DEIM Forum 2022 E41-1

相互的ブートストラッピングによる マイクロブログからの購買報告の効率的な収集

平山　佑夢† 　荘司　慶行† 　Martin J. Dürst†

† 青山学院大学理工学部情報テクノロジー学科　〒 252–5258 神奈川県相模原市中央区淵野辺 5–10–1
E-mail: †hirayama@sw.it.aoyama.ac.jp, ††{shoji,duerst}@it.aoyama.ac.jp

あらまし　本論文では，マイクロブログから「新しいおにぎり GET！うまい！」などの購買報告を効率的に収集するクローリング方法を提案する．SNS 上の他人の購買報告は個人の購買に関する意思決定に影響を与えることが知られている．一方で，これらの購買報告は，SNS 投稿中で占める割合が少なく,「買った」などの直接的に購買を示す語句や，正確な商品名を含まない場合が多いため，網羅的に収集することが困難である．そこで本研究では，購入を示唆する語句からなる辞書と，商品名らしい名詞からなる辞書を，相互再帰的に拡張するブートストラッピング手法を提案する．こうして得られた単語で検索することで，より少ない API アクセス回数で，効率よく購買報告を収集する．実際に Twitter のデータセットを用いた実験から，提案手法が実際により多様な購買報告を，少ないアクセス回数で収集できたことを示した．

キーワード　マイクロブログ，Twitter，ブートストラップ，情報抽出

1 はじめに

近年では，テレビのコマーシャルや新聞や雑誌の広告ではなく，ソーシャルメディア上の投稿をきっかけに商品を購入することが，当たり前になってきている．例えば，85%以上のソーシャルメディア利用者が,「ソーシャルメディア上の情報が商品購入につながった」と答えたという報告がある [1]．くわえて，スーパーマーケットなどにおける日常生活における消費においても，ソーシャルメディアの影響が大きいことも指摘されている [2]．

一般的なマーケティングの研究分野においても，ソーシャルメディアの購買行動への影響は重要視されており，人の購買モデルとして旧来の AIDMA（Attention，Interest，Desire，Memory，Action）モデルに対して，ソーシャルメディアを考慮した新しいモデルが登場してきている．たとえば近年提案されている購買モデルである AISAS（Attention，Interest，Search，Action，**Share**）や AISECAS（Attention，Interest，Search，Comparison，Examination，Action，**Share**）などでは，ソーシャルメディア上での「Share（共有）」行動が，購買行動において強い影響を持つとしている．

このように，現代人が何か商品を購入するかを決定するうえで，ソーシャルメディア上での誰かの意見を参考にすることは，一般的である．例えば，実際に購入した人や，購入をためらった人の意見は，購入していない人には知ることのできない考えを含む．そのため，買うかどうかの判断の参考になる場合が多い．インターネット上でのレビューサイトや，ソーシャルメディアが普及し，誰もが利用者の立場で情報発信をできるようになった現代では,「口コミ」や「レビュー」はウェブ上にあふれかえっている．一方で，これらの膨大な口コミ情報は，様々なソーシャルメディア上に散発的に存在しており，商品ごとにまとめて閲覧したり，網羅的に閲覧することが困難である．

例えば，家電などの電化製品は，様々な広告で大々的に売り出されているため，ネットショッピングで購入する人も多い．そのため，Amazon[3]や楽天 [4]などの通販サイトでも多くのレビューを見ることが可能である．しかし，ネットショッピングで買うことのできないもののレビューは探すことが難しい．例えば，コンビニエンスストアで販売されるスイーツなどの新商品のレビューは通販サイトで見ることは難しい．

こういった実際に購入した消費者の意見と出会う場として，レビューサイトはもとより，マイクロブログも一般的である．SNS が広く普及している現代において，Twitter や Facebook などのマイクロブログは日常的に多くの人によって使用されている．例えば Twitter には,「新しい服を買った！」のような，日常的な購買報告が多数投稿されている．このような購買報告には，通販サイトや，単独のレビューサイトには含まれない商品への感想も多く含まれている．例えば，コンビニエンスストアの新商品などの，実店舗限定の商品は，通販サイトのレビューに登場することはない．また，多くのレビューサイトでは，家電などの，型番を持つような製品を主に対象としている．100 円ショップで見つけた新商品や，正式な名前もついていないような衣類などに関する情報は，レビューサイトだと少ない．

1：株式会社ジャストシステム：Marketing Research Camp 「SNS 広告は商品購入に影響する？最も効果が高いのは Instagram?LINE?メルマガ？」 https://marketing-rc.com/article/20180319.html 2018 年

2：一般社団法人 全国スーパーマーケット協会：スーパーマーケット白書 2017 第 3 章 http://www.super.or.jp/wp-content/uploads/2017/02/hakusho2017-3_4.pdf 2017 年

3：https://www.amazon.co.jp

4：https://www.rakuten.co.jp

このような投稿は，偶然目に触れると役に立つ場合が多いが，自分から探そうとすると，困難である．たとえば，Twitter で，購買報告を多く収集しようとする場合を考える．Twitter にはキーワード検索機能があるが，このような機能を用いても，購買報告を発見することは難しい．この際，購買報告を探そうとして,「買った」などの直接的なキーワードで検索しても，多くの購買報告は発見できない．これは，商品を買った人が皆「○○を買った」と直接的な表現をするわけではないためである．加えて，特定の商品に関する購買報告を探そうとして，商品名で検索した場合も，同様である．個人が日常的なつぶやきの中で，正式な商品名を省略や間違いなしに投稿することは稀である．

そもそも Twitter は，日常の出来事に関する何気ない呟きをするユーザが多いため，口語的な表現を用いた投稿が多い．具体的な例を挙げると，新商品を買った際に,「セブンイレブンで,『シビれる辛さと香り引き立つ麻婆豆腐焼そば』を買った．美味しい！」というように，商品の正式名称や「買った」というキーワードを含んだ Tweet を投稿することは少ない．実際には,「セブンの新しい焼きそば，うま！」などの，曖昧な書きかたをする場合が多い．そのため，単純にキーワード検索を用いても，購買報告を網羅的に収集することができない．

購買報告を収集するうえで，検索によるアプローチと，抽出によるアプローチが考えられる．検索によって多くの購買報告を得ようとすると，検索回数の制限によって，網羅性に限界が生じる．多くのマイクロブログサイトでは，検索回数にシステム的な上限を設けている場合が多い．例えば，Twitter では，キーワード検索を 15 分間に 180 回しか実行できない [5]．そのため,「買った」,「購入した」などの購買を示唆するキーワードで次々と検索していくと，すぐに検索可能な回数の上限に達する．一方で，抽出によるアプローチでも，量的な限界が存在する．高度にトレーニングされた分類器を用いて，あるツイートが購買を示唆するかを判断することは比較的容易である．しかしながら，そもそも，こういった購買報告を含む Tweet は，全 Tweet の 1%にも満たない．そのため，無作為に投稿を集め，それを後から抽出したとしても，十分な数の購買報告を得ることは難しい．

そこで本研究では，検索的なアプローチと抽出的なアプローチを組み合わせ，効率的に購買報告を収集する方法を提案する．このアプローチでは，購買報告を含む投稿を多く取れそうなキーワードで検索を行い，その検索結果から購買報告を抽出する．そして，その抽出結果を用いて，次に検索する際の検索キーワードを決定する．この際に，購買を示唆するキーワードからなる辞書と，商品名などのキーワードからなる辞書を，相互再帰的に交互に更新する，相互的なブートストラップ法を用いる．

ここで提案する相互的なブートストラップ法は，購買フレーズと商品名の 2 つの辞書を用いて，相互的に両方の辞書を構築していく手法である．ブートストラップ法は，知識抽出などの分野において古典的に用いられる，少数のシードから大量の語彙を自動獲得するための手法である．従来のブートストラップを用いた辞書構築では，一般的に 1 つの辞書の語彙を増やしていく．一方で，本手法では，購買を示唆しそうなキーワードとは別に,「何を買ったのか」を表しそうなキーワードについても同時に扱う．2 つの辞書を相互的に更新していくことで，より効率良く辞書構築を行うことを目指す．

このような相互的なブートストラッピングによる購買報告収集を，実際に Twitter に対して行った．実際の実験では，2018 年からランダムに収集したおおよそ 20 億件の Tweet を，実際の Twitter と同様にキーワード検索できるようにして，シミュレーションを行った．実際の分類器として，購買報告の判別のためにファインチューニングした BERT （Bidirectional Encoder Representations from Transformers）分類器を用いた．この分類器では，$F_1 = 0.91$ の精度で購買報告を分類することができる．これらを組み合わせて，実際に，購買を示唆するフレーズと，商品名の辞書を作成する実験を行った．そして，時間あたりの収集効率や，実際に収集できた購買報告の質や種類について分析した．また，実際に Twitter の API を利用して，現在の新しい購買報告 Tweet を収集可能であるかについても実験予定である．

本論文の構成を述べる．第 2 節では，関連研究を紹介する．第 3 節では，実際の購買報告の収集手法について説明する．第 4 節では，収集精度，収集量に関する評価実験について説明する．第 5 節では，評価実験の結果について考察する．最後に，第 6 節では本研究のまとめと今後の課題について述べる．

2 関 連 研 究

本研究は，Twitter からの情報収集に関する研究と，効率的な収集手法に関する研究と深く関係する．また，本研究では購買報告を Twitter から収集するため，マイクロブログにおける購買情報の研究ともいえる．そのため，第 2.1 節では Twitter からの情報収集について，第 2.2 節では効率的な情報収集法，について，第 2.3 節では SNS における購買行動について書く．

2.1 Twitter からの情報収集

SNS の普及により，Twitter を代表とするマイクロブログから情報収集を行う研究が増えてきている．本研究における「購買報告」のように，あるトピックに関する投稿を網羅的に収集して分析する研究は広く行われている．例として，Allison ら [1] は，Twitter から電子たばこに関する投稿を収集することで，電子たばこに関する洞察やテーマを明らかにしている．また，Mrityunjay ら [2] は，Twitter からコロナウイルスに関する投稿への「いいね数」,「RT 数」などを収集することでセンチメント分析を行なっている．このように，近年では，Twitter は社会の動向を反映しているため，Twitter を情報源として活用する研究が一般的になってきている．一方で，これらの研究では，データ収集の方法自体は，キーワードを含んだ Tweet を

5：Twitter「Rate limits — Docs — Twitter Developer」：
https://developer.twitter.com/ja/docs/basics/rate-limits

利用するという，単純な方法を用いている．

また，本研究では，辞書を用いることで，任意の語彙を含む投稿を抽出している．語彙を利用し，Twitter でのセンチメント分析を行う研究は今までにもあった．様々な語彙表現を含む辞書を用いることで，Twitter の情報をより高度に活用する研究も行われている．センチメント辞書に注目した例として，Lei ら [3] は，Twitter を対象とした新しいエンティティレベルの感情分析手法を提案している．製品レビューのような大規模なコーパスを対象としていた現在の感情分析手法とは違い，語彙ベースのアプローチを採用し，エンティティレベルのセンチメント分析を行った．さらに，再現率を向上させるために，語彙ベースの手法の結果に含まれる情報を利用して，意見を述べている可能性の高い追加のツイートを自動的に特定するという手法を用いている．このように，Twitter でのセンチメント分析における手法は研究され続けている．

また，本研究では，ブートストラップ法を用いることで辞書を構築し，それを用いて非明示的な購買報告を収集する．本研究ではブートストラップ法を用いることで辞書を更新する手法を用いたが，別の方法で網羅的にソーシャルメディアから情報を収集し，分析している研究もある．例として，James ら [4] は，時間軸上に整列した文書群から同じトピックに関する文書を見つけるために Topic Detection and Tracking タスクを提案している．また，Yan ら [5] は同じトピックに関する投稿のクラスタリングやトピックに抽出を行うために，ある組織に関する投稿から，その組織に関する新しいトピックを特定するという手法をとった．Padmanabhan ら [6] は，あるツイートを入力とし，それに関連する投稿を余弦類似度を用いることで算出するという方法を用いている．

これらの研究は，どれも何かのキーワードに対して，非明示的に関連する投稿を収集する手法を提案している．しかし，api の制限回数から，常に収集したい投稿を効率的に収集し続けるということはできずにいた．本研究では，購買報告と思われる投稿を，非明示的な投稿も含めて精度よく収集し続けることを目的とする．しかし，Twitter にある投稿というのは口語的な表現が用いられやすく，その投稿が何に関する投稿なのかを判断することが難しい．

そこで，Edgar ら [7] は，Twitter における投稿が何についてのものであるかを明らかにする手法を提案した．任意の投稿について，意味論的に関連する概念を自動的に特定し，対応する Wikipedia の記事へのリンクを生成した．これによってソーシャルメディアのマイニングなどに各投稿を利用することができる．このように，Twitter における情報抽出，データ分析の技術は研究され続けている．

2.2 効率的な購買報告の収集手法

本研究では，マイクロブログから効率的に購買情報を収集するために，検索キーワードからなる辞書を構築する．効率的に語彙を収集する手法として，ブートストラップ法が挙げられる．ブートストラップ法とは，知識抽出などの分野において用いられており，最初に入力として与えた少数のシードをもとにして，大量の語彙を自動獲得することができる手法である．本研究では，商品名と購買フレーズの 2 つの辞書を用いて，投稿を収集する．

Sergey ら [8] は著者, タイトルの小さなペアのシードセットを作成し，ウェブ上でそれらの書籍の出現箇所を探した．その箇所から書籍の引用に関するパターンを認識し，それを繰り返すことで多くの書籍とそれを見つけるためのパターンリストを獲得した．著者とタイトルという 2 つの要素を拡張していく点が本研究と似ているが，本研究では，商品名と購買フレーズのそれぞれで別の辞書を作成することで，相互的なブートストラップ法を可能にしている．

河合ら [9] は，Web 知識を利用したブートストラップによる辞書増殖を行なった．従来の手法では大量の学習データや言語的な知識などの事前準備や事前知識が必要であったが，この手法では，Web ページから語彙を抽出するパターンを単純化し，尚且つブートストラップ法を採用することで少量の語彙から大量の語彙を獲得した．また，CESS フレームワークにおけるブートストラップ式辞書構築アルゴリズムを実装し，その有効性を示した．

このように，ブートストラップ手法を用いることで，少数の語彙から大量の語彙を獲得することを実現できる．しかし，本研究のように，マイクロブログからデータを収集する場合は，口語的な表現や絵文字が用いられることも考慮する必要がある．

例えば，Xiaomei ら [10] は，Twitter でのセンチメント分析を行う際に，新たな感情分析モデルを提案した．Xiaomei らは，従来のマイクロブログを独立した同一分布だとする考え方ではなく，マイクロブログはネットワークデータであると主張し，社会的文脈とトピック文脈を組み合わせた新しい手法を提案した．これにより,「(笑)」や「lol」などの表現スタイルが違う語彙の疎通を可能にした．

また，Fei ら [11] はマイクロブログを分析するにあたって，単語やマイクロブログの投稿を絵文字空間に投影することでマイクロブログ環境における主観性，極性，感情の識別を行う絵文字空間モデルを提案した．従来のセンチメント分析では，いくつかの絵文字以外のほとんどの絵文字はノイズとして扱われ，感情的意味を持っていないとされていた．Fei らはこれを解決する新たなモデルを提案した．このように，マイクロブログでの情報収集の効率性を高めるためには，効率的にデータ収集を行うだけでなく，収集したデータを効率良く活用する手法が必要である．

2.3 SNS における購買行動

本研究は，マイクロブログ上の購買情報にフォーカスした研究と深く関連する．

Yue ら [12] は，マイクロブログのコンセプトと特徴を考察し，それに基づいてマイクロブログにおけるマーケティング戦略を分析した．マイクロブログを活用したマーケティングへの示唆を提示することで，マーケティングにおけるマイクロブログの有用性を主張した．

このように，マイクロブログをマーケティングに活かす研究

はされ続けている．実際に，マイクロブログ上の意見を収集し，商品に反映させている企業も存在している．しかし，そういった企業の数はまだ多くなく，オンラインマーケティング環境の整備が重要視されている．

Weishi ら [13] は，企業のオンラインマーケティング環境を分析するための，ソーシャルネットワーク分析技術の重要性を主張した．また，マイクロブログ・マーケティングの利点とビジネス価値を分析し，情報収集や情報発信の観点から，企業がマイクロブログで戦略的にマーケティングを実施するためのフレームワークを提案した．

本研究では，購買行動にフォーカスした研究を行なっているが，マイクロブログ上の情報を，イベントマーケティングに利用した研究も存在する．

Xin ら [14] は，博物館，美術館，展示場などのイベント主催者が，効果的なイベントマーケティング戦略を求めてソーシャルメディアツールを利用するようになっていることに着眼し，マイクロブログのデータを分析することで，実際に北京市における美術館の Weibo 利用による集客効果を評価した．

本研究では，マイクロブログ上の購買報告を高精度で収集する．本研究では行わないが，その後，収集した購買報告の中身を分析することで，初めてマーケティングに活かすことができる．

Eduard ら [15] は，良い評価をされてる商品は売れやすく，悪い評価が多い商品は売れにくいという考え方からウェブ上の製品やサービスに関するレビューやコメントから，単語の極性を推定した．既知の極性を持つ単語を入力とし，極性を持つ同義語の集合を生成する推論ルールを導入することで，他の単語の極性を推論し，マイクロブログに存在するレビューの極性を判断することを可能にした．

このように，マイクロブログとマーケティングは密接に結びついている．マイクロブログの力を最大限に活かすために，ソーシャルネットワーク分析における技術は研究され続けている．

3 提案手法

本節では，Twitter から効率的に購買報告を収集する手法について提案する．提案手法では，相互的なブートストラップ法を用いて，購買報告を抽出するための辞書を拡張していく．ブートストラップ法とは，少数の単語リストと文書からパターンを自動作成し，このパターンを使って文書から単語を抽出し，抽出した単語を使ってさらにパターンを自動作成する手法である．このような処理によって，少数の入力の単語リストを雪だるま式に大量に増やすことができる．

本研究における提案手法は，実際に収集した Twitter データにあわせたものである．対照するデータの注意として，収集したデータはリツイート（再投稿）やリプライ（返信）を取り除いた，日本語で書かれた投稿である．リツイートを除外したのは，同じ文面の投稿が複数存在すると，購買報告の分類に過学習が生じるためである．また，リプライに関しては，事前実験から購買報告がほとんど含まれないことが示唆されたので，対象から除外した．

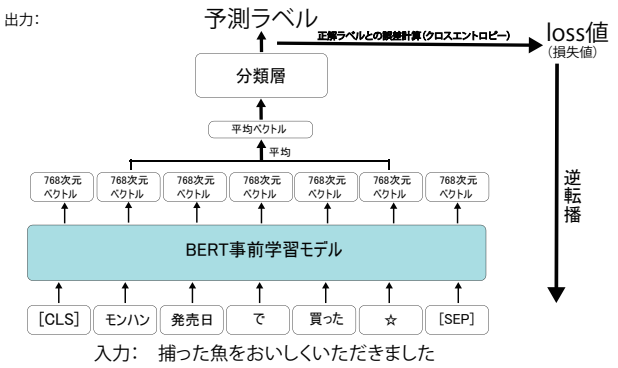

図 1　分類器学習の流れ

3.1 手法の概要

本手法では，シードとなる初期辞書をもとに，2 つの辞書を同時に拡張してゆく：

- **購買フレーズ辞書** :「買った」「ゲット」などの購買を示唆する単語からなる辞書，および
- **商品名辞書** :「新作」「ソフト」などの買ったものを示唆する単語からなる辞書．

それぞれの辞書に単語を追加する際に，購買フレーズで検索した検索結果から商品名の候補を，商品名で検索した検索結果から購買フレーズ候補をそれぞれ発見する．

本手法では，発見された単語を辞書に加えるかの判断に，購買報告かどうかを分類するためにファインチューニングされた BERT 分類機を用いる．このようなアプローチをとった理由として，購買報告が，分類することは容易だが，収集することは困難な性質を持っていることが挙げられる．Tweet の 1 件 1 件を，BERT 分類機で，それが購買報告であるかを分類することは極めて容易である．事前実験の結果,「買った」などの直接的な表現を含まない Tweet を含むデータセットでも，ファインチューニングされた BERT 分類機は $F_1 = 0.91$ の精度で購買報告を抽出可能であった．

一方で，購買報告は，Twitter に投稿される Tweet のうち，たかだか 0.85%であることが確認された．このような状況下では，Tweet を無作為に抽出して，それらから購買報告を抽出するアプローチでは，十分な量の購買報告を収集できない．そこで，購買報告をより多く含む検索結果を得られそうな単語を使って検索を行い，その検索結果から分類機で実際の購買報告を抽出し，得られた購買報告から次の検索キーワードを抽出するというアプローチをとる．

3.2 分類機の学習

あるキーワードで検索した Tweet 群があった際に，その検索結果にどれだけの購買報告が含まれているかを判断するために，BERT を用いて 2 値分類を行う分類機を学習した．図 1 は分類器の学習を示す．この分類機は，ある 1 件の Tweet の本文を入力すると，その Tweet が購買報告であるかを 0 か 1 かで

分類する．この分類機は BERT を用いて作成されており，今回の分類タスクに向けてファインチューニングを施した．

学習のために，無作為に抽出した Tweet に対して，人手で購買報告であるかを 0 か 1 かでラベル付けした．そして，日本語の学習済み BERT モデル [6]に対して，人手で作成した正解データをもとにファインチューニングを行った．

学習時には，誤差計算にロス値としてクロスエントロピー $H(p.q)$ を

$$H(p,q) = -\sum_{x}^{p}(x)\log(q(x)) \qquad (1)$$

として計算した．こうして計算されたロス値をモデルに逆伝播することで，予測の精度を上げた．何度か学習を行なったのち，一番精度の高かったモデルを分類器用のモデルとして採用した．

3.3 シードとなる辞書の人手での構築

シードとなる辞書の人手での構築について説明する．本手法では，商品名辞書，購買フレーズの 2 つを，相互的なブートストラップ法を用いて拡張する．そのため，初期段階で少数のシードとなる単語を用意し，辞書を構築する必要がある．

そのために，実際に人手で Twitter 上でキーワード検索を行ない，収集した投稿に対して，購買報告か否かラベル付けを行った．一定以上の割合で購買報告と思われる投稿を収集することができていた場合，それを有効なキーワードと考え，シードに用いた．

3.4 購買フレーズを用いた商品名辞書の拡張

購買フレーズを用いた商品名辞書の拡張方法について説明する．はじめに，購買フレーズ辞書で検索を行い，投稿を収集する．その後，収集した購買フレーズを含む投稿を，BERT 分類器を用いて分類し，購買報告と判断された投稿のみを抽出する．購買報告と判断された投稿に対して，形態素解析を行い，形態素ごとの頻出度を算出する．頻出度が上位 500 の名詞をまとめ，商品名辞書に追加する単語の候補を作成する．この際，単語の頻出度のみを考慮した場合，一般的な日本語に多く含まれやすい指示語などが上位に出現する．また，Twitter の投稿に高頻度に見られる単語も現れる．これらを取り除くため，あらかじめ Twitter での商品名収集用のストップワードを作成する．

商品名収集用ストップワードを作成するために，ランダムに収集した投稿約 300,000 件のテキストから絵文字や改行を削除する前処理を行い，形態素解析を行う．その後，頻出度が上位 1,000 個の名詞のみを，購買報告収集用のストップワードとした．

商品名辞書に追加する単語の候補の中で，あらかじめ作成した専用のストップワードに含まれるものを取り除く．残った名詞それぞれで投稿を抽出し，それらを BERT 分類器によって購買報告か判断する．抽出した投稿の内，購買報告の割合が 50% 以上だった名詞のみを辞書に追加する．

また商品名に関しては，単純な名詞ではないものもある．例えば,「耳うどん」のような商品名は 1 つの名詞ではない．こういった単語を収集するために，商品名と思われる連続語を辞書に追加する．BERT 分類器によって購買報告と判断された投稿に対して形態素解析を行う．その後，連続する 2 品詞の中に助詞，助動詞が含まれない，内 1 つ以上が名詞である，かつ，2 文字以上の品詞が，購買辞書内の単語の前後にある連続語を候補とする．また，その連続語の中に「する」と含むものは 1 つの単語を示しているとは考えにくいため，取り除く．こうして作成した連続語で検索を行い，収集できた投稿数によって辞書に追加する連続語かの判断を行う．投稿数が 5 件未満の場合は辞書に追加せず，5 件以上 20 件以下の場合は，候補となっている連続語と購買辞書内の単語が共起する割合によって判断する．共起する割合が 50%以上かつ，分類器による購買報告の割合が 50%以上の連続語を辞書に追加する．21 件以上の時は，共起する割合は問わず，分類器による購買報告の割合が 50%以上の連続語を辞書に追加する．

3.5 商品名を用いた購買フレーズ辞書の拡張

商品名を用いた購買フレーズ辞書の拡張方法について説明する．はじめに，商品名辞書で検索を行い，投稿を抽出する．購買フレーズを用いた商品名辞書の作成方法と同じように，抽出した投稿を BERT 分類器を用いて分類し，購買報告と判断された投稿のみに対して，形態素解析を行い，頻出度を算出する．商品名は名詞だが，購買フレーズは動詞であるため，頻出度が上位 500 の動詞に絞り，購買フレーズ辞書に追加する単語の候補を作成する．購買フレーズを用いた商品名辞書の作成方法と同じように，あらかじめ Twitter での購買フレーズ収集用のストップワードを作成する．これも，商品名辞書の拡張と同様に行い，頻出度が上位 300 個の動詞をストップワードとする．商品名辞書を拡張する際と異なる点は，名詞ではなく動詞を扱うため，ストップワードの数を減らしている．

購買フレーズ辞書に追加する単語の候補の中で，あらかじめ作成した専用のストップワードに含まれるものを取り除く．残った動詞それぞれで投稿を抽出し，それらを BERT 分類器によって購買報告か判断する．抽出した投稿の内，購買報告の割合が 50%以上だった動詞のみを辞書に追加する．

4 評価実験

提案した購買報告収集方法の有効性を明らかにするために，Twitter からあらかじめランダムサンプリングしたデータを使って実験を行った．実験の際には，被験者による抽出結果のラベル付けも行った．実験は，実際にこのような購買報告を収集する仕組みを持つクローラを運用した際のことを考えて，1 日ごとに検索を行い，辞書を随時更新してゆく形式にした．また，抽出量や精度を比較するために被験者実験を行う．

4.1 データセット

実験では，マイクロブログの実際のサービス例として，Twitter のデータを用いた．2018 年 4 月から 2020 年 4 月にかけて，Twitter Streaming API 経由で約 20 億件の投稿を集めた．こ

6：京都大学 黒橋研究室 BERT 日本語 Pretrained モデル
https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

れらは，Apache Solr で管理を行なった．実際の分類器として，購買報告の判別のためにファインチューニングした BERT 分類器を用いた．850 件の投稿に対して人手でラベル付けを行い，学習して確認したところ，この分類器では $F_1 = 0.91$ の精度で購買報告を分類することが可能である．実際の実験では，2019 年 4 月の 1 か月分のデータを用いた．

4.2 比較手法

比較のために，購買報告の抽出手法を複数用意した．具体的には，

- **提案手法**：「商品名の辞書」,「購買フレーズの辞書」の 2 つを用いる
- **購買辞書のみ**:「購買フレーズの辞書」のみを用いる
- **商品名辞書のみ**:「商品名の辞書」のみを用いる
- **ランダム**: 辞書を用いず，全てのツイートからランダムに投稿を抽出する
- **特定の語**:「買った」という単語のみを用いる

という 5 つの購買報告収集用のプログラムを，実際に作成した．

4.3 実験タスク

本実験では，実際にブートストラッピングでクエリを拡張しながら，1 か月間クローラを運用した場合を想定したシミュレーションを行った．実験には 4 月 1 日から 4 月 30 日までの Tweet を対象にした．1 日毎に投稿の検索，収集，辞書拡張を行う．実際の Twitter で検索を行う際には，1 日当たりの API アクセス数に回数の上限が設定されている．こうした条件下での運用を再現するため，本実験では，1 日に 100 回という検索回数の上限を設定した．また，1 アクセスにつき，収集可能な投稿数を 100 件までとする．

実験では，提案手法を含む 5 つの手法について,「辞書に追加された語彙数」を 1 日ずつ算出した．そして，収集された購買報告の候補である Tweet について，人手でそれぞれ購買報告であるかを判定した．

全手法について,「辞書に追加された語彙数」を 1 日ごとに,「収集可能な購買報告の数」を 1 週間ごとに算出する．ただし，この際,「ランダム」,「特定の語」については，辞書を用いないため,「辞書に追加された語彙数」は算出しない．また，提案手法は 2 つの辞書を用いるため,「辞書に追加された語彙数」に関しては，商品名辞書，購買フレーズ辞書の合計を算出する．

「購買辞書のみ」,「商品名辞書のみ」の 2 つの手法については，辞書の拡張に関して，提案手法である相互的なブートストラップを行うことができないため，異なった方法をとる．購買辞書を拡張する場合，提案手法では商品名辞書を用いて投稿を抽出したのに対し，実験において，購買辞書のみを拡張する際は，購買辞書を用いて投稿を抽出する．商品名辞書のみを拡張する際も同様に，商品名辞書を用いて投稿を抽出する．

実際に収集した購買報告の候補となる Tweet について，人手で実際に購買報告であるかを判定した．人手で判別した購買報告の割合を算出するために，被験者によるラベル付けを行った．3 名の被験者が，各手法の結果からランダムに抽出した 5250 件の投稿に対して，購買報告か否かを人手でラベル付けした．こうして，各期間ごとに，どの程度の量の購買報告を実際に収集できたかを推定可能である．

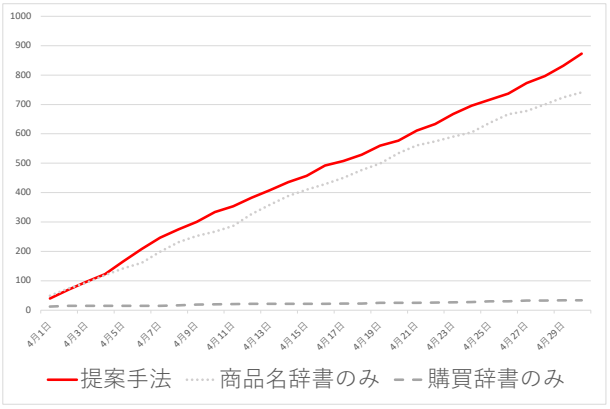

図 2　辞書の単語数．ここでは単語数のみを算出する．

4.4 実　　装

実験のために，実際に購買報告を収集可能なプログラムを作成した．実際のプログラムは Python を用いて実装した．

Twitter からの情報収取は Twitter Streaming API を用いた．Tweet の収集期間は 2018 年 4 月から 2020 年 4 月のものを用いた．Tweet の総数はおおよそ 20 億件である．過去の Tweet をキーワード検索可能にするために，検索インデクサである Apache Solr を用いた．

購買報告の分類器としてファインチューニングした BERT 分類器を用いた．BERT の事前学習モデルとして，京都大学の公開している Wikipedia から学習した言語モデルを用いた．この BERT 分類器は RTX 2080ti を搭載したマシン上で動作させた．

4.5 実験結果

Twitter から収集した実データを用いた，1 か月間の購買報告を収集する実験の結果を示す最初に，辞書に追加された単語数についての結果を示す．各日の辞書の単語数を図 2 に示す．どの手法でも，日が経つにつれて辞書内の単語数が増えていくが,「購買辞書のみ」では，購買フレーズ辞書の単語数があまり増加しなかった．続いて,「商品名辞書のみ」,「提案手法」の順に追加された単語数が多くなった．

次に，実際に人手で分類した購買報告の数を図 3 に示す．人手で分類した結果では，購買報告の数が多い順に「提案手法」「購買辞書のみ」,「特定の語」,「商品名辞書のみ」,「ランダム」の順であった．

5 考　　察

本説では，評価実験の結果について考察し，手法の有用性について議論する．はじめに辞書に追加された単語数について考察する．

購買フレーズ辞書に追加される単語数は，商品名辞書に追加される単語よりも大幅に少なかった．これは，そもそも商品名

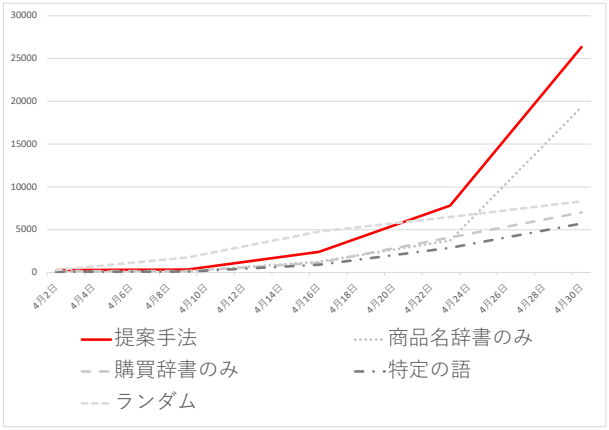

図 3　収集可能な購買報告の数

は無数に存在するが，購買フレーズは数が限られるからである．「買う」や「手に入れる」,「届く」などの意味を持つ単語はそれほど多くない．そのため，辞書に追加される単語数が相対的に少なくなったと考えられる．続いて，追加された単語数が多い順に「提案手法」,「商品名辞書のみ」となった．これは,「提案手法」では商品名辞書だけでなく購買フレーズ辞書も用いるからいう理由だけでなく，辞書を相互的に用いることで効率良く単語を追加することができていることも要因だと考えられる．

収集可能な購買報告の数について考察する．

人手で分類した結果では，購買報告の数が多い順に「提案手法」,「商品名辞書のみ」,「ランダム」,「購買辞書のみ」,「特定の語」の順であった．

「特定の語」では多くの購買報告を収集できなかった理由は,「買った」という単語のみで検索してるからである．この場合,収集精度は高くなるが，収集可能な投稿数が限られるため，有効な手法ではない．

同様に,「購買フレーズ辞書のみ」でも，精度は高いが，収集できる投稿数が少ないため，多くの購買報告を獲得することはできなかったのだと考えられる．

次に,「ランダム」でも収集可能な購買報告の数は 8331 件と，少ない結果になった．これは，Twitter 上の投稿において，購買報告の占める割合は高くないからである．

「商品名辞書のみ」では，比較的多くの購買報告を収集可能であった．この原因は，辞書に追加することのできる単語数が多いため，多くの投稿を収集可能であるからと考えられる．しかし，商品名辞書に追加する単語を商品名辞書で検索しているため，精度が低いと考えられる．そのため,「提案手法」では,購買フレーズで商品名を検索するため精度が保たれるが,「商品名辞書のみ」では著しく精度が落ちたため,「提案手法」と比較すると，多くの購買報告を収集することができなかったのだと考えられる．

次に,「提案手法」について考える.「購買辞書のみ」では，多くの購買報告を収集することができなかった．これは，精度を高く保つことは可能だが，多くの単語を辞書に追加することができないため，収集可能な投稿数には限度があるからと考えられる．そもそもとして，購買フレーズで検索した投稿に関しては購買報告である場合が多い．しかし，商品名で投稿を検索した場合，購買報告の割合が高いとは考えづらい.「iPad」という商品名を例に用いて考える．例えば,「iPad 欲しいな」という投稿や,「iPad 壊れちゃった」という投稿なども多く存在する．そのため，購買フレーズ辞書と商品名辞書の両方を扱う提案手法では，精度を高く保つことが難しい．しかし,「提案手法」が最も多くの購買報告を収集可能となった．これは,「提案手法」が,精度を大きく落とすことはなく，尚且つ最も多くのクエリを獲得できる手法だからだと考えられる.「購買辞書のみ」や「特定の語」のように，精度が高いだけでは，獲得できるクエリの数が少ないため，購買報告を網羅的に収集することはできない．また,「商品名辞書のみ」のように，獲得できるクエリの数が多いだけでは，精度が低く，効率的に購買報告を収集することはできない.「辞書に追加された単語数」,「収集可能な購買報告の数」の結果から,「提案手法」が最も効率的かつ網羅的に，多くの購買報告を獲得することが可能だと考えられる．

また，提案手法を用いることで，従来手法では収集することのできなかった投稿を収集することもできた．例を挙げると,「カラコンめちゃくちゃもれるやーん？そして安かった」という投稿や,「ようやく本体まで全部揃った．新作耳うどんのノイキャン神」といった投稿がある．これらの投稿は,「買った」という直接的な購買示唆となる単語や「正確な商品名」を含んでいない．そのため，従来の手法や，Twitter のキーワード検索では収集することが難しい投稿である．これらの結果からも,提案手法は有効だとか考えられる．

6　おわりに

本研究では，マイクロブログから購買報告を効率的に収集するクローリング方法を提案した．購買報告は,「買った」などの直接的に購買を示す語句や，正確な商品名を含まない場合が多いため，網羅的に収集することが困難である．そこで本研究では，購入を示唆する語句からなる辞書と，商品名らしい名詞と連続語からなる辞書を，相互再帰的に拡張するブートストラッピング手法を提案した．

提案手法の有効性を評価するために，比較手法を 5 つ用意し,「辞書に追加された単語数」,「収集可能な購買報告の数」を比較する実験を行った．2018 年 4 月から Twitter Streaming API 経由で集めた投稿をデータセットとして扱い，計 5250 件の投稿に対して,「購買報告か否か」のラベル付けを被験者にさせた．手法ごとの「辞書に追加された単語数」,「購買報告が占める割合を比較し，収集精度，収集量について考察したところ，提案手法の有効性が証明された．

本研究で達成できなかった，今後の課題となる要素も多数発見された．提案手法の購買報告収集精度が，本研究における課題である．提案手法の有効性を示すことはできたが，収集精度は他手法より高くならなかった．高精度化を実現するために,分類器の高精度化が必要である．

また，本研究では購買報告を網羅的に収集することを目的としたが，商品名を網羅的に収集する上で課題が存在する．提案

手法では分類器を用いるため，収集した投稿に対して形態素解析を行った．そのため，3 つ以上の形態素で構成される長い商品名は正確に収集することができなかった．例えば,「噛むほど旨い焼きするめげそ」のような商品名を 1 つの商品名として収集することも今後の課題である．これらの課題を解決することで，マイクロブログ上の投稿から，網羅的に購買報告を収集するという，本研究の目的を実現できると考えられる．

謝　　辞

本研究は JSPS 科研費 18K18161（代表:荘司慶行），21H03775（代表：大島裕明）の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Allison J Lazard, Adam J Saffer, Gary B Wilcox, Arnold DongWoo Chung, Michael S Mackert, and Jay M Bernhardt. E-cigarette social media messages: a text mining analysis of marketing and consumer conversations on twitter. *JMIR public health and surveillance*, Vol. 2, No. 2, p. e6551, 2016.

[2] Mrityunjay Singh, Amit Kumar Jakhar, and Shivam Pandey. Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, Vol. 11, No. 1, pp. 1–11, 2021.

[3] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, Vol. 89, , 2011.

[4] James Allan, Jaime G. Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[5] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 43–52, 2013.

[6] Padmanabhan Deepak and Sutanu Chakraborti. Finding relevant tweets. In *International Conference on Web-Age Information Management*, pp. 228–240. Springer, 2012.

[7] Edgar Meij, Wouter Weerkamp, and Maarten De Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 563–572, 2012.

[8] Sergey Brin. Extracting patterns and relations from the world wide web. In *International workshop on the world wide web and databases*, pp. 172–183. Springer, 1998.

[9] 河合英紀, 水口弘紀, 土田正明, 國枝和雄, 山田敬嗣ほか. ブートストラップ式同位語辞書構築における検索効率の向上. 情報処理学会論文誌データベース (TOD), Vol. 1, No. 1, pp. 36–48, 2008.

[10] Xiaomei Zou, Jing Yang, and Jianpei Zhang. Microblog sentiment analysis using social and topic context. *PloS one*, Vol. 13, No. 2, p. e0191163, 2018.

[11] Fei Jiang, Yi-Qun Liu, Huan-Bo Luan, Jia-Shen Sun, Xuan Zhu, Min Zhang, and Shao-Ping Ma. Microblog sentiment analysis with emoticon space model. *Journal of Computer Science and Technology*, Vol. 30, No. 5, pp. 1120–1129, 2015.

[12] Yue Sui and Xuecheng Yang. The potential marketing power of microblog. In *2010 Second International Conference on Communication Systems, Networks and Applications*, Vol. 1, pp. 164–167. IEEE, 2010.

[13] Weishi Zeng, Yunru Huang, and Lili Jiang. The study of microblog marketing based on social network analysis. In *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, Vol. 3, pp. 410–415. IEEE, 2011.

[14] Xin Li and Bingruo Duan. Organizational microblogging for event marketing: a new approach to creative placemaking. *International Journal of Urban Sciences*, Vol. 22, No. 1, pp. 59–79, 2018.

[15] Eduard C Dragut, Clement Yu, Prasad Sistla, and Weiyi Meng. Construction of a sentimental word dictionary. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1761–1764, 2010.