法は単純であり、図示すると次のようになる。

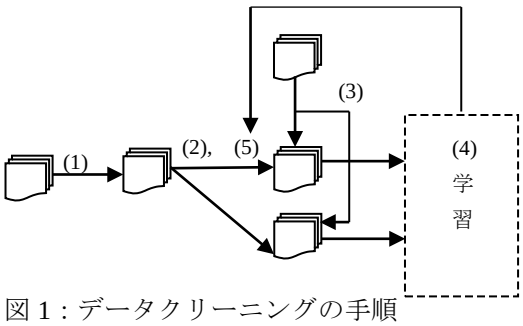

図 1：データクリーニングの手順
出典：文献[4]，p.21，図表 3

この図にもとづいて、以下で説明していく。第 1 に、収集したブログ記事から無作為に人手で処理できる程度のブログ記事(文献[4]では 100 記事)を初期学習データとして抽出する（図中の(1)）。第 2 に、この抽出した記事（初期学習データ）を目視によりクリーンエントリかスパムエントリかに分類する（図中の(2)）。この分類された記事に、過去の研究においてクリーンエントリあるいはスパムエントリと判断されたブログ記事を加えることで、学習データの量をクリーンエントリとスパムエントリの比率を維持しつつ 10 倍に大きくする（図中の(3)）。

第 3 に、ナイーブベイズにより初期学習データを学習する（図中の(4)）。第 4 に、収集したブログ記事からランダムに記事を抽出して、学習に基づいてその記事がクリーンエントリの確率と、同じその記事がスパムエントリである確率を計算する。各記事は、確率の比較によりクリーンエントリかスパムエントリに分類される（図中の(5)）。第 5 に、次のブログ記事の抽出と分類の前に、分類された記事は学習データとして学習される（図中の(4)）。

第 6 に、全ての記事について、記事毎のクリーンエントリの確率とスパムエントリの確率が計算された後で、それら確率をブログオーサー単位で算術平均する。この平均確率の閾値以上に大きいブログオーサーがスパムを発生させるブログオーサー（スパムオーサーと呼ぶ）と判定される。

実際に実行すると、収集されたブログ記事の中で、スパムオーサーはほぼゼロという結果になった。図中の(4)において、次のブログ記事の分類の前に、分類された記事を学習データとしている。これにより、学習データへ強く適合した可能性が考えられる。

2.3　特性キーワードの選択方法

KIP は、ステップ 2 と 3（特性キーワードの選択）において、商品特性キーワードとオーサー特性キーワードを選択している。商品特性キーワードの選択方法として、tf・idf 値を要素とした単語・ブログ記事行列において、ターゲット・キーワードと各単語との間のコサイン類似度の大きい単語がよりターゲット･キーワードに近い単語として選択されている。だが、よりターゲット・キーワードに近い単語の選択方法は tf・idf 値やコサイン類似度以外にもあり得る。

例えば、tf・idf 値には様々あり、文献[6]で言及されている residual idf（残差 idf）のような統計的な方法も考えられる。文献[6]によると、residual idf（残差 idf）はポアソン分布が一般語に対しては当てはまり、キーワードに対しては当てはまらないという考えを利用した方法である。また、ターゲット・キーワードとの関連性は、コサイン類似度以外にも例えば text network analysis など他の指標も候補となり得る。

これらその他の方法を使う場合と比較して、tf･idf 値とコサイン類似度のほうが適切な方法なのか理論的根拠は検討されていない。これまでのところ広く一般に tf・idf 値とコサイン類似度が用いられているという判断にもとづいてこれらを利用してきた。だが、他の基準を最初から排除する必要はない。これからターゲット・キーワードに近い単語を選択する指標として適切な方法を検討すべきである。

2.4　特性キーワードの選択基準

前節と同じく、ステップ 2 と 3（特性キーワードの選択）で設定している特性キーワードの選択基準も再考の対象とすべきである。ここで再考の対象とすべきは 1 つに商品特性キーワードの閾値の設定基準であり、もう 1 つにオーサー特性キーワードの基準である。

商品特性キーワードは、次のような基準で決定されている。tf・idf 値を要素とした単語・ブログ記事行列において、ターゲット・キーワードと（収集されたブログ記事から形態素解析により取り出された）各単語の間のコサイン類似度を計算する。この各単語のコサイン類似度を大きい方から順に並べる。その順で、コサイン類似度の累積相対度数が 0.8 以上の単語までを商品特性キーワードとしている。

なお、初めて累積相対度数が 0.8 を超えた単語を商品特性キーワードに含めている。加えて、累積相対度数が 0.8 を初めて超えた単語と同じコサイン類似度の単語も商品特性キーワードに含めている。しかし、累積相対度数で 0.8 以上という基準は何ら理論的根拠を持っていない。

オーサー特性キーワードの個数は、これまでのところ適切な方法が考えつかず、無根拠に商品特性キーワードと同一の個数としている。しかしこれは不適切な決定方法であるから、オーサー特性キーワードの個数を理論的に決定しなければならない。

2.5　複数のクラスター数で自動的に分析

KIP は、ステップ 4 あるいは 5（オーサークラスタリング）において、ブログオーサーを SOM により 4 つの集団に分割している。だが、この 4 という数字はパラメータである。その根拠は理論的ではなく現実的な理由による。ここで現実的な理由とは、次の 2 つである。1 つに、人間が認識しやすい程度の数である。2 つに、あまり多くの集団数にすると、各集団に含まれるブログオーサーの人数が少なくなりすぎる。こうした現実的な理由で、適切と思える個数として 4 としている。

だが、繰り返しておけば、パラメータであり理論的根拠はない。よって、理論的根拠を持って設定できるならばすべきである。もしも仮に理論的根拠を設定するのが困難であるならば、4 つだけでなく複数の集団の数で分析して、後から集団の数を選べるようにしておくべきと考える。アイデアとしては、現行の 4 つに加えて、自動的に 8 個の集団、16 個の集団､そして 32 個の集団に分割するようにプログラムを作成する。あるいは、分析を開始する時点で、集団数を引数で指定できるようにプログラムを作成する。このような改変が考えられる。

2.6　オーサー特性キーワードでのクラスタリングを全ての集団に対して行う

KIP は、ステップ 4 あるいは 5（オーサークラスタリング）において、ブログオーサーを SOM により 2 度に亘り集団に分割している。まず、商品特性キーワードの使用頻度の近いブログオーサーが同じ集団へと、tf・idf 値を要素とした行列のコサイン類似度によりバッチ型 SOM (Self-Organizing Map)の Batch Map で、4 つの集団に分割される。次の図は全ブログオーサーを 4 つの集団に分割したことを示している。

ブログオーサー集団 a	ブログオーサー集団 b	ブログオーサー集団 c	ブログオーサー集団 d

図 2：全ブログオオーサーのクラスタリング
出典：著者作成

その後に、この分割した集団毎に入れ子に、人物特性キーワードで SOM によりブログオーサーを集団に分割する。この入れ子のクラスタリングは、現在のところロイヤルティの最高と最低の 2 つの集団のみに対して行っている。ここでロイヤルティの最高の集団と最低の集団は、tf・idf 値を要素としたブログオーサー・単語行列をもとに、次の 3 ステップによる相対比率の平均値の計算により決定している。

ステップ 1 は、各ブログオーサーがブログ記事で用いた単語の中で、商品特性キーワードと一致する単語の tf･idf 値の相対比率を求める。ステップ 2 は、4 つの集団ごとに、そのブログオーサーごとの相対比率の値の高い方から 25%を算術平均する。ステップ 3 は、4 つの集団の中で、この平均値の最大の集団をターゲット・キーワードへのロイヤルティの最高の集団とした。平均値の最小の集団がロイヤルティの最低の集団とした。これらである。次の図は、仮にロイヤルティの最高の集団を集団 b、最低の集団を集団 d としたときに、それぞれの入れ子のクラスタリングを示している。

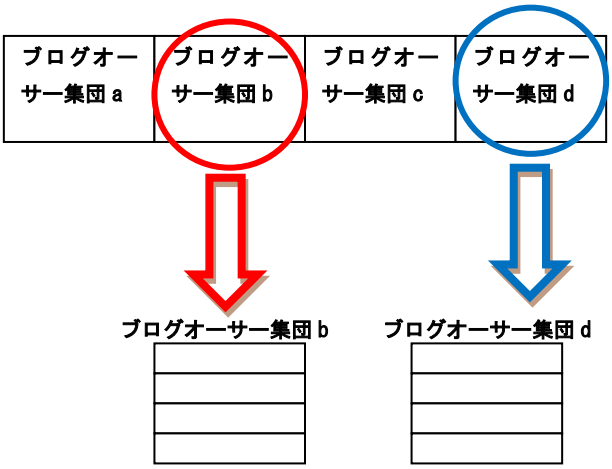

図 3：入れ子のクラスタリング
出典：著者作成

これら 2 つの集団のみへ入れ子分析の実行は、次のような想定を背景にしている。ここでの想定とは、商品特性キーワードで集団に分割したときに、各集団に属するブログオーサー数がロングテイルの分布になっている。言い換えれば、ロイヤルティが最高あるいは最低の集団に属するブログオーサーの人数が他の集団に属するブログオーサーの人数と比較して多いという想定である。

だが実際には、このような分布になっているのかは、実際のデータを見てみなければ判断できない。にもかかわらず現行の KIP は自動的にロイヤルティの最高と最低の 2 つの集団のみに対して入れ子のクラスタリングを実行している。仮に、正規分布のような分布であれば、多くのブログオーサーを分析の対象から除外していることになる。これは適切とはいえない。よって自動的な分析を想定するならば、商品特性キーワードで集団に分割した後、全ての集団に対してオーサー特性キーワードでの入れ子分析を行う。このようにするプログラムへ改変する。

2.7　主成分分析と軸の解釈に関連した課題

KIP は、ステップ 6（ラベリング）において、各主成分軸の解釈を行っている。主成分軸の解釈は、次のような課題がある。1 つに、あまりに多数の単語が出てきても人間の力では解釈が困難になることと、プログラムの計算にも時間がかかることから、主成分分析を行う前に対象となる単語をカイ二乗値により絞っている。だが、せっかく多数の単語を収集・分析しているのであるから、これらの単語を少なくするよりも活用する方向を探りたい。

2 つに、カイ二乗値を計算するという手順の流れから、頻度データを主成分分析に使うのではなく、カイ二乗値を主成分分析のデータとして用いている。だが、主成分分析のデータとして、カイ二乗値の使用を正当化する理論的な根拠はない。したがって、頻度データを主成分分析のデータとすべきである。

3 つに、主成分軸の解釈に固有ベクトルの構成数値を用いている。一般に主成分分析により求められた固有ベクトルの構成数値の絶対値の大きさは、新しい軸への影響の大きさを示している。よって、絶対値が大きいほど、新しい軸へ大きな影響を与えていると理解して軸の解釈を行う。ただし、その固有ベクトルの構成数値の符号には、プラスにポジティブそしてマイナスにネガティブな意味はなく軸の反対方向という意味しかない。

よって、軸の解釈には次のようなステップを踏むことになる。まず、マイナス（あるいはプラス）方向から絶対値の大きな単語を手がかりにして、マイナス（あるいはプラス）方向の軸の解釈を行う。次に、プラス（あるいはマイナス）方向から絶対値の大きな単語を手がかりにして、プラス（あるいはマイナス）方向の軸の解釈を行う。最後に、これら双方の軸の解釈から、軸全体のメタな解釈を行う。こうして軸の解釈を進めていくべきである。しかし実際には、プラス方向あるいはマイナス方向どちらかの解釈をするだけでも困難であり、メタな解釈までできていない。

そこでこの状況を解決すべく工夫しなければならない。その試みの一つとして、要約技術の活用が考

えられる。具体的には次のように考える。まず、プラス方向から絶対値の大きな単語を任意の数（例えば、50 語）だけ抽出する。同様に、マイナス方向からも同数の任意の数（例えば、50 語）だけ選択する。こうして選び出された単語（例えば、合計で 100 語）について、元のブログ記事から、これら選ばれた単語（例えば 100 語）のそれぞれをより高い使用頻度で含んでいるブログ記事を選び出す。

ここで一つの工夫として、使用頻度だけでなく、固有ベクトルの構成数値の絶対値の大きい単語ほどウェイトをかけることが考えられる。選ばれたブログ記事は、人手で処理するにはあまりに多いと思われる。そこで要約技術の活用により、この選ばれたブログ記事を短い文章へと要約する。そして、要約された比較的短い文章を読むことにより、新しい軸の解釈を行う。このような方法によれば、多数の単語から主成分軸の解釈を行うという困難な作業から比較的短い文章の意味を理解するという作業へ転換できる。

3. まとめ

最後に、本研究の目的に立ち返り、本研究をまとめる。本研究の目的は、「市場創造メカニズムの解明に向けて提案された KIP の手順開発および経験研究の蓄積過程で明らかとなった課題を整理すること」であった。この目的の下で、本研究では KIP の各手順に残された課題を 7 点に整理した。

ふり返ると、(1)特に商品名と企業名を中心とした、形態素解析の辞書の充実、(2)データクリーニング法の再考、(3)コサイン類似度以外の方法を含めた、特性キーワード選択方法の再考、(4)特性キーワード選択基準の再考、(5)オーサー分析でのオーサークラスター数の再考、(6)入れ子のクラスタリングの対象の再考、(7)主成分分析の軸の解釈に関連した課題である。

これら 7 つの課題の一つひとつに、今後の研究で取り組まなければならない。それぞれの課題は、これまでの研究でその都度指摘されてきた。だがそのままでは散逸しかねなかった。そこで、本研究において改めて整理し、今後これらを解決する基礎とする。

参考文献

[1] 加藤淳一: ブログテキストマイニングによる海外観光都市に関する消費者ニーズの探索的調査：モナコ公国を事例に, つくば国際大学研究紀要, Vol. 19, pp. 35-50, (2013)

[2] 加藤淳一: マーケティング学説史における KIP（加藤・石川　手順）の位置づけ, つくば国際大学研究紀要, Vol. 21, pp. 1-25, (2015)

[3] 加藤淳一: ブログテキストデータの探索的な調査による市場創造研究の胎動：　事業の定義と KIP の関連を中心に, 統計数里研究所共同研究リポート, Vol.360, pp. 18-28, (2016)

[4] 加藤淳一, 今西衛: ディズニーランドに対する消費者ニーズの探索的探求, 経営情報学会全国研究発表大会要旨集, pp. 19-22, (2012)

[5] Kato, J., Imanishi, M.: An Analysis of Customers' Needs for Smartphone Markets by Using Blog Data and Improvements of Kato & Ishikawa's Procedure from Bayesian Viewpoint, 2012 年秋季研究発表会アブストラクト集(日本オペレーションズ・リサーチ学会), pp.70-71, (2012)

[6] 遠藤雅樹, 横山昌平, 大野成義, 石川博: 特定地域に限定しない観光キーワードの自動抽出, DEIM Forum 2014,　(2014)

---

i https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md　最終アクセス日：　2016 年 8 月 14 日

ii http://directory.rakuten.co.jp/あるいは https://webservice.rakuten.co.jp/api/ichibaitemsearch/ 最終アクセス日：　2016 年 8 月 14 日