

博士論文

ソーシャルメディアからの人気情報を用いた
商品の販売予測に関する研究

指導教員 松尾 豊 特任准教授

東京大学大学院 工学系研究科
技術経営戦略学専攻

野中尚輝

概要

商品の販売予測は、企業が意思決定を行う際に重要となる。近年、消費者は自身の嗜好にそった商品を購入するようになっており、消費が多様化している。そのため、商品の販売予測を行う場合に、個人の嗜好を考慮することの重要性が増している。また人気となる商品は、その時々流行に応じて変化するため、継続的に調査を行うことが重要となる。しかしながら、多様な個人の嗜好を捉えることや継続的な調査を行うことは、アンケート調査といった従来の手法を用いる場合、コストの面で問題があった。

一方で近年、インターネットやスマートフォンの普及が進んだことから、ソーシャルメディアが幅広く利用されるようになった。ソーシャルメディアの利用者数は全世界において拡大が続いており、オンラインレビューやオンライン百科事典、ソーシャルネットワークサービスなどさまざまな形で存在している。ソーシャルメディアの普及により、これまで情報を受け取る側であった消費者が、自身の考えや商品に対する感想といった情報を発信することが可能になった。各ユーザによりソーシャルメディアに投稿される内容は、各ユーザの嗜好を反映しており、これらを分析することで消費者の嗜好を捉えることができる可能性がある。また、ソーシャルメディアへの投稿は継続的に行われるため、人気の変化も捉えることが可能であると考えられる。

このような背景から、ソーシャルメディアを用いた商品の販売予測の研究が行われてきた。特に、多くの人々が鑑賞し、個人の嗜好により対象への評価が異なることからソーシャルメディアにおいて言及されることの多い、映画が予測の対象として扱われてきた。ソーシャルメディアによる映画の予測としては、Twitterのデータを用いる研究やブログから得た情報を用いた研究、レビューの情報を用いた研究が知られている。映画以外のドメインにおけるウェブ上のデータを用いた販売予測の研究としては、入力される検索クエリの回数から自動車の販売数

をはじめとする経済動向を推測した研究や検索クエリの入力回数，Twitter における言及数，Wikipedia のリンク構造を用いてアニメ，漫画といったマルチメディアタイトルの人気予測を行っている研究が存在する．このように映画を中心として，個人の嗜好が対象の評価に反映されるためソーシャルメディアに情報が上がりやすい傾向にある商品の予測が行われてきた．これらの先行研究において，ソーシャルメディアから得られた販売予測の際に用いられる特徴量には，対象となる商品が言及された回数や対象の商品についてのレビューの投稿回数といった指標が用いられてきた．これらの言及された回数やレビューの投稿回数といった特徴量は，予測を行う側のもつ外部知識に基づいて，ヒューリスティックに選択されている．

しかし，ヒューリスティックに特徴量を定めることが難しい場合がある．例えば，入力データが時系列の変動データや対象の商品名が出現しないような文章であるような場合，用いるべき特徴量をヒューリスティックに定めることは難しい．ヒューリスティックに特徴量を選択できない場合，得られる全特徴量を直接予測に用いることも考えられるが，予測対象のデータの量が少ない場合，過学習が起りやすくなり予測が難しくなる．このような背景のもと本研究では，ラベルデータの数が少なく，ソーシャルメディアから直接特徴量を獲得することが難しい場合における，ソーシャルメディアによる商品の販売予測の精度を向上させることを目的とした手法を提案した．具体的には，まずソーシャルメディアから得られた素性を用いて，商品の販売数に関連する指標を予測する問題を設定したその後，設定した問題を解く上で有用な素性および特徴量を人気情報として保持し，得られた人気情報をラベルデータの数が少ない商品の販売数の予測問題において用いる手法を提案した．

本研究では，商品の販売予測の中でも，ソーシャルメディアが予測において有用であると考えられる商品を対象として販売予測を行った．ソーシャルメディアが予測において有用であると考えられる商品は，大量生産されるため潜在的には誰もが入手可能であり，各ユーザの嗜好が商品の評価に影響するためソーシャルメディア上に各ユーザの意見が投稿されやすく，また流行が存在するため継続的に調査を行う必要のある商品である．本研究では，予測においてソーシャルメディアが有用であると考えられる商品から，予測対象として重要と考えられる自動車およびマルチメディアタイトルを対象として検証実験を行なった．

3章では，日本の輸出総額に占める割合でトップの商品である自動車の販売予測を行なった．実験では，記述される内容に主観的な口コミ情報が含まれており，記述内容に対応する商品が

明確であるという理由からオンラインレビューのデータを用いた。まず、レビューサイトに投稿されたレビュー文章に対して感情分析を行い、感情予測タスクにおいて予測への寄与度の大きい単語素性を特定し、人気情報とした。続いて、人気情報とした単語素性を自動車の販売数予測タスクに用い、単語素性を用いない場合、単語素性をランダムに選択した場合、レビュー文中に含まれる単語素性全てを用いた場合および既存の極性辞書をもとに単語を選択した場合の予測精度と比較を行なった。実験結果から、感情予測タスクにて予測に有用であった単語素性を用いることにより、予測精度が向上することを確認した。

次に、4章、5章および6章にて、低価格であることが多く、商品の販売予測を行う上で対象とした商品に対する注目の度合いが重要になると考えられるマルチメディアタイトルについての販売予測を行なった。実験では、マルチメディアタイトルの網羅性が高いという点から Wikipedia のデータを用いた。また、販売数のデータとして、公開される頻度が高く、複数のカテゴリの情報を収集できるという理由からオンラインショッピングサービスのデータを用いた。オンラインショッピングサービスにて公開される販売に関するデータは、販売ランキング上位のデータのみである。そのため入手できるラベルデータは、存在するページ数と同数のデータが得られる Wikipedia のラベルデータと比較して入手できるデータの量が少ない。また、オンラインショッピングサービスの販売データは、ユーザの個人の購買行動と紐づいていないため、ユーザの嗜好を反映できない。そこで本研究では、4章、Wikipedia の編集履歴からユーザの嗜好に関するベクトル表現を抽出する手法について検証し、5章において、データ数が少ない場合における予測精度を向上させる手法を提案した。続いて、4章および5章で得られた知見に基づいて、6章にて販売に関連すると考えられる Wikipedia のページ閲覧数を対象とした予測問題を解き、得られた人気情報を用いて販売予測を行なった。

最後に、上記の研究に基づいた提案手法の限界と応用可能性について考察し、結論を述べる。

目次

| | | |
|-------|---------------------------|----|
| 第 1 章 | 序論 | 1 |
| 1.1 | 背景 | 1 |
| 1.1.1 | 消費予測 | 1 |
| 1.1.2 | ソーシャルメディアの普及とそれによる予測 | 3 |
| 1.1.3 | ソーシャルメディアによる消費予測 | 6 |
| 1.2 | 研究目的 | 7 |
| 1.3 | 本論文の構成 | 10 |
| 第 2 章 | 関連研究 | 13 |
| 2.1 | 消費に関わる研究 | 13 |
| 2.1.1 | 消費に関わる研究の起こり | 13 |
| 2.1.2 | 消費者の行動に関する研究 | 15 |
| 2.1.3 | 消費と企業に関する研究 | 18 |
| 2.2 | 商品の販売を予測する研究 | 21 |
| 2.3 | ソーシャルメディアを用いた予測を行う研究 | 24 |
| 2.3.1 | ソーシャルメディアを用いた選挙の予測 | 24 |
| 2.3.2 | ソーシャルメディアを用いた商品の消費予測を行う研究 | 25 |
| 第 3 章 | 感情分析による人気情報の抽出と自動車の販売予測 | 27 |
| 3.1 | 本章の背景 | 27 |
| 3.2 | 関連研究 | 29 |
| 3.2.1 | 感情分析 | 29 |

| | | |
|-------|-------------------------------|----|
| 3.2.2 | ソーシャルメディアによる販売予測 | 30 |
| 3.2.3 | 属性選択 | 31 |
| 3.3 | 提案手法 | 32 |
| 3.3.1 | 感情分析 | 32 |
| 3.3.2 | 有用な素性の抽出 | 33 |
| 3.3.3 | 販売予測 | 34 |
| 3.4 | 実験・結果 | 35 |
| 3.4.1 | データ | 35 |
| 3.4.2 | データに対する前処理 | 36 |
| 3.4.3 | 感情分析 | 37 |
| 3.4.4 | 販売予測 | 39 |
| 3.5 | 考察 | 43 |
| 3.5.1 | 感情分析 | 43 |
| 3.5.2 | 販売予測 | 44 |
| 3.6 | 本章の結論 | 46 |
| 第 4 章 | ユーザの嗜好を反映したベクトル表現の獲得と人気予測 | 48 |
| 4.1 | 本章の背景 | 48 |
| 4.2 | 関連研究 | 50 |
| 4.2.1 | Wikipedia における編集行動 | 50 |
| 4.2.2 | 多面的情報をもつベクトル表現の獲得 | 51 |
| 4.2.3 | リンク構造による指標 | 52 |
| 4.3 | 提案手法 | 52 |
| 4.3.1 | コンテンツベクトルの取得 | 52 |
| 4.3.2 | 人気予測 | 53 |
| 4.4 | データ | 53 |
| 4.4.1 | Wikipedia からのデータの取得 | 54 |
| 4.4.2 | データに対する前処理 | 55 |
| 4.5 | 予備実験 | 56 |

| | | |
|--------------|--|-----------|
| 4.5.1 | 編集者系列の定性分析 | 56 |
| 4.5.2 | コンテンツベクトルの取得 | 58 |
| 4.6 | 実験 | 60 |
| 4.6.1 | 人気予測 | 60 |
| 4.6.2 | 予測期間と予測精度 | 62 |
| 4.6.3 | 予測精度の精度の高いクラスター | 62 |
| 4.7 | 考察 | 64 |
| 4.7.1 | 編集履歴の系列 | 65 |
| 4.7.2 | ベクトル化 | 65 |
| 4.7.3 | 人気予測 | 66 |
| 4.7.4 | 予測期間と予測精度 | 67 |
| 4.7.5 | 予測精度の精度の高いクラスター | 68 |
| 4.8 | 本章の結論 | 68 |
| 第 5 章 | 事前学習による特徴量抽出と人気予測 | 70 |
| 5.1 | 本章の背景 | 70 |
| 5.2 | 関連研究 | 72 |
| 5.2.1 | 深層学習の事前学習を用いたソーシャルメディアの研究 | 72 |
| 5.2.2 | Wikipedia の多言語性に着目した研究 | 73 |
| 5.3 | 提案手法 | 73 |
| 5.4 | データ | 74 |
| 5.4.1 | Wikipedia データ | 75 |
| 5.4.2 | 検証用データの準備 | 75 |
| 5.5 | 予備実験：予測に用いるニューラルネットワークの構造の決定 | 77 |
| 5.6 | 実験 | 78 |
| 5.6.1 | ニューラルネットワークへのデータ入力 | 79 |
| 5.6.2 | 事前学習に用いるデータと予測結果の関係の分析 | 82 |
| 5.7 | 考察 | 86 |
| 5.7.1 | ニューラルネットワークの構造 | 86 |

| | | |
|--------------|---|------------|
| 5.7.2 | ニューラルネットワークへのデータ入力 | 86 |
| 5.7.3 | 事前学習に用いるデータと予測結果の関係の分析 | 88 |
| 5.8 | 本章の結論 | 89 |
| 第 6 章 | Wikipedia から得た人気情報による販売予測 | 91 |
| 6.1 | 本章の背景 | 91 |
| 6.2 | 提案手法 | 93 |
| 6.2.1 | 事前学習による特徴量の抽出 | 93 |
| 6.2.2 | ユーザ嗜好を考慮した販売予測モデルの学習 | 93 |
| 6.3 | データ | 94 |
| 6.3.1 | 実販売データ | 94 |
| 6.3.2 | Wikipedia データ | 95 |
| 6.3.3 | 実験データの前処理 | 95 |
| 6.4 | 予備実験：販売ランキングデータとページ閲覧数の関係 | 96 |
| 6.4.1 | 販売ランキングとページ閲覧数の定性的分析 | 96 |
| 6.4.2 | 販売ランキングとページ閲覧数の定量的分析 | 98 |
| 6.5 | 実験：販売予測の精度検証 | 98 |
| 6.5.1 | ページ閲覧数データにより事前学習したモデルを用いた予測 | 99 |
| 6.5.2 | ユーザ嗜好の情報を加えたモデルによる予測 | 101 |
| 6.5.3 | 事前学習したモデルとユーザ嗜好の情報をを用いた予測 | 103 |
| 6.6 | 考察 | 103 |
| 6.6.1 | 販売データとページ閲覧数の関係 | 103 |
| 6.6.2 | ページ閲覧数データにより事前学習したモデルを用いた予測 | 105 |
| 6.6.3 | 編集履歴によるベクトル表現を用いた予測 | 105 |
| 6.6.4 | 事前学習したモデルと編集履歴によるベクトル表現を用いた予測 | 106 |
| 6.7 | 本章の結論 | 106 |
| 第 7 章 | 考察 | 108 |
| 7.1 | ソーシャルメディアを用いた販売予測における各論の考察 | 108 |

| | | |
|-------|-------------------------------------|-----|
| 7.1.1 | 感情分析による人気情報の抽出と自動車の販売予測 | 108 |
| 7.1.2 | ユーザの嗜好を反映したベクトル表現の獲得と人気予測 | 109 |
| 7.1.3 | 事前学習による特徴量抽出と人気予測 | 111 |
| 7.1.4 | Wikipedia から得た人気情報による販売予測 | 112 |
| 7.2 | 全体を通しての議論と今後の課題 | 113 |
| 7.2.1 | ソーシャルメディアを用いた予測を行う上での課題 | 113 |
| 7.2.2 | ソーシャルメディアを用いる販売予測の適用可能性 | 118 |
| 第 8 章 | 結論 | 120 |
| | 参考文献 | 122 |
| | 謝辞 | 134 |
| | 発表文献 | 136 |

目次

| | | |
|-----|---|----|
| 1.1 | 日本語版 Wikipedia の記事数推移 | 4 |
| 1.2 | 日本における SNS ユーザ数の推移 (http://ictr.co.jp/report/20160816.html を元に著者が作成) | 5 |
| 1.3 | ソーシャルメディアから取得した人気情報による販売予測 | 8 |
| 1.4 | 本論文の構成 | 11 |
| 3.1 | オンラインレビューを用いた自動車の販売予測の概略図 | 29 |
| 3.2 | 自動車に関するオンラインレビューの投稿 (例) | 36 |
| 3.3 | 単語素性を用いない場合, 全単語素性を用いる場合と提案手法の比較 | 39 |
| 3.4 | ランダムに選択された単語素性を用いる手法と提案手法の比較 | 40 |
| 3.5 | 極性辞書を用いる手法, 感情分析の結果を直接用いる手法と提案手法の比較 | 41 |
| 3.6 | 単語数に対する平均 MAPE 値 | 45 |
| 4.1 | Wikipedia の編集履歴から学習したベクトル表現による人気予測の概略図 | 51 |
| 4.2 | 編集回数と編集人数の関係 | 54 |
| 4.3 | コンテンツベクトルのクラスタリング結果 | 58 |
| 5.1 | 事前学習を行なったモデルによる人気予測の概略図 | 72 |
| 5.2 | 各予測モデルの概要 | 78 |
| 5.3 | スケールを低下したデータ (例) | 82 |
| 5.4 | ページ閲覧数の推移 (例) | 87 |
| 6.1 | Wikipedia から得た人気情報によるコンテンツ作品の販売予測の概略図 | 92 |

| | | |
|-----|---|----|
| 6.2 | 販売スコアと Wikipedia のページ閲覧の可視化 (例) | 97 |
| 6.3 | 販売スコアと Wikipedia のページ閲覧数の相関 | 98 |
| 6.4 | 各予測モデルの概要 | 99 |

表目次

| | | |
|-----|-------------------------------------|-----|
| 1.1 | ソーシャルメディアの種類 | 3 |
| 3.1 | 感情分析に寄与する素性 | 38 |
| 3.2 | 開発セットにおける精度 | 42 |
| 4.1 | コンテンツ作品についての編集履歴 (例) | 57 |
| 4.2 | クエリとしたコンテンツ作品に隣接するコンテンツ作品 | 59 |
| 4.3 | コンテンツ人気予測の結果 | 61 |
| 4.4 | 予測期間と予測精度の関係 | 63 |
| 4.5 | クラスごとの MAPE | 64 |
| 5.1 | 各言語版 Wikipedia データの詳細 | 74 |
| 5.2 | 各言語に正規化における最大値 | 76 |
| 5.3 | 各言語におけるクラスバランス | 76 |
| 5.4 | モデルの構造と分類精度 | 77 |
| 5.5 | 入力データの与え方と分類精度 | 80 |
| 5.6 | 日本語データのスケール低下処理 | 81 |
| 5.7 | 再学習を行わない場合の精度 | 81 |
| 5.8 | 事前学習したモデルによる分類精度 | 85 |
| 6.1 | 各データにおけるクラスバランス | 96 |
| 6.2 | 事前学習したモデルによる分類精度 | 100 |
| 6.3 | ユーザ嗜好ベクトルを用いたモデルによる分類精度 | 101 |

| | | |
|-----|---|-----|
| 6.4 | 事前学習とユーザ嗜好ベクトルを組み合わせたモデルによる分類精度 | 102 |
|-----|---|-----|

第1章

序論

本章では本論文の背景，目的及び章構成について述べる。

1.1 背景

本節では本研究に関する背景について述べる。商品の販売予測は，企業が意思決定を行う際に重要となる。予測を行う際には，過去の販売データをはじめとする統計指標のみでなく，消費者の意見や会社・製品に対する感情，消費者の嗜好といった点を考慮することが重要になる。企業は，アンケート調査といった方法で消費者の意見を収集したり，様々なメディアを通じた広告や社会的活動を行うことで自社および自社の商品の認知向上とイメージの向上を試みたりしてきた。一方，近年ではインターネットやスマートフォンが普及したことから，一般消費者が様々な形で情報を発信することが可能になった。また，ソーシャルメディアを通じて，個人個人の嗜好を取得することも可能になり，それらを用いた消費予測も行われている。

1.1.1 消費予測

消費は，衣・食・住といった生活領域全般にわたって，我々の日常生活と密接に結びついている。消費される対象は，食品や日用雑貨のように使用すればなくなる非耐久消費財から，家電製品や自動車のように中長期間にわたって使用できる耐久消費財，有形ではなく用役や経験という形で提供されるサービスなど様々な領域に広がっている。われわれは，このような多種多様な製品やサービスを消費しながら生活している。

第二次世界大戦以降、先進国を中心に工業化が進み大量生産が可能になるとともに大量消費の時代となった。大量生産により生産が消費を上回る状況下において、企業にとって顧客のニーズに合わせた商品を生産することが重要となった。このような背景のもと、顧客を理解することの必要性と重要性が認識されるようになり、それに伴って消費者の行動や企業の活動について研究する領域は活発に研究されるようになった。米国マーケティング協会の定義によれば、消費者行動とは、「製品やサービスの市場における消費者または意思決定者の行動を指す。しばしば、このような行動を理解し、記述しようとする学際的な学問分野である」*1と定義される [AMA]。消費者行動の研究は、「人々がなぜ購買するか」という購買動機や「人々は何を購買するか」という購買行動についての研究に端緒をなし、「人々がなぜ消費するのか」や「どのように消費するのか」といった消費行動の研究にも拡大してきている。消費者行動の理解には、多面的に消費者の行動を分析する必要があるため、心理学、社会学、社会心理学、文化人類学といった関連領域の理論や方法論を積極的に援用していく必要がある。消費者行動の研究はこれらの関連する諸分野の影響を受けながら発展してきた。またマーケティングにおいては、消費者の購買行動についての理解とともに、消費者による購買行動の結果である商品の販売予測が計画を立てる上で重要となる。

近年、個人の嗜好に基づき消費行動が多様化している。そのため、商品の購買における個人の嗜好を考慮して販売予測を行うことが重要になっている。また、各個人の嗜好は時代とともに変化していくため、継続的に個人の嗜好を調査することが必要となる。継続的に調査を行う必要があることから、調査のコストも重要となる。個人の嗜好を捉えるために有用と考えられるデータとしては、アンケート調査により収集されるデータ、自社の商品の販売データ、ソーシャルメディアのデータがあげられる。アンケート調査は、目的に合わせて調査項目が設定できる反面、毎回の調査にコストがかかる。また、自社の商品の販売データは、競合する商品の情報が存在しないため、全体を通じた比較が困難になる。そこで本研究では、調査のコストと商品の網羅性という観点から、ソーシャルメディアの情報を用いて商品の販売予測を行う。

*1 “The behavior of the consumer or decision maker in the market place of products and services. It often is used to describe the interdisciplinary field of scientific study that attempts to understand and describe such behavior” を和訳

表 1.1 ソーシャルメディアの種類

| 形態 | 事例 |
|----------------------|-------------------|
| ブログ | 各種ブログサイト |
| マイクロブログ | Twitter |
| ソーシャルネットワークワーキングサービス | Facebook, Google+ |
| 仮想ゲーム世界 | GREE, Mobage |
| コンテンツコミュニティ | Youtube, ニコニコ動画 |
| 共同プロジェクト | Wikipedia |

1.1.2 ソーシャルメディアの普及とそれによる予測

ソーシャルメディアの発展

インターネットの機能の拡張に伴い、2000年代中頃には、ティムオライリーにより Web2.0 の考え方が提唱された。ソーシャルメディアは、この Web2.0 の概念に基づいて、「Web 2.0 の観念的および技術的基盤上に構築され、ユーザ生成コンテンツの作成と交換を可能にするインターネットベースのアプリケーション群。」^{*2}と定義される [Kaplan 10]。このようにソーシャルメディアは、ユーザが情報を生成および発信することを可能にするメディアである。

ソーシャルメディアには、図 1.1 に挙げるような形態が存在する。具体的には、ブログ、Facebook^{*3}や Google+^{*4}のような SNS (Social Networking Service; ソーシャルネットワークワーキングサービス)、Twitter^{*5}に代表されるマイクロブログ、Wikipedia^{*6}に代表される共同プロジェクト (Collaborative projects)、Youtube^{*7}やニコニコ動画^{*8}のようなコンテンツコ

^{*2} “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content” を和訳

^{*3} <https://www.facebook.com/>

^{*4} <https://plus.google.com/>

^{*5} <https://twitter.com/>

^{*6} <https://www.wikipedia.org/>

^{*7} <https://www.youtube.com/>

^{*8} <http://www.nicovideo.jp/>

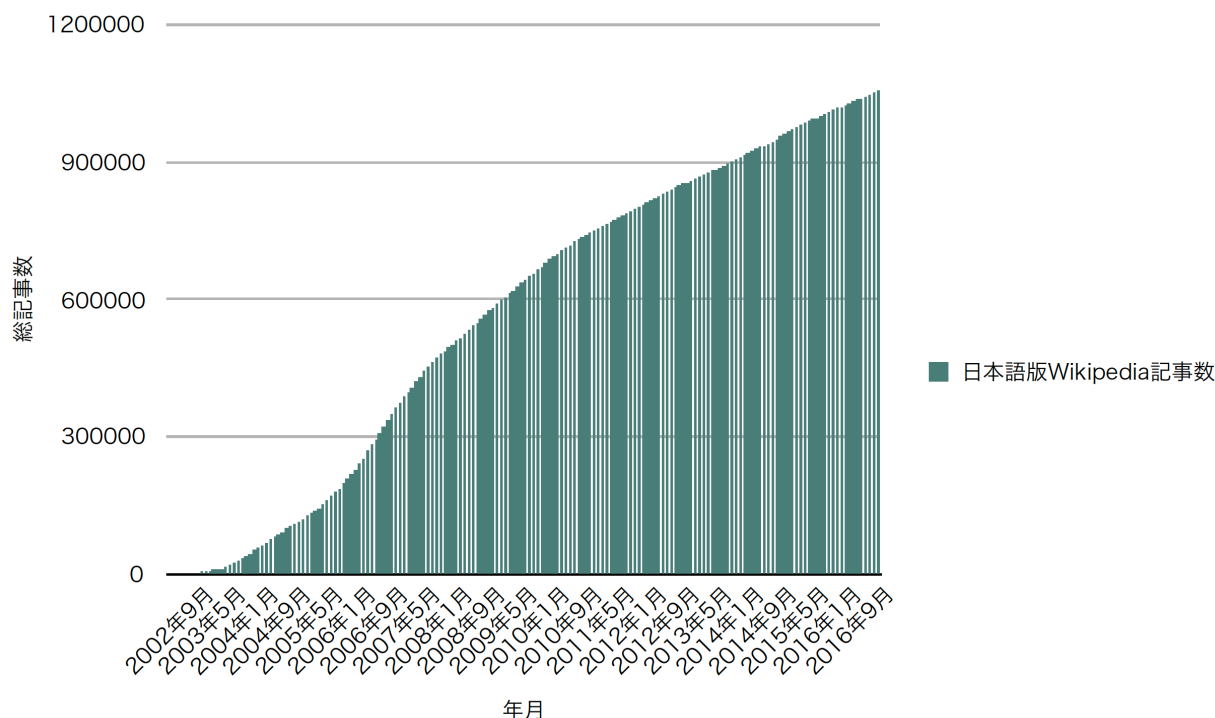


図 1.1 日本語版 Wikipedia の記事数推移

コミュニティ、Mobage^{*9}やGREE^{*10}といったソーシャルゲームが存在する。

ソーシャルメディアの普及の度合いを示す具体的な数字を示す。代表的な共同プロジェクトである Wikipedia には、日本語版において 1,066,231 件の記事が存在し、最も規模の大きい英語版では 5,433,584 件の記事が存在する。また、100 万件以上の記事を有する言語が 13 存在し、全体では 45,372,848 件の記事が存在する。図 1.1 に、日本語版 Wikipedia の記事数の推移を示す。この図から Wikipedia では 2002 年のサービス開始から現在に至るまで記事数の増加が続いていることがわかる。また代表的な SNS である Facebook では、2016 年 12 月時点での国内における月間アクティブユーザ数が 2700 万人となっており、ユーザ数は 1 年間で 200 万人増加している。また全世界でも月間アクティブユーザは増加しており、2017 年には 20 億人に到達している。最後に、図 1.2 に日本全体における SNS ユーザの推移を示す。このデータからも日本における SNS の利用者は年々増加しており SNS の普及が進んでいることが

^{*9} <https://www.mbga.jp/>

^{*10} <http://gree.jp/>

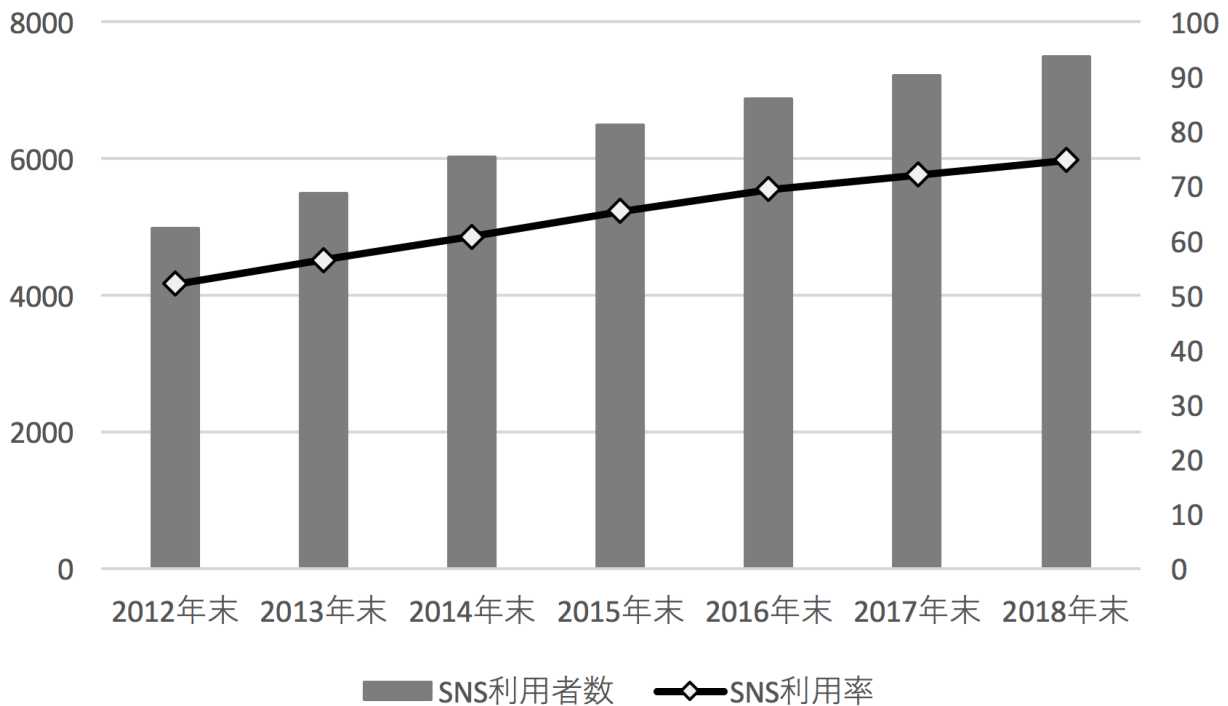


図 1.2 日本における SNS ユーザ数の推移 (<http://ictr.co.jp/report/20160816.html> を元に著者が作成)

わかる。このように世界全体および日本において、ソーシャルメディアの普及は進んでいる。

ソーシャルメディアと実世界

ソーシャルメディアの普及とともに、人々が自身の日常生活、周囲の環境、自身の体験したこと、様々なものごとに対する意見や感想がソーシャルメディア上に数多く投稿されるようになった。この結果、ソーシャルメディア上にはユーザの思想や意見についての情報が大量に投稿・蓄積されている。またユーザが投稿した情報以外にも、サービス上での各ユーザのページ遷移や Wikipedia におけるページの編集履歴といった情報もトラッキングを行うことが可能であり、蓄積されている。近年、これらのデータの収集、分析を通じて実世界で起きていることを分析したり予測したりすることが試みられている。

榊 [榊 14] では、ソーシャルメディアの情報をもとに実世界を観測する「ソーシャルセンサ」という概念を提唱し、その実例を複数挙げている。例えば、マイクロブログである Twitter に投稿される地震に関連する情報を抽出し、ユーザの位置情報と投稿時刻から地震の広がりを検

出ることが行われている [Sakaki 10]. 地震以外にも、交通情報に関する情報を取得した研究も存在する. [Sakaki 12] では、ソーシャルメディア上に投稿から交通に関する情報を抽出した後、位置情報と関連付け、リアルタイムで交通状態を抽出する手法を提案している. また、検索エンジンにおける検索履歴と位置情報からインフルエンザの流行を検出することをしている研究 [Ginsberg 09] も存在する. この研究に関連して、Twitter を用いたインフルエンザの流行検出も行われている [Aramaki 11], [Culotta 10]. このように、人が観察、体験した後、ソーシャルメディアに投稿された内容を分析して、実世界をリアルタイムで観測する研究が行われている.

1.1.3 ソーシャルメディアによる消費予測

ソーシャルセンサでは、ソーシャルメディア上に投稿された内容および時間から実世界で起こっていることをリアルタイムで検出することを試みていた. 一方、ソーシャルメディアを用いた予測を行う研究では、ソーシャルメディア上の情報から得た素性をモデルに組み込むことで予測を行なっている.

近年、消費が多様化していることから商品の販売予測を行う場合、個人の嗜好を考慮することが重要となっている. すでに先行研究において、個人の嗜好に基づいた意見や商品に対する感情といった口コミ情報が商品の販売と相関することが示されている [Chevalier 06], [Dellarocas 07]. 口コミ情報をはじめとした、消費者の嗜好が含まれ商品の販売予測において有用と考えられるデータとしては、アンケート調査のデータ、購入したユーザと商品を紐づけた販売記録のデータ、ソーシャルメディアのデータが挙げられる. アンケート調査は、調査を行うコストがかかり、また設定した質問により結果に偏りが生じる場合も考えられる. また、販売記録のデータは商品を実際に販売している企業以外では、入手することが難しいという問題点がある. 一方、スマートフォンやインターネットの普及拡大に伴い、ユーザの増加が続いているソーシャルメディアからの情報は、多くの場合、公開されているため十分なデータを収集することが可能であると考えられる.

このような背景から、ソーシャルメディアを用いた商品の販売予測の研究が行われてきた. 特に、多くの人々が鑑賞し、個人の嗜好により対象への評価が異なることから、ソーシャルメディアにおいて言及されることの多い映画が予測の対象として扱われてきた. ソーシャルメディア

を用いた映画の予測について研究としては、Twitter のデータを用いる研究 [Asur 10]、ブログから得た情報を用いた研究 [Mishne 06] が知られており、レビューの情報を用いた研究としては [Dellarocas 07] がある。映画以外のドメインにおけるウェブ上のデータを用いた販売予測の研究としては、入力される検索クエリの回数から自動車の販売数をはじめとする経済動向を推測した [Choi 12] の論文や検索クエリの入力回数、Twitter における言及数、Wikipedia のリンク構造を用いてアニメ、漫画といったマルチメディアタイトルの人気予測を行っている研究 [保住 14] が存在する。このように映画を中心として、個人の嗜好が対象の評価に反映されることから、ソーシャルメディアに情報が上がりやすい傾向にある商品の予測が行われてきた。これらの先行研究において、ソーシャルメディアから得られた販売予測の際に用いられる特徴量は、対象となる商品が言及された回数や対象の商品についてのレビューの投稿回数といった指標が用いられてきた。これらの言及された回数やレビューの投稿回数といった特徴量は、予測を行う側のもつ外部知識に基づいて、ヒューリスティックに選択されている。

しかし、ヒューリスティックに特徴量を定めることが難しい場合も考えられる。例えば、入力データが時系列の変動データや対象の商品名が出現しないような文章であるような場合、用いるべき特徴量をヒューリスティックに決めることは難しい。ヒューリスティックに特徴量を選択できない場合、得られる全特徴量を直接予測に用いることも考えられるが、予測対象のデータの量が少ない場合、予測を行うモデルの学習が難しくなるという課題が存在する。

1.2 研究目的

このような背景のもと本研究では、ラベルデータの数が少なく、ソーシャルメディアから直接特徴量を獲得することが難しい場合における、ソーシャルメディアによる商品の販売予測の精度を向上させることを目的とした手法を提案した。より具体的には、図 1.3 に示すように、まずソーシャルメディアから得られた素性を用いて、商品の販売数に関連する指標を予測する問題を設定する。続いて、設定した問題を解く上で有用な素性または特徴量を人気情報として保持する。そして、得られた人気情報をラベルデータの数が少ない商品の販売数の予測問題において用いる。

本研究では、商品の販売予測の中でも、ソーシャルメディアが予測において有用であると考えられる商品を対象として販売予測を行う。ソーシャルメディアが予測において有用であると

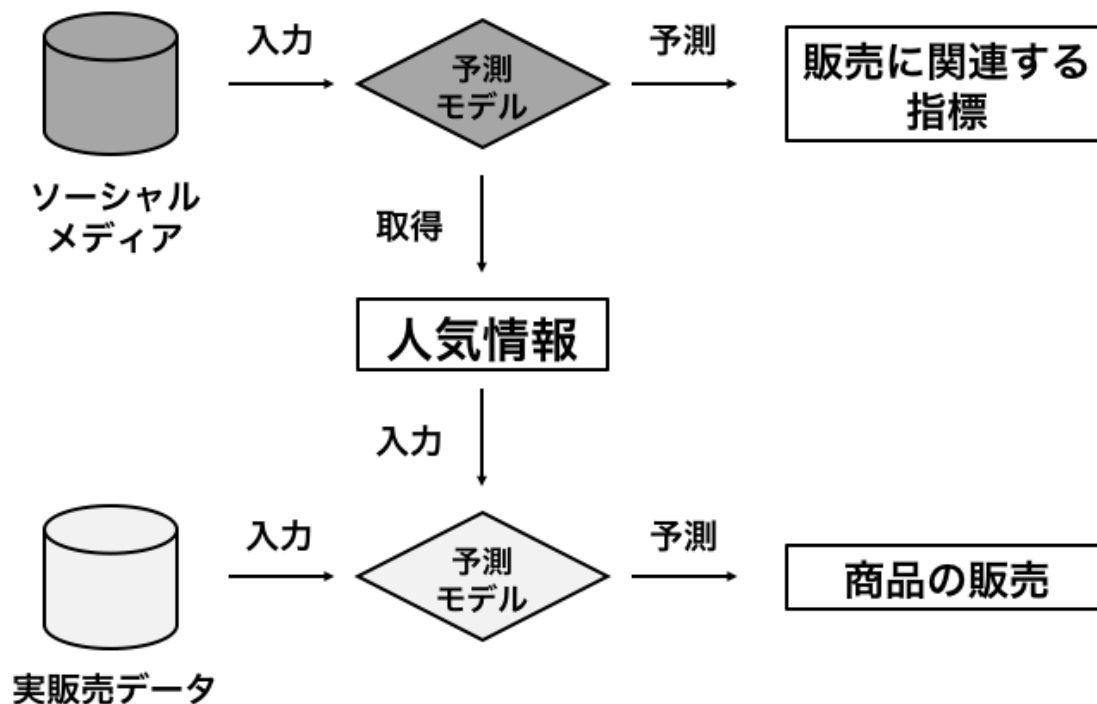


図 1.3 ソーシャルメディアから取得した人気情報による販売予測

考えられる商品は、大量生産されるため潜在的には誰もが入手可能であり、ユーザの嗜好が商品の評価に影響するためソーシャルメディア上に各ユーザの意見が投稿されやすく、また流行が存在するため継続的に調査を行う必要のある商品である。そこで本研究では、ソーシャルメディアが予測において有用であると考えられる商品から、予測対象として重要と考えられる自動車およびマルチメディアタイトルを対象として検証実験を行なった。

自動車は、日本の輸出総額に占める割合でトップの商品であり、その販売予測は重要な課題である。一般に自動車は高価であるため、自動車の購入時に消費者は複数の候補を比較・検討すると考えられる。そこで本研究では、比較・検討の際に用いられると考えられる情報に注目した。比較・検討の際に消費者が用いると考えられる情報としては、自動車関連雑誌をはじめとする記事に記載されている内容、メディアにおける広告、ソーシャルメディア上の情報など

が考えられる。本研究では、データ入手の容易さと消費者の主観が含まれるという点からソーシャルメディアのデータを用いて販売予測を行なった。ソーシャルメディアによる自動車の販売予測では、ブログ記事を用いる手法、Wikipediaを用いる手法、オンラインレビューを用いる方法などが考えられる。本研究では、販売予測で有用であることが知られている商品に対する口コミ情報が記述されており、記述内容と商品の対応関係が明確なオンラインレビューの情報を用いて自動車の販売予測を行う。オンラインレビューに投稿される口コミ情報に含まれる単語には、販売予測に寄与する単語が含まれると考えられる。しかし、入手できる販売データの数と比較してレビューに含まれる全語彙数は非常に大きいため、レビューに含まれる語のうち予測に有用な素性を選択する必要がある。そこで本研究ではオンラインレビューのデータを用いて、販売に関連すると考えられるユーザの感情を予測する問題を解き、予測に有用な素性を抽出した後、販売予測を行なった。

マルチメディアタイトルは、音声、動画、静止画、文書などを組み合わせたマルチメディアコンテンツからなる著作物であり、テレビ、雑誌、ゲーム、インターネットといった幅広いメディア・形態を通じて展開される。マルチメディアタイトルの市場規模は拡大を続けており、代表的なマルチメディアコンテンツであるアニメ産業の市場規模は、2015年に前年比12%増という成長を記録している [増田 16]。また、マルチメディアタイトルはそれ自身がコミックやアニメの放映により収益をあげるだけでなく、その知名度や人気を活用する形で広告やイベントグッズなどにおけるタイアップが可能であり、他の産業に対する波及効果も期待できる。加えて近年、海外に対する輸出の規模も拡大が続いており、クールジャパンプロジェクトにおいても注目されている重要な商品である。

マルチメディアタイトルに関連する商品は、自動車や家電製品といった商品と比較して低価格であることが多い。そのため、商品の販売予測を行う上で、対象とした商品に対する注目の度合いが重要になると考えられる。このような注目の度合いを表す指標として、TVにおける視聴率、検索エンジンにおけるクエリの入力回数、ソーシャルメディア上における指標が考えられる。本研究では、データの入手の容易さからソーシャルメディア上における指標を用いた。ソーシャルメディアを用いたマルチメディアタイトルの販売予測では、オンラインレビューのデータを用いる手法、Wikipediaのデータを用いる手法、ブログ記事を用いる手法などが考えられる。本研究では、マルチメディアタイトルの網羅性という点から Wikipedia の

データを用いて販売予測を行なった。

また本研究では、マルチメディアタイトルに対する注目の度合いが実際の販売予測において有用であることを、実際の販売データを用いて確認した。マルチメディアタイトルの販売に関するデータとしては、商品カテゴリごとに公表される販売データ、オンラインショッピングサービスにおける販売データ、企業の公表する知的財産別売上高のデータを用いることが考えられる。この中から本研究では、公開される頻度が高く、複数のカテゴリの情報を収集できるという理由からオンラインショッピングサービスのデータを用いて検証を行った。オンラインショッピングサービスにて公開される販売に関するデータは、販売ランキング上位のデータのみであるため、入手できるラベルデータは、存在するページ数と同数のデータが得られる Wikipedia のラベルデータと比較して入手できる少ない。また、オンラインショッピングサービスの販売データは、ユーザの個人の購買行動と紐づいていないため、ユーザの嗜好を反映できない。そこで本研究では、第二研究において、Wikipedia の編集履歴からユーザの嗜好に関するベクトル表現を抽出した。次に第三研究において、データ数が少ない場合における予測精度を向上させる手法について検討した。続いて、第二研究および第三研究で得られた知見に基づいて、販売に関連すると考えられる Wikipedia のページ閲覧数を予測する問題を解き、予測に有用な素性を抽出した後、販売予測を行なった。

1.3 本論文の構成

本論文の構成について述べる。図 1.4 に示すように、次章にて関連研究について述べたのち、各ソーシャルメディアを用いた販売予測における課題について取り組み、最後に全体の考察と結論を述べる。まず、2 章では関連研究を俯瞰し、関連する概念を整理することで本研究の学術的位置付けを明確にする。

3 章（第一研究）では、高価であるため購入時に消費者が、複数の候補を比較・検討すると考えられる自動車の販売予測の問題に取り組んだ。実験では、記述される内容に主観的な口コミ情報が含まれており、記述内容に対応する商品が明確であるという理由からオンラインレビューのデータを用いた。まず、レビューサイトに投稿されたレビュー文章に対して感情分析を行い、感情予測タスクにおいて予測への寄与度の大きい単語素性を特定し、人気情報とした。続いて、得られた人気情報を自動車の販売数予測タスクに用い、単語素性をを用いない場

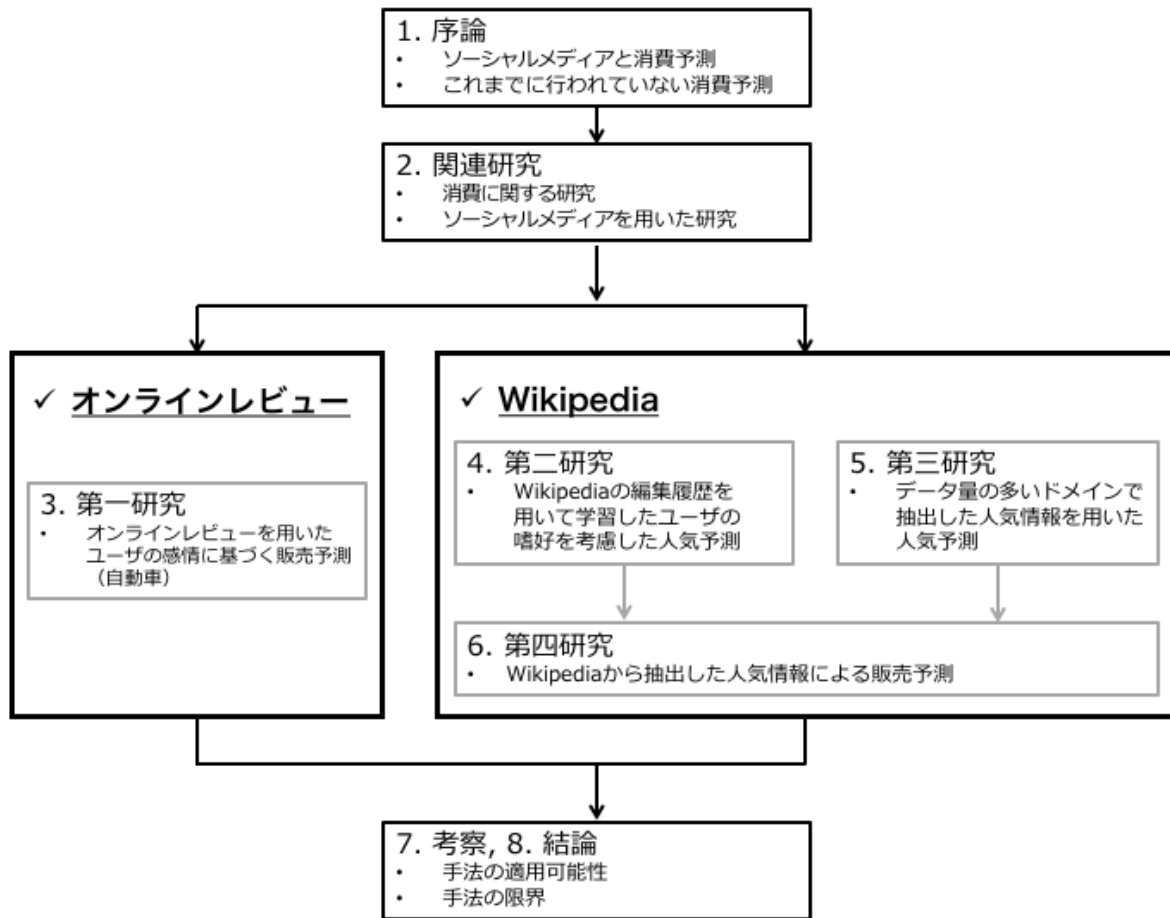


図 1.4 本論文の構成

合、単語素性をランダムに選択した場合、レビュー文中に含まれる単語素性全てを用いた場合および既存の極性辞書をもとに単語を選択した場合の予測精度と比較を行なった。実験結果から、感情予測タスクにて予測に有用であった単語素性を用いることにより、予測精度が向上することを確認した。

4章(第二研究)、5章(第三研究)および6章(第四研究)では、低価格であることが多く、商品の販売予測を行う上で対象とした商品に対する注目の度合いが重要になると考えられるマルチメディアタイトルについての販売予測の問題に取り組んだ。実験では、マルチメディアタイトルの網羅性が高いという点とメディアをまたいだデータであるという点から Wikipedia のデータを用いた。また、販売数のデータとして、公開される頻度が高く、複数のカテゴリの情報を収集できるという理由からオンラインショッピングサービスのデータを用いた。オンラ

インショッピングサービスにて公開される販売に関するデータは、販売ランキング上位のデータのみであるため、入手できるラベルデータは、存在するページ数と同数のデータが得られる Wikipedia のラベルデータと比較して入手できる少ない。また、オンラインショッピングサービスの販売データは、ユーザの個人の購買行動と紐づいていないため、ユーザの嗜好を反映できない。そこで本研究では、第二研究において、Wikipedia の編集履歴からユーザの嗜好に関するベクトル表現を抽出する手法について検証し、続いて第三研究において、データ数が少ない場合における予測精度を向上させる手法について検討する。最後に第二研究および第三研究で得られた知見に基づいて、第四研究にて販売に関連すると考えられる Wikipedia のページ閲覧数を予測する問題を解き、得られた人気情報を用いて販売予測を行なった。

以上を踏まえ 7 章にて、販売に関連する指標を予測する問題を解くことで獲得した人気情報による販売予測の手法について概観し、その限界と応用可能性について議論する。最後に、8 章で本研究の総括を行う。

第2章

関連研究

本章では、ソーシャルメディアから販売予測に関連するタスクにより予測に有用な情報を得る提案手法の関連研究について述べる。まず、大量生産と大量消費を行う社会の出現とそれに伴う消費に関わる研究の起こりについて述べる。続いて、商品の販売を予測する研究について、消費者についての情報を得る情報源の変遷と関連研究について説明する。最後に、ソーシャルメディアを用いた消費予測に関わる既存研究について述べる。

2.1 消費に関わる研究

本節では、消費に関わる研究の起こりと消費者の行動および企業の活動と消費に関わる研究について説明する。

2.1.1 消費に関わる研究の起こり

本節では、消費に関わる研究の歴史について述べる。第二次世界大戦以降、先進国を中心に工業化が進んだことから、大量生産・大量消費を行う社会が出現した。このような社会においては、生産が消費を上回り、消費者が選択に悩むほどモノが存在した。その結果、企業にとって「生産されたものが必ず売れた」それ以前のモノ不足の時代から「売れるモノ」を作る必要のある時代へと変わった。モノが不足した時代からモノが余る時代への移行は、「作ったモノを売る」という生産志向から「作ったモノを売り切る」という販売志向へ移り、その後、「売れるモノを作る」という顧客志向へと企業の経営姿勢を変化させた。

この顧客志向のマーケティングの考え方は、モノの供給過剰が生じた1950年代の米国にて登場し、その後、マーケティング・マネジメント論として体系化されるようになった。マーケティングにおける著名な学者の一人であるコトラーは、「マーケティング・マネジメントとは、標的市場を選択し、優れた顧客価値の創造、伝達、提供を通じて、顧客を獲得、維持、育成する技術である」と述べている。この考え方に基づいて企業にとって効果的なマーケティングの戦略を整理すると、5つのステップに分割される。まず、自社を取り巻く市場環境を中心としたさまざまな環境について十分なリサーチ（R）を行う。続いて、リサーチにより得られた情報に基づいて、比較的類似したニーズをもつ消費者グループを発見し、それをセグメント化（セグメンテーション；S）を行う。これに続いて、標的とするセグメントを決定（ターゲティング；T）し、標的セグメントの消費者に対して、自社の提供物の価値を他所のそれよりも高く評価してもらうため、効果的なポジショニング（P）を行なう必要がある。このセグメンテーション・ターゲティング・ポジショニングの一連の過程はSTPと呼ばれ、戦略的マーケティングの柱とされる。その後、STPにより確定されたポジショニングを前提に、具体的なマーケティング活動に落とし込むマーケティング・ミックスが置かれる。マーケティング・ミックスでは、製品、価格、流通、広告の4つの要素が考慮される。これら四要素が決定された後、計画の実行段階に移り、得られる結果を監視、評価することで次の戦略と戦術を決定する。

以上のように行われるマーケティング・マネジメントの基本プロセスにおいて、柱とされるSTPの段階における消費者のニーズが重要であるとされる。企業が市場の中で存続し、成長を続けるためには、「顧客価値を創造し提供する能力」と「顧客を獲得し維持する能力」が必要となる。そのためには、競合となる他社と比較して消費者のニーズをよりよく充足する「市場提供物」を創造し、それを競争企業と比較してより好まれる方法で提供する必要がある。そのためには、消費者が自社の提供物やその提供方法について、他社との差異を創出し、消費者のニーズや評価基準という観点において、他社より優れている状況を作り出す必要がある。したがって、マーケティングにおける競争優位の確立には、消費者のニーズや行動を理解し、消費者に対して効率的かつ効果的に働きかける必要がある。

このような背景のもと、消費に関わる研究は、消費者の行動の理解に主眼を置いた消費者行動に関する研究と企業の活動に主眼を置いたマーケティングの研究の2つが行われてきた。以降の節では、それぞれの観点での研究についてまとめる。

2.1.2 消費者の行動に関する研究

消費者の行動を理解することは、そのニーズを理解し、企業の成長と存続に繋がるため重要な課題である。消費者の行動の研究は分析課題という観点から、「なぜ買うか」を分析する購買動機に関する研究、「何を買うか」といった購買行動に関わる研究、意思決定に伴う内的なプロセスの解明に関する研究の3つに大別される。本節では、それぞれについて研究の概要を述べる。

購買動機に関わる研究

第一の研究は、消費者の購買動機を研究するモチベーション・リサーチである。マーケティングを行う上で、消費者のニーズを解明し、その購買動機を把握することは非常に重要であり、消費者行動分析において最も基本的な課題である。一般に、動機づけとは、人を行動へと駆り立て、その行動を方向づけ、維持する心理的なメカニズムやプロセス全般のことを指す。また、上のメカニズムの中で、特定の行動を駆動し、方向付け、維持する内的な要因や状態を指す概念が「動機」である。未充足のニーズによって生じた緊張状態が引き金となり、緊張状態の緩和やニーズの充足につながる行動が生起される。このとき、行動の駆動力となる内的状態を動機または「動因」と呼び、行動を生起する外的要因を「目標」または「誘因」と呼ぶ。動機や動因によって、目標の達成や誘引の獲得につながる行動が駆動され、その結果、ニーズの充足と緊張状態の低減がもたらされる。ここで、「ニーズ」は、行動を発言させる生理的ないし心理的な未充足状態のことを指し、生理的なものと心理的・社会的なものに大別される。動機づけの心理的メカニズムにおいては、内的要因として未充足のニーズと外的要因として目標が動機により結び付けられた場合に行動が生起される。

購買行動の文脈では、期待や経験を通じて購買動機が形成される。マーケティングにおいては、この認知プロセスに働きかけることで、購買動機と購買行動に影響を与えられる可能性が出てくる。他方、購買動機の把握は直接的な観察や測定が困難である。モチベーション・リサーチの研究は、精神分析学や質的調査技法を用いることで、購買動機についての探求を行っている。

供給が需要を上回る社会において、企業の経営姿勢は、生産志向から顧客志向へと変化し

た。このような背景の中で、消費者の購買意欲を刺激する方法や購買行動の要因についての研究が注目された。購買行動の要因を分析する研究では、フロイト流の精神分析学に基盤を置き、潜在意識や無意識のもとに行われる非合理的に見える動機や情動的な動機が積極的に扱われた。典型的な事例として、ディヒターらの研究では、自動車が単に移動や運搬の道具ではなく、社会的地位や経済力の象徴であることが示された [Dichter 60]。また、パッカードらは、自動車の車種ブランドは、その所有者の人となりや理想像を語る一番の手段であると指摘した [Packard 57]。自動車のみならず、衣服や装身具についても同様の研究が結果が示されている。

購買行動に関わる研究

未充足のニーズにより動機づけられた消費者は、ニーズを充足するために製品・サービスを購入する。その際、消費者は複数の選択肢の中から購入する対象を選択する。この製品・サービスの選択において、購買された対象を記述し、購買される対象を予測する研究について述べる。

購買行動における選択に関わる研究では、買物を行う際に購入した商品を記入した買物日記パネルから得られる購買履歴データを用いたブランド・ロイヤルティ研究に端をなし、多くのブランド選択モデルの研究が行われた。購買動機の研究が観察することのできない内的な意識といった要素に注目する一方、購買行動の研究は、顕示的な行動として観察されるブランド選択を分析した。パネル調査の結果得られた購買履歴に関するデータをもとに、オペレーションズ・リサーチや経営科学などの分野で開発された確率モデルの手法や考え方を用いて分析が行われた。ブラウンの研究では、「シカゴ・トリビューン」紙の消費者パネル 100 世帯文の購買履歴を分析している [Brown 53]。その結果、ブランド選択のパターンが、同じブランドばかり購入する「完全ロイヤルティ」型、規則性のない「ロイヤルティなし」型、両者の中間の「分割ロイヤルティ」型または「不安定ロイヤルティ」型に分類できることを明らかにした。また、カニングガムらはブランド・ロイヤルティの指標として、最も購入されるブランドの購入割合を購買履歴データを用いた分析から提案している [Cunningham 56]。

購買行動を分析した研究は、購買履歴データの分析を起点として行われてきた。その後、価格や広告といった刺激と購買行動の関係を研究する、刺激-反応アプローチと呼ばれる研究が行われた。このアプローチでは、消費者の行動を観察可能な二つの側面、「刺激」とそれに対する「反応」を分析する。刺激-反応アプローチは、刺激を受けてから反応に到るまでの内的

プロセスをブラックボックスとして扱う反応注目型モデルと内的プロセスを解明することを試みる構造明示型モデルに分けられる。前者の反応注目型モデルでは、ブランド選択の確率を高い精度で当てることを目的としており、ベルヌーイモデルや多項ロジットモデルといった手法が存在する。後者では、消費者が特定の選択に至った内的・心理的なプロセスを解明することに重点が置かれ、刺激と反応を結びつける媒介変数を導入することで明示的な内部構造を説明することを目的としている。

内的プロセスの解明に関わる研究

消費者の行動動機と実際の行動の間に存在する消費者の内部における情報処理のプロセスに注目したのが、内的プロセスに関わる研究である。価格や広告といった刺激と選択や購買といった反応を結ぶ媒介変数として、態度が注目された。

社会心理学者のオルポートによると、態度とはある対象に対して特定の方法で反応しようとする傾向性、ないし行動の準備状態を指す。また、フィシュバインらは態度を「ある対象に対して好意的あるいは非好意的に一貫して反応する学習された先有傾向」と定義している。消費者行動の研究において、態度により購買行動を説明・予測できると考えられるため、注目された。

1960年代における研究では、刺激と反応を結ぶ媒介変数として態度が注目されたが、その後、消費者が行う情報処理のプロセス全体に注目が移った。これは、ニューウェルとサイモンにより提唱された「人間の問題解決行動」モデルを契機として、認知心理学の「情報処理モデル」に依拠する形で登場した。このモデルでは、消費者を情報処理システムの一つとして捉え、情報を探索・取得・解釈・統合する内的なプロセスに焦点を当てる。一連の研究は、ベットマンにより「消費者選択の情報処理理論」として体系化されている。

消費者情報処理理論では、消費者行動を能動的な問題解決行動として捉え、問題解決の手段としての製品やサービスを選択・購買するために、消費者が自ら進んで必要な情報を探索・取得・解釈・統合するプロセスに焦点を当てている。ベットマンのモデルは、消費者の情報処理プロセスを描写した概念モデルであり、情報処理能力、モチベーションと目標階層、スキャナーと中断、注意と知覚符号化、情報取得、記憶、意思決定過程、消費と学習過程という8つの段階に分けられる。近年では、このモデルにさらに、情報を処理する「動機づけ」、それを可能にする「能力」と情報処理の「機会」を加えた枠組みが用いられている。

2.1.3 消費と企業に関する研究

消費は、生産された製品やサービスを購入する消費者とそれを提供する企業が存在することによって成立する。ここでは、製品やサービスを提供する企業の研究についてマーケティング・サイエンスの研究を中心に述べる。マーケティング・マネジメント論に基づいた情報収集、セグメンテーション・ターゲティング・ポジショニング、マーケティング・ミックスの決定、市場におけるテストとコントロールの各段階についてそれぞれ述べる。

情報収集

企業のマーケティング活動において、意思決定を行う際の情報は重要である。マーケティングにおいて有用となる情報源には、社内における日常業務により蓄積される社内記録、企業の意思決定のために必要となる情報を組織的、系統的に収集、分析し利用する活動であるマーケティング・インテリジェンス活動、マーケティング上の特定の課題に対応するために行われる一連の活動であるマーケティング・リサーチの3つがあげられる。

社内記録の代表的な例としては、小売業者やサービス業者におけるPOSデータがあげられる。販売や取引に関連する多くの情報が電子化されるようになったため、これらのデータは自動的に蓄積されるようになっている。また、近年では最終消費者のデータベースである、顧客データベースの重要性が高まっている。顧客データベースには、購買履歴データ、ユーザ登録に関するデータといったデータが含まれる。消費者個人の購買履歴データは、マイレージサービスやポイントカードのサービスに代表されるフリクエンシー・プログラムにより収集される。この仕組みにより企業は、各消費者個人についての購買の傾向といった情報を把握することができる。このように収集されたデータをもとに、商品の推薦といったことも行われている。これまで述べてきた社内記録は、定量的なデータが中心であるが、消費者から寄せられるクレーム情報や過去に営業先に行なった提案内容といった定性的な情報も社内記録として有用となる場合がある。

マーケティング・インテリジェンス活動は、企業の意思決定のために必要となる情報を組織的、系統的に収集、分析し利用する活動である。この活動においては、取引先や顧客と直接接する営業や販売担当が重要な情報源となる。また、モニタリング活動を通じて収集された顧

客からの意見も重要な情報源となりうる。この他にも、取引先から提供されるデータや商用データ提供期間から得られるデータも有用となりうる。近年では、複数の小売機関から購買履歴データを収集し、それらを整理して提供するサービスが始まっている。

マーケティング・リサーチは、特定の課題に対応するために行われるという点で、日常的な情報の収集活動であるマーケティング・インテリジェンス活動や日常業務の結果として蓄積される社内記録とは異なる。マーケティング・リサーチでは、まず意思決定課題の特定と調査目的の明確化を行う。その後、調査計画の策定を行い、必要な情報を収集する。このとき用いられる情報としては、質問・観察・実験の結果得られる一次データと社内外の情報源から得られる二次データがあげられる。二次データには、社内記録といった内部データと、取引先や政府機関から得られるデータやインターネット上から収集される口コミ情報がある。一次および二次データが収集されたのち、得られた情報をもとに分析を行い、その結果をもって意思決定につなげる。

セグメンテーション・ターゲティング・ポジショニング

マーケティングにおける具体的な行動のための意思決定を行うには、明確なマーケティング戦略を決定する必要がある。マーケティング戦略は、セグメンテーション、ターゲティング、ポジショニングの3つの段階に分けられる。

セグメンテーションとは、性別や所得、年齢といった尺度により、規定された市場における顧客を分割することを指す。セグメンテーションでは、顧客の異質性を考慮して全体市場を複数の同質的な市場へ細分化する。細分化された各市場に含まれる顧客は、製品やサービスに対して同質的な認識、選好、動機を有し、別のセグメントとは異質であることが求められる。分割の尺度としては、地理的変数、人口統計的変数、心理的変数、購買行動変数といったものが用いられる。セグメンテーションは過剰に分割を行なった場合、企業にとってのコストが増大するというデメリットも存在するため、セグメント数の増加に伴うコストの増加と顧客対応の向上による収益の増加のバランスを考慮する必要がある。

セグメンテーションにより市場を分割した後、ターゲティングによって対象とする顧客を決定する。ターゲットとしたセグメントに含まれる顧客とそれ以外の顧客の行動や嗜好の違いを分析し、アプローチの手段を考察する。また、消費者個人に対して直接アプローチを行うダイレクト・マーケティングでは、企業はターゲットとしたセグメント内で顧客獲得を行うために

消費者の属するセグメントを予測する必要がある。ターゲティングを行なった後、対象とする顧客に対して提供する価値を戦略ポジショニングの決定により明確化する必要がある。明確なポジショニングを行うことにより、販売する製品が標的の顧客に対して提供する価値が規定される。このとき、価格のみでなく、機能的な側面と感情的な側面を含む品質の両面から価値をとらえることが重要となる。

これらの意思決定がなされた後、マーケティング・ミックスの決定が行われる。

マーケティング・ミックスの決定

マーケティング戦略が決定された後には、マーケティングにおける戦術に相当するマーケティング・ミックスを決定する必要がある。マーケティング・ミックスは、製品（プロダクト）、価格設定（プライシング）、流通、広告（プロモーション）の4つの要素からなる。これらの要素を決定することで、具体的な行動が決定される。

製品のデザインでは、顧客の製品選択行動をモデル化において属性アプローチがしばしば用いられる。これは、顧客の欲望と製品の属性を結びつけることで、製品が顧客に提供する価値を考えるアプローチである。このアプローチの基本的な考え方は、製品そのものではなく製品を構成する属性が顧客に満足感をもたらすという考えである。製品は顧客に対して満足をもたらす複数の属性からなっており、顧客はそこから得られる総合的な満足度とコストを比較し、最も価値の高いものを選択すると考える。製品が複数の属性から成ると考えることによって、多次元の属性空間上に製品を位置付けることが可能になり、競合する製品を含めて一つの共通空間上に製品を位置付けることも可能になる。これを製品ポジショニングといい、この属性空間上で自社製品の相対的なポジショニングを認識することで、顧客に提供できる便益や競合製品に対する競争優位の獲得方策を考察することができる。

製品価格の決定は、価格の上昇による消費者の購買意欲の減退と価格の低下による利益率の低下との間のトレードオフを考慮して行われる。このような状況下における価格設定の方法の一つとして、消費者心理を直接聞く、PSM(Price Sensitivity Measurement) 分析が存在する。この手法は、消費者に質問を行うことで、価格の目安を見つけようとする手法である。消費者の心理を探った後、利潤を最大化する価格を決定する。多くの企業において、原価に一定のマージンを上乗せする、マークアップ・プライシングという価格設定方式が採用されている。マージン率は販売量の変化率と価格の変化率により定義される価格弾力性に大きく依存してい

る。価格弾力性が高いほど消費者は価格に敏感に反応する。消費者が製品間の相違を意識しないものをコモディティと呼び、価格弾力性が高くなる。利益率を高めるためには、価格弾力性を下げることが必要となり、競合製品に対する差別化を明確にする必要がある。

プロモーションの機能は、製品の認知・情報を高めることで購買を促すことで、その活動は、広告、パブリシティ、販売促進、人的販売の4つに分けることができる。これらの活動は消費者を自社製品に引っ張るプル戦略と製品を認知し興味を示した顧客に対して購買を働きかけるプッシュ戦略に大別される。最後に、自社の製品が消費者の手に届くよう適切な流通販路を確保し、購買意欲のある消費者が実際に商品を購入できるようにする必要がある。

テストとコントロール

市場導入を行なった結果、期待した通りに製品が市場に受け入れられているかを時系列で監視することが重要となる。マーケティング・ミックスの決定に基づいて採択された製品仕様のチェックまたは試作品段階でのチェックは、実際にフルスケールで生産に入り市場導入する前に行うことが望ましい。ここでは、製品の試験的導入段階（テスト）と製品の市場導入後に関わる知見（コントロール）について触れる。

製品のテストとしてしばしば行われる手法は、地域を限定して販売し、得られた知見をもとにマーケット・シェアを予測する手法である。ここでは、新製品を試しに購入する消費者の割合（試用購入）や試用購入した消費者のうち製品に満足し繰り返し購入する消費者の割合を観察することで、フルスケールでの導入のシミュレーションを行う。実際の店舗を用いずに仮装の店舗で購買の模擬実験を行う手法も存在し、プリ・テスト・マーケティングと呼ばれる。代表的な事例として、Silk らの ASSESSOR モデルがあげられる。

2.2 商品の販売を予測する研究

企業の経営姿勢が顧客志向に変わるとともに、商品の販売を予測することの重要性が増加した。商品の販売を予測する上で、消費者に関する情報を考慮する方法が重要となる。本節では、用いる消費者に関する情報という観点から商品の販売を予測する研究について整理する。

まず、商品の販売予測に用いる消費者の情報として、パネルデータをはじめとする消費者調査の結果を用いた研究について述べる。インターネットやスマートフォンが普及する以前

には、個人の消費者が情報を発信することは難しかったため、個人の消費者の購買行動や商品に対する評価といった情報を入手する方法は限られていた。消費者についての情報を入手するために用いられた方法として、パネル調査のデータを用いる方法があげられる。Fourtらの研究 [Fourt 60] では、パネル統計を用いて、新しい日用品についての予測を行なっている。また、Massyの研究 [Massy 69] では、コンビニエンスストアにおける商品の需要をパネルデータを用いて予測している。パネル調査以外のデータを用いた事例としては、購入意識調査を行なった結果に基づいて商品の販売予測を行なっている研究が存在する。Morwitzらの研究 [Morwitz 92] では、ユーザを分割して、購入意識調査と購入意識の度合いに応じた過去の購買行動のパターンを用いて販売予測を行なっている。また、Armstrongらの研究 [Armstrong 00] では、複数の対象にて検証を行い、購入意識調査の結果を用いることで、過去の販売の変動のみに基づく予測と比較して予測精度が向上することを示している。このように先行研究において、パネル調査や購入意識調査の結果を用いる手法の有効性が示されているが、これらの手法ではデータの収集においてコストがかかるという点が問題となる。

また映画を中心とした収益の予測では、パネル調査や購入意識調査の結果以外のデータを用いている研究も存在する。消費者についての調査結果を用いない研究では、メディアに掲載される批評家によるレビューが消費者の行動と関連しているという仮定のもと、雑誌や新聞といったメディアに掲載される批評家によるレビューを用いて予測を行なっている。Litmanの研究 [Litman 83] では、映画の興行収益において批評家によるレビューの採点が影響を与えていることを示している。また、Eliashbergらの研究 [Eliashberg 97] では、相関分析から批評家のレビューにより映画の収益を予測できることが示されている。この他にも、Basuroyらの研究 [Basuroy 03] において、批評が映画の興行収益に影響を与えることおよび興行収益の予測に用いることができることが示されている。近年、オンラインレビューが普及したことを背景に、消費者個人の口コミ情報を取得できるオンラインレビューの分析が行われるようになった。

続いて、ブログやオンラインレビューサイトといったソーシャルメディアから商品の販売予測に用いる情報を取得した研究について述べる。インターネットの普及とともに利用者の増加したブログやオンラインレビューには、以前には取得することが難しかった消費者の発する商品への意見や感想といった口コミ情報が記録される。ブログやオンラインレビューサイトに投

稿される情報は、商品の販売予測を行う上で重要な消費者の発信する情報として多くの研究にて用いられている。先行研究では、映画を中心として、音楽、テレビゲーム、書籍といった対象についての販売予測が行われている。Dellarocas らの研究 [Dellarocas 07] では、オンラインレビューの指標を既存のモデルに加えることで予測の精度が向上することを示している。Liu の研究 [Liu 06] では、口コミ情報の量が映画の公開初期の興行収益について説明力をもつことを示している。Duan らの研究 [Duan 08] においても、オンラインレビューへの投稿数が映画の興行収入に影響を与えることが示されている。映画以外についての研究では、Chevalier の研究 [Chevalier 06] において書籍の販売でのオンラインレビューの影響が分析されている。この研究では、2つのレビューサイトを分析対象として、対象についてのレビューの評価の上昇が販売数の増加につながることを示している。また、低い評価のレビューが高い評価のレビューよりも影響力が大きいことを示している。また、Zhu らの研究 [Zhu 10] では、テレビゲームのレビューを分析対象として、レビューが販売に与える影響は人気度の低いゲームおよびインターネットにおける経験が豊富なユーザに対して大きいということが示されている。音楽に関連する販売予測では、Dhar らの研究 [Dhar 09] において、アルバムについてのブログの投稿数および雑誌に掲載されるレビューのいずれもが将来のアルバムの販売数と相関することが示されている。

また、ソーシャルメディアと商品の販売に関わる研究では、既存のメディアや広告の影響とソーシャルメディアの関係を分析した研究が存在する。Trusov らの研究 [Trusov 09] では、ソーシャルネットワーキングサービスにおける加入者の情報を分析することで、口コミ情報の効果と伝統的なマーケティング手法の比較を行なっている。Onishi らの研究 [Onishi 12] では、既存のメディアとソーシャルメディアの関係を分析し、両者の関わりが販売に与える影響について分析している。また、Kim らの研究 [Kim 17] では、映画およびテレビゲームを対象として発売前および発売直後の時期における広告と口コミ情報の影響を分析している。

さらにソーシャルメディアへのユーザの投稿以外にも、特定の語についての検索回数といった情報を用いて販売予測を行う研究も存在する。検索回数の情報は、一部のユーザの投稿によってデータが蓄積されるブログやオンラインレビューと異なり、利用する全てのユーザの行動が蓄積されるため、収集されるデータの規模がブログやオンラインレビューと比較して大きくなる。Choi らの研究 [Choi 12] では、Google における検索回数から近い将来の自動車の販売数、失業率といった指標を予測できることを示している。また、Goel らの研究 [Goel 10] で

は、検索クエリの量を用いることで、公開から一週間の映画の興行収益、発売から 1 ヶ月間のテレビゲームの販売数、音楽のランキングを予測できることが示されている。

オンラインレビューやブログを用いたこれらの予測の研究では、予測に用いる素性として、投稿の数やレビューに付与されている採点結果といった情報が用いられている。これらの指標は、その分野における知識に基づいて、モデルを作成する際にあらかじめ選択されている。

2.3 ソーシャルメディアを用いた予測を行う研究

本節では、ソーシャルメディアからユーザの行動や感情に関わる情報を取得し、予測に用いている研究について述べる。具体的には、ソーシャルメディアを用いた選挙予測、商品の販売予測について述べる。

2.3.1 ソーシャルメディアを用いた選挙の予測

まず、ソーシャルメディアを用いた選挙の予測を行なった研究について述べる。近年、選挙におけるソーシャルメディアの利用が解禁されたため、選挙活動において政治家が積極的にソーシャルメディアを活用している。このような背景から、ソーシャルメディアを用いた選挙結果の予測および支持率の調査についての研究が各国で行われている。

O'Connor [O'Connor 10] らは、Twitter に投稿された内容に含まれる感情を示す語が世論調査の結果と相関することを示した。この研究では、調査項目ごとにあらかじめ指定した語を含む投稿内容を取得し、その中に含まれる肯定的な語と否定的な語の出現回数の比をもとに世論を数値化している。Wang [Wang 12] らの研究では、Twitter に投稿される内容に対して感情分析を適用することで、2012 年のアメリカの選挙における候補者に対する世論をリアルタイムで分析するシステムを提案した。また、Tumasjan [Tumasjan 10] の研究では、ドイツの国政選挙を分析対象とし、Twitter の投稿に含まれる政党名を数えあげることによって選挙結果を予測できると主張している。しかしながら、Andranik らの研究に対しては Jungherr ら [Jungherr 12] が否定的な研究結果を示している。この他にも選挙の予測を行なった研究として、2010 年に行われたイギリスの選挙を複数のソーシャルメディアの情報を用いて予測する研究 [Franch 13] や 2011 年にニュージーランドにて行われた選挙の結果を試みる研究 [Cameron 16]、日本における参議院選挙の結果の予測を試みた研究 [那須 15] が行われてい

る。これらの研究では、候補者や政党名が投稿に出現した回数や Twitter におけるフォロワー数、ツイートの投稿数といった、これまでの知見に基づいて有効と考えられる指標をヒューリスティックに選択した素性が予測に用いられている。またヒューリスティックに選択された特徴以外では、感情分析を行なった結果を予測に用いる場合や、あらかじめ定められた感情を示す語の出現回数を用いて予測を行なっている。

2.3.2 ソーシャルメディアを用いた商品の消費予測を行う研究

続いて、ソーシャルメディアから取得したデータを用いて消費に関わる予測を行なっている研究について述べる。ソーシャルメディアのデータによる消費予測の研究は、消費者個人に注目し、個人が購入・消費する商品を予測する研究および消費者全体に注目し商品の販売数を予測する研究の二通りに大別できる。

個人の購買行動といった消費者個人に着目した消費予測に関連する研究について述べる。ソーシャルメディアの情報を用いて個人の消費者が特定の商品を購入するか、または選択肢の中から選択する商品を予測する研究が存在する。消費者個人が特定の商品を購入するかを予測することは、商品の推薦を行う上で重要である。ここでは、予測の際に用いるソーシャルメディアのデータの種類ごとに関連研究を述べる。

個人の消費者の購買行動を特定する際に用いられるソーシャルメディアの情報として、Twitter や Sina Weibo といったマイクロブログの情報を用いる研究が存在する。[Ding 15] では、Sina Weibo の投稿から購入意思が含まれる投稿を CNN により発見することを試みている。また、[Tsuboi 15] では、Twitter においてユーザが過去に投稿した内容を考慮してデジタルカメラおよびパーソナルコンピュータを購入するか予測している。また、[Korpusik 16] では、RNN を用いて Twitter に投稿された内容から特定の商品の広告を受容するユーザかを推定することを試みている。[Lo 15] においても、Twitter における各ユーザの投稿内容から潜在的な顧客を特定することを試みている。

ソーシャルメディアから投稿内容以外の情報を得て予測に用いる研究では、ユーザのプロフィール情報を用いて購買を予測する研究が存在する。[Zhang 13] では、SNS 上に記載されたプロフィールの情報とオンラインショッピングにおける行動の関係を分析している。[Xiao 16] では、プロフィールの情報に加えて、投稿内容とユーザの経済状態からレストラン

における消費を予測している。また、ユーザの購買行動の予測を行うため、異なるサービスの間のユーザ情報を紐づける手法を提案している研究も存在する [Wang 15]。

次に、ソーシャルメディアを用いた商品の消費される量、すなわち販売数の予測に関連する研究について述べる。本節では、ソーシャルメディアを用いた販売数予測の研究について述べる。

販売予測を行う際に用いられるソーシャルメディアとして、ブログや Twitter, オンラインレビューなどが用いられてきた。販売数予測の対象としては、豊富なデータ数と明確な利用目的という観点から映画や株価のデータを予測する研究が活発に行われてきた。映画の予測では、Twitter のデータを用いた [Asur 10] らの研究、ブログから得た情報を用いた [Mishne 06] らの研究、レビューの情報を用いた [Dellarocas 07] の研究が知られている。また、Baek らの研究 [Baek 14] では、Twitter, Yahoo!Movies, Youtube, ブログという 4 つの異なるソーシャルメディアからの口コミ情報がそれぞれ映画の売上における影響を分析している。

また、映画に関する予測では、ソーシャルメディア上から取得したデータを直接販売予測に用いるのではなく、感情分析を行った後に販売予測に用いる研究も存在する。[Yu 12] では、映画の売上予測を行う際にブログに対して行った感情分析の結果を素性として加える事で予測精度が向上することを示している。また、レビューにおける商品の言及回数といった単純な素性だけでなく、消費者の感情を素性として加えることで販売数予測の精度を向上できることが知られている [Ahn 14]。

映画以外の商品についての予測としては、Twitter の投稿から iPhone の販売を予測した [Lassen 14] 研究や Twitter のデータからスマートフォンのモデルごとの販売数を予測した [Tuarob 13] 研究が存在する。また Dewan らの研究 [Dewan 14] では、伝統的なメディアおよびソーシャルメディアの音楽業界における売上との関係が分析されており、単体の商品ではないものの Boldt らの研究 [Boldt 16] では、Facebook 上の企業に関する Like 数から、企業の商品の販売数を予測することを行っている。この他にも、Twitter や Wikipedia, Google トレンドの情報を組み合わせて、コンテンツ作品についての予測を行なった保住らの研究 [保住 14] が存在する。

第3章

感情分析による人気情報の抽出と自動車の販売予測

3.1 本章の背景

本章では、販売予測においてソーシャルメディアが有効であると考えられる商品から、予測対象として重要と考えられる自動車の販売予測を行なった。自動車の販売予測において用いるソーシャルメディアとしては、ブログ記事、Wikipedia、オンラインレビューなどが考えられる。オンラインレビューには、販売予測において有用であることが知られている商品についての口コミ情報が記されており、各レビューとレビュー内で記述されている商品の対応関係が明確であるという特徴がある。そこで本章では、商品に対するユーザの主観的な意見や感想が記述されるオンラインレビューの情報から予測に有用な素性を抽出し、抽出した素性を用いることで自動車の販売予測を行った。

オンラインレビューをはじめとするテキスト情報は、インターネットやスマートフォンが普及したことにより、ニュース記事、ユーザレビューといった様々な形でウェブ上に大量に存在するようになった。こうしたウェブ上のテキストデータは、収集や解析が比較的容易であるため将来予測における素性として用いられており、多くの研究でその有用性が示されている [Dellarocas 07, Bollen 11]。株価の予測では特定の語の出現頻度 [Wuthrich 98] やソーシャルメディアにおける特定のハッシュタグの出現頻度をはじめとする素性 [Ruiz 12] が有用であることが知られている。[Wuthrich 98] では、ウェブ上の記事に含まれる特定の単語の出現回数

と株価が相関することが示されている。しかしながら、株価との相関が考えられる単語の選別は専門家によってなされており、辞書の構築コストという問題点があった。

一方、[Caruana 98] や [Silver 07] において、特定のタスク（元タスク）で学習された特徴表現を別のタスク（目標タスク）に用いることで、単体で学習したときと比較して学習の性能が向上することが示されている。特に [Silver 07] では、別タスクで学習された特徴表現を目標タスクで用いることで、目標タスクにおける予測精度を向上できることが示されている。また、商品の販売予測においては、レビューをはじめとする口コミ情報から得られるユーザの感情を過去の変動パターンと組み合わせて用いることで、対象とする商品の販売予測の精度を向上できることが知られている [Ahn 14, Yu 12]。口コミ情報に対する感情分析タスクでは、感情を予測する際の素性としてレビューに含まれる単語を用い、感情の予測に寄与する単語素性を特定することが可能である。以上のことから、感情分析タスクにおいて学習された単語素性の中から、感情予測に対する寄与度を指標として単語素性を選別し、販売予測タスクに用いることで予測精度を向上できると考えられる。

そこで、本章では、ウェブ上のデータを利用した商品の販売予測の精度向上を目指し、ユーザの感情予測に有用であった単語素性を利用する方法を提案する。本手法は、図 3.1 に示すように、口コミ情報に対する感情分析において学習された素性を販売予測タスクで利用する。具体的には、レビューに対して行った感情分析から得られた素性と季節性を考慮した過去の販売情報を用いて、自動車の月次の販売数を予測した。実験では、自動車に関するレビューサイトからユーザの投稿を収集し、28 の車種について 2011 年 10 月から 2015 年 7 月までの期間のデータを用いて販売予測を行った。実験の結果として、予測精度の向上する車種と向上しない車種が存在したものの、全体としての予測精度は向上することを確認した。

本章の貢献は、自動車のレビューに対する感情分析において消費者感情の予測に寄与する素性が、自動車の販売予測において有用であることを示したことである。

本章の構成は以下の通りである。まず、3.2 節にて関連研究について述べ、3.3 節にて提案手法について述べる。続いて、3.4 節にて検証実験において用いるデータの取得手順と実験について詳細を説明する。その後、3.5 節にて実験結果を考察し、最後に 3.6 節にて本章の結論を述べる。

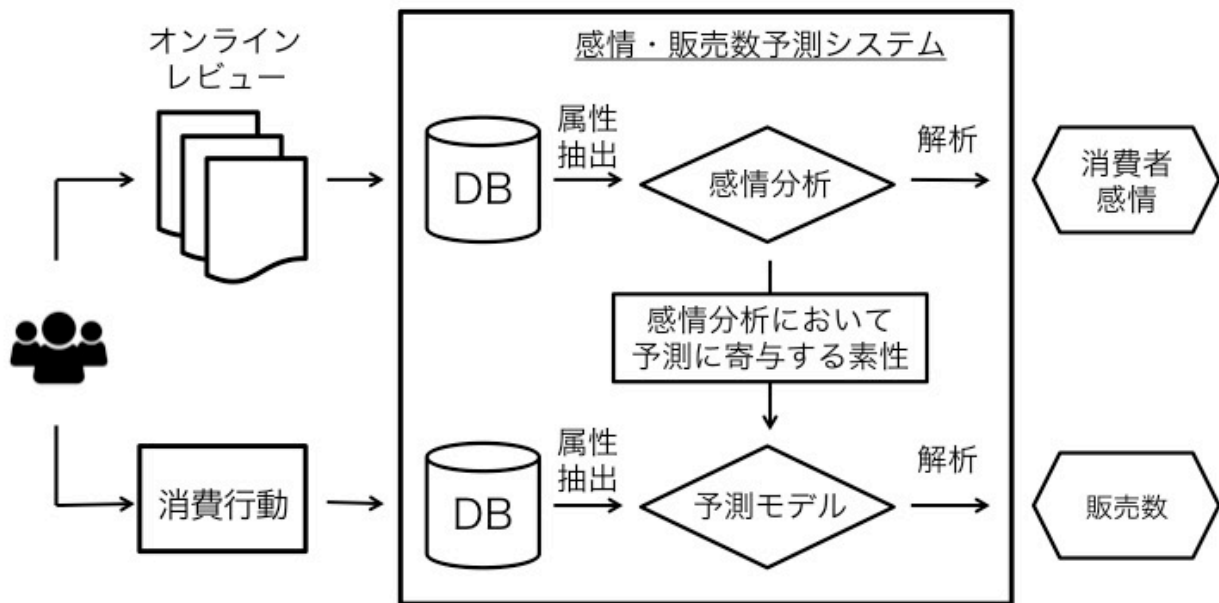


図 3.1 オンラインレビューを用いた自動車の販売予測の概略図

3.2 関連研究

3.2.1 感情分析

感情分析は、特定の対象物に対する書き手の意見を見つけるタスクと定義される [Feldman 13]. 商品の購入をはじめとして、人々は意思決定を行う過程において、他者の意見に影響される。そのため商品に対して消費者が抱く感情を知ることは、売り手にとって重要である。こうした背景から感情分析に関連する研究は、近年大きな注目を集めてきた。

感情分析のアプローチとしては、辞書を用いる手法とラベル付きデータから学習する手法に大別される。辞書を用いる手法では、あらかじめ肯定的および否定的な語を登録しておき、それらを用いて与えられた文章の感情を測定する [Taboada 11]. 既存の辞書を用いる場合、収集したデータに対して容易に適用できることが多いが、そのような辞書には一般的な語しか含まれないため、適用するドメインに特異的な語を評価することはできない。適用するドメインに

特異的な語を考慮して感情分析を行う場合、タスクに適した形で辞書を作成する必要があり、そのコストは非常に大きい。一方、ラベル付きデータから学習する手法では、収集したデータに対して機械学習を行い、感情分析を行う [Pang 02, Gamon 04, Yessenalina 11, Sharma 12]. 機械学習により感情分析を行う場合、学習を行う際に感情を表すラベルが必要となり、問題設定によってはラベルデータの入手が困難な場合もある。しかしながら、多くのレビューサイトではユーザの投稿するレビュー文章と合わせてユーザによる対象物に対する採点結果が公開されている。ユーザによる採点は、商品に対する評価を数値化したものであるため、それを感情分析における教師ラベルとすることが可能であると考えられる。そこで、本章ではレビューサイトから収集したデータに含まれるユーザ採点をラベルとして機械学習により感情分析を行った。

感情分析を販売予測や株価の予測に応用した研究として、映画のレビューに対して感情分析を適用し各映画の売上を予測する研究 [Yu 12] や、Twitter の文章に対して感情分析を適用し、その結果をもとに株価を予測する研究 [Bollen 11], 自動車に関するウェブ上の情報を収集し、収集したデータに対して感情分析を適用した後、販売予測に用いる研究 [Ahn 14] が存在する。これらの先行研究では、あらかじめ学習された分類器またはあらかじめ作成された辞書を用いて感情分析を行った後、販売予測を行っている。しかしながら、辞書を用いる手法は必ずしも辞書を用いず機械学習により感情分析を行う場合と比較して精度が優れているわけではないことが示されている [Pang 02]. また分析対象のドメインに特異的な語を含む辞書の作成には、コストがかかるといった問題点がある。そこで本章では、レビューデータに含まれるユーザ採点の結果をラベルとして機械学習による感情分析を行い、販売予測を行った。

3.2.2 ソーシャルメディアによる販売予測

商品の販売予測は、企業が意思決定を行う際に重要となる。予測を行う際には、過去の販売データをはじめとする統計指標のみでなく、消費者の意見や会社・製品に対する感情といった口コミ情報を考慮することが重要である。すでに複数の研究において、口コミ情報が商品の販売と相関することが示されている。口コミ情報の収集は、アンケートをはじめとする手法が一般的であったが、インターネットやスマートフォンの普及とともに、ウェブ上からデータを取得することが可能になっている。中でも、オンラインレビューサイトや各種 SNS が普及した

ことから、それらの情報を用いて商品の販売数を予測する研究が数多く試みられている。

販売予測を行う際に用いられるデータとしては、ブログや Twitter, レビューなどの個人の発信する情報, 検索クエリの回数といったユーザの行動履歴などがある。このようなデータを用いて販売予測を行う研究としては、特に豊富なデータ数と明確な利用目的という観点から、映画および株価のデータを利用した研究が活発に行われてきた。Twitter のデータを用いる研究 [Asur 10], ブログから得た情報を用いた研究 [Mishne 06] が知られており、レビューの情報を用いた研究としては [Liu 06] がある。映画以外のドメインにおけるウェブ上のデータを用いた販売予測の研究としては、検索クエリの回数から自動車の販売数をはじめとする経済動向を推測した [Choi 12] の論文やブログにおける商品の言及回数を素性として、その商品の販売数を予測する研究が存在する。株価の予測を行う研究としては、企業の有価証券報告書をはじめとする公式発表された書類をデータソースとする研究が存在する [Nakatoh 13]。

ウェブ上から取得したデータを直接販売予測に用いるのではなく、感情分析を行った後に販売予測に用いる研究も存在する。[Yu 12] では、映画の売上予測を行う際にブログに対して行った感情分析の結果を素性として加える事で予測精度が向上することを示している。また、レビューにおける商品の言及回数といった単純な素性だけでなく、消費者の感情を素性として加えることで販売予測の精度を向上できることが知られている [Ahn 14]。このように感情分析により消費者の感情を定量的に評価し、販売予測に用いることで予測の精度が向上することが先行研究にて示されている。感情分析の結果が販売予測に有効であることから、感情分析において予測に対する寄与度の高い素性は販売予測においても有効であると考えられる。しかしながら、先行研究における感情分析はあらかじめ用意した辞書を用いていたため、予測に用いられる素性は一般的な語でしかなかった。本章では、ウェブ上から収集したレビューデータに対して機械学習を用いた感情分析を行い、感情の予測に寄与する素性を用いて販売予測を行ったことがこれまでの研究と異なっている。

3.2.3 属性選択

本章では、レビュー文に対する感情分析を行い、感情の予測において有用であった単語素性を用いて販売予測を行なった。本節では予測に用いる素性の選択と感情分析の関わりについて述べる。

文章分類における属性選択では、[Forman 03] らによるサーベイにおいて様々な属性選択の手法と文章分類精度の関係が調べられている。文章分類の中でも、特に感情分析に関わる素性の選択の研究には、[Abbasi 08] や [Duric 12] といった研究が存在する。[Abbasi 08] では、情報利得と遺伝的アルゴリズムを組み合わせた手法により感情分析に用いる素性の選択を行なっている。また、[Duric 12] では、レビュー文に含まれる主観的な表現に注目し、素性の選択を行なっている。これらの研究は、感情分析を行う際に有用な単語素性の選択を行う手法を提案している。

一方、[Caruana 98] では、特定のタスク（元タスク）で学習された特徴表現が別タスク（目標タスク）にて用いる枠組みを示されている。また [Silver 07] は、実際にニューラルネットワークの中間層における特徴表現が別タスクにおいて予測に有用であることを示した。本章の提案手法では、感情分析タスクにおいて有用だった素性を別タスクである販売予測タスクにおいて用いる点がこれまでの研究と異なっている。

3.3 提案手法

本節では、商品の販売数と相関関係にある口コミ情報に対して感情分析を行い、その結果を用いて販売数を予測する提案手法について説明する。先行研究から口コミ情報を用いることで販売予測の精度が向上する [Dellarocas 07] ことが知られている。したがって、口コミ情報には商品の販売を予測する上で有用な素性が含まれることが考えられる。提案手法では、感情分析において予測に対する寄与度の高い因子を抽出し、商品の販売数を予測する。

3.3.1 感情分析

まず、レビューデータに対する感情分析について述べる。感情分析とは特定の対象物に対する書き手の意見を分けるタスクであり [Feldman 13]、感情分析を行うことにより大規模なレビューデータから消費者が商品に対して抱いている感情を明らかにすることが可能である。対象とする商品を絞って感情分析を行う場合、対象とした商品に関連する特徴的な単語が感情分析において重要になると考えられる。そこで提案手法では、与えられたレビュー中に含まれる情報を教師ラベルとして分類器を学習するアプローチを採用する。提案手法では、感情分析において予測に対する寄与度の高い因子を販売予測において入力として与える。そこで、ラベル

の予測に対する各入力素性の寄与度が算出される分類器を用いた機械学習の手法を採用する。本章では単純ベイズ分類器を用いて感情分析を行う。

対象とするレビューに含まれる単語 $\{w_1, w_2, \dots, w_{|V|}\}$ の集合を V とする。全レビュー数を N 、 j 番目のレビューを d_j 、 $n_i(d_j)$ をレビュー d_j に単語 w_i が出現する回数とする。以上を用いて、レビュー d の bag-of-words 表現 $\mathbf{d} := (n_1(d), n_2(d), \dots, n_{|V|}(d))$ を得る。単純ベイズ分類器では、ある文章 d が与えられた場合のクラスラベル $c^* = \operatorname{argmax}_c P(c|d)$ を求める。ベイズの定理より、

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (3.1)$$

が得られる。各単語 w_i の出現は独立であるという仮定をおくことで、 d のクラス分類を以下のように表わせる。

$$P(c|d) = \frac{P(c) \left(\prod_{i=1}^{|V|} P(w_i|c)^{n_i(d)} \right)}{P(d)} \quad (3.2)$$

ユーザ採点をもとにレビュー d_j が肯定的か否定的かを表すラベル $y_j \in (0, 1)$ を得る。 $X = \{d_1, d_2, \dots, d_N\}$ および $y = \{y_1, y_2, \dots, y_N\}$ として、分類器に (X, y) を入力として与える。

学習された分類器は、入力として与えられた文章の bag-of-words 表現 \mathbf{d} に含まれる単語の出現回数 $(n_1(d), n_2(d), \dots, n_{|V|}(d))$ をもとに感情を表すクラスラベル y を予測する。分類器において学習されるクラスラベル c が与えられた際の各単語についての条件付確率 $P(w_i|c)$ から対数オッズ比を算出する。対数オッズ比の大きい単語素性は、ユーザの肯定的感情に付与されたラベルである正例の予測に対する寄与度が大きい。本章では、これら対数オッズ比の大きい単語素性を、消費者の感情を示す素性であると考え、販売予測に用いる。

3.3.2 有用な素性の抽出

本節では、感情分析タスクにおいて学習した分類器から予測に有用な素性を抽出する方法について説明する。

入力として与えられたレビューを肯定的と否定的の二値に分類する分類器 F を用いる。アルゴリズム 1 に示すように、学習に用いた全語彙の集合 V の中から、学習済みの分類器 F において予測に対する寄与度の大きい単語を K 個選択し集合 W_K を得る。続いて、ある月

Algorithm 1 Feature Extraction From Classifier**Input:** trained classifier F , vocabulary set $V = [w_0, \dots, w_{|V|}]$ **Output:** selected words W_s **for** $i = 0$ to $|V|$ **do** $p_i \leftarrow$ extract coefficient values for w_i from F **end for** $W_s \leftarrow$ top K words order by p value**return** W_s

$t \in T$ における対象とする商品のレビュー中に存在する $u \in W_K$ の出現回数を算出し、 $c_u^{(t)}$ とする。 W_K に含まれる単語 u それぞれについて $c_u^{(t)}$ を算出し、販売予測における入力とする。

感情分析において肯定的なレビューの予測に対する寄与度の高い因子は、ユーザの肯定的な感情を表す因子であると考えられる。したがってこれらの因子がより頻繁に現れる場合、ユーザはその商品に対してより強く肯定的な感情を抱いていると考えられる。ユーザによる評価が肯定的な商品はよく売れると考えられるので、抽出した因子の出現回数を販売予測における素性として用いることは効果的であると考えられる。そこで感情分析タスクにおいて学習した分類器 F の中で予測に対する寄与度の大きさをスコア付けし、スコア上位 K の単語を選択し、販売予測に用いる素性とする。

3.3.3 販売予測

続いて、対象とする商品の販売予測について述べる。商品の売り手にとって、商品のその後の売れ行きを予測することは、マーケティング戦略の決定をはじめとする意思決定における重要な判断材料となる。したがって、商品の販売予測は重要な問題である。一般的に、販売数の推移は連続的であり、将来の販売数を予測するタスクにおいては直近の販売数が重要な素性になる。そこで、予測対象とする商品についての直近の販売数 S を素性として加える。また商品によっては、時期に応じた販売数の変動が見られる場合もあるため、月情報を示す素性 M を追加する。ある月 t における直近の販売数 $S^{(t)}$ および月情報 $M^{(t)}$ 、感情分析における寄与度の大きい素性上位 K 件の情報 $C_K^{(t)}$ を入力として与え、 t から i ヶ月後の販売数 $S^{(t+i)}$ を予測する。

$$S^{(t+i)} = f(S^{(t)}, M^{(t)}, C_K^{(t)}) \quad (3.3)$$

$S^{(t)}$ は実際の販売数を用い、月情報 M は 12 次元の 1-of-k ベクトルにて表す。 $C_K^{(t)}$ は、ある月 t に投稿されたレビューに含まれる感情分析における寄与度の大きい単語上位 K 件、それぞれについての出現回数 $c_n^{(t)}$ を要素とするベクトル $C_K^{(t)} = [c_1^{(t)}, c_2^{(t)}, \dots, c_K^{(t)}]$ とする。ある月 t における $S^{(t)}$, $M^{(t)}$, $C_K^{(t)}$ を結合し入力素性とし、 i ヶ月後の販売台数 $S^{(t+i)}$ を正解ラベルとして学習モデルに与える。

3.4 実験・結果

本節では、提案手法の有効性を示すための実験を行う。実験では、自動車のレビューに対する感情分析から得られた素性を販売予測に適用し、素性を用いない場合と比較することで提案手法の有効性を示す。

3.4.1 データ

自動車販売数データ

自動車についての月次の販売台数データは、日本自動車販売協会連合会（自販連）^{*1}より、毎月の販売台数上位 30 位以内の自動車について取得した。また自販連サイトにて入手できない車種については、car-lineup.com に記載されている販売台数を用いた。なお、両サイトにおいて情報が入手可能な例については、両サイトに記載されている販売台数が一致することを確認した。販売台数のデータは、2011 年 11 月から 2015 年 8 月までの期間のデータを収集した。なお、各月の販売台数は、翌月の中旬に公開されるため、例えば 2011 年 11 月のデータは 2011 年 10 月 1 日から 31 日までの期間に販売された当該車種の合計台数となる。

オンラインレビュー情報

自動車のレビュー情報は、国内最大規模の自動車レビューサイトである GooNet^{*2}から取得した。レビューデータは、サイトにて閲覧可能な 2011 年 11 月からデータ収集を行った 2015 年 7 月末日までの期間にユーザにより投稿された 57,466 件を対象とした。

図 3.2 に投稿されるレビューの具体例を示す。ユーザの投稿には、レビュー対象の車種名、

^{*1} <http://www.jada.or.jp/>

^{*2} <http://www.goo-net.com/index.html>



図 3.2 自動車に関するオンラインレビューの投稿 (例)

ユーザ名、ユーザにより付与されたタグ、投稿日時、ユーザによる採点結果、ユーザによる車に対する感想が含まれる。本章では、ユーザによる採点結果とユーザによる車に対する感想のデータを用いた。

3.4.2 データに対する前処理

解析対象とする車種の選択

本章にて販売予測の対象とした車種は、28 車種である。選択する基準として以下の 4 つを考慮した。

- 分析の対象とした期間の始めである 2011 年 11 月時点ですでに販売されており、かつ分析期間の終わりである 2015 年 8 月時点でも新車として継続的に販売されている車種
- 分析期間における月次の販売台数の平均が 1,000 台以上の車種
- 予測モデルの精度検証を行う期間 (2015 年 1 月から 2015 年 8 月) にモデルチェンジを

行っていない車種

- 分析期間内におけるレビューの総投稿数が 300 件以上ある車種

これは分析対象期間にモデルチェンジが行われた場合、販売数の傾向が大きく変化し、またモデルチェンジは販売メーカー側の都合に大きく左右される現象であるため、今回の研究で検証する対象の範囲外であると考えたためである。また、レビューの投稿数が少ない場合、今回の問題設定において十分に検証が行えないと判断したためである。これら 4 つの条件に照らし合わせた結果、解析の対象となる車種は、ALPHARD, COROLLA, CROWN, CUBE, DEMIO, eK WAGON, ELGRAND, ESTIMA, FIT, FREED, LEGACY, MARCH, MIRA, MOVE, NOAH, NOTE, ODYSSEY, PASSO, PRIUS, SERENA, SIENTA, SWIFT, TANTO, VITS, VOXY, WAGON R, WISH, X-TRAIL であった。

レビューデータに対する前処理

レビューデータに対して行った前処理の概要を説明する。収集したレビュー文に対して MeCab により形態素解析を行い、名詞および形容詞のみを抽出し bag-of-words 形式で表現した。続いて、[Yang 97] にならい、あらかじめ設定した文章頻度に基づき出現回数が閾値以下の語を除いた。閾値は、[Huang 08] を参考に使用される語彙数が同程度となるように決定した。閾値を 30 とし文章頻度の低い語を除いた結果、全語彙集合 V に含まれる語数は 2,265 であった。

3.4.3 感情分析

感情分析の精度

レビューデータに対する感情分析について説明する。収集したレビューデータは前処理を行った後、本章で販売予測の対象期間とした 2014 年 12 月より前までに投稿されたレビューを車種による区別なく感情分析器の学習に用いた。

前処理を行ったレビューデータをユーザによる採点が高い順に並べ、採点の上位 3 割および下位 3 割のレビューを選択し、感情分析用のデータセットとした。ユーザ採点の結果が上位のデータには正のラベルを付与し、下位のデータには負のラベルを付与した。その後、データセットを訓練データとテストデータに 7:3 の割合で分割した。感情分析に用いる分類器として

表 3.1 感情分析に寄与する素性

| rank | word | score | rank | word | score |
|------|-------|-------|------|--------|-------|
| 1 | プレミアム | 11.04 | 10 | 考え | 5.91 |
| 2 | 従来 | 9.07 | 10 | 基準 | 5.91 |
| 3 | サイト | 8.28 | 10 | おとなしく | 5.91 |
| 4 | cx | 7.88 | 10 | 音響 | 5.91 |
| 5 | 辺り | 7.10 | 10 | 久しぶり | 5.91 |
| 6 | 記録 | 6.70 | 10 | 中止 | 5.91 |
| 6 | 無限 | 6.70 | 17 | タコメーター | 5.72 |
| 6 | デジタル | 6.70 | 18 | 周囲 | 5.52 |
| 9 | アクア | 6.11 | 18 | 携帯 | 5.52 |
| 10 | no | 5.91 | 18 | 機敏 | 5.52 |

は、先行研究にて用いられてきた分類器のうち、入力として与えられた素性の予測に対する寄与度を算出できる単純ベイズ分類器を用いた。訓練データを用いて単純ベイズ分類器を学習した後、テストデータに適用し分類精度を検証した。分類の精度の評価は F 値にて行い、分類精度は 0.89 であった。得られた結果は、その後の販売予測タスクを行う上で十分な精度であると考え、感情分析において有用であった素性を販売予測に用いた。なお、単純ベイズ分類器の実装は scikit-learn [Pedregosa 11] を用いて行い、パラメータは [Pedregosa 11] の値をそのまま使用した。

予測に寄与する単語素性の抽出

感情分析に用いた分類器から正例の予測に寄与度の大きい素性を抽出した。予測に対する寄与度は、正例および負例が与えられた際の素性ごとの条件付き確率から対数オッズ比を算出し用いた。正例の予測に対する寄与度の高い単語を表 3.1 に示す。寄与度の大きな単語は「プレミアム」、「従来」であった。それぞれが使われていた代表的な事例を列挙する。「プレミアム」については、「プレミアム感のある外装や装備は満足度が高いです。」といった肯定的な感情を直接的に表現している場合が多かった。「従来」との場合は、「従来よりも燃費と経済性が優れていると感じます。」というようにこれまでのモデルや車種と比較して優れているという文脈で用いられ、レビュー文全体として肯定的になっていた。感情分析と予測に寄与していた因子

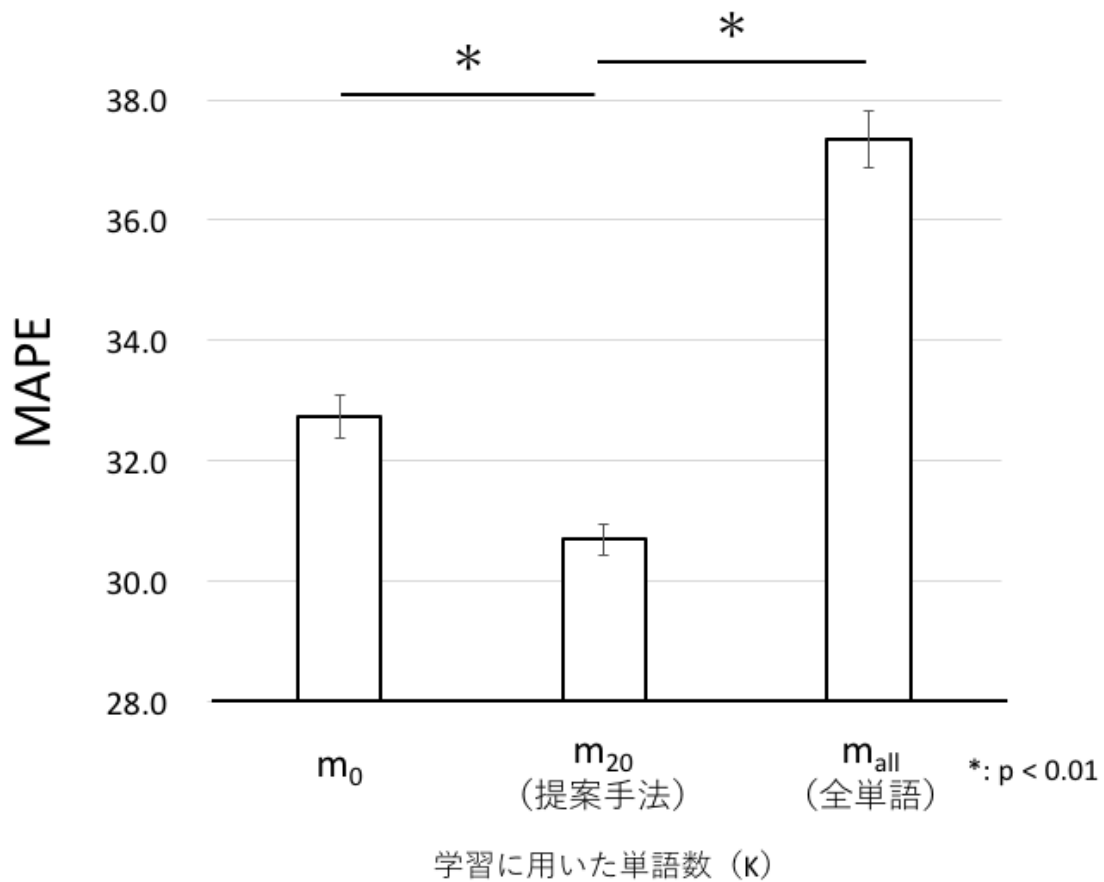


図 3.3 単語素性を用いない場合，全単語素性を用いる場合と提案手法の比較

についての分析から，消費者の感情を強く表すと推測される素性が，感情分析において正例の予測に強く寄与しており，それらの素性を特定・抽出することができることが示唆された。

3.4.4 販売予測

次に，抽出された素性を用いて販売数の予測を行い，提案手法の有用性を検証した。予測モデルの構築は対象とした 28 車種それぞれについて行い，車種ごとに予測期間における実際の販売台数と予測される販売台数の誤差を平均絶対パーセント誤差 (Mean Absolute Percent Error; 以下 MAPE) にて評価した。

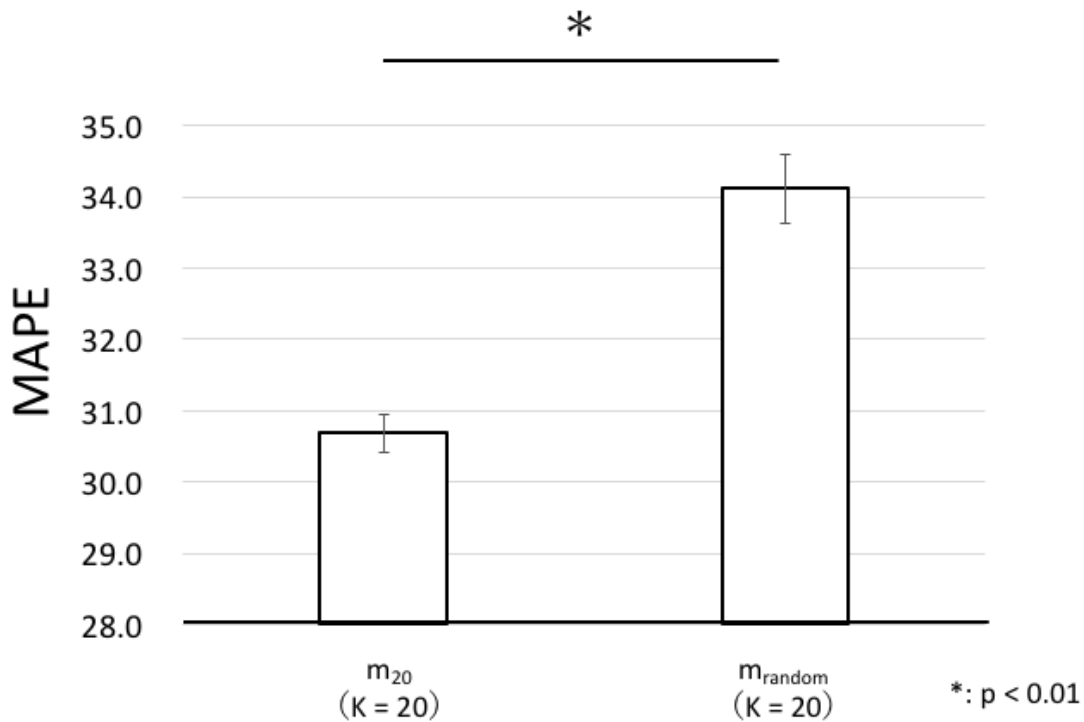


図 3.4 ランダムに選択された単語素性を用いる手法と提案手法の比較

$$MAPE = \frac{1}{T} \sum_{t=1}^T \frac{|Pred_t - True_t|}{True_t} \quad (3.4)$$

ここで、 T は学習期間に含まれる月の数であり、 $Pred_t$ は学習モデルから出力される予測値、 $True_t$ は実測値である。MAPE は時系列データに対して用いる精度の指標の一つである。予測値と実測値の差の絶対値を実測値で除し、予測期間での合計値を算出した後、予測期間の長さで除すことで算出される。これによりパーセント誤差が算出され、車種ごとの誤差を比較できるようになる。

予測期間は、データが入手可能であった期間のうち最後の 8 ヶ月間である 2015 年 1 月から 2015 年 8 月までとした。予測は直近 24 ヶ月間のデータを用いて学習したモデルにより、1 ヶ月後の販売台数を予測することを繰り返すスライディングウィンドウ方式にて行った。つま

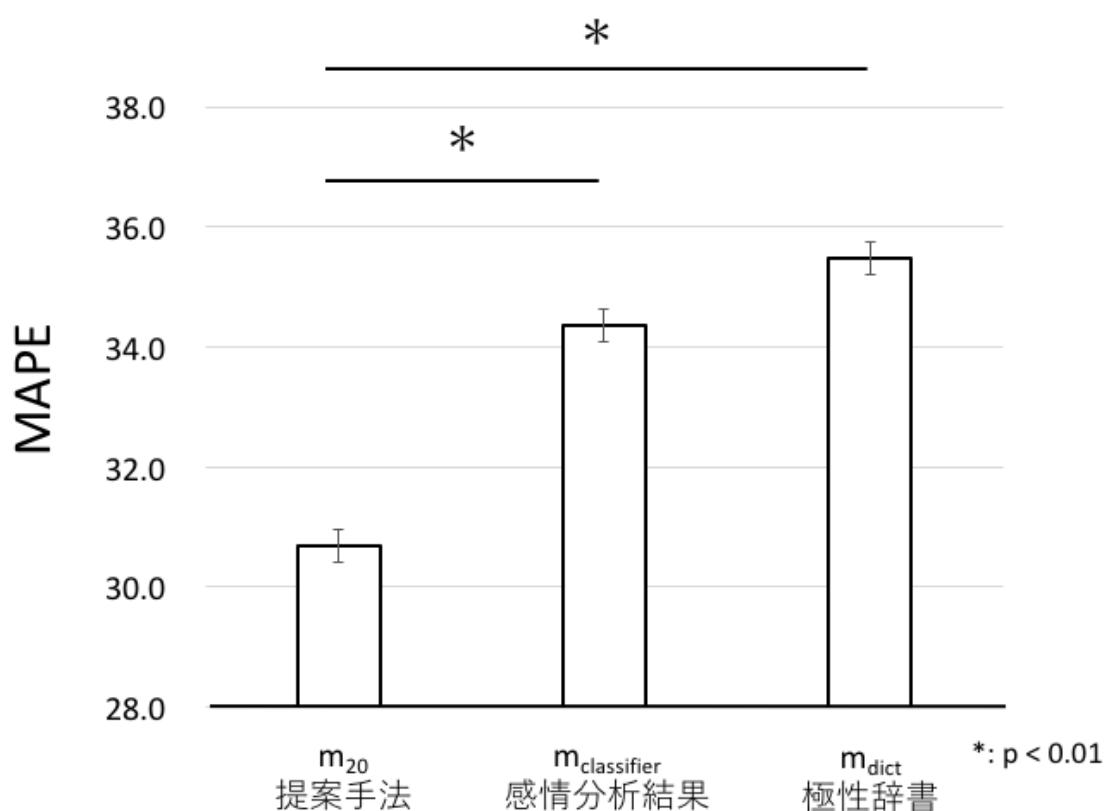


図 3.5 極性辞書を用いる手法，感情分析の結果を直接用いる手法と提案手法の比較

り，まず 2012 年 12 月から 2014 年 12 月までのデータを用いて学習を行い，2015 年 1 月の販売数を予測する．続いて，2013 年 1 月から 2015 年 1 月までのデータを用いて，2015 年 2 月の販売数を予測する．このようにして 2015 年 1 月から 2015 年 8 月までの予測を行った．

予測に用いる学習器としては，ランダムフォレスト回帰 [Breiman 01] を用いた．これは，学習に用いた各素性の重要度が算出できるためである．学習器の実装には，scikit-learn [Pedregosa 11] を用い，用いる木の数を 50 とした以外，全てのパラメータで [Pedregosa 11] の RandomForestRegressor におけるデフォルトの値をそのまま用いた．

まず，提案手法において用いる単語数 K の決定を行なった．ランダムフォレスト回帰に対して，当月の販売台数，1-of-k ベクトルにて表される 12 次元の月情報および感情分析における

表 3.2 開発セットにおける精度

| word num (K) | 10 | 20 | 30 | 40 | 50 |
|--------------|-------|--------------|-------|-------|-------|
| Average MAPE | 35.37 | 34.57 | 34.70 | 34.79 | 34.93 |

予測への寄与度の大きい上位 K 単語の出現回数を与え、翌月の販売数を予測した。学習に用いる 36 ヶ月分のデータにおける最初の 24 ヶ月分を学習セット、残りの 12 ヶ月分を開発セットとし、用いる単語数 K の決定を行なった。 K の値としては、10, 20, 30, 40, 50 を試した。独立した試行を 3 回行い、開発セットにおける平均の予測精度を算出した。算出された平均の精度を表 3.2 に示す。開発セットにおける精度は $K = 20$ とした場合に最も高かったため、以後の実験では $K = 20$ の場合の結果を示す。

続いて、提案手法の有効性を示すため、ベースラインとの比較実験を行なった。当月の販売台数および 1-of-k ベクトルにて表される 12 次元の月情報を与えて学習するモデルを m_0 とし、 m_0 および m_0 に素性を加えた条件をベースラインとした。具体的には、単語素性を用いない条件 m_0 および m_0 に対して感情分析に用いた全語彙を加える条件 m_{all} 、極性辞書に含まれる語の出現回数を加える条件 m_{dict} 、感情分析器による予測結果を用いる条件 $m_{classifier}$ 、ランダムに選択された単語を用いる条件 m_{random} を置いた。これらのベースラインと m_0 に感情分析の結果に対する寄与度の大きい上位 20 単語の出現回数を加えた m_{20} による予測精度を比較した。実験は、独立した 3 回の試行により得られた結果を平均し最終的な結果とした。

まず、用いる単語数による精度の変化を調べるため、単語素性を用いない条件 m_0 、全単語素性を用いる条件 m_{all} と m_{20} の精度を比較した。結果を図 3.3 に示す。全単語を用いる条件 m_{all} における MAPE 値および単語をまったく用いない条件 m_0 における MAPE 値は、それぞれ 32.73 および 37.34 であった。一方、提案手法 m_{20} における MAPE 値は 30.69 であった。また、 t 検定による検定の結果、予測の精度は有意水準 1% で有意に向上していた。

続いて、感情分析の結果得られた単語素性を用いることの重要性を示すため、ランダムに選択された語のレビューにおける出現回数を用いる m_{random} と m_{20} の比較を行なった。結果を図 3.4 に示す。全語彙集合に含まれる語からランダムに提案手法と同数の単語 ($K = 20$) を選択し、選択された語の出現回数を素性として予測を行なった。その結果、 m_{random} における MAPE 値は 34.11 となり、 m_0 での予測精度は m_{random} と比較して有意水準 1% で有意に向

上していた。

最後に、感情分析器から得た単語素性を用いることの有用性を示すため、極性辞書に含まれる語の出現回数 m_{dict} および感情分析の結果をそのまま用いる $m_{classifier}$ と m_{20} を比較した。 m_{dict} では、レビューにおける感情分析に用いられる極性辞書に含まれる単語の出現を素性とした。極性辞書は [高村 06] のものを用いた。極性辞書に登録された単語のうち感情極性実数値が正の単語を選択し、その出現回数を予測における素性とした。 $m_{classifier}$ では本章で用いた感情分析器に対して、レビュー文章を与え、その分類結果における肯定的なレビューの結果と否定的なレビューの比を素性として用いた。得られた結果を図 3.5 に示す。 m_{dict} における MAPE 値は 35.48 であり、 $m_{classifier}$ による予測についての MAPE 値は 34.35 であった。 t 検定による検定の結果、提案手法 m_{20} は両ベースラインと比較して有意水準 1% で有意に予測精度が向上していた。

3.5 考察

3.5.1 感情分析

まず、オンラインレビューに対する感情分析にて得られた結果について考察する。感情分析における評価の高いレビューと低いレビューの分類精度は、F 値で 0.89 であった。分類の対象としたレビューは、得点が上位 3 割および下位 3 割のレビューとした。これはユーザごとに採点基準が異なるため全てのレビューを用いた場合に、レビュー文に記されたユーザの評価が大きく異なる一方で点数が同じであるという状態が生じ、感情分析の評価が低下すると考えたためである。特に車に対する評価が高くない場合の点数（辛口採点の割合）は、ユーザ個人の感覚に依存するため精度が上がらない原因となる。具体的には、あるユーザは「満足度は中程度」という文章と合わせて「4 点」と投稿するが、他のユーザは「3 点」と投稿するといった状況が考えられる。高得点のレビューについては全体的に評価の高さを連想させる語が並ぶが、ユーザ採点が全体の中央値付近のレビューでは文章が高得点のレビューと類似するものと低い得点のレビューに類似するものの両方が混在する。

本章における感情分析では、bag-of-words 形式で表現したレビュー文章を入力として学習器に与えた。そのため、レビュー本文中で各単語が現れる文脈を考慮できていない。例えば、「燃費が悪い」と「悪い部分はない」という 2 つのレビューが与えられたとき、前者は否定的な感情

を表し、後者は肯定的な感情を表すと考えられる。しかし、形態素解析の結果を bag-of-words 形式で表現した場合、前者は「燃費:1, 悪い:1」となり、後者は「悪い:1, 部分:1, ない:1」となり、「悪い」という単語がどちらの例にも含まれることとなる。そのため、bag-of-words 形式では一部の単語の予測に対する寄与度を過小評価してしまう可能性がある。また、表 3.1 において上位にある「従来」といった語は、「従来より〇〇」といった形で用いられることがほとんどであった。こちらについても、語順を考慮しない bag-of-words 形式で表現した場合、失われる情報があると考えられる。文脈を考慮するモデルや n-gram を用いることにより、より感情分析の精度を高めることができる可能性がある。ただし n-gram を用いる場合、語数が膨大になるためその点を考慮する必要がある。

感情分析の結果、正例の予測に寄与する単語は表 3.1 のような単語であった。正例の予測には、「プレミアム」や「無限」といった評価の高さを連想させるような単語の寄与度が高かった。また、「アクア」や「cx」といった個別の車種を指す単語も含まれていた。これらは、当該車種についてのレビューに含まれるだけでなく、他の車種のレビューにも出現していた。例えば、「アクア」という単語は、FIT や PRIUS といった類似の車種のレビューに多く出現しており、しばしばレビュー中でこれらの車種と比較されていた。また「no」という単語は、「no1」といった用いられ方が多かったため、評価の高いレビューの予測に対する寄与度が大きくなったと考えられる。

3.5.2 販売予測

続いて、販売予測を行い提案手法とベースラインを比較した結果および提案手法において用いられる単語数と精度の関係について考察する。

開発セットによるパラメータサーチの結果、本実験における最適な単語数 K の値は 20 となった。 K と精度の関係を調べるため、その他の単語数 ($K = 10, 30, 40, 50, 100, 200, 500, 1000, 2000$) についてもテストセットでの精度を検証したところ、図 3.6 に示す結果を得た。得られた図から、今回実験で用いたデータにおいて単語数 K は 20 が最適であること、および $K = 20$ 以降では K が増加するとともに MAPE 値は悪化していることが明らかとなった。この結果から、適切な単語数を選択することで精度の向上が期待できる一方、用いる単語が過剰になると予測精度が低下することが明らかとなった。

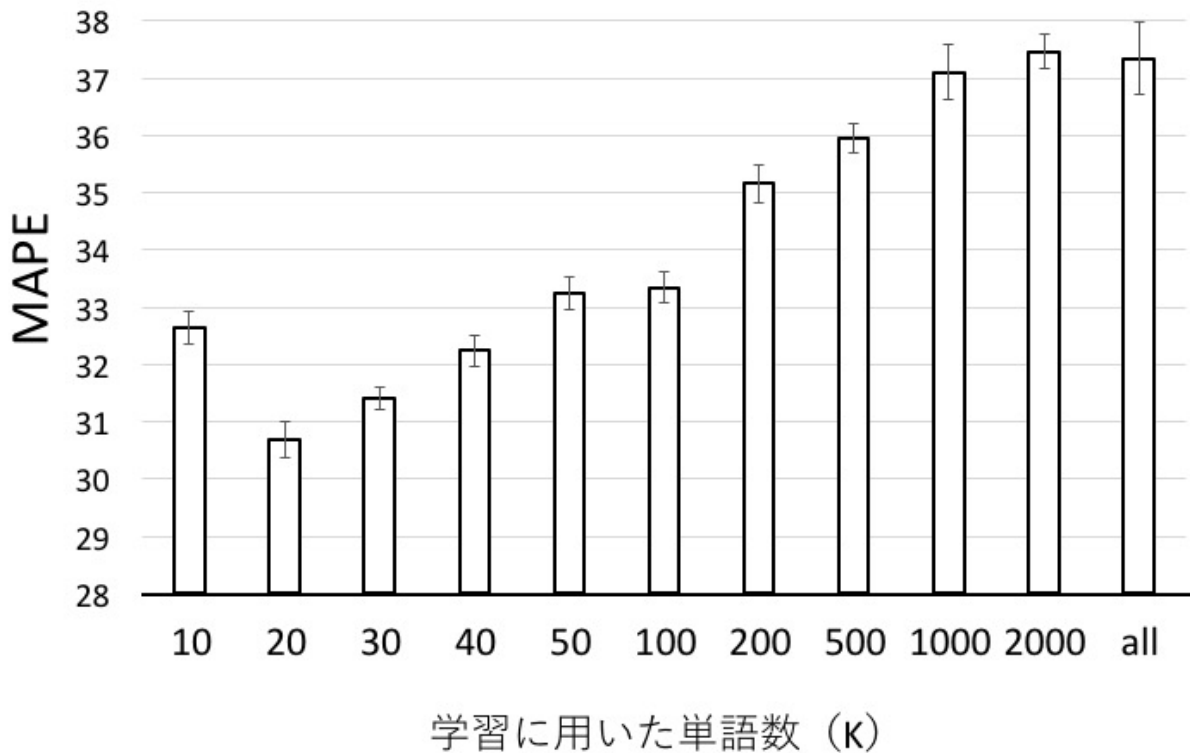


図 3.6 単語数に対する平均 MAPE 値

続いて，ベースラインとの比較を行なった．その結果から，適切に設定された単語数のもとで提案手法による予測精度が有意に向上することが明らかとなった．ベースラインとしては，単語素性を用いない条件 m_0 および m_0 に対して感情分析に用いた全語彙を加える条件 m_{all} ，極性辞書に含まれる語の出現回数を加える条件 m_{dict} ，感情分析器による予測結果を用いる条件 $m_{classifier}$ ，ランダムに選択された単語を用いる条件 m_{random} を置いた．それぞれの提案手法と各ベースラインとの比較結果について考察する．

まず，単語素性を用いない条件 m_0 および全語彙を用いる条件 m_{all} との比較を行なった．この結果から，適切な数の単語を選択することで精度が向上することが示された．適切な単語数を用いた場合，テストセットにおける MAPE 値は m_0 と比較して 6.6%， m_{all} と比較して 21.68% 改善した．また， t 検定を行なった結果， p 値は m_0 に対して 1.4×10^{-3} ， m_{all} に対

して 3.1×10^{-5} であり、統計的に有意に精度が向上していることが示された。本実験の結果から、適切な単語数に対して用いる単語数が少ない場合、対象とするレビューに出現する予測に有用な単語素性の出現回数が少なくなり、予測精度が十分に向上しないことが考えられる。逆に適切な単語数よりも多い単語数の場合、データラベルに対して入力データの次元数が多くなるため、モデルが学習データに対して過学習しやすくなり、結果として予測精度が低下していくことが考えられる。実際に、適切な単語数を超過して用いる単語数が増えるほど、予測精度が低下していることは図 3.6 に示す結果とも一致している。

また、ランダムに選択された単語素性を用いる条件 m_{random} との比較結果からは、感情分析を用いて単語を選択する重要性が示唆された。提案手法と m_{random} において用いる単語素性の数 K を同一の値 ($K = 20$) に設定し、それぞれの手法における予測結果の統計的有意差を検証した結果、有意水準 1% 以下で提案手法による予測の精度が向上していた。この結果から、ランダムな単語素性を追加するだけでは予測精度を向上させることが難しく、感情分析により選択した結果を用いることが重要であることが示唆された。

最後に、極性辞書を用いる条件 m_{dict} および感情分析の結果を用いる $m_{classifier}$ と提案手法の比較結果から、レビューにおける感情分析タスクから得られる単語素性を用いることの重要性が示唆された。極性辞書を用いた条件 m_{dict} との比較結果から、予測に用いる単語素性は極性辞書に記載された一般的な語ではなく、用いるデータから得られる語の方が効果的であることが考えられる。また、 $m_{classifier}$ との比較結果から、感情分析の結果を素性として予測に用いる手法と比較して、感情分析タスクにおいて予測に寄与する単語素性を抽出し、販売予測に用いる提案手法の精度が有意に向上することが明らかとなった。

以上の結果から、感情分析タスクにおいて予測に寄与する単語素性を適切に選択された単語数 K のもと用いることで、有意に予測精度を向上できることが示された。

3.6 本章の結論

本章では、オンラインレビューにおける消費者の感情を分類するタスクにて予測への寄与度の大きい素性を用いることで、販売予測タスクの精度を向上する手法を提案した。提案手法の有用性についての検証は、国内における自動車の販売予測にて行なった。まず、レビューデータに対する感情分析タスクを行い、ユーザの感情の予測に対する寄与度の大きい単語素性を特

定した。続いて、感情分析タスクにおいて特定された単語素性を用いて自動車の販売予測を行った。販売予測における有用性の検証では、レビューに対する感情分析により選択した単語素性が予測において有用であることを示すためのベースラインとして、単語素性を用いない条件、レビューに含まれる全単語の出現回数を用いる条件、ランダムに選択された単語素性を用いる条件、極性辞書に含まれる単語の出現回数を用いる条件および感情分析タスクにて学習された分類器による分類結果を用いる条件を置いた。その結果、適切に設定された単語数のもとでは、提案手法がいずれのベースラインに対しても有意水準 1% で有意に精度向上していることが明らかとなった。また単語数と予測精度の関係から、単語数がある一定以上の数になると精度が低下することが示唆された。

本章では、これまでの研究で知られていた、口コミ情報が商品の売上に重要であること [Dellarocas 07] に着想を得て販売予測モデルを提案した。感情分析によりレビューから抽出した単語素性は、商品に対する消費者の感情を内包していると考えられ、それらを用いることで予測の精度が向上することは先行研究からの知見に合致する。本章の提案手法は、自動車におけるモデルチェンジといったイベントを考慮して予測することが難しいといった側面があるものの、商品に対するレビューが十分に存在しユーザの感情が抽出できるという条件を満たせば、自動車のみならず幅広く用いることが可能であると考えられる。

第4章

ユーザの嗜好を反映したベクトル表現の獲得と人気予測

4.1 本章の背景

本章では、低価格であることが多く、商品の販売予測を行う上で対象とした商品に対する注目の度合いが重要になると考えられるマルチメディアタイトルについての販売予測の問題に取り組んだ。マルチメディアタイトルは、音声、動画、静止画、文書などを組み合わせたマルチメディアコンテンツからなる著作物であり、テレビ、雑誌、ゲーム、インターネット、カードゲームといった幅広いメディア・形態を通じて展開される。このようなマルチメディアタイトルの市場規模は拡大を続けており、代表的なマルチメディアコンテンツであるアニメ産業の市場規模は、2015年に前年比12%増を記録している [増田 16]。また海外におけるマルチメディアタイトルの市場規模も拡大が続くことが見込まれており、クールジャパンプロジェクトにおいても注目されている [経済 16]。加えて、マルチメディアタイトルはそれ自身がコミックやアニメの放映により収益をあげるだけでなく、その知名度や人気を活用する形で広告やイベントグッズなどにおけるタイアップが可能であり、他の産業に対する波及効果も期待できる。

ソーシャルメディアを用いたマルチメディアタイトルの販売予測では、オンラインレビューのデータを用いる手法、Wikipediaのデータを用いる手法、ブログ記事を用いる手法などが考えられる。この中から本章では、マルチメディアタイトルの網羅性という点から Wikipediaのデータを用いて販売予測を行なった。

先行研究において消費予測を行う上で、予測対象に対する消費者の口コミ情報を用いることで精度が向上することが知られている [Dellarocas 07]。しかしながら、Wikipedia 上に記述される内容は、百科事典としての性質から客観的な事実が大半である。そこで本研究では、商品の販売予測を行う上で消費者の口コミ情報と同程度に、消費者の嗜好をもとにした商品カテゴリーをはじめとする複数の尺度についての類似度を考慮した多面的な情報を考慮することが重要であると考え、それらを用いて予測を行なった。商品の推薦において代表的な手法である協調フィルタリングは、商品とユーザのベクトル表現をもとに推薦を行うなど、ユーザの嗜好を反映したモデルである [Resnick 94], [Shardanand 95]。このように推薦問題においては、消費者の嗜好を考慮するモデルが大きな成果をあげている。また人気予測タスクでは、口コミ情報から得た特徴量を用いて映画の人気予測を行う研究 [Mishne 06] や予測対象から抽出した特徴量を用いて人気予測を行う研究 [Khosla 14] が存在したが、対象についてユーザの嗜好をもとにした複数の尺度についての類似度が考慮された事例は少なかった。

そこで本研究では、Wikipedia よりユーザの嗜好を抽出し、それに基づくコンテンツ作品のジャンルや流行した年代といった多面的な情報を得ることを試みる。Wikipedia は、幅広い内容を網羅しており、多くのユーザにより利用、編集されるソーシャルメディアの一つである。Wikipedia において、ユーザが自身の興味関心の高い項目に関するページを編集するため、各ユーザが編集する項目はそのユーザの嗜好を反映していると考えられる。したがって、編集履歴から得られる様々なユーザの嗜好を考慮することで、対象とするコンテンツ作品のもつジャンルや流行した年代といった複数の尺度について類似度を考慮した多面的な情報を得られると考えられる。

本研究では、図 4.1 に示すように、ウェブ上におけるコンテンツ作品に関するユーザの行動履歴の系列データから、消費者の嗜好に基づくコンテンツ作品の多面的な情報を学習し、それらを用いてコンテンツ作品の人気予測を行った。より具体的には、コンテンツ作品についての Wikipedia の編集履歴に対し、自然言語処理の分野で大きな注目を集める Word2vec[Mikolov 13b] を応用し、コンテンツ作品のベクトル表現を得たのち、Wikipedia の被リンク数を指標として人気予測を行った。被リンク数は、ブログ [Wu 06] やウェブページ [Page 99] の人気や重要度の推定に用いられているため、本研究では各コンテンツ作品の人気指標として Wikipedia におけるリンク構造から得た被リンク数を用いた。また、編集履歴よ

り学習されたベクトル表現についての定性的分析も合わせて行なった。

本研究の貢献は以下である。

- ウェブ上でのユーザの行動履歴からコンテンツ作品のベクトル表現を学習できることを示した。
- コンテンツ作品の人気予測において、編集履歴から学習したベクトル表現を用いることで予測精度が向上することを示した。
- コンテンツ作品についてクラスタリングを行うことにより、予測精度が向上する場合と低下する場合の違いを分析した。

本章は以下のように構成される。4.2 節にて関連研究を示し、4.3 節にて提案手法について説明する。4.5 節では提案手法の前提条件を検証する予備実験、4.6 節では提案手法の有効性を検証する実験について述べ、4.7 節にて実験の考察と手法の発展性について議論する。最後に結論を 4.8 節に記す。

4.2 関連研究

本節では、本章での提案手法に関連する研究について述べる。

4.2.1 Wikipedia における編集行動

オンライン百科事典 Wikipedia は、幅広い内容を網羅しており、多くのユーザにより利用および編集されるソーシャルメディアであり、多くの研究がなされている。具体的には、大量の知識データおよびリンク構造によるそれらの関係性を知識ベースとして用いる研究 [Milne 08], [Strube 06] や Wikipedia に関わる社会的側面を分析した研究 [Welser 11], [Butler 08] などが存在する。

Wikipedia に関する研究の中でも、特に編集や編集行動に焦点を当てた研究も存在している。Wikipedia には、登録ユーザと未登録の一般ユーザが存在しており、[Nov 07] では、Wikipedia の登録ユーザに対し、アンケート調査を行い、彼らのモチベーションを調べている。また、[Welser 11] では、登録者の編集行動における社会的な役割を分析している。人気予測の文脈では、ページの編集回数を予測における素性として用いた、[保住 14] が存在する。

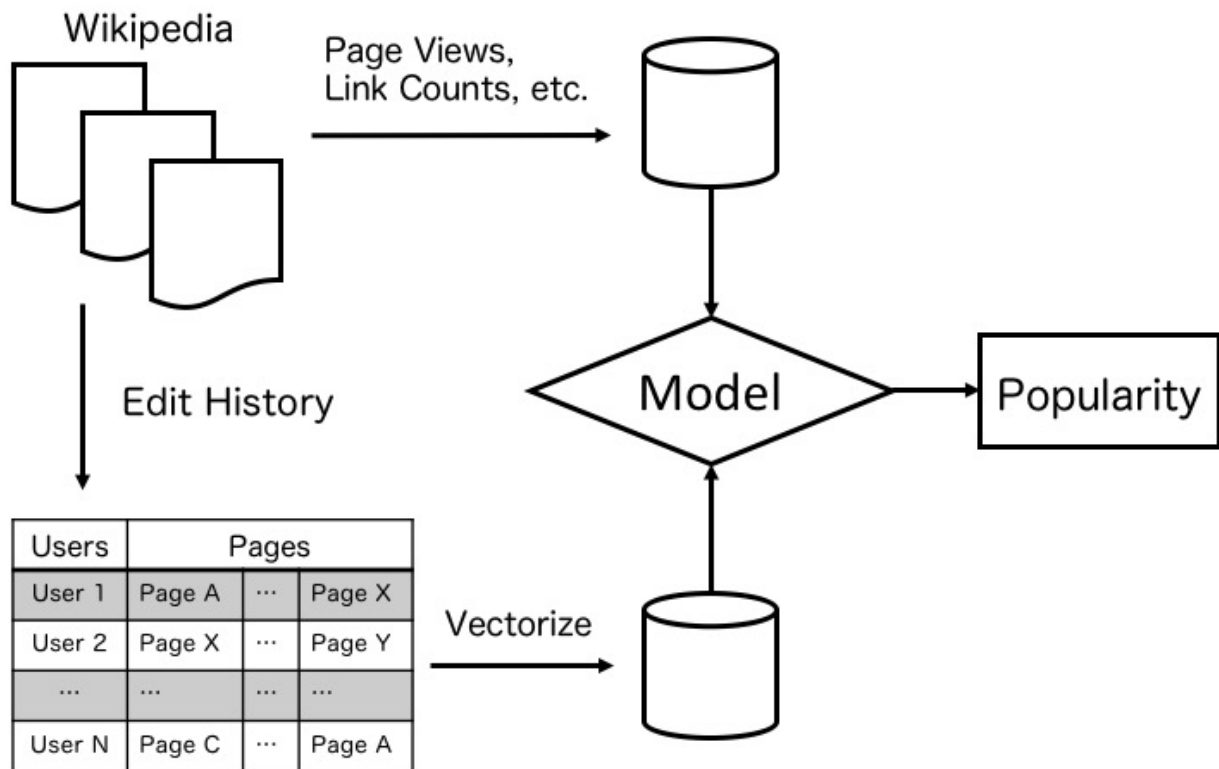


図 4.1 Wikipedia の編集履歴から学習したベクトル表現による人気予測の概略図

4.2.2 多面的情報をもつベクトル表現の獲得

系列データを入力として，その各要素について単語における意味表現のような多面的な情報を取得する手法が注目されている．代表的な事例としては，単語の分散表現を獲得する Word2vec[Mikolov 13b] およびその研究から派生し，文章の表現を獲得する [Le 14] が存在し，近年自然言語処理の分野で大きな成果をあげている．Word2vec により学習されたベクトル表現は，単語の意味表現を保持したものとなったことが示されており，学習されたベクトル表現は様々なタスクに利用されている．

自然言語処理以外の分野においても，ベクトル表現を学習する手法が提案されている．一例として，グラフ構造を入力として与え，グラフ内の各ノードのベクトル表現を獲得する

[Grover 16] らの Node2vec がある。また [Barkan 16] では、ユーザの商品閲覧履歴から商品に関するベクトル表現を学習し、推薦問題に用いている。

4.2.3 リンク構造による指標

ウェブページのリンク構造から各ページの重要度、質、人気を推定する方法が提案されている。代表的な手法は、被リンク数の多いページまたは重要なページからリンクが貼られているページを重要なページであると考え、上位に位置づけるページランク [Page 99] である。ページランクは、ウェブページの人気・注目度の指標として、Google の検索エンジンに導入され、大きな成果を上げた。

リンク構造を用いて人気・重要度の指標とする事例はウェブページ以外でも存在する。ブログにおけるリンク構造は、冪乗則に従う [Leskovec 07] ことが知られており、[Wu 06] では、リンク構造を用いてブログのランク付けを行なっている。また、[Kliegr 08] では Wikipedia における各ページの被リンク数を Wikipedia における人気の指標として扱っている。

4.3 提案手法

本節では、Wikipedia 上でのユーザの編集履歴をもとにコンテンツ作品のベクトルを取得し、コンテンツ作品の人気予測に用いる提案手法について説明する。

4.3.1 コンテンツベクトルの取得

Wikipedia の編集履歴からコンテンツベクトルを学習する手法について述べる。ユーザは自身の興味関心に基づいてページを編集すると考えられる。また各ユーザの編集履歴において隣接する Wikipedia の記事は、ユーザの興味が似ているコンテンツ作品に対応する Wikipedia の記事であると考えられる。

そこでユーザごとに編集履歴を時系列に並べ、あるコンテンツ作品の前後に出現するコンテンツ作品が与えられた場合に元のコンテンツ作品を予測するように学習を行う。より厳密には、あるコンテンツ作品の系列 $c_1, c_2, c_3, \dots, c_T$ が与えられた際に、 c_t をその前後に存在するコンテンツ作品 c_{t-k}, \dots, c_{t+k} により予測できるように各コンテンツ作品 c のベクトル表現を学習する。ここで、各コンテンツ作品 c を単一のベクトルにマッピングする行列を

C とする。マッピングを行う行列 C の学習は、Continuous Bag-of-Words(CBOW) および Skip-gram[Mikolov 13a] により行う。

4.3.2 人気予測

人気予測は、Wikipedia における各コンテンツ作品に対応する Wikipedia の記事における被リンク数を人気の指標として、Wikipedia の各記事から得られる素性およびコンテンツ作品のベクトル表現をモデルに与えることで行う。Wikipedia の各記事から得られる素性として、ページ閲覧数、編集回数、被リンク数を用いる。ページ閲覧数および編集回数は日次のデータとして取得し、その平均値を月次データとする。一方、被リンク数は日次の変動が小さいため当該月の初日の値を記録し用いる。学習時には、コンテンツ作品のベクトル表現と月次の素性からの特徴量を合わせて人気の予測を行う。

月次の素性およびコンテンツベクトルからの特徴量の抽出には、多層ニューラルネットワーク (MLP) を用いる。あるコンテンツ作品 c に対応する Wikipedia 記事についてのある月 t における月次の編集回数を e_c^t 、ページ閲覧数を v_c^t 、被リンク数を l_c^t とする。月次の素性 $\mathbf{X}_{c,M}^t$ を

$$\mathbf{X}_{c,M}^t = [e_c^t, v_c^t, l_c^t] \quad (4.1)$$

とし、MLP に与える。

また、コンテンツ作品に対応する Wikipedia の記事についての編集履歴を用いて学習された C から得られる、 c についてのベクトル表現を T_c とする。月次の素性を与える MLP とは別に T_c を入力として与える MLP を作成する。それぞれの MLP から得られた出力を、新たな MLP への入力とし、最終的な出力を得る。

モデルに対して、 $\mathbf{X}_{c,M}^t$ および T_c 、正解ラベル y を与え、学習を行う。

4.4 データ

本節では、実験に用いたデータの取得および取得したデータの前処理について述べる。

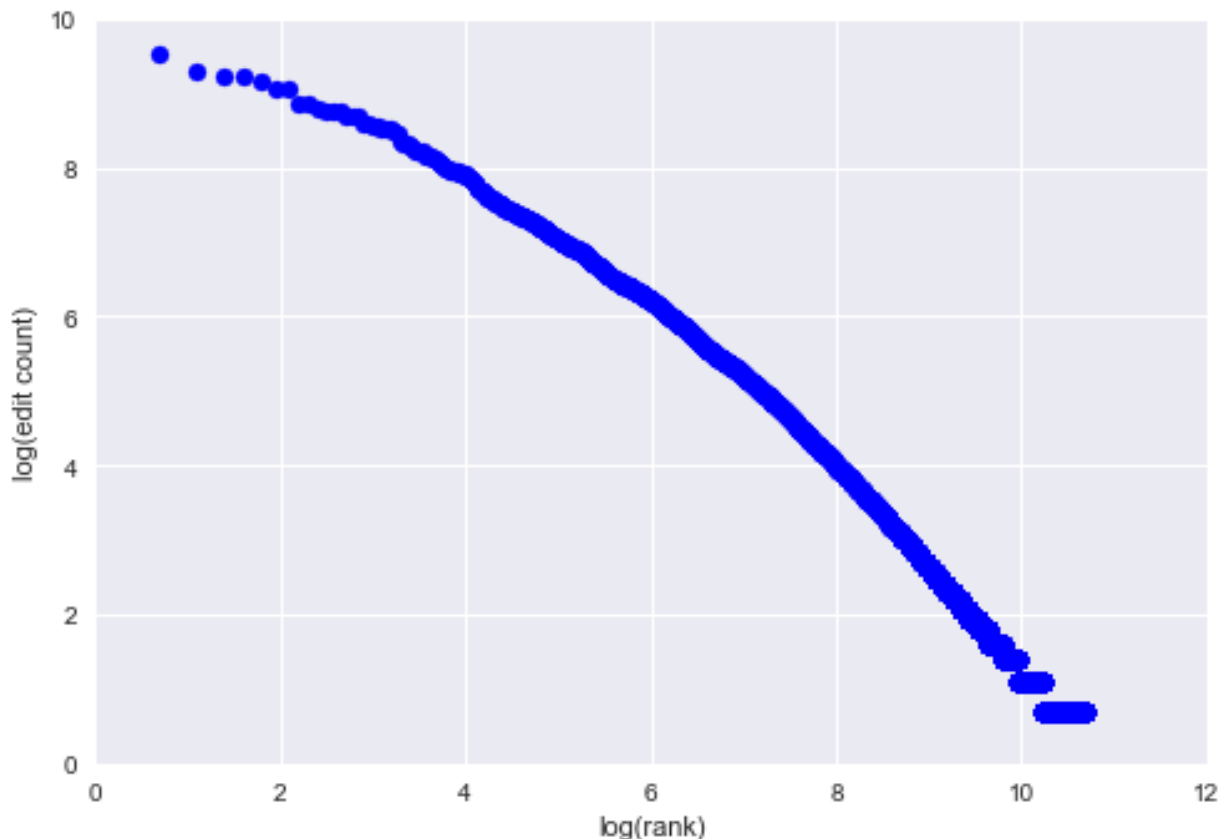


図 4.2 編集回数と編集人数の関係

4.4.1 Wikipedia からのデータの取得

Wikipedia のデータは、MediaWiki^{*1}から取得した。まず、日本語版 Wikipedia に存在するページについての 1 時間ごとのページ閲覧数を取得した。続いて、取得した 1 時間ごとのデータをもとに 24 時間分のページ閲覧数の合計値を算出し、日次のページ閲覧数のデータとした。また編集履歴の情報から各ページが編集された回数を日次で算出した。

月次のデータとして、当該月の前月におけるページ閲覧数の平均値および平均の編集回数を算出した。加えて、毎月初日における各ページへの被リンク数を月次のデータとして算出した。すなわち 2015 年 10 月の各素性の値として、2015 年 9 月 1 日から 30 日までのページ閲覧数および編集回数の平均値から平均ページ閲覧数および平均編集回数を算出し、2015 年 10

^{*1} <https://dumps.wikimedia.org/>

月 1 日時点での被リンク数を算出した。

編集者によるページの編集履歴は以下のように取得した。コンテンツベクトルの学習に耐えうる十分に長い系列長を確保するため、編集履歴は 2015 年 8 月末までにアニメ・漫画・ゲームカテゴリに属するページの編集回数が多く、Wikipedia の編集者として登録されているユーザの編集履歴を選択した。対象としたアニメ・漫画・ゲームカテゴリに属するページの編集者と各編集者の編集回数は、図 4.2 に示すようにジップの法則に従って分布していた。そこで、編集回数が全体の上位 5 % にあたる 2,500 人の編集者を選択し、解析の対象とした。選択されたユーザによる最低の編集回数は、86 回であった。

選択されたユーザの編集履歴を用いてコンテンツベクトルの学習を行った後、作成した日次および月次のデータを用いて、将来の人気予測を行った。

4.4.2 データに対する前処理

コンテンツ作品ページの統合

長期間連載されている作品の Wikipedia ページでは、メインページ以外にもキャラクターページや関連作品ページなどが存在する。元々は一つのページに記述されていた内容が複数のページに分散されて記載されるため、そのようなコンテンツ作品に関するメインページは、関連ページに情報が分散しページの閲覧数や被リンク数が減少する。そのため、見かけ上のページ閲覧数や被リンク数が少なくなってしまうという問題が生じる。この問題を解決するため、Wikipedia の関連ページを大元のページに対する構造をユーザが記述する Path Navi を利用した。解析対象とした全ページの中で Path Navi が存在するページがある場合、Path Navi を解析しメインページとの対応を取得した。Path Navi が存在する場合、予測に用いる編集回数、ページ閲覧、被リンク数をメインページの値に加えることで、見かけ上の値が減少する問題に対処した (Path Navi が存在しない場合、統合はなされない)。

解析対象とする作品の選択

本研究では、アニメ・ゲーム・漫画を人気予測の対象とした。予測対象とするコンテンツ作品に対応する Wikipedia 記事の取得は、二段階に分けて行った。

まず、Wikipedia のカテゴリ「アニメ」、「漫画」、「ゲーム」の下に存在する Wikipedia 記事のうち「○○の一覧」といったページを除いて取得した。続いて、コンテンツ作品に対応する

Wikipedia 記事の統合を行い、各コンテンツ作品ごとに複数存在する Wikipedia 記事のページごとの指標を一つにまとめた後、対象とした期間の被リンク数の平均上位 2,000 件を取得した。その後、低次元にベクトル化でき、対象とした期間に編集が一回以上行われているという条件を満たすコンテンツ作品に対応する統合後の Wikipedia 記事に絞り、さらに各月における被リンク数の変動が前月比 10% 以内の作品を選択した。これは、被リンク数に急激な変動が起こる場合は、今回のモデルによる予測対象外としたためである。絞り込みの結果、解析対象となった作品は 1,547 件となった。

4.5 予備実験

本節では、提案手法を用いる際に検証すべき点の前提条件についての分析を行う。実験は日本語版の Wikipedia を用いて行い、人気予測対象はアニメ、漫画、ゲームなどのコンテンツ作品を対象とした。対象としたデータを用いて、提案モデルを用いて人気予測を行う際の前提としている、Wikipedia の編集履歴においてユーザの嗜好が反映されていること、編集履歴を用いて学習したコンテンツベクトルがジャンルや流行した年代といった多面的な情報をもつことを検証した。

4.5.1 編集者系列の定性分析

まず、Wikipedia の編集者による編集履歴についての分析を行なった。本章では、登録されている編集者のうち 2015 年 10 月以前の編集回数上位 2,500 人に着目した。ベクトル化を行う前提として、編集履歴にユーザの嗜好が反映され、何らかの形で類似するコンテンツ作品に対応する Wikipedia の記事が、編集履歴上にて隣接することが求められる。そこで、今回対象とした編集者からランダムに 5 名を選択し、その編集履歴の一部について定性的な分析を行った。

ベクトル化の際には、あるコンテンツ作品のベクトル表現を得るために編集履歴上にて周辺に存在するコンテンツ作品に対応する Wikipedia 記事 10 件を用いている。そこで系列長 10 の編集履歴の断片を選択し、分析を行った。分析の対象とした編集者の中からランダムに編集者およびその編集者の編集履歴における位置を選択し、選択された位置以降に編集された 10

表 4.1 コンテンツ作品についての編集履歴 (例)

| | 編集者 1 | 編集者 2 | 編集者 3 |
|----|-----------------|-------------|----------|
| 1 | ルパン三世 | ワンパンマン | 七色いんこ |
| 2 | 金田一少年の事件簿 | チャゲチャ | ネオ・ファウスト |
| 3 | ルパン三世 VS 名探偵コナン | 斉木楠雄のΨ難 | 火の鳥 |
| 4 | ルパン三世 (映画) | チャゲチャ | バンパイヤ |
| 5 | ルパン三世 VS 名探偵コナン | 斉木楠雄のΨ難 | ピピちゃん |
| 6 | ルパン三世 (映画) | 青の祓魔師 | マグマ大使 |
| 7 | ルパン三世 VS 名探偵コナン | 斉木楠雄のΨ難 | マコとルミとチイ |
| 8 | ルパン三世 (映画) | 暗殺教室 | マイクロイド S |
| 9 | 金田一少年の事件簿 | 斉木楠雄のΨ難 | どろろ |
| 10 | ルパン三世 (映画) | NARUTO (映画) | 鉄腕アトム |

| | 編集者 4 | 編集者 5 |
|----|-----------------|----------------------------|
| 1 | コードギアス (キャラクター) | MARVEL VS. CAPCOM (テレビゲーム) |
| 2 | コードギアス (キャラクター) | ジョジョの奇妙な冒険 (テレビゲーム) |
| 3 | コードギアス (アニメ) | ジョジョの奇妙な冒険 (漫画) |
| 4 | コードギアス (テレビゲーム) | [削除されたページ] |
| 5 | コードギアス (キャラクター) | さくらがんぼる! |
| 6 | コードギアス (キャラクター) | 大江戸ロケット |
| 7 | コードギアス (漫画・小説) | スパイダーマン |
| 8 | コードギアス (キャラクター) | X-MEN VS. STREET FIGHTER |
| 9 | コードギアス (キャラクター) | マーベル・スーパーヒーローズ |
| 10 | コードギアス (キャラクター) | エックス・メン・チルドレン オブ ジアトム |

のコンテンツ作品を抽出した結果を表 4.1 に示す。

編集履歴についての系列を 5 つ選択した結果，系列 1 では，「ルパン三世」の関連作品および「金田一少年の事件簿」といった作品が並んだ。系列 2 では，「ワンパンマン」や「斉木楠雄のΨ難」など 2010 年代のギャグ漫画が並んだ。系列 5 では，「ジョジョの奇妙な冒険」関連作品とアメコミの関連のゲームが多く見られた。また系列 3 では，手塚治虫の作品が並び，系列 4 では「コードギアス」シリーズの関連ページが並んでいた。

以上をまとめると，編集履歴において隣接する Wikipedia 記事に対応するコンテンツ作品は，ジャンル，作者，流行した年代，商品としてのカテゴリーなどが類似している傾向が見られ，ユーザの嗜好を反映していることが示唆された。この結果から，各ユーザの編集履歴にお

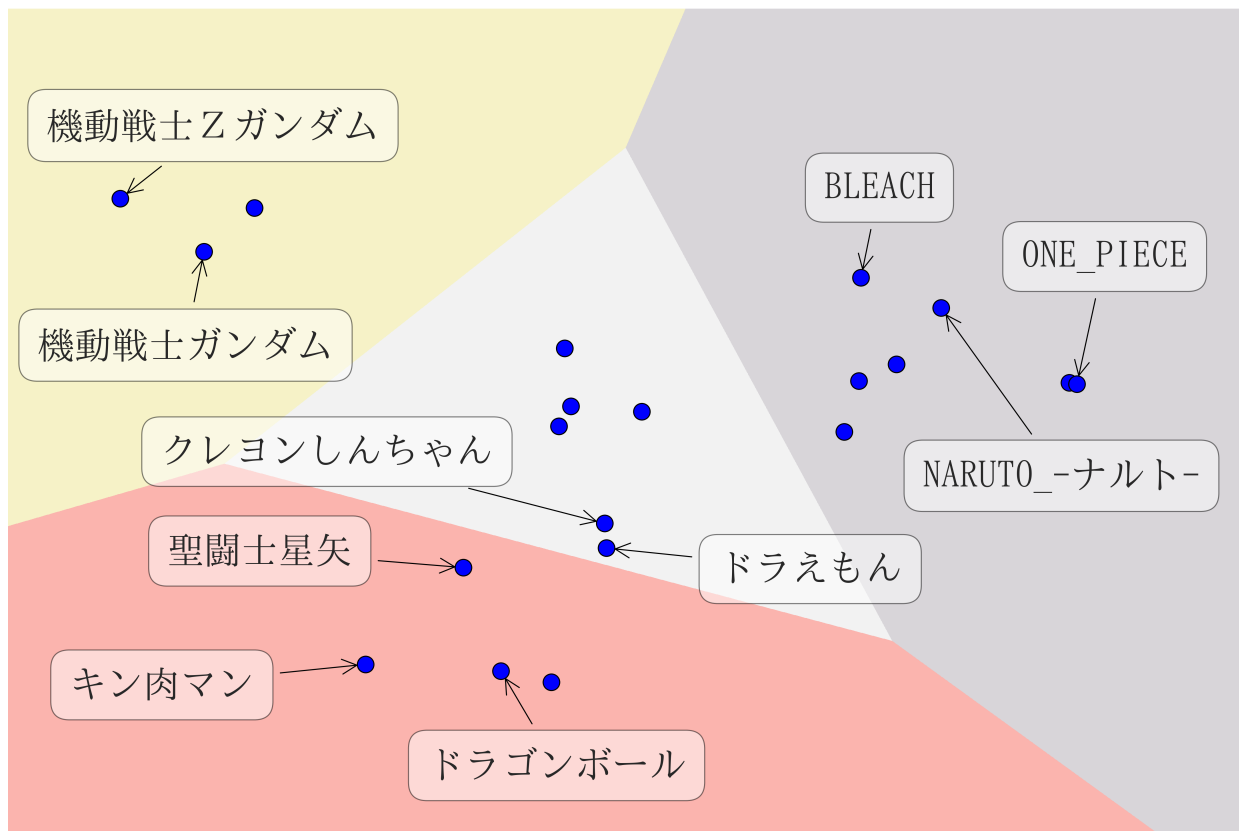


図 4.3 コンテンツベクトルのクラスタリング結果

いて隣接する Wikipedia 記事に対応するコンテンツ作品のベクトル表現により対象とするコンテンツ作品のベクトルを表現できることが示唆された。

4.5.2 コンテンツベクトルの取得

続いて、Wikipedia の編集者による編集履歴から各コンテンツ作品のベクトル表現を学習する方法について述べる。人気予測の対象としたカテゴリに属するページの編集回数が多い編集者の編集履歴を用いた。編集回数が少ない編集者のデータからの学習は難しいため、全 50,000 人の上位 5% に当たる編集回数上位 2,500 人の編集履歴をベクトル化の対象とした。選択された編集者の編集回数は 86 回～20,793 回であった。

各コンテンツ作品を単語と見なし、編集系列においてあるコンテンツ作品の前後に現れるコンテンツ作品を予測するように学習した。エポック数は 50,000 回とし、ウィンドウサイズ

表 4.2 クエリとしたコンテンツ作品に隣接するコンテンツ作品

| | | クエリとした作品 | |
|--------|---|-------------|-----------------------|
| | | SLAM DUNK | ドラゴンボール (アニメ) |
| 隣接する作品 | 1 | 幽遊白書 | ドラゴンボール (漫画) |
| | 2 | タッチ (漫画) | ドラゴンボール (アニメ ; スペシャル) |
| | 3 | Dr. スランプ | ドラゴンボール (映画) |
| | 4 | 気まぐれオレンジロード | Dr. スランプ |
| | 5 | I's | ドラゴンボール (アニメ ; スペシャル) |

| | | クエリとした作品 | |
|--------|---|-------------|------------|
| | | NARUTO(漫画) | ドラえもん |
| 隣接する作品 | 1 | NARUTO(ゲーム) | ドラえもん (映画) |
| | 2 | NARUTO(映画) | クレヨンしんちゃん |
| | 3 | FAIRY TAIL | ドラえもん (映画) |
| | 4 | NARUTO(ゲーム) | ドラえもん (映画) |
| | 5 | D.Gray-man | 21 エモン |

は 10, 編集履歴における出現数が 10 回以上の Wikipedia 記事に対応するコンテンツ作品をベクトル化した。ベクトル化の際のモデルには CBOW および Skip-gram[Mikolov 13a] を用いた。

学習を行った後, 得られたベクトルについて, 定性的な分析を行った。学習されたベクトルの有効性を検証するため, コンテンツ作品を 4 件選択し, それぞれのコンテンツ作品のベクトルとベクトル空間において最も近接しているコンテンツ作品上位 5 件を取得した。上位 5 件には, クエリとしたコンテンツ作品に関連する作品や同時期に放映・連載された作品, 似た傾向にある作品が選ばれた。選択したコンテンツ作品と類似のコンテンツ作品をまとめたものを表 4.2 に示す。

また, 図 4.3 に被リンク数上位 20 件のコンテンツ作品に対して主成分分析を行い, 二次元に射影した結果を示す。右の領域には, ポケットモンスター, ONE PIECE, NARUTO といった 10 代後半から 20 代にかけて人気のあるコンテンツ作品が並んだ。中央の領域には, クレヨンしんちゃんやドラえもんといった子供向けのコンテンツ作品が多く見られた。左上の領域には, 機動戦士ガンダムおよび機動戦士 Z ガンダムとガンダムシリーズの作品が位置し, 左

下の領域には、ドラゴンボール、聖闘士星矢、Dr. スランプなど 80 年代から 90 年代にかけて流行したコンテンツ作品が位置した。各領域ごとに対象年齢や放送していた年代といった性質が類似するコンテンツ作品が置かれていた。この結果から学習されたベクトルは各作品のジャンルや流行した年代といった情報を内包していることがわかった。

4.6 実験

本節では予備実験の結果、ジャンルや流行した年代といったコンテンツ作品の多面的な情報を捉えていることが示唆されたコンテンツベクトルを用いて、コンテンツ作品の人気予測を行う。また、ベースラインとしてコンテンツベクトルを用いない場合の精度を算出し、コンテンツベクトルを用いることにより予測精度が有意に向上することを示す。

4.6.1 人気予測

人気予測は、Wikipedia における各コンテンツ作品に対応する Wikipedia 記事の被リンク数を人気の指標として、Wikipedia から得られた素性とコンテンツベクトルを入力として与えたモデルにより行う。被リンク数は、ウェブ上のページの重要度の指標 [Page 99] やブログの人気度の指標 [Wu 06] として用いられる。そこで本研究では、[Kliegr 08] をもとに、Wikipedia における各ページの被リンク数を Wikipedia 記事に対応するコンテンツ作品の人気指標とした。

予測モデルは 2 つの多層ニューラルネットワーク (MLP) からなり、それぞれの MLP に月次データ $\mathbf{X}_{c,M}^t$ およびコンテンツベクトル T_c を入力として与えた。コンテンツベクトルとしては、Skip-gram により学習を行なった T_{sg} および CBOW により学習を行なった T_{cbow} の 2 つのいずれかを用いた。また、ベースラインとしてコンテンツベクトルを用いないモデルによる予測を置き、提案手法と比較した。

月次の素性を扱う MLP は、各層のユニット数を 8 とし、2 層の構造とした。また、コンテンツベクトルを扱う MLP は、各層のユニット数を 32 とし全 2 層の構造で、各層間にドロップアウト層を挿入した。両 MLP の出力を結合し、ユニット数 24 の MLP 層に入力として与え、最終的な出力を得た。最終層を除く層の活性化関数には ReLU を用い、最終層では線形関数を用いた。なお、コンテンツベクトルを用いない場合には、上記のモデルからコンテンツベ

表 4.3 コンテンツ人気予測の結果

| | MAPE ($\times 10^2$) | | |
|-----------------------|------------------------|---------------------|----------|
| | Proposed T_{sg} | Proposed T_{cbow} | Baseline |
| 1 | 0.153 | 0.148 | 0.150 |
| 2 | 0.147 | 0.148 | 0.154 |
| 3 | 0.146 | 0.149 | 0.150 |
| 4 | 0.147 | 0.148 | 0.151 |
| 5 | 0.147 | 0.147 | 0.151 |
| Average | 0.148 | 0.148 | 0.151 |
| p-value (vs Baseline) | 0.004 | 0.051 | - |

クトルを扱う部分を除いた MLP を用いた。

学習は誤差逆伝播法にて行い、最適化は RMSprop を用いた。学習率は 0.0001 とし、エポック数は 800 として、早期終了を用いて実験を行った。

解析の対象としたコンテンツ作品について、2015 年 10 月から 2016 年 3 月までの 6 ヶ月分のデータを学習データ、2016 年 4 月から 7 月までの 3 ヶ月分のデータをテストデータとし、学習データを用いてモデルの学習を行った。その後、テストデータをモデルに与えコンテンツ作品ごとに 3 ヶ月分の予測を行い、実際の値からのずれを平均絶対パーセント誤差 (MAPE) にて評価した。代表的なコンテンツ作品であるアニメの放映は、3 ヶ月を 1 クールとしているため、本研究では予測期間を 3 ヶ月と設定した。

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_{true} - y_{pred}}{y_{true}} \right| \quad (4.2)$$

コンテンツベクトルを用いた場合と用いない場合について、全コンテンツ作品についての MAPE の値の平均値を表 4.3 に示す。実験は独立した 5 回の試行を行い、得られた結果の平均を最終的な結果とした。コンテンツベクトルを用いた場合、 T_{cbow} および T_{sg} のいずれについても、用いない場合と比較して平均の MAPE の値は 2.0 % 低下していた。また、 t 検定による p 値は T_{cbow} で 0.004, T_{sg} で 0.051 であり、 T_{cbow} を用いた場合に有意水準 1% で有意に MAPE の値が低下していた。以降の結果は、 T_{cbow} による結果を示す。

4.6.2 予測期間と予測精度

続いて、予測期間の長さや提案手法による予測の精度の関係を調べた。0, 1, 2, 3ヶ月後の被リンク数を予測し、 T_{cbow} およびベースライン手法の予測精度を算出した。データは、2015年10月から2016年9月までのデータを用い、そのうち2016年7月から9月のデータを各予測におけるテストデータとした。実験は5回の独立した試行を行い、その平均値を最終的な結果とした。

表4.4に予測期間の長さや提案手法の精度の関係を示す。提案手法およびベースライン手法のいずれにおいても、予測期間が長くなるほど予測精度が低下していた。また、0ヶ月後の予測では両手法の間における予測精度の違いは小さく、 t 検定を行なった結果、提案手法とベースラインの間に有意な差は認められなかった。1, 2, 3ヶ月後の予測においては、予測期間が長くなるほど p 値は小さくなるものの、有意水準1%で有意差が認められたのは、3ヶ月後の予測のみであった。

4.6.3 予測精度の精度の高いクラス

MAPEを算出した結果、コンテンツベクトルを用いない場合と比較して精度が上がるコンテンツ作品と上がらないコンテンツ作品が存在した。両者の違いを分析するため、コンテンツベクトルによるクラスタリングを行い、各クラスごとの平均のMAPE値を算出し、コンテンツベクトルを用いる場合と用いない場合の変化を比較した。表4.5に、各クラスに含まれるコンテンツの数とコンテンツベクトルを用いない場合に対する平均のMAPEの改善を百分率で示した。クラスタリングはKMeansにて行い、セントロイドの数は全コンテンツ作品数1,547に対して15とした。

クラスタリングを行った結果、各クラスに含まれるコンテンツ作品数はクラス6を除いて22から318の間であった。クラスに含まれるコンテンツ数が50以上であり、コンテンツベクトルを用いた場合と用いない場合を比較して平均のMAPE値の改善が大きい5つのクラス1, 2, 5, 8, 11に着目し、それぞれに含まれる被リンク数の多いコンテンツ作品20件を調べた。

表 4.4 予測期間と予測精度の関係

| | MAPE ($\times 10^2$) | | | |
|-----------------------|------------------------|----------|------------|----------|
| | month 0 | | month 1 | |
| | T_{cbow} | baseline | T_{cbow} | baseline |
| 1 | 1.31E-02 | 8.12E-03 | 7.61E-02 | 7.78E-02 |
| 2 | 2.11E-02 | 2.77E-02 | 7.71E-02 | 7.86E-02 |
| 3 | 1.45E-02 | 9.96E-03 | 7.63E-02 | 8.47E-02 |
| 4 | 3.93E-02 | 1.93E-02 | 8.05E-02 | 8.00E-02 |
| 5 | 8.45E-03 | 1.41E-02 | 8.03E-02 | 7.97E-02 |
| Average MAPE | 1.93E-02 | 1.58E-02 | 7.81E-02 | 8.02E-02 |
| p-value (vs baseline) | 0.607 | - | 0.211 | - |

| | MAPE ($\times 10^2$) | | | |
|-----------------------|------------------------|----------|------------|----------|
| | month 2 | | month 3 | |
| | T_{cbow} | baseline | T_{cbow} | baseline |
| 1 | 1.19E-01 | 1.21E-01 | 1.48E-01 | 1.50E-01 |
| 2 | 1.19E-01 | 1.22E-01 | 1.48E-01 | 1.54E-01 |
| 3 | 1.19E-01 | 1.21E-01 | 1.49E-01 | 1.50E-01 |
| 4 | 1.23E-01 | 1.21E-01 | 1.48E-01 | 1.51E-01 |
| 5 | 1.19E-01 | 1.30E-01 | 1.47E-01 | 1.51E-01 |
| Average MAPE | 1.20E-01 | 1.23E-01 | 1.48E-01 | 1.51E-01 |
| p-value (vs baseline) | 0.137 | - | 0.004 | - |

クラスタ 1 には「機動戦士ガンダム」シリーズの作品のみが含まれており、コンテンツベクトルを用いることで 19.73% 精度が向上していた。クラスタ 2 では、「ひぐらしのなく頃に」や「エウレカセブン」といったコンテンツ作品が含まれており、MAPE の改善は 5.27% であった。クラスタ 8 では、「とある魔術の禁書目録」や「鋼の錬金術師」などのコンテンツ作品が含まれ、改善は 3.20% であった。クラスタ 11 では、「ゲゲゲの鬼太郎」、「銀河鉄道 999」、「サイボーグ 009」を始めとするコンテンツ作品が含まれ、3.29% 向上していた。最後にクラスタ

表 4.5 クラスタごとの MAPE

| cluster id | ratio | number of contents |
|------------|--------|--------------------|
| 1 | 19.73 | 61 |
| 2 | 5.27 | 318 |
| 3 | 1.00 | 119 |
| 4 | 17.72 | 22 |
| 5 | -5.64 | 120 |
| 6 | 18.92 | 1 |
| 7 | 2.66 | 132 |
| 8 | 3.20 | 192 |
| 9 | 4.36 | 31 |
| 10 | -36.09 | 22 |
| 11 | 3.29 | 183 |
| 12 | -2.07 | 31 |
| 13 | 11.96 | 31 |
| 14 | 1.36 | 246 |
| 15 | 4.81 | 23 |

5では、「進撃の巨人」や「妖怪ウォッチ」といった作品が含まれ、5.64% 予測精度が低下していた。

4.7 考察

この章では、編集者ごとの編集履歴の分析、Wikipedia 記事に対応するコンテンツ作品のベクトル化を行なった結果の可視化、人気予測およびクラスタごとの人気予測の精度の変化について、得られた実験結果から考察する。

4.7.1 編集履歴の系列

コンテンツ作品をベクトル化する際の前提として、入力される系列において隣接するコンテンツ作品は類似していることが求められる。Wikipedia の編集者は、自身が詳しいページについて編集を行うため、編集履歴において隣接する Wikipedia 記事に対応するコンテンツ作品は何らかの形で類似していると考えられた。解析の対象とした編集者の編集履歴系列からランダムに一部を取り出し、実際に隣接する Wikipedia 記事に対応するコンテンツ作品が類似しているかを定性的に分析した。

その結果、隣接する Wikipedia 記事に対応するコンテンツ作品は、同一コンテンツ作品の関連ページである場合やジャンルが同一であるコンテンツ作品、作者が同一であるコンテンツ作品などの Wikipedia 記事であり、両者の間には関連があることが明らかとなった。この編集履歴を用いることにより学習されるコンテンツ作品のベクトル表現は、ジャンルや作者についての情報、流行した年代といった情報を保持するように学習されていた。

4.7.2 ベクトル化

Wikipedia の編集履歴から得られたコンテンツベクトルについて、定性的な分析を行った。表 4.2 に示すように、学習されたベクトル空間において、あるコンテンツ作品の周囲に位置するコンテンツ作品は、何らかの関わりがあった。最も多い関係性が、対象としたコンテンツ作品とその関連作品という関係性であった。また、作者が同一人物である場合もベクトル空間上の近い位置に存在した (Dragon Ball と Dr. スランプ, ドラえもんと 21 エモン)。加えて、同時代に連載・放映されたコンテンツ作品が近くに位置していた (SLAM DUNK と 幽遊白書)。その他にも、ジャンル・対象年齢が近いコンテンツ作品が選択される傾向にあった。

また、図 4.3 に対象としたコンテンツ作品のうち、被リンク数上位 20 件のコンテンツ作品のベクトルを 2 次元に射影した図を示した。各コンテンツ作品について主成分分析により得られた第一成分および第二成分の値を用いて KMeans クラスタリングを行った。その結果、一つ目のクラスタにはポケットモンスターや NARUTO, ONE PIECE といった 10 代後半から 20 代男性を中心に幅広い層に人気のある作品が位置した。二つ目のクラスタは、クレヨンしんちゃんやドラえもんといった広い年齢層に親しまれ、比較的古くから存在し近年にもアニメ

が継続的に放映されているまたは近年映画化した作品が位置した。3つ目のクラスタには、機動戦士ガンダムに関連するシリーズが位置した。ガンダムシリーズは、他のクラスタからの距離も大きく、ユーザ層が異なることが示唆された。実際にガンダムシリーズは、中高年の男性からの支持が大きいことが知られている。最後のクラスタには、ドラゴンボールや聖闘士星矢など80年代から90年代初頭にかけて流行したコンテンツ作品が位置した。各クラスタごとにそれぞれ流行した時代、主な支持層による違いが顕著に現れていると考えられる。以上の結果から、編集履歴を用いて学習したコンテンツベクトルは、ユーザの嗜好に基づく、コンテンツ作品のもつジャンルや流行した年代といった複数の尺度についての類似度を考慮したコンテンツ作品の多面的な情報を表現していると考えられる。

4.7.3 人気予測

解析対象としたコンテンツ作品について、対応する Wikipedia 記事における3ヵ月後の被リンク数を予測した。モデルには月次の入力素性とともに、編集者履歴から CBOW および Skip-gram それぞれの手法で学習したコンテンツベクトルを入力として与えた。その結果、CBOW および Skip-gram のいずれの場合においてもコンテンツベクトルを使用しない場合と比較して、予測精度は2.0%向上した。これは、対象としたコンテンツ作品に関する情報をコンテンツベクトルとして入力することで、対象のコンテンツ作品に対するユーザの嗜好がモデルに組み込まれたからであると考えられる。コンテンツベクトル以外のモデルに与えるページ閲覧数の推移や編集回数、当月の被リンク数といった特徴量は、それぞれ短期的な変動を捉えていると考えられる。しかしながら、対象としたコンテンツ作品を支持するユーザ層や長期にわたり形成されるコンテンツ作品に対するユーザの嗜好を考慮することができていない。コンテンツベクトルは Wikipedia の編集履歴から学習されており、長期にわたり形成されてきたコンテンツ作品に対するユーザの嗜好や対象のコンテンツ作品を支持するユーザ層の情報をモデルに与えていると考えられる。その結果、将来の人気予測においてコンテンツベクトルをモデルに与えることで予測精度が改善したと考えられる。

4.7.4 予測期間と予測精度

続いて、提案手法による予測精度の向上と予測期間の関係について調べた。0, 1, 2, 3 ヶ月後の予測を行なった結果、ベースラインと提案手法の両方において、予測期間が長くなるほど予測精度が低下していた。この結果から予測期間が長いほど予測が難しくなることがわかる。これは予測期間が長くなるほど、予測対象の変動は大きくなるためであると考えられる。

次に、コンテンツベクトルを用いた予測を行う提案手法とベースライン手法の比較した。0 ヶ月後（当月）の予測では、提案手法とベースラインの手法の間で MAPE 値に有意な差は認められなかった。当月の予測において、モデルに入力として与えられる特徴量のひとつと出力ラベルが一致している。このため、モデルは入力された特徴量のうちのひとつをそのまま出力すれば良いということになる。出力されるべき情報が入力データに含まれているため、コンテンツベクトルの有無に関わらず高い精度で出力を予測できると考えられる。一方、コンテンツベクトルを与えた場合ベースラインと比較して、モデルに入力として与えられる特徴量が多くなる。その結果、モデルが複雑になり、結果として精度が低下したと考えられる。

また、1, 2, 3 ヶ月後の予測では、予測期間が長くなるほどベースラインの手法と比較して提案手法による予測精度が上回っていた。両手法の間の統計的有意差についても、予測期間が長くなるほど p 値が小さくなり、3 ヶ月後の予測において提案手法は有意水準 1% でベースライン手法を上回っていた。本研究で予測対象とした Wikipedia の被リンク数は、ページ閲覧数といった指標と比較して短期的な変動が小さい傾向にある。そのため予測期間が短い場合には、コンテンツベクトルの有無によらず、入力データに含まれる当月の被リンク数の情報をもとに高い精度で予測ができると考えられる。一方、予測期間が長くなると当月の値からの変動が大きくなる。この変動は予測対象のコンテンツ作品の特徴やジャンルといった情報に影響されると考えられる。コンテンツベクトルは、コンテンツ作品の持つ多面的な情報を内包しているため、予測期間が長くなった場合、提案手法による予測がベースラインと比較して高い精度を示すようになると考えられる。

また、4 ヶ月以上の期間の予測については、本研究の対象であるコンテンツ作品の中でも多数を占めるアニメ作品の放映が 3 ヶ月ごとであり、新作の発表が 3 ヶ月以上前に行われないことがあるため行わなかった。

4.7.5 予測精度の精度の高いクラスタ

最後に得られたコンテンツベクトルに基づき、コンテンツ作品をクラスタリングし、予測精度が上がるクラスタと低下するクラスタを調べた。精度の変化が大きいクラスタのうち50以上のコンテンツ作品を含むクラスタに着目し、各クラスタに含まれる被リンク数の多い作品上位20件を確認することで分析を行った。精度が最も向上したクラスタ1に含まれる作品上位20件すべてが「機動戦士ガンダム」に関連する作品であった。ガンダムシリーズは、40年代の男性を中心に非常に強い人気がある。シリーズ作品が一つのクラスタを形成しており、またコンテンツベクトルを与えることで予測精度が大幅に向上したことから、このクラスタには固有の傾向があると考えられる。クラスタ2および8には、少年漫画および少年向けアニメの人気作品が数多く含まれていた。これらの作品は、10代後半から20代の男性を中心としたユーザ層に支持されていると考えられる作品である。コンテンツベクトルを用いることで予測精度は、両クラスタそれぞれ5.27%および3.20%向上していた。クラスタ内に存在する作品は、クラスタ内の作品すべてがガンダムシリーズであったクラスタ1ほど傾向が似ていないため予測精度の向上幅が小さくなったのではないかと考えられる。またクラスタ11には「サイボーグ009」や「銀河鉄道999」といった昭和期の作品が主に含まれていた。これらの作品は、中高年からの支持を得ていると考えられ、コンテンツベクトルによりそのような情報を与えられたことで精度が向上したと考えられる。最後に、精度が大幅に低下したクラスタ5について述べる。クラスタ5に含まれる作品の多くは、テレビアニメ、映画、ゲーム、漫画と幅広く様々なメディアに展開された作品であった。このため、ゲーム、映画など各メディアにおける支持層が広く存在し、結果として予測が難しくなったのではないかと考えられる。

以上をまとめると、コンテンツベクトルを用いる場合、作品のジャンルや支持するユーザ層がコンテンツベクトルに強く含まれるほど予測精度は向上する。逆に、複数のメディアにて展開され、支持するユーザ層が幅広くなった場合には予測精度が低下すると考えられる。

4.8 本章の結論

本研究ではウェブ上でのユーザの行動履歴を用いて、ユーザの嗜好に基づくジャンルをはじめとする複数の尺度についての類似度を考慮したコンテンツ作品の多面的な情報を学習し、そ

れらを用いてコンテンツ作品の人気予測を行った。具体的には、Wikipedia におけるユーザによる編集履歴に対して、Word2vec を応用した低次元ベクトル表現の獲得手法を適用した。その後、新たなニューラルネットワークに学習したベクトル表現を与えることで、人気予測の精度が向上することを示した。提案手法による予測精度は、予測期間が長い場合に有意に向上していた。これは予測期間が長い場合、入力として与えられる現在の人気だけでなく、作品についての情報が重要になるためであると考えられる。またクラスタごとの予測精度の比較結果から、獲得したベクトル表現に作品のジャンルや支持するユーザ層といった要素が強く含まれるほど予測精度が高くなることが示唆された。

今回の実験では、ウェブ上でのユーザの行動履歴として Wikipedia での編集履歴を用いた。しかし、提案手法はユーザが興味を持っていると考えられる対象物についての系列データを入力することができれば適用可能である。したがって編集履歴に限らず、レビューサイトにおけるレビュー文章の投稿履歴やオンラインショッピングサイトにおけるページの閲覧履歴などの幅広いユーザの行動履歴に対して適用可能であると期待される。

第5章

事前学習による特徴量抽出と人気予測

5.1 本章の背景

本章では、予測モデルの学習に用いるデータが十分に入手できない場合における予測手法について検証した。販売予測を行う場合に、予測モデルを学習するためのデータが十分に入手できるとは限らない。一方、目的の予測に関連しているデータは大量に入手できる場合が考えられる。例えば、日本における販売データの入手は容易であるが、販売を取り扱う企業が異なるといった理由から海外における販売データの入手は困難であるといった場合が考えられる。そこで本章では、グローバルに展開される商品に関する予測を例に、学習に用いることのできるデータが少ない場合における予測手法について検証した。

企業にとって自社商品の市場を拡大することは重要な課題である。近年、企業のグローバル化が進んでおり、商品を海外に展開する機会は増加している。そのため、海外市場の開拓を行うことの重要性は増している。海外市場を開拓し拡大するためには、各地域の市場ごとに適切な意思決定を行う必要がある。

このような背景のもと、海外市場における対象とする商品の販売数予測は重要な課題である。商品の販売数や将来の人気を予測することは、企業の意思決定において重要であり、これまでに多くの研究が行われている。これらの研究では、ソーシャルメディア上の情報を用いて、販売数を予測することを行なっている。しかし、海外の市場に関するデータを入手するこ

とは困難であり、十分なデータが入手できるとは限らない。また市場調査などにより入手可能なデータを増やす場合、自国内で調査を行う場合と比較してコストがかかる。

一方、近年大きな注目を集めている深層学習を用いた画像処理では、大規模なデータセットにおいて事前に学習したモデルを、目的のデータセットで再学習することで少ないデータでも高い精度を実現することが可能であることが知られている [Oquab 14]。例えば、画像中の物体を検出し、その物体のクラス分類を行う一般物体検出といったタスクでは、大規模なデータセットにて学習済みのニューラルネットワークの重み情報を用いて学習するニューラルネットワークを初期化している [Liu 16]。事前に学習されたニューラルネットワークにより特徴量を獲得することが可能になるため、少ないデータであっても学習を効率的に進めることが可能であることが知られている。

そこで本研究では、データ量が豊富な地域において得られるデータを、データ量の少ない地域における予測に活用する深層学習を用いた手法を提案する。本手法は図 5.1 に示すように、大量のデータが入手できる地域にて入手したデータを用いて学習したニューラルネットワークを入手可能なデータが少ない地域のデータに適用する。検証は、各地域における自然環境といった物理的な要因に影響されることの少ない情報財の中で、長期にわたって継続されている財であるマルチメディアタイトル（アニメ、ゲーム、マンガ）にて行う。データは、マルチメディアタイトルが幅広く網羅されており、各言語における紐付けがなされている Wikipedia のデータを入力として用い、Wikipedia のページ閲覧数の増減を予測対象とする。

本研究の貢献は以下である。

1. データ量の多い地域のデータを用いてモデルの事前学習を行うことでデータ量の少ない地域における予測精度が向上することを示した
2. 事前学習をする際に、データのスケールを低下させることで再学習後のモデルによる精度が向上する場合があることを示した

本章は、以下のように構成される。まず、5.2 節にて関連研究について述べ、続いて 5.3 節にて提案手法について説明した後、5.4 節にて実験において用いたデータについて記す。続いて、5.5 節にて、予備実験について述べ、その後、5.6 節にて実験の概要と結果について記す。5.7 節にて、実験の結果について考察し、最後にこの章における結論を述べる。

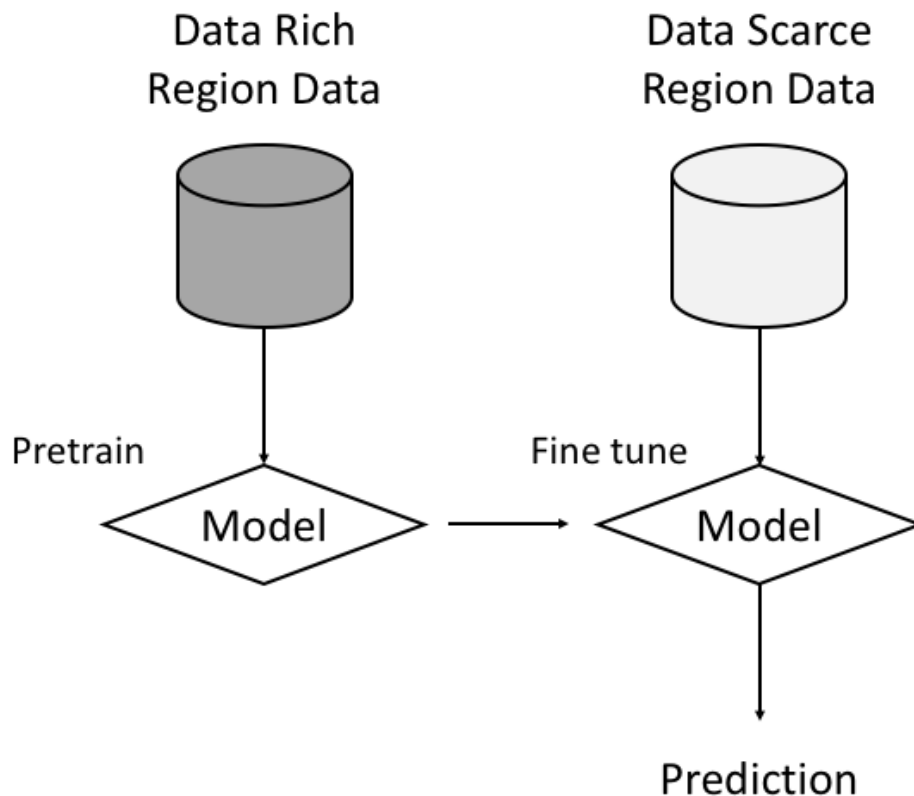


図 5.1 事前学習を行なったモデルによる人気予測の概略図

5.2 関連研究

5.2.1 深層学習の事前学習を用いたソーシャルメディアの研究

深層学習は、画像処理、音声認識、自然言語処理、強化学習の領域で大きな成果をあげている [LeCun 15], [Simonyan 14], [Hinton 12], [Sutskever 14], [Mnih 15]. 画像処理の分野では、大規模なデータセットにより事前学習したモデルを、小規模なデータしかない別のタスクに適合するように再学習することが有効であることが知られている [Oquab 14].

画像以外のデータでは、オンラインレビューに対する感情分析タスクにおいて、深層学習により複数のカテゴリにて学習した特徴表現を用いることで感情分析の予測精度が向上すること

が示されている [Glorot 11]. また, [Ding 15] では, Twitter から取得したデータに対して, 畳込みニューラルネットワークを適用し, 学習されたニューラルネットワークの中間表現を別のタスクにて用いることで精度が向上することを示している.

本研究では, 画像, テキストといった領域で適用され, 効果があることが示されてきた深層学習による事前学習の手法をソーシャルメディアから取得したカウントデータに適用し, 予測精度が向上するかを検証している.

5.2.2 Wikipedia の多言語性に着目した研究

本研究では, 各国のコンテンツ作品に関するユーザの嗜好に基づいた多面的な情報を Wikipedia の編集履歴から取得している. また, Wikipedia 内の言語リンクを用いて, 異なる言語間でのコンテンツ作品の紐付けを行っている. ここでは, Wikipedia の多言語性に着目した関連研究について述べる.

Wikipedia の多言語性に注目した研究では, 多言語にわたる記述内容を用いているものが存在する. 例えば, [Potthast 08] や [Sorg 12] では, 多言語にわたる情報探索に Wikipedia の本文を用いている. また [Ni 09] では, Wikipedia に対してトピックモデルを適用し, 多言語にわたる「トピック」を取得しており, [Ni 11] ではさらに学習されたトピックをもとにテキスト分類を行っている.

Wikipedia の記述内容以外の多言語にわたる点に注目した研究には, 多言語にわたる固有表現抽出を学習する [Nothman 13] やタクソノミーの抽出を行う [Melo 10] 研究が存在する. この他にも [Bao 12] では, Wikipedia の言語間での情報の偏りを解消するための施策を試みている. また [保住 14] では, 言語リンクを用いて言語間でのコンテンツの紐付けを行い, 各国におけるコンテンツの人気予測を試みている.

5.3 提案手法

本節では, データ量の多い地域のデータを用いて事前学習を行い, そのモデルをデータ量の少ない地域のデータにより再学習する提案手法について説明する.

本研究では, 系列として得られる日次の特徴量をニューラルネットワークに与える. ある地域 r におけるアイテム i について, ある月 m の t 日目に得られる日次の特徴量を, $x_{t,i}^{m,r}$ と

表 5.1 各言語版 Wikipedia データの詳細

| data | Titles | PV | | | Edit | | |
|------|--------|---------|---------|--------|------|------|--------|
| | | Max | Mean | Median | Max | Mean | Median |
| ja | 166 | 80,551 | 1143.01 | 569.0 | 81 | 0.76 | 0 |
| en | 148 | 290,684 | 1381.55 | 624.0 | 98 | 0.60 | 0 |
| zh | 139 | 923,145 | 492.06 | 241.0 | 67 | 0.51 | 0 |
| fr | 119 | 68,163 | 196.67 | 72.0 | 73 | 0.16 | 0 |
| ko | 111 | 785 | 24.85 | 11.0 | 126 | 0.17 | 0 |
| ru | 106 | 25,361 | 249.49 | 74.0 | 108 | 0.12 | 0 |
| th | 78 | 1,115 | 33.84 | 13.0 | 130 | 0.09 | 0 |
| id | 69 | 1,034 | 24.29 | 11.0 | 11 | 0.04 | 0 |
| vi | 60 | 1,673 | 41.44 | 15.0 | 77 | 0.11 | 0 |

する。 $x_{i,i}^{m,r}$ は、日次で得られる各変数を要素としてもつベクトルである。ある月 m における一ヶ月のデータを系列に並べ、結合した特徴量を $X_i^{m,r}$ とする。ある地域 r において対象とする期間 M に含まれる全ての月 m 、対象とする全てのアイテム i について得られる特徴量から、 $X^r = \{X_i^{m,r} | \forall i, m \in M\}$ を得る。また、ある地域 r の各月 m におけるアイテム i についてのクラスラベル $y_i^{m,r}$ から $y^r = \{y_i^{m,r} | \forall i, m \in M\}$ を得る。

事前学習は、データが豊富な地域 $r1$ におけるデータ (X^{r1}, y^{r1}) をパラメータ W_{r1} および b_{r1} をランダムに初期化したニューラルネットワークに与えることで行う。続いて、データが少ない地域 $r2$ における予測のための学習を行う。再学習は、事前学習にて得られたモデルのパラメータ W_{r1} および b_{r1} により初期化されたニューラルネットワークにデータ (X^{r2}, y^{r2}) を与えることで行う。

5.4 データ

本節では、実験に用いる Wikipedia のデータおよびその前処理について説明する。

5.4.1 Wikipedia データ

Wikipedia のデータは、MediaWiki^{*1}から取得した。取得対象とした言語は、日本語 (ja)、英語 (en)、中国語 (zh)、フランス語 (fr)、韓国語 (ko)、ロシア語 (ru)、タイ語 (th)、インドネシア語 (id)、ベトナム語 (vi) の 9 言語である。実験では、ページ編集履歴およびページ閲覧数の情報を用いた。ページ閲覧数は、各言語について 1 時間ごとに算出されるデータを取得し、24 時間分のページ閲覧数を合計し日次のページ閲覧数とした。ページの編集回数は、各言語版 Wikipedia のダンプデータに含まれるページ編集履歴をもとに日次で算出した。月次の平均ページ閲覧数は、各月の日次ページ閲覧数を合計し当該月の日数で割ることで算出した。データは 2015 年 12 月 15 日から 2017 年 3 月 31 日までの期間のデータを対象とした。

PathNavi を用いて関連するページの統計量を足し合わせた。日本語版 Wikipedia と各言語版 Wikipedia におけるページの紐付けは、各ページに存在する言語リンクを用いて行なった。対象とした言語に含まれるページのうち日本語版のページとの言語リンクが存在しない場合、実験の対象から外した。

検証実験の対象としたマルチメディアタイトルは、日本のオンラインショッピングサイトにおける 2016 年 1 月 1 日から 12 月 31 日までの期間の販売ランキング上位に位置した 166 のタイトルとした。各言語版 Wikipedia におけるページ閲覧数、編集回数についての統計量は、表 5.1 のようになった。

5.4.2 検証用データの準備

日次のデータについて過去 14 日間の移動平均を算出し、その後、表 5.2 に示す各言語について定めた値により割ることで正規化を行なった。なお、正規化後の値が 1 を超える場合は、1 とした。その後、各月初日から末日までのデータを系列として扱い、入力データを作成した。系列長が 31 未満の場合は、系列長が 31 となるようにゼロパディングを行なった。

予測対象の値は、算出された月次の平均ページ閲覧数について、当月および翌月の値を用いて変化率を計算し、変化率に基づいてクラスラベルを付与した。クラスラベルは、変化率

^{*1} <https://dumps.wikimedia.org/>

表 5.2 各言語に正規化における最大値

| lang | Daily PV | Daily Edit | Monthly PV |
|------|----------|------------|------------|
| ja | 50,000 | 60 | 10,000 |
| en | 100,000 | 60 | 25,000 |
| zh | 25,000 | 50 | 10,000 |
| fr | 5,000 | 30 | 2,000 |
| ko | 600 | 35 | 220 |
| ru | 12,500 | 40 | 6,000 |
| th | 800 | 40 | 330 |
| id | 640 | 10 | 300 |
| vi | 1,250 | 40 | 640 |

表 5.3 各言語におけるクラスバランス

| data | label | Class1 | Class2 | Class3 | Class4 | Class5 |
|------|-------|-------------|-------------|-------------|-------------|-------------|
| ja | 1992 | 467 (23.4%) | 498 (25.0%) | 334 (16.8%) | 290 (14.6%) | 403 (20.2%) |
| en | 1776 | 282 (15.9%) | 497 (28.0%) | 407 (22.9%) | 338 (19.0%) | 252 (14.2%) |
| zh | 1668 | 343 (20.6%) | 352 (21.1%) | 286 (17.1%) | 250 (15.0%) | 437 (26.2%) |
| fr | 1428 | 242 (16.9%) | 386 (27.0%) | 312 (21.8%) | 271 (19.0%) | 217 (15.2%) |
| ko | 1332 | 329 (24.7%) | 348 (26.1%) | 190 (14.3%) | 199 (14.9%) | 266 (20.0%) |
| ru | 1272 | 191 (15.0%) | 348 (27.4%) | 304 (23.9%) | 241 (18.9%) | 188 (14.8%) |
| th | 936 | 215 (23.0%) | 216 (23.1%) | 176 (18.8%) | 145 (15.5%) | 184 (19.7%) |
| id | 828 | 148 (17.9%) | 234 (28.3%) | 155 (18.7%) | 154 (18.6%) | 137 (16.5%) |
| vi | 720 | 141 (19.6%) | 154 (21.4%) | 147 (20.4%) | 133 (18.5%) | 145 (20.1%) |

が 20% 以上減少する場合 (Class1), -20% から-5% の場合 (Class2), -5% から 5% の場合 (Class3), 5% から 20% の場合 (Class4), 20% 以上の場合の 5 つの場合 (Class5) に分割した。各言語ごとのクラスラベルの割合を表 5.3 に示す。

表 5.4 モデルの構造と分類精度

| Wikipedia language | lr | MLP1 | MLP2 | 1DCNN |
|--------------------|-------|-------|-------|--------------|
| ja | 0.384 | 0.498 | 0.510 | 0.552 |
| en | 0.351 | 0.425 | 0.496 | 0.539 |
| zh | 0.371 | 0.492 | 0.487 | 0.525 |
| fr | 0.369 | 0.454 | 0.438 | 0.510 |
| ko | 0.355 | 0.414 | 0.401 | 0.459 |
| ru | 0.381 | 0.492 | 0.470 | 0.497 |
| th | 0.423 | 0.499 | 0.498 | 0.515 |
| id | 0.406 | 0.431 | 0.422 | 0.446 |
| vi | 0.328 | 0.423 | 0.418 | 0.434 |
| average | 0.398 | 0.459 | 0.453 | 0.497 |

5.5 予備実験：予測に用いるニューラルネットワークの構造の決定

本節では、予測に用いるニューラルネットワークの構造を決定するための実験を行なった。モデルは、ロジスティック回帰 (*lr*)、3層の多層パーセプトロン (*MLP1*)、4層の多層パーセプトロン (*MLP2*) および1次元畳み込みニューラルネットワークを用いた構造 (*1DCNN*) の4種類を検討した。最適化はいずれのモデルについても RMSprop にて行い、学習時のエポック数は1500とし、早期終了を25に設定し、学習を行なった。入力データは、ページ閲覧数と編集回数の2次元、系列長31の31×2次元のデータとして1次元畳み込みニューラルネットワークに与えた。ロジスティック回帰および多層パーセプトロンでは、31×2次元の行列を62次元のベクトルに変換した後、入力として与えた。予測は、当月と1ヶ月後の平均ページ閲覧の変化率についてを対象として行なった。学習は、2016年1月から9月までの9ヶ月分のデータを用いて行い、2016年10月から12月までの3ヶ月間のデータを用いてモデルの精度評価を行なった。学習データとして用いるデータの中からランダムに選択された10%のデータを学習時の検証データとした。

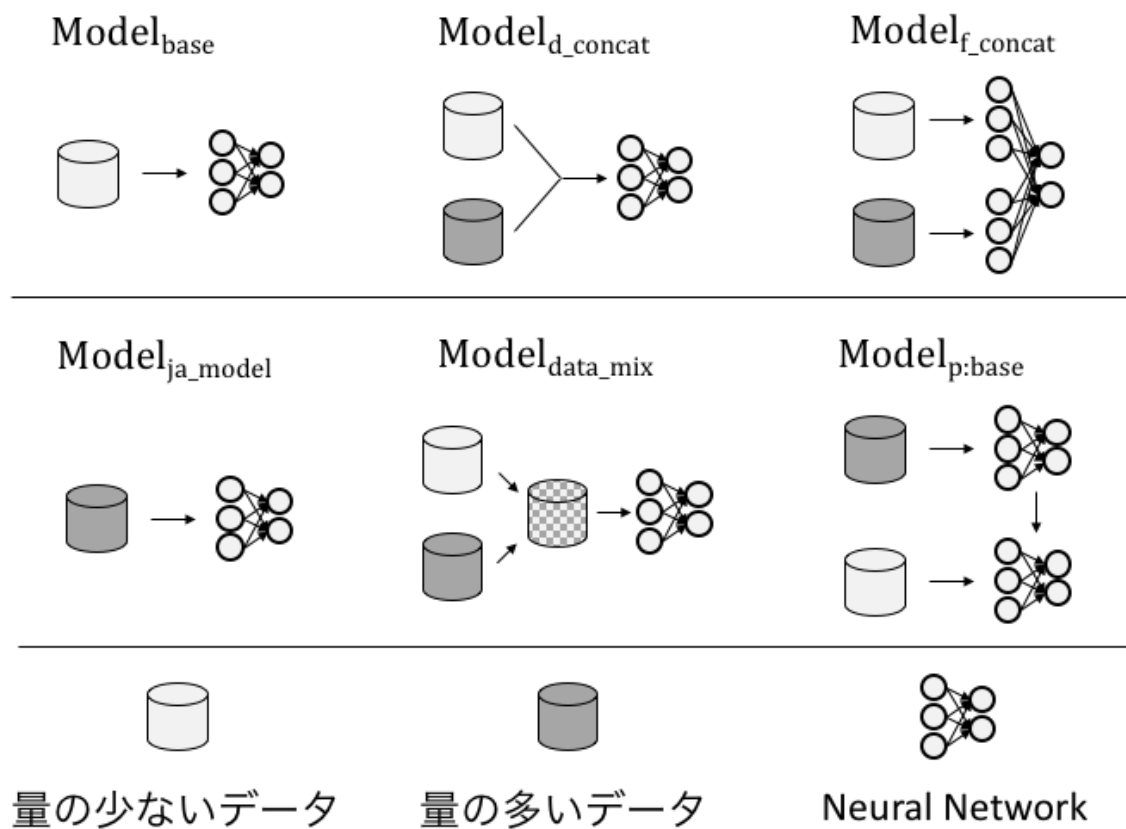


図 5.2 各予測モデルの概要

各言語について4種のモデルで予測した結果を表5.4に示す。予測結果はいずれのモデルについても、異なるランダムシードを用いた5回の試行結果の平均値を示している。検証実験の対象とした全9言語において、一次元畳み込みニューラルネットワーク(1DCNN)による分類結果が最も優れていた。これ以後の実験では、一次元畳み込みニューラルネットワーク(1DCNN)による結果をベースラインとして扱う。

5.6 実験

本節では、予備実験にて優れた分類精度を示すことが明らかとなった一次元畳み込みニューラルネットワークに対して、複数の地域におけるデータが存在する場合のデータ入力の形式と最適な事前学習手法の検討を行う。

5.6.1 ニューラルネットワークへのデータ入力

データの少ない地域における予測手法として、予測対象とした地域で得られるデータをそのまま用いることが考えられる。そこで本研究では、予備実験により高い分類精度を示すことが明らかになった1次元畳み込みニューラルネットワークによるモデルに各地域で得られるデータを用いるモデルをベースラインとした。

本研究では、データ量が多い地域のデータを活用して、データ量の少ない地域における予測問題の精度を向上することを目的としている。データ量の多い地域にて得られるデータは、単にデータの量が多いだけでなく、ユーザ数が多いことから細かな変動を反映している。そこで、ベースラインモデルに対して、データ量の多い地域のデータを有効に用いる手法を検証した。検証した手法は、以下の通りである。また各手法の概要を図5.2に示す。なお、データ量の多い地域のデータとして日本語版 Wikipedia のデータを用いた。

Model_{base} データ量の少ない地域で得られるデータのみを用いてモデルの学習を行い、データ量の少ない地域における予測を行う場合

Model_{d_concat} データ量の少ない地域のデータと対応するデータ量の多い地域のデータを結合し、モデルの学習を行い、データ量の少ない地域における予測を行う場合（対応するデータが存在しない場合は、学習データに含めない）

Model_{f_concat} データ量の少ない地域とデータ量の多い地域のデータを別々のニューラルネットワークに入力として与え、抽出された特徴量を結合し学習を行い、データ量の少ない地域における予測を行う場合（対応するデータが存在しない場合は、学習データに含めない）

Model_{ja_model} データ量の多い日本語のデータで学習されたモデルを用いてデータ量の少ない地域における予測を行う場合

Model_{data_mix} データ量の少ない地域とデータ量の多い地域のデータを区別なく混合し、学習を行い、データ量の少ない地域における予測を行う場合

Model_{p_base} データ量の多い日本語のデータで事前学習されたモデルをデータ量の少ない地域のデータで再学習し、データ量の少ない地域における予測を行う場合

表 5.5 入力データの与え方と分類精度

| Wikipedia language | $Model_{base}$ | $Model_{d_concat}$ | $Model_{f_concat}$ | $Model_{ja_model}$ | $Model_{data_mix}$ | $Model_{p:base}$ |
|--------------------|----------------|---------------------|---------------------|---------------------|---------------------|------------------|
| en | 0.539 | 0.531 | 0.561 | 0.534 | 0.603 | 0.606 |
| zh | 0.525 | 0.507 | 0.468 | 0.547 | 0.550 | 0.547 |
| fr | 0.510 | 0.485 | 0.516 | 0.509 | 0.552 | 0.568 |
| ko | 0.459 | 0.426 | 0.444 | 0.469 | 0.482 | 0.479 |
| ru | 0.497 | 0.416 | 0.431 | 0.517 | 0.530 | 0.584 |
| th | 0.515 | 0.536 | 0.530 | 0.583 | 0.559 | 0.593 |
| id | 0.446 | 0.424 | 0.442 | 0.441 | 0.463 | 0.477 |
| vi | 0.434 | 0.431 | 0.421 | 0.471 | 0.486 | 0.485 |
| average | 0.491 | 0.470 | 0.477 | 0.509 | 0.528 | 0.542 |

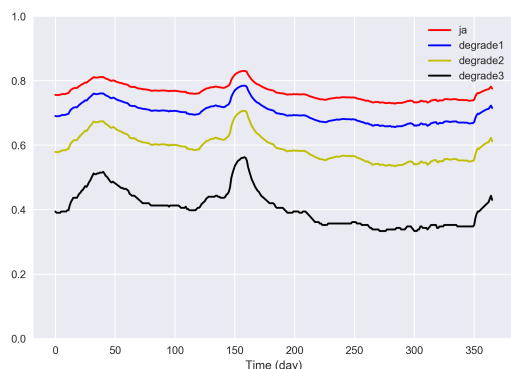
表 5.6 日本語データのスケール低下処理

| Down scale level | PV | Edit |
|------------------|------|------|
| scale1 | 10 | 3 |
| scale2 | 100 | 6 |
| scale3 | 1000 | 12 |

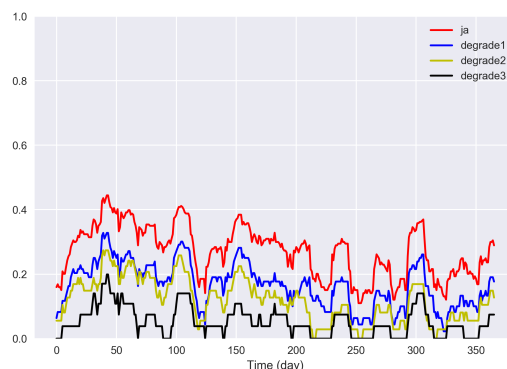
表 5.7 再学習を行わない場合の精度

| lang | <i>ja_full</i> | Limit | | Scale down | | |
|------|----------------|-------|-------|------------|--------------|--------|
| | | 1000 | 500 | scale1 | scale2 | scale3 |
| ja | 0.552 | 0.518 | 0.456 | 0.508 | 0.466 | 0.363 |
| en | 0.534 | 0.514 | 0.456 | 0.522 | 0.505 | 0.402 |
| zh | 0.547 | 0.515 | 0.448 | 0.493 | 0.464 | 0.363 |
| fr | 0.509 | 0.502 | 0.432 | 0.519 | 0.548 | 0.448 |
| ko | 0.469 | 0.457 | 0.403 | 0.478 | 0.485 | 0.421 |
| ru | 0.517 | 0.510 | 0.408 | 0.513 | 0.531 | 0.435 |
| th | 0.583 | 0.568 | 0.503 | 0.570 | 0.577 | 0.458 |
| id | 0.441 | 0.421 | 0.373 | 0.463 | 0.484 | 0.404 |
| vi | 0.471 | 0.472 | 0.434 | 0.494 | 0.526 | 0.428 |

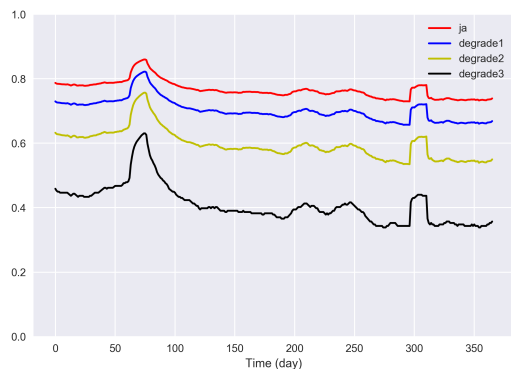
それぞれの手法にて、日本語版を除く 8 言語において予測した結果を表 5.5 に示す。分類精度を検証した結果、全 8 言語の平均ではデータ量の多い日本語のデータで事前学習を行なったモデルを各言語のデータで再学習した $Model_{p:base}$ の精度が最も高く、ベースラインの手法と比較して、10.5% 改善していた。個別の言語に着目すると、中国語 (zh)、韓国語 (ko)、ベトナム語 (vi) において日本語のデータを混合して学習した $Model_{data:mix}$ の精度がわずかに上回っていたが、 $Model_{p:base}$ との差はいずれの場合も分類精度の差で 0.5% 以下と小さかった。



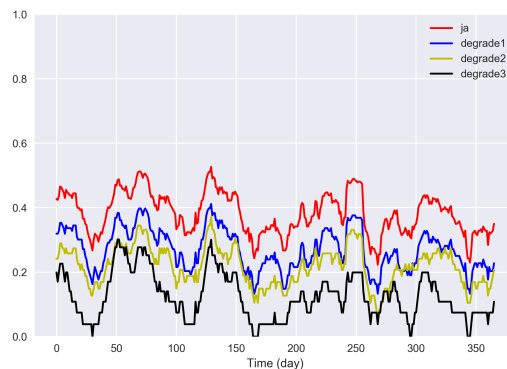
(a) ページ閲覧数 例 1



(b) 編集回数 例 1



(c) ページ閲覧数 例 2



(d) 編集回数 例 2

図 5.3 スケールを低下したデータ (例)

5.6.2 事前学習に用いるデータと予測結果の関係の分析

今回の実験において、データ量の多い地域のデータとして用いた日本語版 Wikipedia のデータは表 5.1 に示すように、他の言語のデータと比較してデータの量が多く、個別のデータのスケールも大きいという特徴を持つ。そこで事前学習において用いるデータの量とスケールが、再学習後の結果に与える影響を明らかにするため、日本語のデータに制限を加えたデータを用いた実験を行なった。まず、制限を加えたデータで学習したモデルによる予測精度を検証した。続いて、制限を加えたデータで事前学習を行なった後、再学習したモデルによる予測精度を検証した。

はじめに、制限を加えたデータについて説明する。データの量が結果に与える影響を調べるため、データ量を制限したデータを準備した。データ量を制限したデータは、学習に用いる日本語版 Wikipedia のデータをランダムに指定した数だけ選択することにより作成した。本実験では、データ数を 500 および 1000 とした。また、データのスケールが結果に与える影響を調べるため、データのスケールを低下したデータを準備した。スケールを低下したデータは、日本語版 Wikipedia データにおけるページ閲覧数および編集回数それぞれの値を、表 5.6 に記された値で割り、小数点以下を丸めることで作成した。スケールを低下したデータは、ページ閲覧数および編集回数を除した値により、 $scale1$ 、 $scale2$ および $scale3$ とした。図 5.3 に元データおよびスケールを低下したデータの例を示す。この図から表 5.6 に示す値によりデータを除すことで、値の大きさに応じてデータのスケールが低下していることがわかる。

制限を加えたデータを用いて学習したモデルにより、各言語のテストデータを分類した結果を表 5.7 に示す。なお、 ja_full のデータの列に示す結果は、表 5.5 における $Model_{ja_model}$ と同一である。実験の結果、データ数を制限した場合、データ数が少なくなるほどテストデータにおける分類精度も低下していた。また、データのスケールを低下させた場合は、日本語、英語、中国語、タイ語の 4 言語で分類精度が低下した一方、フランス語、韓国語、ロシア語、インドネシア語、ベトナム語の 5 言語では一部の条件で精度が向上した。

続いて、データ数の削減およびスケールの低下による事前学習がデータ量の少ない地域におけるデータの分類精度に与える影響を検証した。データ数を減らして事前学習を行い、その後再学習したモデルを、事前学習に用いたデータ数に応じて、 $Model_{p:lim1000}$ および $Model_{p:lim500}$ とした。また、データのスケールを低下して事前学習を行い、その後再学習したモデルを、スケールの低下度合いに応じて $Model_{p:scale1}$ 、 $Model_{p:scale2}$ および $Model_{p:scale3}$ とした。

日本語を除く 8 言語のテストデータにおける分類精度を表 5.8 に示す。分類精度を算出した結果、事前学習のモデルとしてスケールを低下した場合の一つである $Model_{scale1}$ を用いた場合、8 言語の平均として最も高い分類精度を示した。この結果は、各言語のデータを用いて学習を行なった $Model_{base}$ と比較して、13.0% の精度改善であった。事前学習を行う場合に用いるデータ数を減らした場合、再学習後のモデルにおける分類精度は、事前学習に用いたデータ数にしたがって低下していた。一方、事前学習を行う場合に用いるデータのスケールを低下させた場合、再学習後のモデルにおける分類精度は、事前学習に用いたデータ数にしたがって低下していた。

ケールを低下した場合、再学習後のモデルにおける分類精度は、元データを用いて学習した場合 ($Model_{p:base}$) よりも高くなっている場合もあった。なお、 $Model_{base}$ は表 5.4 における $1DCNN$ と同一であり、 $Model_{p:base}$ は表 5.5 における $Model_{p:base}$ と同一である。

表 5.8 事前学習したモデルによる分類精度

| Wikimedia language | $Model_{base}$ | $Model_{p:base}$ | $Model_{p:itm1000}$ | $Model_{p:itm500}$ | $Model_{p:scale1}$ | $Model_{p:scale2}$ | $Model_{p:scale3}$ |
|--------------------|----------------|------------------|---------------------|--------------------|--------------------|--------------------|--------------------|
| en | 0.539 | 0.606 | 0.607 | 0.556 | 0.624 | 0.598 | 0.603 |
| zh | 0.525 | 0.547 | 0.530 | 0.512 | 0.540 | 0.537 | 0.530 |
| fr | 0.510 | 0.568 | 0.567 | 0.560 | 0.568 | 0.564 | 0.546 |
| ko | 0.459 | 0.479 | 0.477 | 0.457 | 0.507 | 0.485 | 0.473 |
| ru | 0.497 | 0.584 | 0.585 | 0.560 | 0.624 | 0.614 | 0.599 |
| th | 0.515 | 0.593 | 0.571 | 0.519 | 0.592 | 0.588 | 0.562 |
| id | 0.446 | 0.477 | 0.486 | 0.460 | 0.498 | 0.498 | 0.514 |
| vi | 0.434 | 0.485 | 0.479 | 0.459 | 0.483 | 0.495 | 0.459 |
| average | 0.491 | 0.542 | 0.538 | 0.510 | 0.554 | 0.547 | 0.536 |
| improvement | - | 10.5% | 9.6% | 4.0% | 13.0% | 11.6% | 9.2% |

5.7 考察

本節では、予備実験および実験の節にて得られた結果についての考察を述べる。

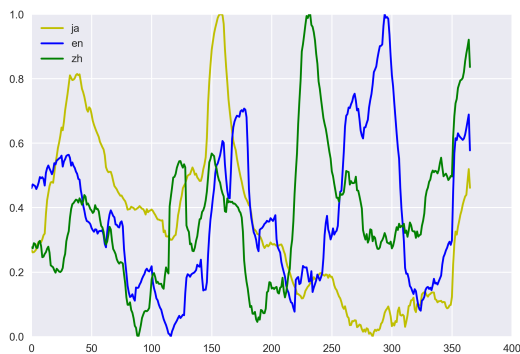
5.7.1 ニューラルネットワークの構造

まず、予備実験として予測の際に用いる学習モデルについて検討を行なった。学習モデルは、ロジスティック回帰 (lr)、多層パーセプトロン ($MLP1$ および $MLP2$) および一次元畳み込みニューラルネットワーク ($1DCNN$) の4種類のモデルを検証した。対象とした全9言語のデータについて、それぞれのモデルを学習し、テストデータにおける分類精度を検証した。その結果、全ての言語において $1DCNN$ を用いた場合に最も精度が高くなった。また、 lr と $MLP1$ または $MLP2$ を比較した場合、全ての言語において多層パーセプトロンの結果が優れていた。

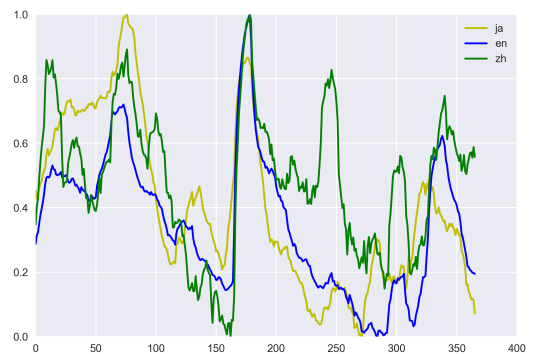
lr と $MLP1$ および $MLP2$ の比較結果から、 lr で捉えることのできなかった特徴量が、隠れ層の導入により $MLP1$ および $MLP2$ では捉えられるようになり、その結果分類精度が向上したと考えられる。また $1DCNN$ では、さらに畳み込み処理により入力データにおける時系列の関係性を捉えるようになったため精度が高くなったと考えられる。

5.7.2 ニューラルネットワークへのデータ入力

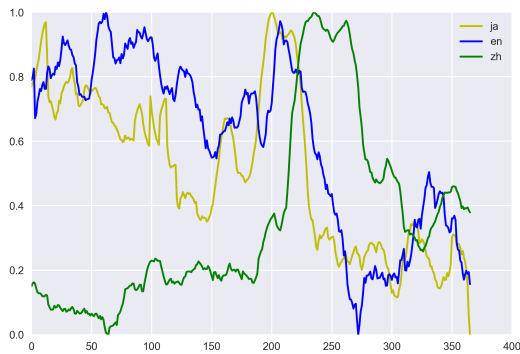
続いて、データ量が多い地域およびデータ量が少ない地域、両方のデータを用いて効果的な学習を行う手法を検証した。検証は、データを複数の入力方法で一次元畳み込みニューラルネットワークを用いたモデルに入力することで行なった。具体的には、対応するデータを結合する場合、対応するデータ同士から抽出した特徴量を結合する場合、日本語データで学習したモデルを用いる場合、データを混合して学習する場合、日本語データで事前学習を行う場合の5つの場合について検証を行なった。その結果、事前学習を行なった場合に分類精度は最も高くなっていた。最終的に得られるモデルの学習に用いたデータ数は、 $Model_{base} = Model_{d.concat} = Model_{f.concat} < Model_{ja.model} < Model_{data.mix} = Model_{p.base}$ となっている。この結果から、学習に用いるデータ数が多いほど分類精度は高くなることが示唆される。また、学習に用いたデータ数が同数である $Model_{data.mix}$ と $Model_{p.base}$ の間で差が生



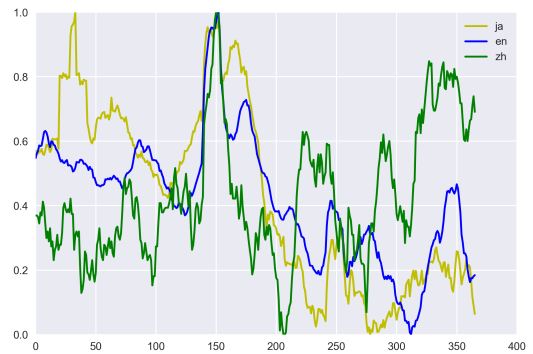
(a) ページ閲覧数比較 1



(b) ページ閲覧数比較 2



(c) ページ閲覧数比較 3



(d) ページ閲覧数比較 4

図 5.4 ページ閲覧数の推移 (例)

じていることから、事前学習の後に再学習する方が各言語に特化したモデルとなり精度が高くなったと考えられる。今回扱ったデータのうち中国語と韓国語のデータでは、単独で学習した場合と事前学習を行なった場合の精度改善の度合いが5%以下と小さかった。韓国語のデータでは、表 5.1 に示すように、ページ閲覧数のスケールが事前学習の際に用いる日本語のデータと大きく異なることが原因として考えられる。また中国語のデータでは、事前学習において用いた日本語のデータと比較して、データの変動パターンが異なっていた。図 5.4 に、日本語、英語、中国語の3つのデータについて、ページ閲覧数の推移をプロットしたものを示す。各言語のデータを0から1の範囲で正規化し可視化した結果、事前学習により精度が10%以上向上した英語のデータでは、ページ閲覧数のピークが日本語のデータと類似する傾向にあった一

方、中国語のデータではピークが日本語と異なることが多かった。中国では、インターネットの規制が強いためページ閲覧数の推移に影響を与えた可能性がある。このように、事前学習に用いるデータと再学習を行うデータのパターンやスケールが大きく異なる場合においては、提案手法の有効性が低下する可能性が示唆される。

5.7.3 事前学習に用いるデータと予測結果の関係の分析

最後に、事前学習を行う場合に用いるデータの量およびスケールが、再学習後のモデルに与える影響について分析した。データの量またはデータのスケールに制限を加えたデータを作成し、作成したデータを用いて学習した場合と制限を加えない場合を比較した。まず、制限を加えたデータで学習したモデルによる分類精度を検証した。その後、制限を加えたデータにより事前学習し、その後再学習したモデルによる分類精度を検証した。

制限を加えたデータで学習したモデルによる分類精度について述べる。学習に用いるデータの量を減らした場合、いずれの言語においても予測精度はデータ量に応じて低下していた。この結果から、学習において用いるデータの量が分類において重要であることが示唆された。これは、データ量が多いほど多様なパターンの入力が存在することとなり、結果としてモデルがより良い特徴量を抽出できるようになったためと考えられる。一方、スケールを低下した場合には、一部の言語で分類精度が向上していた。これは、学習時のデータとテスト時のデータのスケールが近いほど予測が行いやすいためであると考えられる。

次に、制限を加えたデータにより事前学習し、その後再学習したモデルによる分類精度を検証した。その結果、データのスケールを減少させた $Model_{p:scale1}$ および $Model_{scale2}$ の精度が、事前学習において全てのデータを完全な状態で用いた $Model_{p:base}$ を上回っていた。特に $Model_{p:base}$ と比較した場合に改善の度合いの大きかったのは、韓国語、ロシア語、インドネシア語、ベトナム語の4つの言語であった。ここでは、精度改善の大きかった中でも、データ量が比較的多い韓国語とロシア語のデータについて考察する。本実験において入力として用いたページ閲覧数とページの編集回数に注目すると、表 5.1 に示すように、ロシア語と韓国語のデータではページ閲覧数のスケールは大きく異なる。一方、編集回数のスケールは平均値で比較した場合、比較的近い値となっている。またこれらの2言語における編集回数の値は、最も精度が向上した条件である $Model_{p:scale}$ において用いた、スケールを低下した日本語のデー

タの編集回数の値とも比較的近い値となっている。このことから編集回数のスケール調整が予測精度の向上に寄与していることが示唆される。しかしながら、編集回数の平均の値では、ロシア語や韓国語と同程度の規模となっているフランス語のデータでは、スケールを低下した場合の予測精度の向上がみられない。そのため、予測に寄与する素性は編集回数のスケールのみではなく、ページ閲覧数の形状やページ閲覧数のスケールといった要素との関係が影響を与えらるると考えられる。

5.8 本章の結論

本研究では、異なる地域における商品の販売予測問題に取り組んだ。販売数予測を行う場合、各地域にて得られたデータを用いることが考えられるが、地域によっては十分なデータが得られないということが考えられる。そこで本研究では、データ量が多い地域にて得られたデータを用いてデータ量が少ない地域における予測精度を向上させる手法について検証した。予測対象としては、販売地域における自然環境の影響を受けにくい情報財を扱い、検証実験を行なった。検証実験では、はじめにオンラインショップサイトにおける商品の販売ランキングデータと Wikipedia における商品に対応するページの閲覧数に相関関係があることを確認した。続いて、ページ閲覧数予測問題において用いるニューラルネットワークのモデルについて検討した。その結果、実験を行った全 9 言語において次元畳み込みニューラルネットワークを用いる場合に最も精度が高くなることを明らかになった。その後、データ量の多い地域のデータを用いて、データ量の少ない地域における予測精度を向上するモデルについての検討を行った。データ量の少ない日本語を除く 8 言語について実験を行った結果、日本語のデータによる事前学習を行った場合の分類精度が最も高くなっていた。最後に、事前学習をする場合におけるデータの量とデータのスケールの影響を分析した。実験の結果から事前学習に用いるデータと再学習の際に用いるデータが似ている状況において再学習後の分類精度が高くなることが示唆された。本研究の結果から、各国のデータのみで学習を行う場合と比較して、再学習を行う場合平均で 10.5%、事前学習に用いるデータのスケールを低下した場合には平均で 13.0%、分類精度が向上することが明らかとなった。この結果から、データ量が少ない地域における予測を行う場合に事前学習が有効であると考えられる。今回の実験では、スケールの低下をヒューリスティックに決めた値にしたが行なったが、データの類似度をもとに自動で

最適な値を決められるようにすることで、より予測精度を上げることが可能になると考えられる。

第 6 章

Wikipedia から得た人気情報による販売予測

6.1 本章の背景

本章では，コンテンツ作品の販売予測の問題を考える．商品の販売予測を高い精度で行うモデルを学習するためには，学習時により多くのデータを用いる必要があるが，販売数に関するデータを増やすことは難しい．一方，Wikipedia をはじめとするソーシャルメディアは，多くのユーザによって利用されており，データの収集も容易であることが多い．

そこで本章では，コンテンツ作品に関わる Wikipedia のデータから抽出した人気情報を用いて，コンテンツ作品の販売予測を行う．コンテンツ作品の販売予測に用いる人気情報としては，4 章および 5 章において得られた知見に基づいた人気情報を用いる．5 章では，少ないデータ量での予測を行う場合，大量に入手できる関連したデータを用いることで予測精度が向上することを示した．そこで本章では，Wikipedia から取得したページ閲覧数のデータにより深層学習のモデルの事前学習を行い，学習されたモデルを用いて販売予測を行うことにより，予測精度を向上させることを試みる．また，4 章にて提案した Wikipedia の編集履歴から学習したコンテンツ作品のベクトル表現を用いて，予測対象とするコンテンツ作品についての情報をモデルに加えることで予測精度を向上することを試みる．4 章および 5 章で示した手法に基づく提案手法の全体像を，図 6.1 に示す．本章では，オンラインショッピングサイトにおける実販売データを対象として，販売数の増減を予測する．

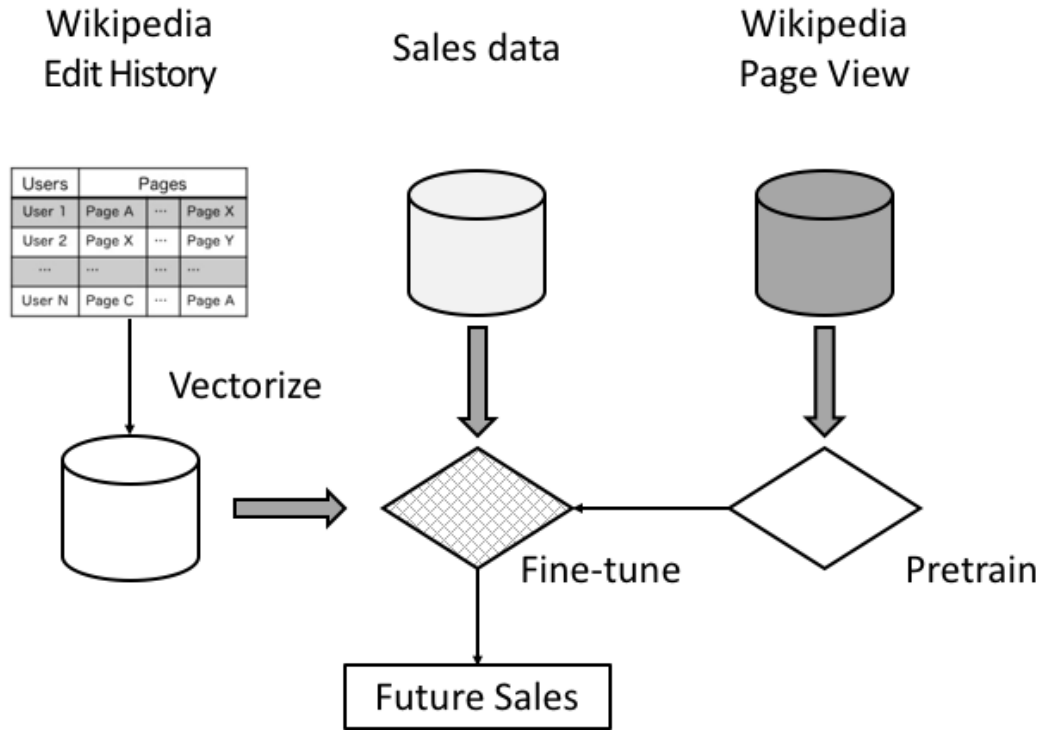


図 6.1 Wikipedia から得た人気情報によるコンテンツ作品の販売予測の概略図

本章の貢献は以下である。

1. Wikipedia 上のコンテンツ作品に関するページの閲覧数を予測する事前学習を行うことで、コンテンツ作品の販売予測の精度が向上することを示した
2. Wikipedia の編集履歴から学習したコンテンツ作品に関するベクトル表現を、Wikipedia のページ閲覧数予測により事前学習したモデルと合わせて用いることで、販売予測の精度がさらに向上することを示した。

本章は、以下のように構成される。まず、6.2 節にて提案手法について説明した後、6.3 節にて実験において用いたデータについて記す。続いて、6.4 節にて、予備実験について述べ、その後、6.5 節にて実験の概要と結果について記す。6.6 節にて実験の結果について考察し、最後

にこの章における結論を述べる.

6.2 提案手法

本章では, Wikipedia からページ閲覧数の予測タスクにより事前学習を行い, 学習したモデルにより販売の予測を行うモデルについて説明する.

6.2.1 事前学習による特徴量の抽出

はじめに, 事前学習による特徴量の抽出について述べる. 画像処理の分野では大規模なデータセットを用いて学習されたモデルを用いることで, 小規模なデータセットのタスクにおける精度を向上できることが知られている [Oquab 14]. そこで本章では, データが大量に入手できる Wikipedia のデータを活用することで, 入手できるデータの少ない販売データに関する予測の精度を向上させることを試みる.

まず, Wikipedia のデータを用いてモデルの事前学習を行う. ここでは, 系列として得られる日次の特徴量を畳み込みニューラルネットワークに与える. Wikipedia におけるコンテンツ作品 i について, ある月 m の t 日目に得られる日次のページ閲覧数を $x_{t,i}^{m,w}$ とする. 次に, ある月 m における一ヶ月のページ閲覧数を系列に並べたデータを $X_i^{m,w}$ とする. Wikipedia において対象とする期間 M に含まれる全ての月 m , 対象とする全てのコンテンツ作品 i について得られる特徴量から, $X^w = \{X_i^{m,w} | \forall i, m \in M\}$ を得る. また, Wikipedia の各月 m におけるコンテンツ作品 i についてのクラスラベル $y_i^{m,w}$ から $y^w = \{y_i^{m,w} | \forall i, m \in M\}$ を得る.

事前学習は, 入手可能なデータの多い Wikipedia におけるデータ (X^w, y^w) を, ランダムに初期化されたパラメータ W_w および b_w をもつ畳み込みニューラルネットワークに与えることで行う. 学習により得られた畳み込みニューラルネットワークのパラメータ W_w および b_w を用いて, 販売データに関する予測モデルの学習を行う.

6.2.2 ユーザ嗜好を考慮した販売予測モデルの学習

本節では, Wikipedia のページ閲覧数を予測するタスクにより事前学習を行なったモデルを用いた販売予測を行うモデルの学習について説明する. Wikipedia のページ閲覧数のデータと同様にコンテンツ作品 i について, ある月 m の t 日目に得られる日次の販売デー

タを $x_{t,i}^{m,s}$, ある月 m における販売データの系列を $X_i^{m,s}$ とする. また, 対象とする期間 M に含まれる全ての月 m , 対象とする全てのコンテンツ作品 i について得られる $X_i^{m,s}$ から $X^s = \{X_i^{m,s} | \forall i, m \in M\}$ を得る. また, クラスラベル $y_i^{m,s}$ から $y^s = \{y_i^{m,s} | \forall i, m \in M\}$ を得る. 事前学習により得られたニューラルネットワークのパラメータ W_w および b_w により初期化されたモデルに (X^s, y^s) を与え, 学習を行う.

ここで, Wikipedia のページ閲覧数および販売スコアのデータは, 個人の行動と紐づいていないため, ユーザの嗜好を考慮できないという問題点がある. そこで本章では, 4章にて用いられたユーザの嗜好に基づくコンテンツ作品のベクトル表現を用いて予測を行う. コンテンツ作品 i についてのユーザの嗜好に基づくベクトル表現を V_i とする. ランダムに初期化したニューラルネットワークに V_i を与え, その出力を事前学習により得られた畳み込みニューラルネットワークの中間層と結合する. 最後に, 事前学習されたモデルからの出力とユーザの嗜好に基づくベクトルを与えたニューラルネットワークの出力を結合し, クラスラベル y_i を予測する.

6.3 データ

本節では, 実験に用いるオンラインショッピングサイトにおける実販売データおよび Wikipedia のデータについて説明する.

6.3.1 実販売データ

実販売データは, オンラインショッピングサイト楽天^{*1}から取得した. 提供されている API を通じて, コミック, テレビゲーム, アニメ (DVD) カテゴリについての日次の販売ランキングを収集した. 販売ランキングのデータは各カテゴリ上位 350 件までについて, 商品名とランキングの項目を対象とし, 2015 年 12 月 1 日から 2017 年 3 月 31 日までの期間のデータを収集した. 続いて, 出現回数上位 166 件の作品について, 取得したデータの商品名を人手により確認し, 日本語版 Wikipedia のページ id と紐づけた.

本章では, Wikipedia のページ id と紐づけた商品の販売ランキングデータを用いて実験を

^{*1} <https://www.rakuten.co.jp/>

行なった。販売ランキングでは、例えば「XXX 1 巻」および「XXX 2 巻」というように、同一の作品について複数の商品がランキングに含まれる場合がある。そのため、ランキングのままでは複数の同一作品が購入されていることを考慮することが難しい。そこで、販売ランキングをもとに販売スコアを算出した。ある商品 i についての販売スコア $Score_i$ は、販売ランキング $Rank_i$ および扱っているデータ全体で最も低いランキングの値 $Rank_{bottom}$ を用いて以下の式により算出した。

$$Score_i = Rank_{bottom} - Rank_i \quad (6.1)$$

この式に従い、作品ごとに日次の販売ランキングから販売スコアを算出し、日次の販売スコアとして実験に用いた。

6.3.2 Wikipedia データ

Wikipedia のデータは、MediaWiki^{*2}から取得した。本章の実験では、2015 年 12 月 15 日から 2017 年 3 月 31 日までの期間における日本語版 Wikipedia における日次および月次のページ閲覧数の情報を用いた。ページ閲覧数は、1 時間ごとに算出されるデータを取得し、24 時間分のページ閲覧数を合計し日次のページ閲覧数とした。続いて、各月の日次ページ閲覧数を合計し当該月の日数で割ることで、月次の平均ページ閲覧数を算出した。また、2017 年までの Wikipedia におけるページ編集履歴を取得した。なお、取得したページ閲覧数の情報は、PathNavi を用いることで関連するページの統計量を足し合わせた。

6.3.3 実験データの前処理

実験に用いるデータの前処理について説明する。まず、モデルに与える入力データに対する前処理について述べる。収集した Wikipedia のページ閲覧数および販売スコアのデータそれぞれについて、14 日間の移動平均を算出した。その後、Wikipedia のページ閲覧数については最大値を 50,000 として 0 から 1 の範囲に、販売スコアのデータは最大の販売スコアに基づいて 0 から 1 の範囲に、それぞれ正規化した。続いて、各月初日から末日までのデータを一次元の系列として扱い、モデルに与える入力データとした。なお、系列長が 31 未満の場合は、系列長が 31 となるようにゼロパディングを行なった。

^{*2} <https://dumps.wikimedia.org/>

表 6.1 各データにおけるクラスバランス

| dataset | Class1 | Class2 | Class3 | Class4 | Class5 |
|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Sales | 365 (38.0%) | 103 (10.7%) | 58 (6.0%) | 78 (8.1%) | 356 (37.1%) |
| Wikipedia | 3114 (15.6%) | 6062 (30.4%) | 4178 (20.9%) | 3744 (18.8%) | 2858 (14.3%) |

次にモデルの出力データの作成について述べる。本章では、月次の販売スコアの変化率をもとにしたクラスラベルを出力とした分類問題として販売予測を行った。出力データは、算出された月次の平均ページ閲覧数または月次の販売スコアに基づくクラスラベルとした。クラスラベルは、当月および翌月の値を用いて変化率を計算し、その変化率に基づいて付与した。それぞれのクラスラベルは、変化率が20%以上減少する場合 (Class1), -20% から-5% の場合 (Class2), -5% から5% の場合 (Class3), 5% から20% の場合 (Class4), 20% 以上の場合の5つの場合 (Class5) とした。表 6.1 に販売スコアおよび Wikipedia のページ閲覧数のデータから得られた1ヶ月後の変化率に基づく各クラスのラベル数を示す。

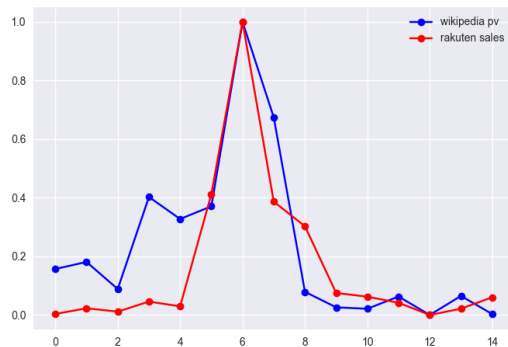
6.4 予備実験：販売ランキングデータとページ閲覧数の関係

本章では、提案手法を用いる際の前提条件である、Wikipedia のページ閲覧数と販売スコアの関連について定性的および定量的な検証を行う。

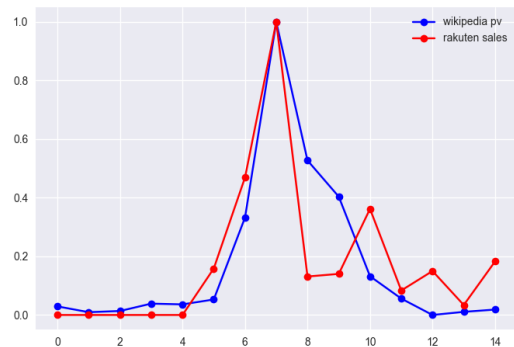
6.4.1 販売ランキングとページ閲覧数の定性的分析

まず、Wikipedia のデータと販売ランキングのデータの定性的な分析を行なった。本章で扱ったコンテンツ作品について、月次の販売スコアおよび月次のページ閲覧数の関係を分析した。対象としたコンテンツ作品の中から4つの作品を選択し、販売スコアおよびページ閲覧数を作品ごとに0から1の範囲に正規化し、変化のパターンを可視化した。可視化した結果を図 6.2 に示す。

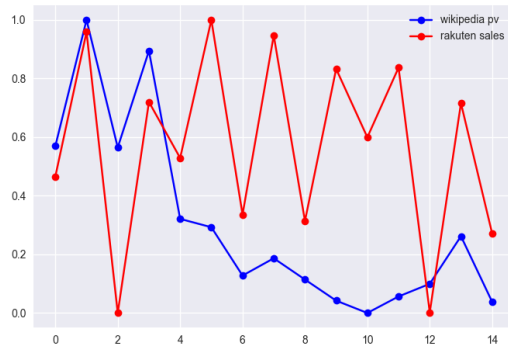
販売スコアと Wikipedia のページ閲覧数の変化のパターンを比較したところ、大きく2つのパターンが見られた。一つは、分析対象としたコンテンツ作品、「Hunter x Hunter」および



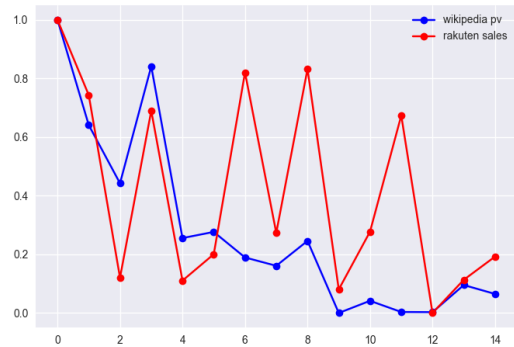
(a) 例 1: Hunter x Hunter



(b) 例 2: ベルセルク



(c) 例 3: フェアリーテイル



(d) 例 4: 監獄学園

図 6.2 販売スコアと Wikipedia のページ閲覧の可視化 (例)

「ベルセルク」において見られた傾向である。このパターンでは、一つの大きなピークが存在し、Wikipedia のページ閲覧数および販売スコアの両方のピークが一致していた。また、ピーク以外の時期についても、両者の間の違いは大きくなかった。もう一つのパターンは、「フェアリーテイル」および「監獄学園」にて見られた傾向である。このパターンでは、販売スコアにおいて 2 または 3 ヶ月の周期性が見られた。Wikipedia のページ閲覧数では、すべてのピークにおける一致ではなかったものの、販売スコアのピークと一致する箇所が複数存在した。

以上の結果から、販売スコアと Wikipedia のページ閲覧数は対象とする作品により変動パターンに差はあるものの、変動の傾向は似ていることが明らかになった。

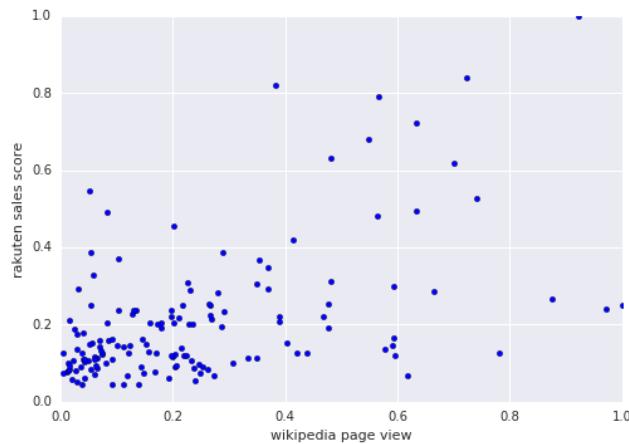


図 6.3 販売スコアと Wikipedia のページ閲覧数の相関

6.4.2 販売ランキングとページ閲覧数の定量的分析

続いて、Wikipedia のページ閲覧数と販売スコアのデータについて定量的な分析を行なった。分析は、予測対象とした Wikipedia のページ閲覧数と販売スコアのデータの間の相関係数を算出することにより行なった。データを収集した全期間にわたるコンテンツ作品ごとのページ閲覧数および販売スコアの値をそれぞれ足し合わせ、合計の販売スコアおよび合計のページ閲覧数を算出した。次に、算出したそれぞれの値から、上位および下位 5 つのコンテンツ作品を外れ値として分析対象から除外した。その後、合計ページ閲覧数と合計販売スコアそれぞれを 0 から 1 の範囲に正規化し、平方根を算出した後、散布図を作成した。作成した散布図を図 6.3 に示す。また、両者の間のピアソンの相関係数を算出したところ、0.52 という値であった。

この結果から Wikipedia のページ閲覧数と販売スコアの間には緩やかな相関関係があることが明らかになった。

6.5 実験：販売予測の精度検証

本節では、販売予測における人気情報の有用性を検証するための実験について説明する。まず、提案手法の有用性を示すため、ベースラインとして販売スコアのデータのみを用いて学習されたモデルを置く。次に、Wikipedia のページ閲覧数を用いた事前学習を行なったモデル、Wikipedia の編集履歴によるベクトル表現を用いたモデルおよび両者を組み合わせたモデル

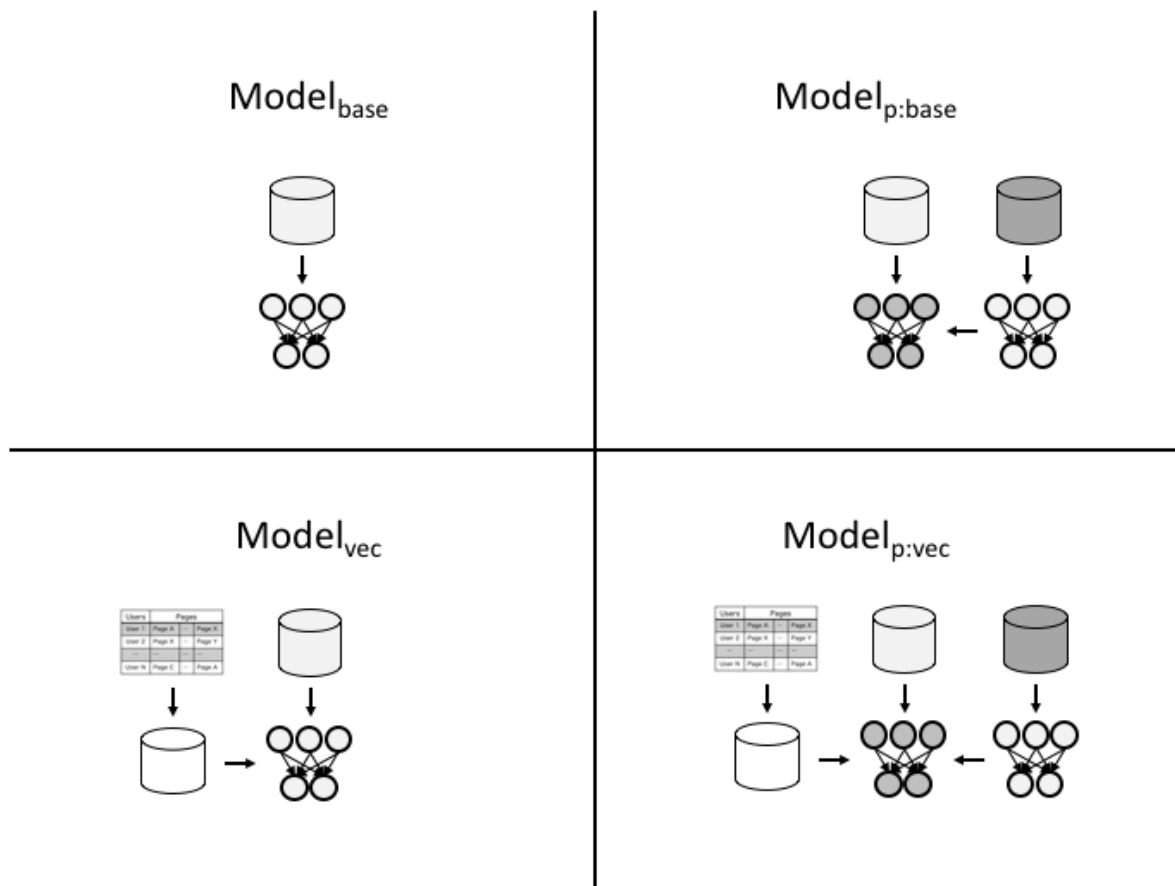


図 6.4 各予測モデルの概要

の3つモデルによる販売予測を行い、ベースラインと比較する。図 6.4 に、本章にて検証したモデルの概要を示す。

6.5.1 ページ閲覧数データにより事前学習したモデルを用いた予測

はじめに、事前学習を行なったモデルを用いて販売スコアの予測を行なった結果について述べる。予測モデルの学習を行う場合、学習に用いるデータの量が多いほど、得られるモデルによる予測精度は高くなると期待される。5章にて、予測対象についてのデータ量が少ない場合、データが大量に入手でき予測対象と関連のあるデータを用いて事前学習を行うことにより、予測精度を向上できることが明らかとなった。そこで本章では、予備実験の結果から販売スコア

表 6.2 事前学習したモデルによる分類精度

| | $Model_{base}$ | $Model_{p:base}$ |
|---------|----------------|------------------|
| exp1 | 0.558 | 0.575 |
| exp2 | 0.538 | 0.588 |
| exp3 | 0.567 | 0.571 |
| exp4 | 0.567 | 0.533 |
| exp5 | 0.525 | 0.579 |
| avg. | 0.551 | 0.569 |
| gain | - | 3.33% |
| p-value | - | 0.183 |

と関連することが明らかになった Wikipedia のページ閲覧数を用いて、深層学習のモデルの事前学習を行い、学習されたモデルを用いて販売スコアの予測を行った。

まず事前学習として、Wikipedia における日次のページ閲覧数のデータを用いて、当月から翌月への平均ページ閲覧数の予測を行なった。事前学習には、販売予測における学習期間と同様の期間である 2016 年 1 月から 9 月までの 9 ヶ月間のデータを用いた。入力データとしては、1 ヶ月分の日次のページ閲覧数を 31 次元の系列データとして、1 次元の畳み込みニューラルネットワークに与えた。最適化は、RMSprop を用いて行い、学習時のエポック数は 1500、早期終了を 10 に設定し学習を行なった。

続いて、事前学習されたモデルを用いて、販売予測を行ない、その精度を検証した。事前学習されたモデルに、日次の販売スコアのデータを与え、当月から翌月への平均ページ閲覧数の変化率の予測を行うことでモデルの再学習を行なった。再学習には、2016 年 1 月から 2016 年 9 月までの 9 ヶ月間のデータを用い、予測精度の検証には 2016 年 10 月から 12 月までの 3 ヶ月間のデータを用いた。入力データとしては、1 ヶ月分の日次の販売スコアを 31 次元の系列データとして、事前学習されたモデルに与えた。再学習における最適化は、事前学習と同様の設定において行なった。ただし、学習率は事前学習の場合の 1/10 とした。

本節で用いたモデルは以下の 2 つである。

$Model_{base}$ 事前学習を行わず、ランダムに初期化を行なった後、販売スコアの予測により学習を行なったモデル

表 6.3 ユーザ嗜好ベクトルを用いたモデルによる分類精度

| | $Model_{base}$ | $Model_{vec}$ |
|---------|----------------|---------------|
| exp1 | 0.558 | 0.583 |
| exp2 | 0.538 | 0.579 |
| exp3 | 0.567 | 0.575 |
| exp4 | 0.567 | 0.504 |
| exp5 | 0.525 | 0.529 |
| avg. | 0.551 | 0.554 |
| gain | - | 0.61% |
| p-value | - | 0.857 |

$Model_{p:base}$ Wikipedia のページ閲覧数予測により事前学習を行なった後、販売スコアの予測により再学習を行なったモデル

これら 2 つのモデルについてそれぞれ 5 回の独立した試行を行い、販売スコアの予測精度を検証した。その結果、表 6.2 に示すように、モデルをランダムに初期化した $Model_{base}$ と比較して Wikipedia のページ閲覧数の予測による事前学習を行った $Model_{p:base}$ は、5 回の試行の平均として 3.33% 予測精度が向上していた。しかしながら、両者の間で t 検定を行なった結果、統計的有意差はみられなかった。

6.5.2 ユーザ嗜好の情報を加えたモデルによる予測

次に、コンテンツ作品についてのユーザの思考を考慮したモデルによる予測精度を検証した。消費者は、自身の好みに基づいて購入するコンテンツ作品を選択すると考えられる。そのため、コンテンツ作品についての販売予測を行う際に、ユーザの嗜好は重要であると考えられる。しかし、Wikipedia のページ閲覧数および販売スコアのデータは、いずれも個別の消費者の行動と紐づいていないため、 $Model_{base}$ および $Model_{p:base}$ では、消費者の嗜好を考慮できていない。一方、4 章において、ユーザの思考を反映したベクトル表現が人気予測において有用であることが示されている。

そこで本章では、ユーザの思考を考慮して販売予測を行うため、4 章の手法を用いた。まず、

表 6.4 事前学習とユーザ嗜好ベクトルを組み合わせたモデルによる分類精度

| | $Model_{base}$ | $Model_{vec,p}$ |
|---------|----------------|-----------------|
| exp1 | 0.558 | 0.600 |
| exp2 | 0.538 | 0.596 |
| exp3 | 0.567 | 0.579 |
| exp4 | 0.567 | 0.596 |
| exp5 | 0.525 | 0.596 |
| avg. | 0.551 | 0.593 |
| gain | - | 7.72% |
| p-value | - | 0.002 |

4章に従い、コンテンツ作品に関する Wikipedia の編集履歴を用いてコンテンツ作品のベクトル表現を学習した。続いて、学習されたベクトル表現を用いて販売予測の精度が向上するかを検証した。

予測に用いるモデルについて説明する。予測モデルへのデータの入力には、2つのニューラルネットワークを用いた。一方には、 $Model_{base}$ で用いたモデルと同様のモデルに日次の販売スコアのデータを与えた。もう一方のニューラルネットワークには、入力データとして与えられるコンテンツ作品に対応するベクトル表現を与えた。続いて、2つのニューラルネットワークの隠れ層の出力を結合し、翌月の販売スコアを予測した。 $Model_{base}$ とコンテンツ作品のベクトル表現を用いたニューラルネットワークを組み合わせた、このモデルを $Model_{vec}$ とした。 $Model_{vec}$ の学習では、2つのニューラルネットワークの学習を同時に行なった。また、モデルの学習条件は $Model_{base}$ と同様とした。

コンテンツ作品に関するベクトル表現を用いた $Model_{vec}$ について、 $Model_{base}$ らと同様に5回の独立した試行を行い、販売スコアの予測精度を検証した。その結果、表 6.3 に示すように、 $Model_{vec}$ による予測精度は $Model_{base}$ と比較して、0.61% という小幅な精度上昇を示した。5回の試行の結果の分散は 1.3×10^{-3} となっており、 $Model_{base}$ の分散 3.5×10^{-4} と比較して、大きな値となっていた。また、 t 検定を行なった結果、統計的有意差はみられなかった。

6.5.3 事前学習したモデルとユーザ嗜好の情報を用いた予測

最後に、Wikipedia のページ閲覧数の予測による事前学習を行なったモデルによる販売スコアの予測において、ユーザの嗜好に基づくコンテンツ作品のベクトル情報を用いた場合の予測精度について検証した。 $Model_{vec}$ と同様の構造のニューラルネットワークについて、販売スコアを入力とする部分のパラメータの初期値に $Model_{p:base}$ と同様モデルを用いた。このモデルを $Model_{vec,p}$ とし、 $Model_{vec}$ と同一の条件で学習を行なった。

5 回の独立した試行を行い、販売スコアの予測精度を検証した。その結果、表 6.4 に示すように、 $Model_{base}$ と比較して、平均の予測精度が 7.72% 上昇した。また t 検定を行なった結果、有意水準 1% で有意な差がみられた。

6.6 考察

本章では、予備実験および実験の章にて得られた結果についての考察を述べる。

6.6.1 販売データとページ閲覧数の関係

まず、販売スコアと Wikipedia のページ閲覧数の関係を分析した結果について述べる。本章では、目的の予測対象である販売に関するデータが十分に入手できないことから、データ量が豊富な Wikipedia の情報を用いて予測モデルの事前学習を行なっている。本提案手法を用いる際の前提条件として、Wikipedia のページ閲覧数と販売に関するデータが関連していることが必要となる。そこで本稿では、まず実際の商品の販売ランキングをもとに算出された販売スコアのデータと Wikipedia における商品に対応するページの閲覧数が関連するかを定性的および定量的に分析した。

はじめに、ページ閲覧数と販売スコアの定性的な分析を行なった。定性的な分析は、予測対象として扱ったコンテンツ作品の中から 4 つの作品を選択し、Wikipedia のページ閲覧数と販売スコアを可視化し、両者の変動パターンを比較することにより行なった。その結果、ページ閲覧数と販売スコアの変動パターンには、大きく二つの場合が存在した。一つ目の場合は、一つの大きなピークが存在し、Wikipedia のページ閲覧数および販売スコアの両方のピークが一致する場合であり、分析対象としたコンテンツ作品「Hunter x Hunter」および「ベルセル

ク」において見られた傾向である。「Hunter x Hunter」は、可視化の対象とした期間に含まれる2016年6月に4年ぶりとなる単行本の新刊が発売されている。そのため、単行本の発売とともに Wikipedia のページ閲覧数と販売スコアが大幅に増加したと考えられる。また、単行本の発売と連載の再開は2016年3月に公表されていたため、Wikipedia のページ閲覧数は販売スコアに先行して増加していたと考えられる。「ベルセルク」は、2016年6月に単行本の新刊が発売されており、また2016年7月よりテレビアニメの放映が行われていた。そのため、2016年6月から販売スコアが増加し、アニメの公開とともにより大きな注目を集め、さらに販売スコアが増加したと考えられる。二つの目の場合は、販売スコアにおいて2または3ヶ月の周期性が見られ、その一部と Wikipedia のページ閲覧数が一致するという場合であり、「フェアリーテイル」および「監獄学園」にて見られた。これら2つのコンテンツ作品はいずれも2から3ヶ月の周期で単行本の新刊を発売している点で共通する。販売スコアは、単行本の新刊の発売に合わせて大きくなり、次の発売月の前月まで低下するという周期性が見られる。一方、Wikipedia のページ閲覧数は周期的に変化しているものの、販売スコアの変動と比較するとその変動幅は小さい傾向にあった。

次に、分析の対象とした全コンテンツ作品を対象として、定量的な分析を行なった。Wikipedia のページ閲覧数および販売スコアのデータを対象とした全期間について足し合わせ、両者の間の相関係数を算出した。ピアソンの相関係数を算出したところ、両者は0.52という値で相関していた。今回、相関係数の算出は販売スコアとページ閲覧数それぞれ全期間の合計値を用いた。これは、短期的な販売スコアを算出した場合、定性分析における「フェアリーテイル」や「監獄学園」において見られるように、単行本の新刊の発売日といった外的要因による変動が大きくなると考えられたためである。相関係数が0.52という値であったことから、Wikipedia のページ閲覧数と販売スコアは緩やかに相関していると言える。

以上の定性分析および定量分析の結果から、Wikipedia のページ閲覧数は実際の販売に関するデータと関連していると考えられ、Wikipedia のページ閲覧数によりモデルの事前学習を行うことで販売予測モデルの精度を向上できる可能性が示唆された。

6.6.2 ページ閲覧数データにより事前学習したモデルを用いた予測

続いて、販売スコアと関連すると考えられる Wikipedia のページ閲覧数のデータを用いた事前学習により、販売予測の精度が向上するかを検証した。実験では、Wikipedia のページ閲覧数の予測による事前学習を行なったモデル ($Model_{p:base}$) および事前学習を行わずパラメータをランダムに初期化したモデル ($Model_{base}$) を比較した。5 回の独立した試行による予測精度の平均を算出した結果、事前学習を行なった $Model_{p:base}$ による予測精度が $Model_{base}$ による予測精度を 3.33% 上回っていた。しかしながら、2 つのモデルによる予測結果に対して t 検定を行った結果、統計的有意差はみられなかった。

$Model_{base}$ と比較した場合の精度の向上は、予測対象である販売スコアと関連のある Wikipedia のページ閲覧数の予測による事前学習にて、入力として与えられる系列データの変動パターンから予測に有用な特徴量を獲得できたためであると考えられる。しかしながら、Wikipedia のページ閲覧数と販売スコアについての定性分析の結果から明らかになったように、販売スコアの変動パターンには、Wikipedia のページ閲覧数とは異なるパターンで変動する場合があった。そのため、Wikipedia のページ閲覧数の予測において獲得できなかったパターンが存在した可能性が考えられる。

6.6.3 編集履歴によるベクトル表現を用いた予測

次に、Wikipedia の編集履歴を用いて学習した、ユーザの嗜好に基づいたコンテンツ作品のベクトル表現を用いて販売スコアの予測を行なった。販売スコアのもとである販売ランキングのデータは、ある商品が売れた量に基づいて算出されているため、個別の消費者の購買行動と結びついていない。また、Wikipedia のページ閲覧数も同様に個別のユーザの行動と直接結びついていない。そのため、販売予測において重要と考えられる個別の消費者の嗜好をモデルに取り入れることができていない。そこで本章では、4 章にて人気予測に有用であることが示されたコンテンツ作品のベクトル表現を用いたモデル ($Model_{vec}$) により、販売予測を行い $Model_{base}$ の結果と比較した。

コンテンツ作品のベクトル表現を用いて販売予測を行い、 $Model_{base}$ による予測と比較した結果、販売予測の精度の改善は 0.61% と限定的であった。これは、販売予測に用いるデータ

数が少なく、販売スコアを入力とするニューラルネットワークの学習が安定しなかったことから、ユーザ嗜好に基づくコンテンツ作品のベクトルを入力とするニューラルネットワークにおいて、予測に有用な素性が抽出できなかつたためと考えられる。

6.6.4 事前学習したモデルと編集履歴によるベクトル表現を用いた予測

最後に、Wikipedia のページ閲覧数の予測による事前学習を行なったモデルによる販売スコアの予測において、ユーザの嗜好に基づくコンテンツ作品のベクトル情報を用いたモデル ($Model_{vec,p}$) による予測精度を検証した。その結果、 $Model_{vec,p}$ による販売スコアの予測精度は、 $Model_{base}$ と比較して 7.72% 向上していた。また、Wikipedia のページ閲覧数の予測による事前学習を行なった $Model_{p:base}$ と比較しても、販売スコアの予測精度は 4.2% 向上していた。これは、Wikipedia のページ閲覧数予測において学習された変動のパターンにより精度が向上したことに加えて、Wikipedia のページ閲覧数および販売スコアのデータに含まれていない、消費者の嗜好に関する情報が加わったためであると考えられる。また、販売スコアを入力とするニューラルネットワークの学習が事前学習により安定した結果、ユーザ嗜好に基づくコンテンツ作品のベクトルを入力とするニューラルネットワークにおいて、予測に有用な素性を抽出することが容易になったためであると考えられる。

6.7 本章の結論

本章では、Wikipedia のデータを用いたコンテンツ作品の販売予測問題に取り組んだ。販売予測のためのモデルを構築する場合、販売に関するデータが多いほど予測精度は向上すると考えられる。しかしながら、販売に関するデータを大量に集めることは難しい。一方で、Wikipedia をはじめとするソーシャルメディアは、多くのユーザによって利用されており、データの収集も容易であることが多い。そこで本章では、Wikipedia から販売予測に有用な情報を抽出し、販売予測の精度を向上することを試みた。まず Wikipedia のページ閲覧数とオンラインショッピングサイトにおける販売データが関連することを確認した。続いて、深層学習の手法を用いて販売予測を行なった。深層学習は、大規模なデータによる事前学習を行うことで、関連する小規模なデータにおける学習の精度が向上することが知られているため、本章においても大量に入手可能な Wikipedia のデータを用いて事前学習を行なった。具体的には、

Wikipedia のページ閲覧数を予測するタスクにより予測モデルの事前学習を行い、学習されたモデルを用いて販売予測を行なった。加えて、先行研究によりユーザの嗜好を反映することが示唆された Wikipedia の編集履歴に基づくコンテンツ作品のベクトル表現を用いた予測を行なった。その結果、事前学習を行ったモデルにてコンテンツ作品のベクトル表現を用いた場合、ベースラインの手法と比較して、販売予測の精度が 7.72% 向上することが明らかになった。この結果から、コンテンツ作品の販売予測において、Wikipedia のページ閲覧数を用いて事前学習を行なったモデルを用いることが有用であることが示唆された。本章では予測対象としてコンテンツ作品を扱ったが、Wikipedia にはコンテンツ作品以外のページも存在するため、コンテンツ作品以外の商品でも同様の手法が適用できることが期待される。

第7章

考察

本章では本研究の成果について整理した上で、本研究の貢献についてまとめる。その上で本研究を通じて得られた成果の限界について論じ、今後の課題について議論する。

7.1 ソーシャルメディアを用いた販売予測における各論の考察

7.1.1 感情分析による人気情報の抽出と自動車の販売予測

3章では、オンラインレビューのデータを用いて代表的な耐久消費財である自動車の販売予測を行なった。オンラインレビューから得られる素性の次元数は、販売数のデータと比較して次元数が多い。そのため予測を行う際には、全ての素性を用いるのではなく、販売予測に有用な素性を選択する必要があった。本研究では、有用な素性を選択する手法として、オンラインレビューの情報を用いた感情分析タスクを行い、タスクを行う上で有用な素性を用いて販売予測を行なった。

提案手法の有用性を検証するため、日本国内で継続的に販売されている28の車種について販売予測を行なった。提案手法とベースラインの手法を比較した結果、提案手法による予測精度が有意に向上していた。一方、単純に感情分析の結果を用いる手法による予測精度はベースラインの手法を下回っていた。また、一般的に感情分析に用いられる極性辞書に含まれる単語を用いた場合の予測精度およびランダムに選択した単語を追加した場合の予測精度も提案手法を下回っていた。

本研究では、商品についての口コミ情報が含まれており、また商品と投稿される内容との対

応関係が明確であることからオンラインレビューを用いて予測を行なった。実験結果から耐久消費財の販売予測においては、口コミ情報の中に含まれる消費者の感情予測に寄与する素性が重要であることが示唆された。このような情報は、ユーザの作成するブログ記事やツイッターのようなマイクロブログからも取得できる可能性がある。ツイッターに投稿されるデータは、数の上では膨大であり、レビューの投稿数が少なく解析できなかった商品についても解析できるようになると考えられる。しかしながら、それらの情報は言及されている対象の商品が必ずしも明確でなく、表記の揺らぎも多く存在すると考えられるため、商品との対応付けの点で困難が生じる。

3章での提案手法は、ユーザによって投稿されたレビューの数が少ない車種には適用できない点や自動車の販売数に大きな影響を与えるモデルチェンジといったイベントを考慮できていないといった限界が存在する。ユーザの投稿数が十分に期待できる商品は、代表的な商品であることが多く、マイナーな商品についての予測は難しいという問題点が存在する。また、新規に発売された商品についてもレビューの情報が十分に集まらないことから予測を行うことが難しいという点が手法の限界としてあげられる。加えて、モデルチェンジといったイベントは、企業のマーケティング戦略により決定されるため予測を行うことが難しい。特にオンラインレビューでは、消費者の購入後の感想が記入されることが多いため、販売に関するイベントを考慮することが難しい。このような問題に対しては、ニュースや企業からのプレスリリース情報を分析し、モデルチェンジといったイベントの情報をモデルに与えることが予測精度を向上する上で必要になると考えられる。

7.1.2 ユーザの嗜好を反映したベクトル表現の獲得と人気予測

4章では、Wikipediaにおけるページの編集履歴からユーザの嗜好を反映したベクトル表現を獲得した。その後、獲得したベクトル表現を用いて、Wikipediaの被リンク数を指標とした人気予測を行なった。獲得したベクトル表現とWikipediaにおける月次の平均ページビュー数、合計編集回数といった情報をニューラルネットワークに与えることで、コンテンツ作品の人気予測を行なった。提案手法をベースラインモデルと比較した結果、短い期間の予測では有意差が認められなかったものの、3ヶ月後の人気予測にて提案手法はベースラインと比較して有意に予測精度が向上していた。また、獲得されたベクトル表現をもとにクラスタリングを行

い、予測精度が向上するクラスと低下するクラスについて分析を行なった。その結果、コンテンツベクトルを用いる場合、作品のジャンルや支持するユーザ層がコンテンツベクトルに強く含まれるほど予測精度は向上することが明らかになった。逆に、複数のメディアにて展開され、支持するユーザ層が幅広くなった場合には予測精度が低下することを示唆する結果も得られた。また、獲得したベクトル表現についての定性的な評価から、ユーザの行動をもとに獲得したベクトル表現は、ジャンルや作者、流行した年代といった情報を含んでいることが明らかになった。加えて、学習されたベクトル表現を二次元に射影した結果、類似すると考えられる作品が近くに存在し、クラスを形成していることが明らかになった。

4章における提案手法は、客観的な記述から成り立つ Wikipedia の情報のみを用いている。そのため、消費者の主観的な意見が記された口コミ情報を用いることができない場合における予測手法として有用である可能性がある。本研究では、編集履歴を用いて各商品についてのベクトル表現を取得したが、編集履歴の情報はユーザが訪れたページの系列やレビューの投稿履歴の系列で置き換えられる可能性がある。ページ遷移のデータは、ページの編集回数やレビューの投稿回数といったデータよりも規模が大きいと考えられる。そのため、ユーザが訪れるページからページに対応する商品のベクトル情報を取得することができれば、レビューにおける言及回数が少ないような商品についても予測できると期待される。

一方、提案手法の限界として、一般のサイトにおいて同様の手法を適用しようとした場合には、ユーザの編集履歴に相当する情報を入手することが困難であることがあげられる。また、Wikipedia における記事の編集やオンラインレビューの投稿といった行動は、全てのユーザが等しく行うのではなく、一部の熱心なユーザによって行われていると考えられる。そのため、このような情報から学習された結果には、バイアスがかかっている可能性が存在する。加えて、今回用いたベクトル化の手法は、長期にわたる系列データを用いており、時期による区別を行っていない。そのため、短期間の変動を捉えることは難しいという点がある。この点については、短期的な変動を捉えることが期待できるページ閲覧数の推移といった情報と組み合わせることで予測精度の向上が期待できる。

7.1.3 事前学習による特徴量抽出と人気予測

5章では、予測対象についてのデータが十分に得られない場合における予測手法を提案した。提案手法では、予測対象と関連し大量に入手できるデータを用いて深層学習のモデルの事前学習を行い、その後、予測対象のデータを用いて再学習を行った。実験では、データが大量に入手できる場合とできない場合の事例として、Wikipediaにおける異なる地域の間でデータを用いた。

予測は、画像処理をはじめとする様々な分野で大きな成果を出している深層学習の手法を用いて行なった。データ量が豊富な地域のデータとデータ量の少ない地域におけるデータを用いて学習する手法を検討した結果、データ量の多い地域のデータを用いて事前学習を行い、その後データ量の少ない地域のデータにより再学習することで予測精度が向上することが示された。また、データのスケールを落とすように加工したデータを用いて事前学習を行なったモデルによる実験結果から、事前学習において用いるデータに適切な加工を行うことで再学習したモデルの精度がさらに向上することを明らかにした。

本研究では、ページ閲覧数およびページの編集回数を入力として予測を行なった。しかし、入力データの形状は必ずしもページ閲覧数のようなデータである必要はない。このことからデータ量の多いデータを用いて事前学習を行い、その後再学習を行うという手法は、オンラインレビューのようなテキストデータにおいても適用可能であると期待される。また、実験の結果から事前学習の際に用いるデータ量が多いほど、予測精度が向上していることが明らかになった。この結果から、複数の地域におけるデータが入手可能な場合、すべての地域におけるデータを用いて事前学習を行い、その後特定の地域におけるデータのみを用いて再学習を行うことで予測精度がさらに向上することが期待できる。加えて、深層学習の研究では、画像とテキストといった異なるモダリティの情報を組み合わせる研究も存在する。ソーシャルメディアにおいても、Twitterにおいて画像付きの投稿がなされたり、写真の撮影・加工・共有が主であるInstagramが流行するなど、テキストと画像の両方の情報をもつサービスが存在する。そのようなソーシャルメディアを分析する際に深層学習のモデルを用いた手法により予測精度が向上する可能性があると考えられる。

一方、深層学習によるモデルでは学習の際に大量のデータが必要になることが知られてい

る。データ量が少ない地域のデータによる再学習を行う際にも、一定以上のデータがない場合は学習が進まない可能性がある。また、事前学習に用いるデータと再学習に用いるデータの分布が異なる場合、再学習の効果は限定的になることが中国語の実験結果から示唆される。

7.1.4 Wikipedia から得た人気情報による販売予測

6章では、4章および5章にて提案し、Wikipediaの指標を用いた人気予測で有用性を明らかにした手法を用いて、実際の販売データの予測を行なった。本研究では販売データとして、オンラインショッピングサービスにおける日次の販売ランキングデータを用い、コンテンツ作品を予測対象とした。

オンラインショッピングサイトにおける販売データでは、合計の販売数のみが公開される。そのため、販売データに個別のユーザの嗜好に基づいた情報は含まれていない。そこで本研究では、4章にて人気予測に用いた、Wikipediaの編集履歴に基づくコンテンツ作品のベクトル表現を用いた。また、入手可能な販売データは少数であったため、5章にて提案した事前学習を行う手法を用いた。事前学習を行う際のデータとしては、販売データと関連することを確認したWikipediaのページ閲覧数のデータを用いた。コンテンツ作品のベクトル表現を用いる手法と事前学習を行う手法を組み合わせる結果、いずれの手法も用いない場合と比較して、予測精度が有意に向上した。

本研究では、販売データと関連するデータとしてWikipediaのページ閲覧数のデータを用いた。事前学習に用いるデータは、大量に入手が可能であり、販売データと関連がある必要がある。このようなデータの例として、検索クエリのログデータを用いた予測も同様に行うことができる。検索クエリのログは、Wikipediaのページ閲覧数と異なり、個別のユーザの行動履歴も合わせて取得可能である。個別のユーザの行動履歴を合わせて考慮することで、より高い精度で予測ができる可能性がある。

一方、予測対象とする商品のカテゴリによっては、Wikipediaのページ閲覧数といった大量に入手できるデータと実際の販売データの関連が小さい場合が考えられる。例えば、自動車の場合は価格が高く購入頻度も低いため、コンテンツ作品と比較すると、提案手法による予測精度が高くない可能性がある。

7.2 全体を通しての議論と今後の課題

本節では、ソーシャルメディアを用いた販売予測全般における限界と課題、今後の可能性について述べる。

7.2.1 ソーシャルメディアを用いた予測を行う上での課題

ソーシャルメディアを用いた販売予測では、ソーシャルメディアに投稿される内容または編集記録といったユーザの活動情報から予測に寄与する特徴量を発見し、予測モデルに組み込むことで予測精度を向上することを試みている。ソーシャルメディアから得られる情報は、アンケート調査をはじめとする現実世界を通じて得られるデータと比較して、広範な地域から大規模に獲得することが可能である。しかしながら、予測を行う上で問題となる点も存在する。本節では、ソーシャルメディアのデータを用いる上で、課題となる点について述べ、手法の限界について考察する。

ソーシャルメディアを利用するユーザの情報

第一の課題として、ソーシャルメディアを利用するユーザに関する情報が得られない点があげられる。アンケート調査と比較して、広範な地域から大規模にデータを収集できる利点が存在する一方、ソーシャルメディアを日常的に利用するユーザに偏りがある点やユーザの居住地といった情報が得られないといった問題がある。

ソーシャルメディアを日常的に利用するユーザの内訳は、年齢や性別、国・地域により大きく異なる。例えば、ソーシャルメディアの一つである LINE は日本において 20 代の 73.0% が利用している一方、60 代では 19.5% と大きな差がある。また KakaoTalk は、全年齢層にわたる日本における利用率が 1.4% であるのに対し、韓国では 75.6% となっており地域間で大きな開きがある [総務 16]。これに加えて、男女間での利用に差がある場合もあることも知られている。Wikipedia における編集者は、女性の割合が増加しているものの、いまだ 70% 以上が男性であることが知られている [Hill 13]。

このようにソーシャルメディアを利用するユーザには偏りがあるため、ソーシャルメディアを用いた販売予測が困難になる場合が考えられる。例えば、ソーシャルメディアを用いた販売

予測を行う上で、予測対象の製品を用いるユーザ層とソーシャルメディアを用いるユーザ層に乖離がある場合、予測は困難になると考えられる。高齢者向けの製品についての情報をソーシャルメディア上から取得することは、アニメや漫画といった若年齢層を対象とする製品についての情報を取得することと比較して難しいと考えられる。

また、個別のユーザについての詳細な情報を得られない場合があることも予測を行う上で問題となる場合がある。現実世界における調査であれば、匿名の調査であったとしても調査を行った日時や地域を記録することが可能である。しかし、ソーシャルメディアを通じて得られるデータは、ユーザが自発的に記入した場合を除いて、居住地をはじめとするユーザの個人情報を得ることは難しい。

このようにユーザが居住する地域といった情報の識別が難しい場合があるため、特定の地域に限定した販売予測といったタスクは困難である。ソーシャルメディア上の情報からユーザの居住地を特定する際に、使用されている言語が有用な場合があるが、これは日本語や韓国語といった、その言語の話者が居住する地域が限定されている場合にしか用いることができない。例えば、英語の場合、全世界的に用いられているためユーザが北米、ヨーロッパ、オーストラリアといった国・地域のいずれから発信しているかを特定することは困難である。また、言語情報からユーザの居住している国を特定することが可能な場合であっても、その国の中でユーザが居住している地域を特定することは難しい。Twitterをはじめとする一部のサービスでは、ユーザの位置情報をプロフィールや投稿内容に含めることができるが、位置情報が含まれる割合は多くない。こういった問題を解消するため、投稿内容から位置情報を推定する方法に関する研究が行われている。将来的には、これらの手法と組み合わせることで、詳細な地域における予測をできるようになる可能性がある。

望ましくないデータの混入

第二の課題として、ソーシャルメディアから得られた情報に望ましくないデータが含まれる場合があげられる。

代表的な事例としてステルスマーケティングがある。ステルスマーケティングとは「人々が宣伝されていると気づかない方法で製品を宣伝する」^{*1}行為とされ、Twitterにおけるツイー

^{*1} “advertising a product in such a way that people are not aware that you are trying to persuade them to buy it” を和訳

トやオンラインレビューサイトの投稿にて行われることがある。また Wikipedia においても、競合する企業の製品について悪意のある情報を書き込むことで、自社の製品を持ち上げる行為に、企業が関与し有罪となった事例も知られている。このようなデータが予測を行う際に用いられるデータに含まれる場合、予測結果にも影響を及ぼすことが考えられる。しかしながら、ステルスマーケティングは人間のユーザにも気づかれにくいように行われるため、投稿される内容だけを見て特定することは困難である。この他にもソーシャルゲームや現実世界でのイベントなどと連動し、ゲームやイベントにおいて報酬を与えることでソーシャルメディアに自社の製品についての投稿を促すようにする場合やアカウントの乗っ取りにより本来のユーザが意図しない情報が投稿・拡散されることもある。ステルスマーケティングが疑われる投稿に対する周囲のユーザの反応といった点や長期間にわたる変動の中で異常値検出を行うことでこの問題に対処できる可能性がある。

予測対象についてのエンティティマッチング

第三の課題として、予測対象とする製品についてのエンティティマッチングの問題がある。本稿にて行なった実験において用いたデータは、オンラインレビューまたは Wikipedia のデータであり、いずれも取得されたデータに対応する製品の特定は容易であった。しかしながら、異なるサービスの間では同一の製品であっても名称が若干異なっていたり、個別ユーザの発信する情報に含まれる製品名は略称である場合や誤字があるなど表記にゆらぎが生じたりしている場合がある。また異なる言語間においても、同一の製品が別の名称を持つ場合がしばしば存在する。

このように異なるソーシャルメディアの間や異なる言語間では、同一の製品に対する名称が異なるためそれらに対応づける作業が必要となる場合がある。本稿の第一研究で扱った自動車についても、オンラインレビューサイトではブランド名と評価対象のモデル名まで含まれる一方、販売台数の統計情報ではブランド名のみでの表記となっていた。また、第三研究で扱ったマルチメディアタイトルの名称についても、オンラインショッピングサイトにおける表記と Wikipedia における表記が異なる場合があったため、エンティティマッチングを行う必要があった。この他にも、「君の名は。」という作品は英語圏においては「Your Name.」、ロシア語圏では「Твоё имя」という作品名であり、それぞれ同一の作品が異なった名称で呼ばれている。この問題を解消するためには、人手による作業が必要となることが多いため、コス

トが大きくなる。特に、異なる言語間でのエンティティマッチングには、言語に関する知識も必要となるためコストがさらに大きくなる。

エンティティマッチングの問題は言語に依存しない表現を介すことで、解消できる可能性がある。異なる言語や表現であってもその表現が指す対象が同一である場合、共通の表現によって異なる表現を紐付けられると考えられる。製品の場合には、実物の製品またはその画像が共通表現になりうる。機械翻訳の分野では、画像を介すことで言語間の翻訳を行なわれている。類似の手法を用いることで言語非依存の表現を介したエンティティマッチングが可能になると考えられる。またエンティティマッチングの問題が解消できた場合、複数のソーシャルメディアを用いた予測が可能になる。ソーシャルメディアごとに利用しているユーザ層や投稿されるデータの傾向に違いがあるため、複数のソーシャルメディアを用いることでより多角的な情報を用いて予測を行うことができる。実際に、保住らの研究 [保住 14] では、Wikipedia 中に含まれる異なる表現をリンク構造をもとに対応づけることで Twitter や Google トレンドにおける情報を組み合わせた予測を行なっている。

予測対象の価格とユーザの所得の影響

第四の課題として、予測対象とする商品の価格とユーザの所得の影響があげられる。ここでは、本稿で予測対象としたコンテンツ作品および自動車の販売予測における商品の価格の影響およびユーザの所得の影響を考察する。

まず、4章から6章で予測対象としたコンテンツ作品の価格が販売数に与える影響について述べる。今回の実験で予測対象としたコンテンツ作品の中心である漫画やアニメの価格は、例えば漫画の単行本であれば400円から700円程度であり、商品の価格としては安価である。そのため、ユーザの所得が予測に与える影響は小さいと考えられる。また商品間の価格の差も小さいことが多いため、商品の価格が予測に与える影響は小さいと考えられる。

一方、3章にて予測対象とした自動車は、一般的に購入される商品の中では、高価な商品である。価格が高い商品は低価格な商品と比較して、消費者の購入意思決定において価格が与える影響が大きくなるため、価格の情報を予測モデルに与える必要があると考えられる。3章の提案手法では、車種ごとに販売予測のモデルを構築しており、また予測対象の車種の前月の販売台数を入力素性に加えている。前月の車種の販売台数は当該車種の価格に影響を受けているため、提案手法において、予測対象の車種の価格情報は関節的にモデルに組み込まれていると

考えられる。しかしながら、税制の変更やキャンペーンによる価格の大きな変動といった要素は考慮できていない。モデルチェンジといったイベントと同様に、これらの外部情報を取り込んだモデルとすることが、予測精度を向上させる点で今後の課題となると考えられる。また、商品の価格が高いため、個別のユーザの所得といった面を考慮することも重要になると考えられる。

本稿にて提案した手法はいずれも、ユーザの所得を考慮することはできていない。ユーザの所得は、特に高額な商品を購入する際に、大きな影響を与えると考えられるため、ユーザの所得を考慮することは重要な課題となる。しかしながら、一般的にユーザがソーシャルメディア上に自身の所得を直接的な形で投稿することは稀であると考えられるため、ユーザの所得情報を抽出することは容易ではない。そのため、間節的な方法でユーザの所得を推測する必要がある。ソーシャルメディアによっては、ユーザのプロフィール情報にユーザの学歴や所属組織およびその組織における役職が記載されている場合があり、この情報から所得を推測できる可能性がある。また、[Preoțiu-Pietro 15]では、プロフィール情報ではなくユーザの投稿内容からユーザの所得を予測している。このようなモデルと組み合わせて用いることでユーザの所得を考慮したモデルを構築できる可能性がある。

学習データに含まれない事象による影響

第五の課題として、学習データに含まれない事象がテスト時に生じることの影響があげられる。具体的には、税制度の変更や自然災害の発生といった起こることが稀な出来事が考えられる。本稿で扱った内容では、3章における自動車のモデルチェンジが該当する。また、4章、5章および6章で扱ったマルチメディアタイトルについても漫画の新刊の発売といったイベントがオンラインショッピングサイトにおける販売ランキングに大きな影響を与えていた。このように起こることは稀であるが販売数に大きな影響を与える事象は、予測を行う上で難しい問題である。

本稿における研究では、販売数に大きな影響を与える事象について、3章ではモデルチェンジを行なった車種を除外することで、4章、5章および6章では長期的な相関を示す間接的な指標を用いることで、それぞれ対処した。しかしながら、実際の予測を行う上で販売に影響を与えうるイベントについてあらかじめ知ることは困難であり、また間接的な指標が必ずしも得られるとは限らない。モデルチェンジといった販売戦略に関わる情報については、ニュースや

企業のプレスリリースといった情報と組み合わせることで予測モデルに組み込める可能性があるが、大規模な地震といった自然現象はそもそも予測を行うことが困難であるため、モデルに組み込むことが難しい。

このような外部要因を考慮することには機械学習の手法の性質上難しい。機械学習の手法は、経験やデータといった過去の有限事象に基づいて予測を行う帰納法に基づいた手法である。そのため、これまでに起きたことのない事象についての予測を行うことは困難である。これは、機械学習の手法を用いる上で本質的な問題であり、販売予測を行う上で注意すべき点である。

7.2.2 ソーシャルメディアを用いる販売予測の適用可能性

本節では、ソーシャルメディアを用いた販売予測の適用可能性について述べる。

予測モデルに用いるデータの種類の拡大

本稿における実験において用いたソーシャルメディアのデータは、オンラインレビューサイトから取得したテキスト情報と Wikipedia 上で記録されたページ閲覧数といった定量的な値であった。最近では画像を投稿することが主な Instagram がユーザ数を伸ばしており、また Youtube やニコニコ動画といった動画共有サービスも多くのユーザを獲得している。このように、ソーシャルメディア上にはテキスト情報や定量的な情報以外にも画像や動画といった情報がユーザから発信されている。これらの情報は、消費の予測を行う上で有用であると考えられるが、これまで画像や動画と形態のデータから情報を抽出することが難しかったため、予測において用いられることは少なかった。

他方、深層学習の分野では、画像中に含まれる物体の検出や音声認識、動画中での行動分析や行動予測といったタスクにおいて大きな成果をあげている。そこで、ソーシャルメディアに投稿される画像や動画に対して、深層学習を適用し有用な素性を発見することで、販売予測に用いることができると考えられる。

IoT デバイスを通じて取得されるデータ

ソーシャルメディアは、それ以前には大規模に収集することが困難であった、消費者個人の製品に対する意見や感想を収集することを可能にした。しかしながら、ソーシャルメディア上

の情報においても言及されることの少ない商品も存在する。ソーシャルメディアにおいてユーザが日常生活を逐一発信することは少ないため、食品のように生活の中で日常的に消費される商品の情報を取得することは困難である。

一方、近年「モノのインターネット」が注目を集めている。IT用語辞典によると、モノのインターネットは「コンピュータなどの情報・通信機器だけでなく、世の中に存在する様々な物体（モノ）に通信機能をインターネットに接続したり相互に通信することにより、自動認識や自動制御、遠隔計測などを行うこと」と定義される。モノのインターネットにおける事例の一つとして、冷蔵庫があげられる。インターネットに接続された冷蔵庫は、ドアの開閉や買い物メモを通じてユーザの生活習慣を学習する。ユーザから得られる買い物メモといった情報からこれまで入手することが難しかった、食品といった日常用品の販売予測の可能性も開ける。このようにモノのインターネットを通じて、日常生活の様々な場面からデータを入手できるようになることが考えられ、得られたデータを用いてより広範な製品についての予測も可能になると考えられる。

データ取得の制限

ソーシャルメディアのデータは消費の予測を行う上で有用であるものの、データを用いる上での制約も存在する。一つの制約として技術的な制約があげられる。本稿で用いた Wikipedia のデータは全てダンプデータとして定期的に公開されるものを収集している。また、オンラインレビューの情報はウェブクロウリングの技術を用いて収集し、オンラインショッピングサイトのデータは API を通じて収集した。サイト側が提供する API を通じたデータの収集およびダンプされたデータの利用は、公式に提供されているものであり収集における問題が生じることは少ない。一方、オンラインレビューといった情報をクロウリングを通じて収集する場合、情報元の講じるクロウリング対策によりデータの収集が困難になる場合がある。

第 8 章

結論

本研究では、ソーシャルメディアのデータを用いて販売予測に関連するタスクを解き、その中で得られた情報を用いて販売予測を行った。

3 章では、オンラインレビューの情報を用いて、代表的な耐久消費財である自動車の販売予測を行なった。収集したオンラインレビューのデータにおける消費者の感情を予測するタスクを解き、感情予測に有用な素性を特定し、販売予測にて用いた。その結果として、オンラインレビューから得られた素性を用いることで販売予測の精度が有意に向上することを示した。

続いて 4 章では、Wikipedia の編集履歴から各記事に対応するマルチメディアのベクトル表現を獲得し、それを用いることでマルチメディアタイトルについての人気予測を行なった。人気予測を行なった結果、学習されたベクトル表現は、Wikipedia の被リンク数を指標とした人気予測タスクにおいて有用であることが明らかになった。

次に 5 章では、予測対象について入手できるデータの量が少ない場合における予測手法を提案した。提案手法では、まず深層学習のモデルを用いて、目的の予測対象に関連し、大量に入手できるデータを用いて事前学習を行なった。続いて、目的の予測対象についてのデータを用いて再学習を行うことで、Wikipedia のページ閲覧数を指標とした人気予測の精度が向上することを確認した。

最後に 6 章にて、4 章および 5 章において Wikipedia の指標を用いた人気予測で有用であることが明らかになった情報を用いて、オンラインショッピングサイトにおける販売データの予測を行なった。その結果、Wikipedia の人気予測において有用である情報を組み合わせて用いることで、販売データの予測精度が向上することが明らかになった。

以上を通じて本研究では、ソーシャルメディアから販売予測に関連するタスクを解くことで、販売予測において有用な情報を獲得し、販売予測を行う手法の提案を行なった。

参考文献

- [Abbasi 08] Abbasi, A., Chen, H., and Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums, *ACM Transactions on Information Systems (TOIS)*, Vol. 26, No. 3, p. 12 (2008)
- [Ahn 14] Ahn, H.-I. and Spangler, W. S.: Sales Prediction with Social Media Analysis, in *2014 Annual SRII Global Conference*, pp. 213–222IEEE (2014)
- [AMA] AMA, : American Marketing Association Dictionary
- [Aramaki 11] Aramaki, E., Maskawa, S., and Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter, in *Proceedings of the conference on empirical methods in natural language processing*, pp. 1568–1576Association for Computational Linguistics (2011)
- [Armstrong 00] Armstrong, J. S., Morwitz, V. G., and Kumar, V.: Sales forecasts for existing consumer products and services: Do purchase intentions contribute to accuracy?, *International Journal of Forecasting*, Vol. 16, No. 3, pp. 383–397 (2000)
- [Asur 10] Asur, S. and Huberman, B. A.: Predicting the future with social media, in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Vol. 1, pp. 492–499IEEE (2010)
- [Baek 14] Baek, H., Oh, S., Yang, H.-D., and Ahn, J.: Chronological Analysis of the Electronic Word-of-Mouth effect of Four Social Media channels on Movie Sales: Comparing Twitter, Yahoo! Movies, YouTube, and Blogs., in *PACIS*, p. 65 (2014)
- [Bao 12] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D.: Omnipedia: bridging the wikipedia language gap, in *Proceedings of the SIGCHI Conference*

-
- on Human Factors in Computing Systems*, pp. 1075–1084ACM (2012)
- [Barkan 16] Barkan, O. and Koenigstein, N.: Item2Vec: Neural Item Embedding for Collaborative Filtering, *arXiv preprint arXiv:1603.04259* (2016)
- [Basuroy 03] Basuroy, S., Chatterjee, S., and Ravid, S. A.: How critical are critical reviews? The box office effects of film critics, star power, and budgets, *Journal of marketing*, Vol. 67, No. 4, pp. 103–117 (2003)
- [Boldt 16] Boldt, L. C., Vinayagamorthy, V., Winder, F., Schnittger, M., Ekran, M., Mukkamala, R. R., Lassen, N. B., Flesch, B., Hussain, A., and Vatrappu, R.: Forecasting Nike’s sales using Facebook data, in *Big Data (Big Data), 2016 IEEE International Conference on*, pp. 2447–2456IEEE (2016)
- [Bollen 11] Bollen, J., Mao, H., and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011)
- [Breiman 01] Breiman, L.: Random forests, *Machine learning*, Vol. 45, No. 1, pp. 5–32 (2001)
- [Brown 53] Brown, G. H.: Brand Loyalty-fact of fiction, *Trademark Rep.*, Vol. 43, p. 251 (1953)
- [Butler 08] Butler, B., Joyce, E., and Pike, J.: Don’t look now, but we’ve created a bureaucracy: the nature and roles of policies and rules in wikipedia, in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1101–1110ACM (2008)
- [Cameron 16] Cameron, M. P., Barrett, P., and Stewardson, B.: Can social media predict election results? evidence from new zealand, *Journal of Political Marketing*, Vol. 15, No. 4, pp. 416–432 (2016)
- [Caruana 98] Caruana, R.: Multitask learning, in *Learning to learn*, pp. 95–133, Springer (1998)
- [Chevalier 06] Chevalier, J. A. and Mayzlin, D.: The effect of word of mouth on sales: Online book reviews, *Journal of marketing research*, Vol. 43, No. 3, pp. 345–354 (2006)
- [Choi 12] Choi, H. and Varian, H.: Predicting the present with Google Trends, *Economic*

- Record*, Vol. 88, No. s1, pp. 2–9 (2012)
- [Culotta 10] Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages, in *Proceedings of the first workshop on social media analytics*, pp. 115–122ACM (2010)
- [Cunningham 56] Cunningham, R. M.: Brand loyalty-what, where, how much, *Harvard Business Review*, Vol. 34, No. 1, pp. 116–128 (1956)
- [Dellarocas 07] Dellarocas, C., Zhang, X. M., and Awad, N. F.: Exploring the value of online product reviews in forecasting sales: The case of motion pictures, *Journal of Interactive marketing*, Vol. 21, No. 4, pp. 23–45 (2007)
- [Dewan 14] Dewan, S. and Ramaprasad, J.: Social media, traditional media, and music sales., *Mis Quarterly*, Vol. 38, No. 1 (2014)
- [Dhar 09] Dhar, V. and Chang, E. A.: Does chatter matter? The impact of user-generated content on music sales, *Journal of Interactive Marketing*, Vol. 23, No. 4, pp. 300–307 (2009)
- [Dichter 60] Dichter, E.: *The strategy of desire*, Transaction Publishers (1960)
- [Ding 15] Ding, X., Liu, T., Duan, J., and Nie, J.-Y.: Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network., in *AAAI*, Vol. 15, pp. 2389–2395 (2015)
- [Duan 08] Duan, W., Gu, B., and Whinston, A. B.: Do online reviews matter?—An empirical investigation of panel data, *Decision support systems*, Vol. 45, No. 4, pp. 1007–1016 (2008)
- [Duric 12] Duric, A. and Song, F.: Feature selection for sentiment analysis based on content and syntax models, *Decision Support Systems*, Vol. 53, No. 4, pp. 704–711 (2012)
- [Eliashberg 97] Eliashberg, J. and Shugan, S. M.: Film critics: Influencers or predictors?, *The Journal of Marketing*, pp. 68–78 (1997)
- [Feldman 13] Feldman, R.: Techniques and applications for sentiment analysis, *Communications of the ACM*, Vol. 56, No. 4, pp. 82–89 (2013)

-
- [Forman 03] Forman, G.: An extensive empirical study of feature selection metrics for text classification, *Journal of machine learning research*, Vol. 3, No. Mar, pp. 1289–1305 (2003)
- [Fourt 60] Fourt, L. A. and Woodlock, J. W.: Early prediction of market success for new grocery products, *The Journal of Marketing*, pp. 31–38 (1960)
- [Franch 13] Franch, F.: (Wisdom of the Crowds) 2: 2010 UK election prediction with social media, *Journal of Information Technology & Politics*, Vol. 10, No. 1, pp. 57–71 (2013)
- [Gamon 04] Gamon, M.: Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, in *Proceedings of the 20th international conference on Computational Linguistics*, p. 841 Association for Computational Linguistics (2004)
- [Ginsberg 09] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L.: Detecting influenza epidemics using search engine query data, *Nature*, Vol. 457, No. 7232, pp. 1012–1014 (2009)
- [Glorot 11] Glorot, X., Bordes, A., and Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach, in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 513–520 (2011)
- [Goel 10] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J.: Predicting consumer behavior with Web search, *Proceedings of the National academy of sciences*, Vol. 107, No. 41, pp. 17486–17490 (2010)
- [Grover 16] Grover, A. and Leskovec, J.: node2vec: Scalable Feature Learning for Networks, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016)
- [Hill 13] Hill, B. M. and Shaw, A.: The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation, *PloS one*, Vol. 8, No. 6, p. e65782 (2013)
- [Hinton 12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Se-

- nior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97 (2012)
- [Huang 08] Huang, A.: Similarity measures for text document clustering, in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, pp. 49–56 (2008)
- [Jungherr 12] Jungherr, A., Jürgens, P., and Schoen, H.: Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welp, im “predicting elections with twitter: What 140 characters reveal about political sentiment” , *Social science computer review*, Vol. 30, No. 2, pp. 229–234 (2012)
- [Kaplan 10] Kaplan, A. M. and Haenlein, M.: Users of the world, unite! The challenges and opportunities of Social Media, *Business horizons*, Vol. 53, No. 1, pp. 59–68 (2010)
- [Khosla 14] Khosla, A., Das Sarma, A., and Hamid, R.: What makes an image popular?, in *Proceedings of the 23rd international conference on World wide web*, pp. 867–876ACM (2014)
- [Kim 17] Kim, H. and Hanssens, D. M.: Advertising and Word-of-Mouth Effects on Pre-launch Consumer Interest and Initial Sales of Experience Products, *Journal of Interactive Marketing*, Vol. 37, pp. 57–74 (2017)
- [Kliegr 08] Kliegr, T., Svátek, V., Chandramouli, K., Nemrava, J., and Izquierdo, E.: Wikipedia as the premiere source for targeted hypernym discovery, *Wikis, Blogs, Bookmarking Tools Mining the Web 2.0*, p. 38 (2008)
- [Korpusik 16] Korpusik, M., Sakaki, S., Chen, F., and Chen, Y.-Y.: Recurrent Neural Networks for Customer Purchase Prediction on Twitter., in *CBRecSys@ RecSys*, pp. 47–50 (2016)
- [Lassen 14] Lassen, N. B., Madsen, R., and Vatrapu, R.: Predicting iphone sales from iphone tweets, in *Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International*, pp. 81–90IEEE (2014)

-
- [Le 14] Le, Q. V. and Mikolov, T.: Distributed Representations of Sentences and Documents., in *ICML*, Vol. 14, pp. 1188–1196 (2014)
- [LeCun 15] LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, *Nature*, Vol. 521, No. 7553, pp. 436–444 (2015)
- [Leskovec 07] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., and Hurst, M.: Patterns of cascading behavior in large blog graphs, in *Proceedings of the 2007 SIAM international conference on data mining*, pp. 551–556SIAM (2007)
- [Litman 83] Litman, B. R.: Predicting success of theatrical movies: An empirical study, *The Journal of Popular Culture*, Vol. 16, No. 4, pp. 159–175 (1983)
- [Liu 06] Liu, Y.: Word of mouth for movies: Its dynamics and impact on box office revenue, *Journal of marketing*, Vol. 70, No. 3, pp. 74–89 (2006)
- [Liu 16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C.: Ssd: Single shot multibox detector, in *European conference on computer vision*, pp. 21–37Springer (2016)
- [Lo 15] Lo, S. L., Chiong, R., and Cornforth, D.: Using support vector machine ensembles for target audience classification on Twitter, *PloS one*, Vol. 10, No. 4, p. e0122855 (2015)
- [Massy 69] Massy, W. F.: Forecasting the demand for new convenience products, *Journal of Marketing Research*, pp. 405–412 (1969)
- [Melo 10] Melo, de G. and Weikum, G.: MENTA: Inducing multilingual taxonomies from Wikipedia, in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1099–1108ACM (2010)
- [Mikolov 13a] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013)
- [Mikolov 13b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)
- [Milne 08] Milne, D. and Witten, I. H.: Learning to link with wikipedia, in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 509–518ACM

- (2008)
- [Mishne 06] Mishne, G., Glance, N. S., et al.: Predicting Movie Sales from Blogger Sentiment., in *AAAI spring symposium: computational approaches to analyzing weblogs*, pp. 155–158 (2006)
- [Mnih 15] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, pp. 529–533 (2015)
- [Morwitz 92] Morwitz, V. G. and Schmittlein, D.: Using segmentation to improve sales forecasts based on purchase intent: Which” intenders” actually buy?, *Journal of marketing research*, pp. 391–405 (1992)
- [Nakatoh 13] Nakatoh, T., Amano, H., and Hirokawa, S.: Prediction of growth rate of operating income using securities reports, in *Advanced Applied Informatics (IIAIAAI), 2013 IIAI International Conference on*, pp. 84–88IEEE (2013)
- [Ni 09] Ni, X., Sun, J.-T., Hu, J., and Chen, Z.: Mining multilingual topics from wikipedia, in *Proceedings of the 18th international conference on World wide web*, pp. 1155–1156ACM (2009)
- [Ni 11] Ni, X., Sun, J.-T., Hu, J., and Chen, Z.: Cross lingual text classification by mining multilingual topics from wikipedia, in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 375–384ACM (2011)
- [Nothman 13] Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R.: Learning multilingual named entity recognition from Wikipedia, *Artificial Intelligence*, Vol. 194, pp. 151–175 (2013)
- [Nov 07] Nov, O.: What motivates wikipedians?, *Communications of the ACM*, Vol. 50, No. 11, pp. 60–64 (2007)
- [O’Connor 10] O’Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A.: From tweets to polls: Linking text sentiment to public opinion time series., *ICWSM*, Vol. 11, No. 122-129, pp. 1–2 (2010)

-
- [Onishi 12] Onishi, H. and Manchanda, P.: Marketing activity, blogging and sales, *International Journal of Research in Marketing*, Vol. 29, No. 3, pp. 221–234 (2012)
- [Oquab 14] Oquab, M., Bottou, L., Laptev, I., and Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724 (2014)
- [Packard 57] Packard, V.: *The Hidden Persuaders*. New York: David McKay Company (1957)
- [Page 99] Page, L., Brin, S., Motwani, R., and Winograd, T.: The PageRank citation ranking: bringing order to the web. (1999)
- [Pang 02] Pang, B., Lee, L., and Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques, in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86 Association for Computational Linguistics (2002)
- [Pedregosa 11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, Vol. 12, No. Oct, pp. 2825–2830 (2011)
- [Potthast 08] Potthast, M., Stein, B., and Anderka, M.: A Wikipedia-based multilingual retrieval model, in *European Conference on Information Retrieval*, pp. 522–530 Springer (2008)
- [Preoțiuc-Pietro 15] Preoțiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., and Altras, N.: Studying user income through language, behaviour and affect in social media, *PloS one*, Vol. 10, No. 9, p. e0138717 (2015)
- [Resnick 94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews, in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186 ACM (1994)
- [Ruiz 12] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A.: Correlat-

- ing financial time series with micro-blogging activity, in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 513–522ACM (2012)
- [Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, in *Proceedings of the 19th international conference on World wide web*, pp. 851–860ACM (2010)
- [Sakaki 12] Sakaki, T., Matsuo, Y., Yanagihara, T., Chandrasiri, N. P., and Nawa, K.: Real-time event extraction for driving information from social sensors, in *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012 IEEE International Conference on*, pp. 221–226IEEE (2012)
- [Shardanand 95] Shardanand, U. and Maes, P.: Social information filtering: algorithms for automating “word of mouth” , in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 210–217ACM Press/Addison-Wesley Publishing Co. (1995)
- [Sharma 12] Sharma, A. and Dey, S.: A comparative study of feature selection and machine learning techniques for sentiment analysis, in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, pp. 1–7ACM (2012)
- [Silver 07] Silver, D. L. and Mercer, R. E.: Sequential inductive transfer for coronary artery disease diagnosis, in *2007 International Joint Conference on Neural Networks*, pp. 2635–2641IEEE (2007)
- [Simonyan 14] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014)
- [Sorg 12] Sorg, P. and Cimiano, P.: Exploiting Wikipedia for cross-lingual and multilingual information retrieval, *Data & Knowledge Engineering*, Vol. 74, pp. 26–45 (2012)
- [Strube 06] Strube, M. and Ponzetto, S. P.: WikiRelate! Computing semantic relatedness using Wikipedia, in *AAAI*, Vol. 6, pp. 1419–1424 (2006)
- [Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to sequence learning with neural networks, in *Advances in neural information processing systems*, pp. 3104–3112 (2014)

-
- [Taboada 11] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M.: Lexicon-based methods for sentiment analysis, *Computational linguistics*, Vol. 37, No. 2, pp. 267–307 (2011)
- [Trusov 09] Trusov, M., Bucklin, R. E., and Pauwels, K.: Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site, *Journal of marketing*, Vol. 73, No. 5, pp. 90–102 (2009)
- [Tsuboi 15] Tsuboi, Y., Jatowt, A., and Tanaka, K.: Product Purchase Prediction Based on Time Series Data Analysis in Social Media, in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on*, Vol. 1, pp. 219–224IEEE (2015)
- [Tuarob 13] Tuarob, S. and Tucker, C. S.: Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data, *ASME Paper No. DETC2013-12661* (2013)
- [Tumasjan 10] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M.: Predicting elections with twitter: What 140 characters reveal about political sentiment., *Icwsn*, Vol. 10, No. 1, pp. 178–185 (2010)
- [Wang 12] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 us presidential election cycle, in *Proceedings of the ACL 2012 System Demonstrations*, pp. 115–120Association for Computational Linguistics (2012)
- [Wang 15] Wang, Y., Li, J., Liu, Q., and Ren, Y.: Prediction of purchase behaviors across heterogeneous social networks, *The Journal of Supercomputing*, Vol. 71, No. 9, pp. 3320–3336 (2015)
- [Welser 11] Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M.: Finding social roles in Wikipedia, in *Proceedings of the 2011 iConference*, pp. 122–129ACM (2011)
- [Wu 06] Wu, Y. and Tseng, B. L.: Important Weblog Identification and Hot Story Summarization., in *AAAI Spring Symposium: Computational Approaches to Analyzing We-*

- blogs*, pp. 221–227 (2006)
- [Wuthrich 98] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., and Zhang, J.: Daily stock market forecast from textual web data, in *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, Vol. 3, pp. 2720–2725IEEE (1998)
- [Xiao 16] Xiao, Y., Wang, Y., Mao, H., and Xiao, Z.: Predicting Restaurant Consumption Level through Social Media Footprints., in *COLING*, pp. 3328–3338 (2016)
- [Yang 97] Yang, Y. and Pedersen, J. O.: A comparative study on feature selection in text categorization, in *Icml*, Vol. 97, pp. 412–420 (1997)
- [Yessenalina 11] Yessenalina, A. and Cardie, C.: Compositional matrix-space models for sentiment analysis, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 172–182Association for Computational Linguistics (2011)
- [Yu 12] Yu, X., Liu, Y., Huang, X., and An, A.: Mining online reviews for predicting sales performance: A case study in the movie domain, *IEEE Transactions on Knowledge and Data engineering*, Vol. 24, No. 4, pp. 720–734 (2012)
- [Zhang 13] Zhang, Y. and Pennacchiotti, M.: Predicting purchase behaviors from social media, in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1521–1532ACM (2013)
- [Zhu 10] Zhu, F. and Zhang, X.: Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics, *Journal of marketing*, Vol. 74, No. 2, pp. 133–148 (2010)
- [経済 16] 経済産業省：コンテンツ産業の現状と今後の発展の方向性, http://www.meti.go.jp/policy/mono_info_service/contents/downloadfiles/shokanjikou.pdf (2016)
- [高村 06] 高村大也, 乾孝司, 奥村学 他：スピンモデルによる単語の感情極性抽出, *情報処理学会論文誌*, Vol. 47, No. 2, pp. 627–637 (2006)
- [榎 14] 榎 剛史：ソーシャルセンサによる観測手法の提案及びその応用に関する研究, 博士論文 (2014)
- [総務 16] 総務省：総務省 情報通信白書, <http://www.soumu.go.jp/johotsusintokei/>

whitepaper/ja/h28/pdf/ (2016)

- [増田 16] 増田 弘道 and 氷川 竜介 and 数土 直志 and 陸川 和男 and 森 祐治 and 長谷川雅弘: アニメ産業レポート 2016 サマリー (日本語版) 1.1, <http://aja.gr.jp/?wpdmdl=991> (2016)
- [那須 15] 那須野薫, 奥山晶二郎, 中西鏡子, 松尾豊 他: Twitter における候補者の選挙地盤に着目した国政選挙の当選者予測, 情報処理学会論文誌, Vol. 56, No. 10, pp. 2044–2053 (2015)
- [保住 14] 保住純, 飯塚修平, 中山浩太郎, 高須正和, 嶋田絵理子, 須賀千鶴, 西山圭太, 松尾豊: Web マイニングを用いたコンテンツ消費トレンド予測システム, 人工知能学会論文誌, Vol. 29, No. 5, pp. 449–459 (2014)

謝辞

本研究を遂行するにあたり，研究室の方々をはじめ，非常に多くの方々からご指導，ご支援をいただきましたことをこの場を借りて心より感謝いたします。

指導教官である松尾 豊特任准教授には，非常にお忙しい中貴重な時間を割いてご指導いただきました。研究の内容に関して直接的にご指導いただいた点のみならず，学内外で活躍される姿から研究のみに留まらず多くのことを学びました。他分野の修士課程からきた自分を快く受け入れてくださり，ご指導ご鞭撻いただきましたことを心より感謝いたします。

本論文の副査である東京大学大学院工学系研究科 和泉 潔教授，武田 史子准教授，鳥海 不二夫准教授，茂木 源人准教授には貴重なお時間を割いて本論文の審査を行なっていただきました。審査を通じていただいた多角的なご指摘やご助言は，自身の研究を深める上で非常に有意義なものでした。この場を借りて，感謝いたします。

松尾研究室の中山 浩太郎特任講師には，研究論文の執筆の際に細かな点までご助言をいただいた他，研究に関わる技術的な面でも数多くのご助言やご支援をいただきました。いただいたご助言やコードレビューをはじめとするご支援なくして研究を遂行することは困難でした。この場を借りて，心より感謝いたします。

また，松尾研究室の皆様方には，研究へのご助言を中心に様々な面でお世話になりました。大澤 昇平特任助教には，研究のプレゼンテーションを中心に細かな点までご指導いただきました。飯塚 修平博士，片岡 泰之氏には，研究論文に対するフィードバックをいただきました。松尾研究室秘書の永本 登代子さん，小泉 恵美子さん，吉田 和美さん，阿部 美和子さん，また退職された中野 佐恵子さん，浪岡 亮子さん，木全 弥栄さんには，書面作成や経理といった事務手続きの面で多大なる支援をいただきました。Alfredo Solano Martinez 氏，Michael Bawiec 氏，Emilio Castillo 氏には研究活動を行う上で欠かすことのできない，サーバをはじ

めとする環境を整えていただきました。曾根岡 侑也氏，大野 峻典氏には，深層学習の実装勉強会の運営をしていただきました。勉強会を通じて得た情報は，自身の研究においても非常に有用なものとなりました。松尾研究室の先輩・後輩の皆様には，様々な研究の内容についての議論など様々な面でお世話になりました。同時期に博士論文に取り組んだ鈴木 雅大氏には，博士論文の執筆に関わる議論や発表に対する助言をいただきました。研究生活が充実したものとなったのも，研究室が居心地の良い，素晴らしい環境だったからだと思います。最後になりましたが，同期入学の岩澤 有祐博士には，研究に留まらず多くの助言をいただきました。博士課程を修了できるのも同期がいてくれたからこそであると思っています。この場を借りて，松尾研究室の皆様に感謝申し上げます。

また，東京大学 社会構想マネジメントを先導するグローバルリーダー養成プログラム (GSDM) には，幅広い分野の先生方から研究に関するご助言やさまざまな機関を訪問する機会をいただき，自身の視野を広げることができました。特に，東京大学大学院新領域創成科学研究科 加納 信吾准教授，City University of Hong Kong, School of Energy and Environment (SEE) 鎗目 雅准教授には，GSDM プログラムにおける副指導教員をお引き受けいただき，研究の推進と GSDM における活動についてご助言をいただきました。この場を借りて，御礼申し上げます。

Stanford 大学の星 岳雄教授には，貴重な留学の機会をいただきました。現地での滞在を通じて，研究に取り組み方や日本における研究環境との違いを肌で感じることができ，大きな刺激を得ることができました。異なる分野の研究に触れることができた点も，自身の研究を別の視点からみるきっかけとなり有益でした。この場を借りて，御礼申し上げます。

最後に，研究を続ける自分を心温かく見守り，様々な形で私を支えてくれた家族全員に感謝します。

発表文献

論文誌（査読あり）

1. 野中尚輝, 中山浩太郎, 松尾豊: オンラインレビューから抽出した消費者の感情を用いた自動車販売予測, 情報処理学会論文誌データベース (TOD), Vol. 10, No. 3, pp.1-10(2017) (採録済み)
2. 野中尚輝, 中山浩太郎, 松尾豊: Wikipedia の編集履歴から学習したベクトル表現によるコンテンツの人気予測, 電子情報通信学会論文誌 (特集号), Vol.J101-D, No.4, pp.-, Apr. 2018. (採録済み)
3. 野中尚輝, 中山浩太郎, 松尾豊: 深層学習によるデータが少ない地域における人気予測, 人工知能学会論文誌 (投稿中)
4. 野中尚輝, 中山浩太郎, 松尾豊: Wikipedia から抽出した特徴量を用いた深層学習によるコンテンツの販売予測, 人工知能学会論文誌 (投稿準備中)

国際学会（査読あり）

1. **Naoki Nonaka**, Kotaro Nakayama, Yutaka Matsuo: “Contents Popularity Prediction by Vector Representation Learned from User Action History.” DATA ANALYTICS 2017. Barcelona, Spain, November 12-16, 2017. (Accepted)

国内口頭発表

1. 野中尚輝, 松尾豊. オンラインレビュー情報の利用による自動車の売上予測手法の提

- 案, 第 29 回人工知能学会全国大会, 北海道, Jun. 2015.
2. 野中尚輝, 中山浩太郎, 松尾豊. RNN を利用したコンテンツ産業の消費トレンド予測, 第 30 回人工知能学会全国大会, 福岡, Jun. 2016.
 3. 野中尚輝, 中山浩太郎, 松尾豊. ユーザの行動履歴から学習したベクトル表現によるコンテンツの人気予測, 第 9 回データ工学と情報マネジメントに関するフォーラム (第 15 回日本データベース学会年次大会), 岐阜, Mar. 2017.
 4. 野中尚輝, 中山浩太郎, 松尾豊. オンラインレビューから抽出した消費者の感情を用いた自動車販売予測, 第 10 回 WebDB Forum2017, 東京, Sep. 2017.
 5. 山内雄太, 村岡恒輝, 野中尚輝, 早矢仕晃章, 今泉允聡, 金善右, 松並研作, Lee H.W., 鎗目雅. 書誌情報を用いた多分野にわたる研究インパクトの評価法の検討, 電子情報通信学会ソサエティ大会 2016, 北海道, Sep. 2016.
 6. 村岡恒輝, 山内雄太, 野中尚輝, 早矢仕晃章, 今泉允聡, 金善右, 松並研作, Lee H.W., 鎗目雅. 多分野にわたる研究インパクトの評価法の検討, 電子情報通信学会 人工知能と知識処理研究会 (AI): データ市場特集 III, 信学技報, Vol.116, No.460, pp.13-17, 岡山, Feb. 2017.
 7. 對間昌宏, 野中尚輝, 竹内啓, 鎌野智樹, 中間雅之, 小池束紗, 貞広幸雄. インテンシティの定量化: ポジションによる相違及び日本と欧州のチームの比較, スポーツアナリティクスと統計科学, 東京, Dec. 2016.
 8. 對間昌宏, 野中尚輝, 竹内 啓, 鎌野智樹, 中間雅之, 小池束紗. インテンシティの定量化による日本サッカーにおける戦術及び練習方針の改善, 情報処理学会 第 79 回全国大会, 名古屋, Mar. 2017.

受賞

1. 第 9 回データ工学と情報マネジメントに関するフォーラム (第 15 回日本データベース学会年次大会), 学生プレゼンテーション賞
2. 第 10 回 WebDB Forum2017, 論文賞 Runners-up
3. 第 10 回 WebDB Forum2017, 学生奨励賞