1H5-1

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

Proposal of method to forecast automobile sales by using online review information

野中　尚輝[*1]　　松尾　豊[*1]
Naoki Nonaka　　Yutaka Matsuo

[*1] 東京大学工学系研究科技術経営戦略学専攻
Graduate School of Engineering, the University of Tokyo

When people buy so-called durable consumer goods, such as cars or electronic goods, they tends to consider for a long time and compare several items, compared to when they purchase non-durable consumer goods such as commodity. In recent years, user review of items posted to online review sites are increasing. Thus, in this paper, we analyzed the data from reviews site and model the set that consumers form. Our experimental results on forecast of automobile sales showed that by modeling items using online reviews increased the accuracy of forecast. Our proposed method is considered to be useful not only to automobiles but also to durable consumer goods.

1. はじめに

少子高齢化の進行とともに，日本の人口は減少しており，それに伴い日本国内における内需は減少することが予測されている。内需の縮小への対応の 1 つとして生産された製品の輸出が挙げられる。東アジア，東南アジア諸国は地理的に日本に近く，また近年の成長が著しいため輸出先として重要となる。輸出を行う際には，現地の市場動向を把握することが重要となる。しかしながら，アジア各国の現地における市場動向は異なるため，個別に市場動向を調査する必要が生じる。

一般に現地における市場調査を行うコストは高いため，複数の国を調査することコストの上昇を意味する。一方，web 上のデータは基本的に場所を問わずアクセス可能である。そこで，web 上のデータを用いて市場動向を把握することができれば，市場調査にかかるコストを大幅に削減することが可能になり，輸出にかかるコストが削減される。その結果，企業の輸出促進と販売戦略の効率化を行うことができると考えられる。こうした背景から本研究では，web 上のデータを用いて，特定の地域における自動車の売上を予測するモデルの構築することを目標とした。まず，日本国内における自動車の売上を web 上のデータを用いて予測するモデルの構築を行った。

経済学において消費財とは，個人や家庭で使用するために購入するものであり，耐久消費財と非耐久消費財に分類される。非耐久消費財は使用期間が短い財あるいは消費されてなくなる財で，衣類や食料品がこれに含まれる。一方，耐久消費財は長期間使用できる財で，これには自動車や大型家電製品などが含まれる。自動車産業は日本国内において重要な位置を占める産業であり，自動車産業の盛衰が国内経済に与える影響は大きい。一般に消費者は非耐久消費財を購入する際には長時間の比較検討を行わないが，耐久消費財を購入する際には長時間の比較検討を行う。[Shocker 91]では，消費者が耐久消費財の購入に際して，入手可能な全商品集合である「ユニバーサル集合」から，その部分集合で名前を知っている「知名集合」を形成した後，目的にかなう商品集合で構成される「考慮集合」を形成するとした。[Shocker 91]や[Roberts 97]において，考慮集合をモデル化することにより，現実をうまく表現できることを示唆している。

本稿では，web 上のレビューサイトの情報から前述の考慮集合をモデル化することで，代表的な耐久消費財である自動車の売上予測を行うモデルの構築を行った。予測にはレビューの情報から得て転移させた素性を用いてアイテムの前処理を行った後，サポートベクトル回帰（SVR）を用いた。本稿で提案する手法は，web 上のレビューから学習した素性を転移して予測を行っており，転移学習の応用であると考えられる。実験の結果，レビュー情報を用いずに売上予測を行った場合と比較して，予測の精度は向上した。

本稿は以下のように構成される。2 章にて関連研究について述べ，本研究の新規性を明確にし，3 章にて手法の評価に利用したデータの概要を記す。その後，4 章にて提案手法の説明，5 章にて実験概要と結果について記述する。6 章にて本研究の応用可能性および限界について述べ，最後に結論と今後の展望を述べる。

2. 関連研究

2.1 非耐久消費財の売上予測

web 上のデータを用いた非耐久消費財の売上予測に関連する研究は数多く存在する。その一つとして，日本製のコンテンツの消費トレンドを予測するシステムとして Asia Trend Map (ATM)が挙げられる[保住 14]。ATM では検索クエリ数，Twitter での言及の数，Wikipedia での言及という 3 種の web 上のデータをもとにサポートベクトル回帰（SVR）にて国内のマンガの売上部数を予測している。この他にも，Twitter におけるツイートの数から映画の売上を予測する研究[Asur 10]などが行われている。

2.2 耐久消費財の売上予測

耐久消費財の売上予測に関する研究としては，web 上のデータを用いない耐久消費財の売上予測に関する研究として[Scott 00]があり，この研究では消費者意識調査に基づいて売上予測が行われている。web 上のデータを用いた研究としては，[Choi 09]が行った特定期間における Google の検索クエリの入力回数を素性として月別の自動車や自動車部品の売上などの表すモデルの設計が挙げられる。また，機械学習を用いた手法としては[Brühl 09]による自動車の売上予測モデルを構築が挙げられる。しかしながら，これら

連絡先: 野中 尚輝，東京大学大学院工学系研究科技術経営戦略学専攻，nonaka@weblab.t.u-tokyo.ac.jp

の研究において予測されているのは自動車全体の売上であり，個別の車名の売上予測は行っていない。耐久消費財の売上予測に関する研究で，個別の商品名ごとの売上予測をweb 上のデータを用いて行った研究は筆者の知る限り存在しない。

2.3 サポートベクトル回帰を行うデータの前処理

SVR にて将来予測を行う研究は様々な分野で行われている。また[Wu 07]や[Lu 09]は，SVR による予測を行う前に用いるデータに対して前処理を施し，予測精度の向上を示している。しかし，これらの研究にて行われている手法はデータ自身の特性を用いて分割しており，外部のデータを用いてアイテムを分割した後に SVR にて予測を行っている本研究とは異なる。

2.4 転移学習との関連

[神嶌 10]によれば，「ある問題を効果的かつ，効率的に解くために，別の関連した問題のデータや学習結果を再利用するのが転移学習である」とある。本稿で提案する手法は，web 上のレビュー情報を用いてアイテムのクラスタリングを行った後，学習した内容を用いて売上予測を行う。この手法は，レビュー情報から素性表現（クラスタ）を学習し，それを新たなタスクである予測問題に転移しているおり，上述の転移学習の枠組みを応用した研究であると考えられる。転移学習に関連する研究では，[Aizenberg 12]がインターネットラジオ局におけるユーザのプレイリスト情報を転移させることで，他のサービスにおけるユーザの行動を推定している。また[Yinqing 14]では，レビューの文章の情報を用いることで，ユーザによるアイテムの採点値を予測している。これらの研究は，web 上の情報から得た素性を転移させている点で本稿での提案手法と類似するが，転移した素性から SVR を用いてアイテムの売上予測を行っているという点においてこれらの研究は提案手法とは異なる。

3. データ概要

本研究では，日本国内における自動車評価サイトの 1 つである Goonet に投稿されたレビューおよび総務省統計局，日本銀行の公表するデータを用いて行った。Goonet は中古車情報誌 Goo の web 版ページであり，中古車市場情報やユーザによる車のレビューが投稿されており，国内では最大規模の車情報サイトである。Goonet 上でユーザは国産車／輸入車を問わず様々な車のページを閲覧し，任意にレビューの投稿および他のユーザの投稿したレビューの閲覧が可能である。

Goonet 上には 2015 年 3 月時点で約 58,000 件のレビューが投稿されている。各ユーザから投稿されるレビューには，レビュー対象となる車名，各種項目（総合，外観，内装，走行性，燃費，乗り心地，装備，価格，満足度）に対するユーザによる採点，項目別タグの付与（利用シーン（計 6 種），オススメ（計 6 種），特徴（計 15 種））および任意のコメントが含まれる。

今回の研究においては，ここに含まれる特徴タグ（計 15 種：カッコいい，荷室，静粛性，視界，ワイルド，落ち着き，広い，加速，安定性，小回り，キュート，安全，操作性，高級感，乗降）を用いて分析を行った。ユーザは 1 つのレビューごとに，上記の 15 種のタグの中から任意の数だけタグを付与することができる。収集したレビューに含まれる車から，20 以上のタグが付与されている車のみを選択し，解析の対象とした。1 つのタグを 1 つのベクトルとみなし，各車を 15 次元のベクトルで表現した。その後，各車についてタグの総数の違いによる影響を除去するため，各次元の値を出現したタグの総数で割り，正規化した。

4. 提案手法

本章では，web 上のデータを用いた耐久消費財の売上予測モデルを構築するための手法の提案を行う。前章までに述べてきたように，消費者は耐久消費財を購入する際に考慮集合を形成することが考えられ，これをモデル化することにより，売上予測の精度が向上すると考えられる。本研究では web 上のレビューサイトに投稿されたレビュー情報を用いることで，考慮集合をモデル化することができると考え，まずレビューを解析し，そこで得た素性を転移して予測問題に用いる。

4.1 Web 上のレビューを用いた売上予測モデルの構築手法

我々は，耐久消費財の購入においてユーザが形成する考慮集合のモデル化を web 上のレビューを用いることで行い，考慮集合に対する売上を予測モデルの構築を行った後，個別の車の売り上げを予測する手法を提案する。この手法は web 上の情報を用いることで耐久消費財の売上予測の精度を向上させる。

実験で予測対象とする耐久消費財は自動車の売上である。自動車は，その利用目的に沿っていくつかの種類に分類できると考えられる。消費者はそれぞれの目的に合致する自動車の集合を形成し，その中から最適なものを購入していると考えられる。本研究では，まずレビューサイトにおけるユーザ投稿のレビューから抽出したタグ情報を素性として，自動車のクラスタリングを行い，消費者が形成する考慮集合のモデル化を行う。クラスタリングには k-means 法または Fuzzy c-means 法のハードおよびソフト 2 つのクラスタリング手法を用いる。最適なクラスタの数は各々の手法について，3 から 9 の間で実験を行い精度の最も高かった数を選択する。クラスタクラスタリングを行った後，得られた結果を素性として予測問題に転移する。その後，各クラスタに対して上述の素性を用いてサポートベクトル回帰にて売上予測を行う。続いて，各クラスタ内で各車種が予測期間の間どの程度のシェアを獲得するのかを，最新月におけるクラスタ内おけるシェアにより算出した。最後に，予測期間における各車種の属するクラスタの売上と各車種シェアの積を算出し，それをもって各車種の売上予測値とする。

なお，売上を予測するために用いる素性は，対象とする月の 12 ヶ月前の数値を用いる。すなわち X 年 1 月の売上を対象とした場合には，X-1 年 1 月の経済指標を素性として用いる。

5. 実験・結果

5.1 実験方法

本章では，提案手法の有効性の評価を行う。Goonet へ投稿されたレビューを用いて自動車のクラスタリングを行った後，4 章で述べたように予測値を算出する。予測値の算出は 12 ヶ月分行い，当該期間の実測値と予測値について二乗平均平方根（RMS）の値を算出する。予測を行った

238 車種について RMS 値を算出し，その合計値を最終的な比較に用いた。比較対象とする手法についても同様に RMS 値を算出した。RMS 値は予測値と実測値のずれであるためこの値が小さいほど優れた手法となる。

まず，4 章で述べた通り，対象サイトに投稿されたレビューを用いて自動車のクラスタリングを行った。続いて，2007 年 1 月から 2012 年 12 月までの自動車の売上データと 3 章で述べた素性を用いてクラスタごとの売上予測モデルの構築を行い，2013 年 1 月から 12 月までの売上の予測を行った。その後，4 章で述べた通りシェアを算出し，2013 年 1 月から 12 月における各車の売上を算出した。それぞれのクラスタリング手法について，最適なクラスタ数を求めるため，クラスタ数 4 から 9 について RMS 値の合計を算出し，比較した（表 1）。なお，アイテム間の距離にはコサイン距離を用いた。

また提案手法の有用性を検証するための比較手法としては，以下を用いた。手法の評価には評価対象とした全車名についての RMS 値の合計および車名ごとの RMS 値の大きさに基づいた手法間での精度の順位の和（Total Rank）を用いた。

表 1: クラスタ数による RMS 値総計の比較

	Number of clusters					
	4	5	6	7	8	9
k-means	142.121	142.080	142.672	142.785	143.786	144.528
c-means	141.442	141.278	141.121	141.409	141.396	141.421

(1) クラスタリングを行わない SVR（Control SVR）

提案手法と同一期間の自動車の売上について，提案手法と同一の素性を用いて SVR にて予測モデルの構築を行う。予測された期間について RMS 値を算出し，他の手法と比較する。

(2) 法律上の車種分類を用いたクラスタリングを行った後の SVR（Car Type）

各車種はその規格により，法律上の分類（乗用車，軽自動車，トラックなど）が行われる。提案手法と同一の期間の自動車の売上について，上述の法律上の売上を用いてクラスタリングを行い，提案手法と同様に SVR にて予測モデルの構築を行い，シェアを算出した後，RMS 値を算出する。

(3) ランダムウォーク（Random Walk）

各車について最新月から 12 ヶ月間の売上をランダムウォークにて予測した後，RMS 値を算出する

(4) 売上が変化しないという仮定（No Change）

各車について，最新月の売上がその後 12 ヶ月間変化しないとして，RMS 値を算出する。

5.2 結果

最適となるクラスタ数については，Fuzzy c-means 法の場合は 6，k-means 法の場合は 5 であった（表 1）。2 つの手法の間で最適となるクラスタ数が異なるのは，ソフトクラスタリングとハードクラスタリングの手法の違いによるものであると考えられる。これらの結果は消費者が形成する自動車についての考慮集合は大きく 5 または 6 になるということを示している。

また，238 車種全てについて，車名ごとの RMS 値の大きさに基づいた精度の順位和では，Fuzzy c-means 法を用いた提案手法（提案 c）を用いた場合が最小となり，続いて k-means 法を用いた提案手法（提案 k）であった（表 2）。同様に RMS 値の合計においても，提案 c を用いた場合が最小となり，続いて提案 k であった（表 3）。順位の和（表 2）および RMS 値の合計（表 3）のいずれの評価基準を採用した場合にも，クラスタリングを用いた場合（提案 c，提案 k）とクラスタリングを用いない手法を用いた場合の結果を比較することで，クラスタリングを行うことにより予測精度が向上していることがわかる。また法律上の登録車種情報を用いたクラスタリングを用いた場合との比較から，レビューを用いることで精度が向上することがわかる。

以上の結果から，web 上のレビューサイトの情報を用いてアイテムのクラスタリングを行う本手法が，自動車の売上予測を行う上で有用であることがわかる。

表 2: 手法ごとの順位の和

	手法					
	提案手法		Control SVR	Car Type	Random Walk	No Change
	k-means	c-means				
Total Rank	530	511	990	844	1003	600

表 3: 手法ごとの RMS 値の総計

	手法					
	提案手法		Control SVR	Car Type	Random Walk	No Change
	k-means	c-means				
Total RMS value	142.08	141.12	265.85	209.92	228.60	146.22

6. 考察

本研究では，web 上のレビューを用いて，消費者の形成する考慮集合をモデル化することで耐久消費財の売上の予測精度の向上を試みた。今回，自動車の売上予測モデルの構築を行ったが，この手法は，家電製品を始めとする耐久消費財の売上予測モデルの構築についても応用可能であると考えられる。本手法においては消費者が購入を考えるアイテムについて考慮集合を形成し，比較検討を行った上で，その中からアイテムを購入するということを前提としている。従って，消費者が考慮集合内のアイテムを複数購入する，または選択を行う際に比較検討を行わない消費財（非耐久消費財）の売上予測には適さないと考えられる。

本手法は，レビューサイトのタグが存在しない場合やそもそもレビューが存在しない場合には，現段階では用いることができない。また，車種によってはクラスタリングを行わずに予測した結果が提案手法と比較して優れている場合も存在した。この結果から，両手法による予測結果を組み合わせて用いることで更に予測精度が向上する可能性が示唆される。

7. まとめ

本稿では，web 上のレビューを用いて耐久消費財の売上予測モデルを構築する手法を提案した。実験の結果，クラスタリングを用いずに売上を予測する手法および法律上の自動車の分類を用いたクラスタリングにより売上予測を行う手法と比較して，提案手法は精度が高かった。本研究で

用いた素性は一般的なデータであり，入手に大きな困難は伴わない。従って，適切なレビューサイトを選択すれば，日本のみならず海外にも応用することが可能であると考えられる。また，転移学習の応用事例として，レビューの情報から学習した素性を転移して，予測に用いることで予測の精度が向上することを示した。今後，6 章で述べたような課題を解決することで，web の情報から海外の市場動向を把握することができるようになり，企業の輸出戦略を決定する際の補助を行うことができるようになるだろう。

8. 謝辞

著者は，文部科学省プログラム「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）」による補助を受けた。この場を借りて，感謝の意を表します。

参考文献

[Shocker 91] Shocker A. D., Moshe Ben-Akiva, Bruno Boccara, Prakash Nedungadi: Consideration set influeces on consumer decision-making and choice: Issues, models, and suggestions, Markting Letters, Springer, Volume2, Issue 3 pp 181-197 (1991)

[Roberts 97] Roberts, John H., James M. Lattin.: Consideration: Review of research and prospects for future insights. *Journal of Marketing Research* pp. 406-410 (1997)

[Asur 10] Asur, S. and Huberman, B. A.: Predicting the Future With Social Media. *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 492-499 (2010)

[Scott 00] Armstrong, J. Scott, Vicki G. Morwitz, V. Kumar: Sales forecasts for existing consumer products and services: Do purchase intentions contribute to accuracy?, International Journal of Forecasting, 16.3 pp. 383-397 (2000)

[Choi 09] Choi, H and Varian, H.: Predicting the Present with Google Trends. *Google Inc. Technical Report* (2009)

[Brühl 09] Berhard Brühl, Marco Hülsmann, Detlef Borscheid, Christoph M. Friedrich, Dirk Reith: A Sales Forecast Model for the German Automobile Market Based on Time Series Analysis and Data Mining Methods, *Advances in Data Mining. Applications and Theoretical Aspects*. Springer Berlin Heidelberg, Volume 5633, pp. 146-160 (2009).

[Aizenberg 12] Aizenberg, N., Koren, Y., Somekh, O.: Build your own music recommender by modeling internet radio streams, In *Proceedings of the 21st international conference on World Wide Web* ACM, pp. 1-10 (2012)

[Yinqing 14] Xu, Yinqing, Wai Lam, Tianyi Lin: Collaborative Filtering Incorporating Review Text and Co-clusters of Hidden User Communities and Item Groups. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, pp. 251-260 (2014)

[Wu 07] Wu, C. L., K. W. Chau, Y. S. Li.: River stage prediction based on a distributed support vector regression, Journal of Hydrology 358.1 pp. 96-111 (2008)

[Lu 09] Lu, Chi-Jie, Tian-Shyug Lee, Chih-Chou Chiu: Financial time series forecasting using independent component analysis and support vector regression, Decision Support Systems 47.2 pp. 115-125 (2009)

[保住 14] 保住 純，飯塚 修平，中山 浩太朗，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)

[神嶌 10] 神嶌 敏弘: 転移学習, 人工知能学会論文誌 Vol. 25, No.4, pp. 572-580 (2010)