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

Consumer Trend Prediction System Using Web Mining

保住 純
Jun Hozumi

東京大学大学院 工学系研究科
Graduate School of Engineering, the University of Tokyo
hozumi@weblab.t.u-tokyo.ac.jp

飯塚 修平
Shuhei Iitsuka

(同 上)
iitsuka@weblab.t.u-tokyo.ac.jp

中山 浩太郎
Kotaro Nakayama

東京大学 知の構造化センター
Center for Knowledge Structuring, the University of Tokyo
nakayama@cks.u-tokyo.ac.jp

高須 正和
Masakazu Takasu

チームラボ株式会社
Team Lab Inc.
takasu@team-lab.com

嶋田 絵理子
Eriko Shimada

経済産業省
Ministry of Economy, Trade and Industry
shimada-eriko@meti.go.jp

須賀 千鶴
Chizuru Suga

(同 上)
suga-chizuru@meti.go.jp

西山 圭太
Keita Nishiyama

(同 上)
nishiyama-keita@meti.go.jp

松尾 豊
Yutaka Matsuo

東京大学大学院 工学系研究科
Graduate School of Engineering, the University of Tokyo
matsuo@weblab.t.u-tokyo.ac.jp

keywords: web mining, prediction, consumption trends

Summary

The Ministry of Economy, Trade and Industry and the corporate world in Japan have recently embraced the 'Cool Japan' policy. They have assisted Japanese content industries in exporting Japanese media contents, such as *manga* (comics), to foreign countries, especially in Asia. However, this overseas expansion has not been successful in producing profits to the expected degree. The main reason for this shortfall is that companies are unable to perceive local consumer consumption trends in those countries easily and at low cost. Consequently, they are unable to select media contents that might be popular in such areas in the future. Herein, we design a consumption trend calculating system that incorporates web mining. It is readily applicable to many countries and contents. Specifically, we use web mining of data elicited from search counts on search engines, tweet counts from Twitter, and article data from Wikipedia to calculate consumption trends based on weekly manga sales data in Japan. Results show that this model can predict consumption trends for six months with high accuracy and that it can be adapted to calculate consumption trends of other contents effectively, such as *anime* (animation) in Japan and manga in France. Moreover, we establish the ASIA TREND MAP web service to inform industries about these calculated consumption trends for Asian countries.

1. はじめに

近年,日本の官公庁や産業界を中心に,日本のマンガやアニメをはじめとしたコンテンツを多く海外に輸出展開することを目指したクール・ジャパン戦略が推進されている。クール・ジャパン戦略の起源は,米国人ジャーナリストの Douglas McGray が 2002 年に投稿した記事 [McGray 02] に対し,官公庁や産業界が反応したことであると言

われている [佐々木 10]。

実際,2011 年に行われたアジア 10 都市における調査^{*1}によると,図 1 に示す通り,アジアの多くの都市において日本のアニメやマンガが高い人気を得ていることが分かる。

*1 博報堂 Global HABIT 調査アジア 10 都市における日・韓・欧米コンテンツ受容性比較
<http://www.hakuhodo.co.jp/uploads/2011/07/20110705.pdf>

図1 アジアにおけるマンガ・アニメ受容性調査の回答割合

しかし、日本製のコンテンツがアジア諸国で評判を得ている一方で、ビジネスの観点から見ると、評判から期待されるほどの収益が十分に得られていないという点において、コンテンツの海外展開は失敗している状況にある [杉山 06]。このような状況に陥った原因の一つに、海外輸出をビジネスとして管理してこなかったという歴史的背景が挙げられる。海外諸国に向けた日本製のアニメの輸出は 1960 年代に始まったが、アニメの知的財産権やロイヤリティの管理がビジネスとして適切に行われていなかったため、その影響が現在も残っていると考えられている [杉山 06]。

一方で、Web マイニングの分野では、これまで本や映画の売れ行きを予測するといった研究やサービスがさまざまな形で行われてきた。例えば、研究としては、Twitter 上でのツイートをもとに映画の興行収入を予測する [Asur 10] や、Yahoo! 上での検索状況をもとにコンテンツの売上状況を予測する [Goel 10] が挙げられる。また、商用サービスとしては、ホットリンク株式会社の提供するクチコミ@係長^{*2}や、NTT コミュニケーションズ株式会社の提供する Biz マーケティング Buzz Finder^{*3} が挙げられる。自社、もしくは自国の製品がどのくらい消費されているか、また人気や潜在需要が上がっているのか、下がっているのかといった消費トレンドを知ることは、企業や国の戦略や政策の策定において重要な意思決定材料となる。

そこで、本研究では、複数の Web 上のデータをもとに、複数の言語圏に関して、統一的な予測モデルを構築する。このモデルを用いて、日本製コンテンツの消費トレンドを予測するシステム ASIA TREND MAP を開発する (図 2)。予測モデルは、日本国内のマンガの売上部数、検索クエリ数、Twitter での言及の数、Wikipedia

図2 消費トレンド予測システムの概念図

での言及という 3 つの Web 上のデータをもとにサポートベクトル回帰 (SVR) を用いて予測するものであり、得られたモデルを日本以外の各国でのトレンドの推定に用いる。それぞれのデータを用いた場合の予測精度を示し、複数のデータをあわせて用いた方が精度が上がることを示す。また、予測モデルが他種のコンテンツや海外のコンテンツにも有効であることを示すために、日本のアニメや、日本のコンテンツについて販売元を横断した信頼性の高い正解データが取得できるフランス語圏のマンガを用いた評価も行う。

本研究の新規性と有用性は、以下のとおりである。まず、検索エンジンと Twitter、Wikipedia という 3 つの Web サービスからなる素性を組み合わせて消費トレンドを予測するモデルを設計している。この 3 つは多くの国や言語圏で一般的に使われているサービスである。これまで単一のデータ (もしくは自社のデータと組み合わせて) 予測する研究はあったが、広く使われている複数の Web 上のデータを使ってモデル化する試みはほとんど行われていない。2 点目に、これまでの研究やサービスでは、基本的に単一の言語圏に対して Web 上のデータの分析が行われていたが、本論文で設計し、精度の検証を行う消費者行動データの予測モデルは、複数の言語圏における適用を前提として設計されている。本研究によって得られる知見は対象とする言語圏に依らないので、従来あまり研究が行われていなかった言語圏の研究においても適用可能である。

本論文は、以下のように構成される。2 章では、本研究と関連性の高いデータ予測の研究や多言語の素性構築に関する研究を紹介する。3 章では、モデルの設計手法について述べる。4 章では、モデルの精度の検証実験について述べる。5 章では、実際に設計したシステムが稼働する Web サービスの概要を述べる。6 章では、本研究の総括と今後の展望を述べる。

2. 関連研究

本章では、本研究に関連する先行研究を紹介する。具体的には、消費者行動によるデータの予測に関する研究

*2 クチコミ@係長 <http://www.hottolink.co.jp/kakaricho>

*3 Biz マーケティング Buzz Finder - NTT コミュニケーションズ <http://www.ntt.com/marketing/bf/>

を、素性となるデータの収集元別に紹介し、続いて、多言語の素性構築に関する研究を紹介する。

2.1 消費者行動によるデータの予測に関する研究

§1 Web 上のデータによる素性を用いない研究

従来行われているデータの予測研究として、企業がもつ過去のデータをもとに予測を行う研究がある。例えば、[Thiesing 97] では、スーパーマーケットの二週間後の売上を、過去の売上状況やキャンペーン活動、祝日の有無をもとに予測している。また、[Giering 08] では、小売チェーン店における各製品の売上を、80 店舗における 600 種類の品物の売上げデータをもとに、店舗別に予測している。しかし、[Thiesing 97] や [Giering 08] のような、企業内のデータを素性に用いる予測は、そのデータが取れない場合には適用できない。本研究は、全てパブリックに取得可能なデータをもとにしているところが特徴的である。

§2 検索エンジンのデータによる素性を用いた研究

検索エンジンにおける検索回数からデータを予測する研究は、これまでに複数行われている。たとえば、[Choi 09] では、Google における特定のクエリでの検索回数を素性とした、月別の自動車や自動車部品の売上、海外の国別渡航者数を表すモデルを設計している。また、[Goel 10] では、Yahoo! における特定クエリでの検索回数を素性として、映画の初週興行収入やテレビゲームの初週売上、音楽の週間ランキングといった、コンテンツの人気を表すデータの予測を行っている。その結果として、検索回数は将来のコンテンツの人気を表すデータ予測に対して効果があるが、[Goel 10] で作られた線形モデルにおいては、予測する期間がヶ月先になると、実際の値と予測値との相関係数が 0.1 ほど下がる傾向にあることを示し、さらに、コンテンツの種類によっても予測の精度に差が生じることを示している。しかし、[Choi 09] や [Goel 10] によるデータの予測実験は、予測する先の範囲が最大 1.5ヶ月後までにとどまっているのに対し、本研究では、最大で 6ヶ月後までの、より期間の長い予測実験を行う。本研究の予測対象が、単発の映画やテレビゲームなどの売上でなく、通常は何年にもわたって継続される漫画の 1 巻から最新刊までを通した人気であるためである。

§3 Twitter のデータによる素性を用いた研究

Twitter から収集されるデータから予測を行う研究も、これまでに多数行われている。たとえば、[Asur 10] では、Twitter 上での上映一週間前における、一時間あたりの映画について言及されたツイート数から、映画の興行収入の予測を行っている。さらに、ツイート内で映画に対して好意的な意見を示しているか否かを、人間が見ることで判定し、その結果を素性として加えることで、Twitter から作られる情報のみであっても、高い精度の予測が行えることを示している。また、[Bollen 11] では、定期

間において収集されたツイートを、文章に込められている感情を分析するツールを用いて分析し、その結果を素性として、ダウ平均株価の推移の予測を行っている。

§4 Wikipedia のデータによる素性を用いた研究

Wikipedia を用いた研究には、そのページ構成や編集履歴から素性を作成し予測を行った研究がある。たとえば、[Kittur 07] では、Wikipedia ページの編集合戦^{*4}が起る要因を、編集回数や編集に携わった人数、ページのサイズやリンク数などを素性とすることで、どの要因が編集合戦を引き起こすかを調べ、ページの編集に携わる人数が多いページでは、編集合戦が起りにくくなることを示している。また、[Loizou 10] では、映画ごとのページに含まれているリンクの数や、登録カテゴリ、ページ本文中の単語を素性とすることで、映画のレビューサイトである IMDb^{*5} における、映画の評価レーティングを予測している。

その他の Web 上のデータを素性に用いた予測実験は、主にブログからデータを収集し、素性として用いているものが多い。たとえば、[Abel 10] では、タイトルが含まれるブログの記事数を素性とすることで、映画の興行収入や、CD アルバムの売上ランキングを予測している。また、[Ishii 12] では、映画について言及されたブログの記事数や広告宣伝費、初日動員数をもとに、映画公開後の一日ごとの興行収入を予測するモデルを作成している。

Wikipedia に関する研究の一部は、Wikipedia に複数の言語版が存在することを前提として行われている。たとえば、[Nastase 10] では、ページ内のリンクやカテゴリ登録状況に注目することで、多言語版の Wikipedia に応用可能な、概念の関連性を表すネットワークの構成手法を示している。また、[Adar 09] では、複数言語間における Wikipedia ページにおける情報量の差を、自動で編集して埋め合わせるシステムを設計している。本研究では、Wikipedia が複数言語版存在することを活かして、複数言語での予測モデルの構築を行っているところが特徴的である。

3. 消費トレンド予測モデル

本章では、本論文の提案手法における前提条件を述べた後に提案手法の詳細を説明し、最後にその評価方法を記す。

3.1 前提条件

本研究で設計し検証を行う予測システムは、世界各国の消費トレンドを統一的に把握することを目的とするため、Web マイニングの対象とする Web サービスや、そこからのデータ収集、素性の作成方法に制約を設ける必要

*4 ページ編集者がお互いに相手の編集結果を繰り返し削除し合い、自分の編集結果を押し通そうとすること。

*5 IMDb - Movies, TV and Celebrities <http://www.imdb.com/>

がある．そこで本研究ではモデルを設計する際に，1 章で議論した以下の条件を設け，その条件下で最適な予測システムを設計することにする．

- 世界中で共通に使用されている Web サービスをマイニングして，知識を得る．
- 素性を作成するとき，品詞や文法構造に注目する自然言語処理技術を用いない．
- コンテンツ特有の知識や性質は用いない．(例: マンガにおける刊行巻数や，作者，主人公や物語の特徴など)

データの収集を行う Web サービスには，検索エンジンと Twitter，そして Wikipedia を採用する．これらを採用した理由は，世界各国で使用されており，これまでの研究でも用いられているためである．なお，検索エンジンと Twitter から収集されるデータは，その API の仕様上，データの取得が行える回数に制限があるため，消費トレンドを算出する際に，あらゆるコンテンツにおいて検索エンジンと Twitter から収集できるデータを使用することはできない．このような背景から，本実験では検索エンジンと Twitter から得られるデータを使用するモデルと，それらを使用せず Wikipedia のみから収集できるデータのみを使用するモデルの，計 2 種類のモデルを設計する．今回設計する予測モデルでは，現在から最長 6ヶ月後までの消費トレンドを予測を行う．予測する期間が伸びることでモデルの精度がどう変化するかを調べるため，消費トレンドの予測を現在 (0ヶ月後) から 6ヶ月後まで 1ヶ月ずつ行う．

また，消費トレンドの元とするデータには，シリーズ別マンガ週別推定売上部数を採用する．マンガを消費トレンドの元とするデータとして採用した理由は，マンガが他のコンテンツに比べて比較的安価で購入しやすいことから販売部数をトレンドとしてとみなすことができ，また，出版社以外の情報収集経路から複数の出版社における売上情報を一括して収集できることである．以下に詳しい内容を説明する．

3.2 データセット

検索エンジンから収集するデータは，検索エンジンにおける，指定したタイトルをクエリとした Google での検索回数とする^{*6}．特定のクエリを 2004 年 1 月から現在までの推定検索回数を国別に集計する．このデータでは検索回数が多いクエリにおいては週単位での推定検索回数が集計されるが，検索回数が極端に少ない場合は月別の推定検索回数だけの集計となることがある．

Twitter から収集するデータは，指定したクエリを含んだツイート数とする．Twitter のツイートは公式 API を使用した場合，最長で二週間前までのツイートしか収集できないので，ツイート数の集計には，日本人ユーザーによるツイートが収集されている，株式会社ホットリンク

の提供するデータベースを使用する．このデータベースはクエリを指定することで，2011 年 6 月から 2012 年 8 月までの任意の期間において，日本人の Twitter ユーザーによってこのクエリを含むツイートがつぶやかれた合計回数を，1 日単位で集計することができる．

Wikipedia から収集するデータは，Wikipedia の指定ページにおける編集履歴データと，ページの内容をパターン処理することで得られる，ページ自体の解析データとする．それらのデータの集計には，Wikipedia から公式に提供されている 2012 年 6 月 3 日時点の Wikipedia 日本語版ダンプデータと 2012 年 6 月 1 日時点の Wikipedia フランス語版ダンプデータに解析を行った，Wikipedia 解析データベースを使用する．また，その解析には著者らが独自に開発したプログラムを用いる．このプログラムはダンプデータ内の各 Wikipedia ページについて，その記法である Wikimedia 記法のパターンに着目することで，各ページから貼られているリンク数やページごとのカテゴリリンク数など公式に提供されているダンプデータには含まれていないデータを，その言語に依らない統一されたアルゴリズムで集計することができる．

消費トレンドの正解データの作成には，オリコン株式会社から提供されるコミックシリーズ別週別売上表を使用する．このデータには 2011 年 6 月 27 日から 2012 年 7 月 1 日までにおけるマンガのシリーズ別週別推定売上部数が，その期間内における合計推定売上部数の上位 300 タイトルについて集計されている．

また，コミックシリーズ別週別売上表に含まれるタイトルについて，MediaWiki API^{*7}と Wikipedia 解析データベースを用いて Wikipedia における該当ページのページ ID を調査する．この処理は Wikipedia 内で同一ページ内で言及されているタイトルを特定し，同一ページ内で言及されているタイトルを以後一つのタイトルとして取り扱うために行われる．例えば『テニスの王子様』シリーズと『新テニスの王子様』シリーズはどちらも Wikipedia において「テニスの王子様」ページに記載されているので，どちらも『テニスの王子様』として扱う．

§1 クエリの前処理

収集した Wikipedia のページ ID から，そのページのタイトルを Wikipedia 解析データベースから自動的に取得する．この方法で収集されるタイトルはデータベースに格納する都合上，本来半角スペースである部分がアンダーバー「_」に変更されている (例:「DEATH_NOTE」)．またページによっては，ページタイトルに丸括弧で括られた情報 (例:「ドリフターズ (漫画)」，「風光る (渡辺多恵子)」) や，普段では言及される際に省かれてしまうようなサブタイトルが付与されていることがある．(例:「AKB49 ~ 恋愛禁止条例 ~」) これらの理由から，収集したページタイトルに以下のパターン処理を行ったものを，検索回数やツイート数を集計する際のクエリとする．

^{*6} Google トレンドを用いる．

^{*7} MediaWiki API <https://www.mediawiki.org/wiki/API/>

表 1 本研究で使用する素性一覧とその説明

素性	記号
月間検索回数 †	$S_0, S_1, S_2, S_3, S_4, S_5$
月間ツイート回数 †	$T_0, T_1, T_2, T_3, T_4, T_5$
月間 Wiki 総編集回数 †	$W_{e(m)}$
月間 Wiki 総編集人数 †	$W_{p(m)}$
月間 Wiki 無登録ユーザ編集回数 †	$W_{a(m)}$
月間 Wiki 1 人あたり編集回数 †	$W_{e/p(m)}$
月間 Wiki 無登録ユーザ編集率 †	$W_{a/e(m)}$
Wiki フォワードリンク数	W_{fl}
Wiki バックワードリンク数	W_{bl}
Wiki バック・フォワードリンク比	$W_{bl/fl}$
Wiki 他言語ページ数	W_{ll}
Wiki 記事ページ長 †	W_{len}
Wiki 記事内セクション数	W_{sec}
Wiki 記事内テンプレート数	W_{tem}
Wiki 登録カテゴリリンク数	W_{cat}

- タイトル中のアンダーバー「_」を半角スペース「 」に置換する。
- 「(」と「)」とに挟まれた部分、2 つの「~」に挟まれた部分、語尾に付与された「シリーズ」、が存在する場合はそれを取り除き、処理後にタイトル語尾に半角スペースが存在する場合はさらにそれを取り除く。ただしこの処理を行った結果クエリの総文字数が 2 文字以下になる場合 (例: 「日常_(漫画)」, 「ラフ_(漫画)」) は、そのクエリがマンガのタイトルとは異なる意味で言及されることが多いことから処理後のクエリを用いることでマンガのタイトルを言及する検索回数やつぶやき回数が適切に取得できない可能性が高まるため、この処理を行わない。

この処理を行うことで「ドリフターズ_(漫画)」と「AKB49 ~恋愛禁止条例~」は、それぞれ「ドリフターズ」と「AKB49」に変換される。

3.3 素性の作成

3.2 節で述べたデータセットから、予測モデルに使用する素性を作成する。作成する素性の一覧を表 1 に記す。素性の語尾に † が付いているものについては、トレンドを予測する月から m ヶ月前の月の素性を X_m とし、さらに、過去 6 ヶ月分の月の平均を表す $\bar{X}$ と、直近 3 ヶ月間の合計値を、さらにその前の 3 ヶ月間の合計値で割った ΔX という素性を作成する。 ΔX は、過去半年間で数値が上昇傾向にあるか減少傾向にあるかを表すものである。また、素性の語尾に † が付いているものは桁数が大きく、そのまま素性として使用すると精度が悪くなってしまう [Goel 10] ので、元の素性に対し自然対数をとったものを素性として使用する。

セクション数はページの章立ての総数、テンプレート数はページに使用されているテンプレートの総数、カテゴリリンク数はそのページに含まれるカテゴリリンクの総数を表す。例えば「ONE PIECE」のページは、「概要」「歴史」「ストーリー」「サバイバルの海 超新星編」等からなる計 46 章から構成され、「継続中の作品」「ネタバレ

図 3 C と Y との対応関係の例

レ」「Infobox animanga/header」等からなる計 49 個のテンプレートが使用され、「ONE PIECE」「週刊少年ジャンプの漫画作品」「漫画作品 わ」「壊れたファイルへのリンクがあるページ」の計 4 つのカテゴリリンクが存在する。このことから「ONE PIECE」のページにおいては、 $W_{sec} = 46$, $W_{tem} = 49$, $W_{cat} = 4$ となる。

3.4 消費トレンドの平滑化

マンガのシリーズ別週別売上部数は、新刊が発売された週に急激に増加し、それ以降は急激に減少するというパルス状の変化をする。このため、マンガの週別売上部数をそのまま消費トレンドとみなすことは適切とは言えず、各週ごとの、その週に新刊が出た場合の売上部数の推定値のほうが、より適切な消費トレンドの表現になると考えられる。

そこで、第 w 週における実際のコミックシリーズ別推定売上部数を C_w から補正済週別マンガ売上部数指標 Y_w を算出し、それをもとに消費トレンドを作成する。まず、前週に比べて C が 5 倍以上になっている週をピーク週とし、ピーク週の集合 P に含める。また、 P の要素数を n とし、期間内のピーク週を早いものから順に $p_1, p_2, \dots, p_n$ と定める。ピーク週は $C = Y$ とし、ピーク週でない週については、 p_1 より前の週の Y は Y_{p_1} を、 p_n より後の週の Y は Y_{p_n} とする。そして、それ以外の週はその週の直前と直後のピーク週における Y を直線で結んだ場合にその週に該当する Y の値とする。この条件を図で示すと図 3 となる。

この方法で求めた補正済週別マンガ売上部数指標 Y_w を 7 で割ることで、1 日のマンガ売上部数指標 Y_d とし、月内における Y_d を合計することで、補正済月別マンガ売上部数指標 Y_M を作成する。

最後に、以上の方法で求められた Y_M を指標として扱いやすいものとするため、その最大値が 100 に近い値となるよう Y_M に 6.75 を掛けた値を消費トレンドとする。

3.5 交差検定とモデルの評価

SVR による予測は、そのパラメータの設定次第で精度が変化する。このため K -fold 交差検定を行って過学習を防ぎながら、SVR のパラメータ C と、RBF カーネル (Radial Basis Function Kernel, 動径基底カーネル) のパラメータである γ を、予測の精度が最も高まるように設定する。

本実験では $C = (2^{-15}, 2^{-13}, 2^{-11}, \dots, 2, 2^3, 2^5)$ と、 $\gamma = (2^{-15}, 2^{-13}, 2^{-11}, \dots, 2, 2^3, 2^5)$ の全ての C と γ の組み合わせ合計 121 組で交差検定を行う。予測精度の評価にはスピアマンの順位相関係数を用い、テスト事例内における順位と、モデルによる予測結果による順位との相関を評価する。その値が最も高くなるものを精度が高いモデルであると判断し、そのときの (C, γ) をモデルに採用する。一般に精度の評価として用いられるピアソンの相関係数ではなくスピアマンの順位相関係数を用いる理由は、実際に消費トレンド予測の結果を実際のビジネスで用いる場合、予測する指標の数値そのものの精度よりトレンドの順位が正しい予測のほうが、予測結果を意思決定の材料として使用する際に、より有益なものとなることである。

3.6 モデルの一般性の検証

設計したモデルが教師事例に対する過学習を防いでいたとしても、それが日本のマンガのみに対して精度が高く、他種のコンテンツに応用できないモデルとなっている可能性がある。そこで、本研究で設計した精度が高いと考えられるモデルが、比較的人気の低いマンガや、アニメ等の他のコンテンツ、そして他国のコンテンツにおいても精度のよい予測が行えるか検証を行う。具体的には、人気コンテンツとそれに該当しない同数の無作為に選出した同種類のコンテンツの消費トレンドを、本実験で作成したモデルを用いて算出し、指標の高い順に並び替えたときに、上位にどれだけ人気コンテンツが含まれるか、その分類精度によってモデルの有用性を検証する。

4. 評価実験

これより、提案手法によって設計されたシステムが消費トレンド予測に対して有効であることを評価実験によって示す。本実験では、予測として使うことができるデータの期間を 6ヶ月間に設定した。そしてその期間内のデータから作られる素性をもとに、0ヶ月後 (現在) から 6ヶ月後までの消費トレンドを SVR を用いて算出した。たとえば、2012 年 1 月から 2012 年 6 月までのデータからは、2012 年 6 月から 12 月までの消費トレンドを月別に算出する。

シリーズ別マンガ週別推定売上部数上位 300 シリーズのタイトルのうち、Wikipedia 解析データベース内にページが存在しなかった 2 タイトルを取り除き、さらに、

Wikipedia 内で同じページに記述されていた 4 タイトルを週別推定売上部数を合計して、計 294 タイトルの消費トレンドを取得した。次に、それらのタイトルについてクエリを作成し、2011 年 1 月～2011 年 11 月における検索エンジンと Twitter と Wikipedia のデータを 3.2 節のデータセットから収集し、表 1 にある素性を作成した。

ただし、Wikipedia 解析データベースからは、Wikipedia の該当ページにおけるフォワードリンク数、バックワードリンク数、他言語ページ数、記事セクション数、登録テンプレート数、カテゴリリンク数を、解析元ダンプデータ作成時点での値しか収集することができないので、それ以前の値を遡って収集するには、その時点において作成された過去のダンプデータを追加で解析する必要がある。そして、その解析には多大な時間とリソースが必要となるので、本実験ではこれらの過去の値を収集することができなかった。そのため、本実験では、これらの素性は Wikipedia 日本語版解析データベースに含まれている 2012 年 6 月 3 日における値を暫定的に 2011 年 1 月から 2011 年 11 月までの値として採用し、そのことが評価に与える影響を抑えるために、これらの素性を用いた予測モデルでは時間経過に伴う平均順位相関の推移は評価せず、6ヶ月後 1 点のみの消費トレンドを算出して評価することにした。

素性ベクトルを作成し、収集した素性に不備があり正確に予測を行うことができないタイトルを取り除き、2011 年 1 月～5 月を起点とした 6ヶ月のデータで予測を行う教師事例 1247 個と、2011 年 6 月から 11 月までのデータで予測を行うテスト事例 253 個を作成した。

さらに、人気コンテンツの一覧が得られた日本のマンガ、日本のテレビアニメ、フランスのマンガについて、予測モデルを適用し、一般性の検証を行った。人気コンテンツの基準は、日本のマンガは、オリコン公式ホームページより 2012 年 11 月の週別コミック売り上げランキング TOP5 に含まれているコミック計 15 タイトル、日本のアニメは、2012 年 12 月週別アニメ DVD 売り上げランキング TOP5 に含まれる TV アニメ計 8 タイトル、フランスのマンガは、コミック情報を提供するフランスの団体である ACBD が 2012 年に発表したレポート^{*8}に掲載のある、2012 年にフランスで最も多く売れたマンガ上位 10 タイトルのうち、Wikipedia 解析データベースからデータが収集できたものとした。また、比較を行うその他のコンテンツについては、日本のマンガは Wikipedia 日本語版の「日本の漫画作品一覧」ページに含まれるマンガ、日本のテレビアニメは「日本のテレビアニメ作品一覧」ページに含まれるテレビアニメ、フランスのマンガは「日本の漫画作品一覧」ページに含まれるマンガのうち、フランス語版 Wikipedia に同マンガのページヘリンクが貼られているマンガから、それぞれのコンテンツにおける人気コンテンツと同数のタイトルを無作為に選出

*8 Bilan 2012 <http://www.acbd.fr/bilan/bilan-2012.html>

図4 モデル S1-S5 の平均順位相関の推移

表2 検索回数のみを用いた予測モデル

モデル	使用素性
S1	$S_0, S_1, S_2, S_3, S_4, S_5$
S2	S_0, S_5
S3	S_0, S_1
S4	$\bar{S}, \Delta S$
S5	$S_0, S_1, S_2, S_3, S_4, S_5, \bar{S}, \Delta S$

した．これらの選ばれたコンテンツについて，Wikipedia 解析データベースから 2011 年 12 月から 2012 年 5 月までのデータを収集して素性を作成し，それらを用いて人気コンテンツ分類精度の検証を行った．

4.1 複数の Web サービスによる素性を用いた予測モデル

まず，検索回数とツイート回数，Wikipedia のページ編集回数から作られる素性の，どれか一種類のみを用いたモデルを設計し，精度の検証を行った．検証に用いたモデルには，期間内のすべての月別素性を導入したモデル (S1, T1, W1) と，[Choi 09] の実験を参考に作成した，直近の月と最も離れた月の素性を使用したモデル (S2, T2, W2)，[Goel 10] の実験を参考に作成した，直近 2ヶ月分の指標を使用したモデル (S3, T3, W3)，期間全体の傾向を示す素性を用いて作ったモデル (S4, T4, W4)，そして，データから作成した全ての素性を使用したモデル (S5, T5, W5) を採用した．

実験の結果，検索回数から作られた素性を用いた予測モデルでは，1ヶ月後まではモデル S4，それ以降はモデル S2 がもっとも高い精度を示した．また，ツイート回数から作られた素性を用いた予測モデルではモデル T3 が，Wikipedia ページ編集回数から作られた素性を用いた予測モデルではモデル W2 が，すべての予測期間において最もよい精度を出していた (表 2, 表 4, 表 6, 図 4, 図 5, 図 6) ．

図5 モデル T1-T5 の平均順位相関の推移

表3 ツイート回数のみを用いた予測モデル

モデル	使用素性
T1	$T_0, T_1, T_2, T_3, T_4, T_5$
T2	T_0, T_5
T3	T_0, T_1
T4	$\bar{T}, \Delta T$
T5	$T_0, T_1, T_2, T_3, T_4, T_5, \bar{T}, \Delta T$

図6 モデル W1-W5 の平均順位相関の推移

表4 Wiki ページ編集回数のみを用いた予測モデル

モデル	使用素性
W1	$W_{e(0)}, W_{e(1)}, W_{e(2)}, W_{e(3)}, W_{e(4)}, W_{e(5)}$
W2	$W_{e(0)}, W_{e(5)}$
W3	$W_{e(0)}, W_{e(1)}$
W4	$W_e, \Delta W_e$
W5	$W_{e(0)}, W_{e(1)}, W_{e(2)}, W_{e(3)}, W_{e(4)}, W_{e(5)}, W_e, \Delta W_e$

図7 モデル C1-C4 の平均順位相関の推移

図8 モデル E1-E5 の平均順位相関の推移

表5 複数サービスのマイニング結果を組み合わせた予測モデル

モデル	使用素性
C1	S_0, S_5, T_0, T_1
C2	$S_0, S_5, W_{e(0)}, W_{e(5)}$
C3	$T_0, T_1, W_{e(0)}, W_{e(5)}$
C4	$S_0, S_5, T_0, T_1, W_{e(0)}, W_{e(5)}$

以上の実験結果を踏まえ、複数の Web サービスから得られるデータの素性を組み合わせてモデルを作成し、その精度を検証した(表5, 図7). 検索回数による素性はモデル S2 で用いた S_0 と S_5 , ツイート回数による素性はモデル T3 で用いた T_0 と T_1 , Wikipedia 編集回数による素性はモデル W2 で用いた $W_{e(0)}$ と $W_{e(5)}$ を採用し、それらを組み合わせて素性として用いた.

作成したモデルの中で最も精度が高くなったモデルは、全ての種類の素性を用いたモデル C4 であった.

4.2 Wikipedia のデータによる素性のみを用いた予測モデル

次に、Wikipedia から得られるデータのみを用いた予測モデルの設計と検証を行う. 先述の実験結果を踏まえ、Wikipedia 編集回数のみを用いたモデルで最も精度の高いモデル W3 に用いた素性 W_0, W_5 を用いて、ページの月別の編集回数 (E1=W1), 編集人数 (E2), 無登録ユーザ編集数 (E3), 一人あたり編集回数 (E4), 無登録ユーザ編集率 (E5) を素性として用いた予測モデルを作成し、モデルの精度検証を行った(表6, 図8).

実験の結果、編集情報から作られる素性のみを用いた予測モデルでは、編集回数をモデル E1(W1) が全ての期間で最も高い精度を示した. また、無登録ユーザに関する素性を用いたモデル E3 と E5 は、他のモデルと比べて精度が低かった.

さらに、Wikipedia 編集履歴情報による素性を用いたモデル E1 ~ E5 の中で最も精度が良かったモデル E1 に、Wikipedia のページ情報から作られる素性を1つ加えたモデルを設計し、精度を検証した(表7). ただし、4章で述べたとおり、これらのモデルにおいて E1 から追加

表6 Wiki ページ編集情報のみを用いた予測モデル

モデル	使用素性
E1	$W_{e(0)}, W_{e(5)}$
E2	$W_{p(0)}, W_{p(5)}$
E3	$W_{a(0)}, W_{a(5)}$
E4	$W_{e/p(0)}, W_{e/p(5)}$
E5	$W_{a/e(0)}, W_{a/e(5)}$

するページ情報による素性は、全ての月において同じ値を暫定的に採用しているため、各モデルにおける現在から6ヶ月後までの予測精度の変化は、その素性を追加する前のモデル E1 の予測精度の変化によるものと考えられる. よって、この組成を用いたモデルについては6ヶ月後の消費トレンドの予測のみを行い、その精度を比較することにした.

実験の結果、ページのリンク数、被リンク数をモデルに素性として追加することで、精度が大きく向上することが分かった.

図7の実験結果から、素性を加えるほど精度が向上することが考えられるので、モデル E1 にリンクに関する素性2つを加えたモデル N9 と、テンプレート数とカテゴリリンク数の素性を加えたモデル N10 を作成した. それら両方加えたモデル N11 と、全てのページ情報による素性を加えたモデル N12 を設計し、精度を検証した(表8).

その結果、モデルは素性の追加に応じて精度が向上し、平均順位相関係数は最大0.93に達した. さらに、最大の精度を示したモデル N12 に、4.1節で最大の精度を示したモデル C4 で用いた検索回数からなる素性 S_0, S_5 とツイート回数からなる素性 T_0, T_1 を加えたモデル A1 を作成し、同様の精度の検証を行った. その結果、モデル A1 はモデル N12 を上回る精度を示した. この結果は4.1節で示した複数の Web サービスによる素性を用いることでモデルの精度が向上するという結果に合致する.

表7 Wikipedia 解析データによる素性を用いた 6ヶ月後の消費トレンド予測モデルの精度

モデル	使用素性	平均順位相関
N1	$W_{e(0)}, W_{e(5)}, W_{fl}$	0.6586
N2	$W_{e(0)}, W_{e(5)}, W_{bl}$	0.5498
N3	$W_{e(0)}, W_{e(5)}, W_{bl/fl}$	0.4080
N4	$W_{e(0)}, W_{e(5)}, W_{il}$	0.4161
N5	$W_{e(0)}, W_{e(5)}, W_{len}$	0.3970
N6	$W_{e(0)}, W_{e(5)}, W_{sec}$	0.4189
N7	$W_{e(0)}, W_{e(5)}, W_{tem}$	0.4671
N8	$W_{e(0)}, W_{e(5)}, W_{cat}$	0.4192

表8 Wikipedia 解析データによる素性を組み合わせた 6ヶ月後の消費トレンド予測モデルの精度

モデル	使用素性	平均順位相関
N9	$W_{e(0)}, W_{e(5)}, W_{fl}, W_{bl}$	0.8523
N10	$W_{e(0)}, W_{e(5)}, W_{tem}, W_{cat}$	0.5157
N11	$W_{e(0)}, W_{e(5)}, W_{fl}, W_{bl}, W_{tem}, W_{cat}$	0.9276
N12	$W_{e(0)}, W_{e(5)}, W_{fl}, W_{bl}, W_{fl/bl}, W_{il}, W_{len}, W_{sec}, W_{tem}, W_{cat}$	0.9380
A1	$S_0, S_5, T_0, T_1, W_{e(0)}, W_{e(5)}, W_{fl}, W_{bl}, W_{fl/bl}, W_{il}, W_{len}, W_{sec}, W_{tem}, W_{cat}$	0.9444

4.3 他コンテンツにおける消費トレンド分類精度の検証

本実験では、3.2 節にて用意したデータセットでは日本のマンガ以外のコンテンツについては検索エンジンと Twitter から検証を行うためのデータを API 制限やデータが存在しないなどの理由で収集できないため、Wikipedia から収集されたデータのみを用いたモデルのみ検証を行った。具体的には、平均順位相関が 0.5 以上であったモデル N1, N2, N9~N12 を検証した (表 9)。

各コンテンツの種類ごとに、4 章冒頭で定めた人気コンテンツとそれと同数の無作為選出した他のコンテンツのトレンドを算出し、そのトレンドが大きいコンテンツ上位 50 % に人気コンテンツがどれだけ含まれるかを数えることでモデルの人気コンテンツ分類精度を調べた。表 9 において、分母は総人気コンテンツ数、分子は算出されたトレンドが大きいコンテンツ上位 50 % に含まれた人気コンテンツ数を示している。

無作為にコンテンツを選出した場合の分類精度は 50% になるが、検証に用いた予測モデルでは、多くのコンテンツにおいて 50% を上回る分類精度を示し、日本とフランスのマンガはいずれのモデルでも 50% を超えた。そして、モデルの精度は日本のマンガトレンド予測における平均順位相関の高さと一致しないことも判明した。

表9 異なるコンテンツによる人気コンテンツ分類精度の検証結果

モデル	日本マンガ	日本 TV アニメ	フランスマンガ
N1	11 / 14	4 / 7	8 / 10
N2	11 / 14	5 / 7	5 / 10
N9	9 / 14	3 / 7	7 / 10
N10	9 / 14	3 / 7	6 / 10
N11	11 / 14	2 / 7	8 / 10
N12	11 / 14	4 / 7	7 / 10

5. ASIA TREND MAP

5.1 サービス概要

1 章で述べた背景を踏まえ、アジアにおける消費トレンド予測情報サービス「ASIA TREND MAP」*9を開発し、2013 年 5 月にリリースした。消費トレンドの予測には、本研究で設計した予測モデルを用いたシステムを使用する。

本サービスでは、マンガ、アニメ、ゲームの各コンテンツごとに、アジア諸国における消費トレンドの推移動向を確認することができる。トレンド算出の対象となるコンテンツは、Wikipedia 日本語版に掲載されているコンテンツであり、アジア諸国における消費トレンドは、その国で標準的に使われる言語版の Wikipedia にそのコンテンツのページが存在するものについて表示される。

図 9 は本サービスのトップページのイメージ図である。ページの上部にはアジアの各国ごとに、現在から 6ヶ月後の消費トレンドが高いと予測されるコンテンツの名称とイメージ画像が、地図上にマッピングされて表示される。また、ページの下部には、国ごとに 6ヶ月後の消費トレンドが大きい上位 5 コンテンツが、一覧形式で表示される。

また、コンテンツ名のリンクをクリックすることで、図 10 で表される、コンテンツ別市場予想ページが表示される。このページではコンテンツのイメージ図と概要が表示され、各国ごとに、そのコンテンツの 6ヶ月後までの消費トレンドの推移動向が、その上昇具合や下降具合に応じて 5 段階の天気アイコンとして表示される。このページを見ることで、コンテンツごとに将来流行しそうな国を調べることができる。

そして、消費トレンドについては、データ収集元である検索エンジンや Twitter の API の制限により、無制限にデータを収集することができないという理由から、ユーザの興味が高いと思われる一部のコンテンツのみ、検索エンジンと Twitter、Wikipedia データによる素性で算出された消費トレンドが表示され、それ以外のコンテンツについては、Wikipedia データによる素性のみで算出された消費トレンドが表示される。

5.2 今後の課題

ASIA TREND MAP では発展途上国を含むアジア諸国をトレンド予測の対象としているが、その国において検索エンジンや Twitter の利用状況が活発でない場合、データが収集できないことがある。また、その国で Wikipedia があまり利用されていない場合、日本のコンテンツに対応したページがほとんど存在しないこともある。Web サービスからデータが収集できない場合は、消費トレンドの算出ができない。この問題はサービス自体で解決できる

*9 ASIA TREND MAP <http://www.asiatrendmap.jp/>

図9 「ASIA TREND MAP」サイトトップイメージ

図10 「ASIA TREND MAP」コンテンツページイメージ

ものではないので、将来これらの Web サービスの利用者が増えることで解決されることを待つことになる。

さらに、精度の高い予測を行うためにはコンテンツ名の同義語を考えなければならない。検索エンジンや Twitter においてコンテンツ名をクエリとして指定してデータを収集する場合は、同時に同義語をクエリとして指定し、そのデータも収集してトレンドの算出に使うことが望ましい。日本語の場合は、同義語を適宜調べて追加することが可能であるが、他言語における同義語を把握することは、それに比べて難しい。他言語における同義語をいかにして収集していくかは、今後の課題として検討していく必要がある。

6. ま と め

本研究では世界各国でのコンテンツ消費トレンドを統一的に把握できるプラットフォームを構築するために、

Web 上のデータマイニングで得られる情報を基に消費トレンドを予測するモデルの設計を行った。Web マイニングの対象は、言語ごとの検索エンジン検索回数と Twitter のツイート数、さらに Wikipedia の解析データとした。本研究の結果、検索回数やツイート回数、Wikipedia 編集履歴情報を組み合わせて素性として用いることで、精度の高い日本のマンガ消費トレンド予測システムが設計できることを示した。また、Wikipedia のみから素性を収集する場合においても、Wikipedia ページ編集履歴情報の他にページの解析情報を素性として用いることで、精度の高い日本のマンガ消費トレンド予測システムが設計できることを示した。また、日本のマンガ以外のコンテンツにおける消費トレンド予測に対して予測モデルが有効であるか検証を行い、そのうち一部のモデルは他のコンテンツの消費トレンド予測に有効であることを示した。そして、アジア各国での日本のコンテンツの消費トレンド予想を提供する Web サービス ASIA TREND MAP を開発し、本研究により設計した予測モデルを搭載した消費トレンド予測システムを、このサービスに実装した。

今後は、さらにモデルの精度を高めていくために、あらゆるコンテンツや言語圏において、消費トレンド予測に有効である一般的な法則を発見し、コンテンツの種類ごとの各素性の影響度の差をなくしていくための適切な素性の補正方法を検討する。そして、今回のデータセットからは収集できなかった、コンテンツ名の同義語や作品中の登場人物、作者名などをクエリとしたデータを収集し、それが予測モデルの精度の向上に貢献するか検討を行うことも、今後の課題とする。このようなモデルの精度の向上に取り組む続けることによって ASIA TREND MAP の利便性を高めていき、クール・ジャパン戦略の成功に貢献したいと考えている。

謝 辞

オリコン株式会社には、マンガのシリーズ別推定売上部数表を、株式会社ホットリンクには、Twitter のツイートデータベースを提供して頂きました。この場を借りてお礼申し上げます。

◇ 参 考 文 献 ◇

- [Abel 10] Abel, F., Diaz-Aviles, E., Henze, N., Krause, D. and Siehn-del, P.: Analyzing the Blogosphere for Predicting the Success of Music and Movie Products. *International Conference on Advances in Social Networks Analysis and Mining*, pp.276-280 (2010)
- [Adar 09] Adar, E., Skinner, G. and Weld, D. S.: Information arbitrage across multi-lingual Wikipedia. *In Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp.94-103 (2009)
- [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)
- [Bollen 11] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market. *Journal of Computer Science*, Vol.2, Issue.1, pp.1-8

(2011)

- [Choi 09] Choi, H. and Varian, H.: Predicting the Present with Google Trends. *Google Inc. Technical Report* (2009)
- [Giering 08] Giering, M.: Retail sales prediction and item recommendations using customer demographics at store level. *ACM SIGKDD Explorations Newsletter*, Vol.10, No.2, pp.84-89 (2008)
- [Goel 10] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. and Watts, D. J.: Predicting consumer behavior with Web Search. In *Proceedings of the National Academy of Sciences of the United States of America*, Vol.107, No.41, pp.17486-17490 (2010)
- [Ishii 12] Ishii, A., Arakaki, H., Matsuda, N., Umemura, S., Urushidani, T., Yamagata, N. and Yoshida, N.: The 'hit' phenomenon: a mathematical model of human dynamics interactions as a stochastic process. *New Journal of Physics*, Vol.14, Issue.6 (2012)
- [Kittur 07] Kittur, A., Suh, B., Pendleton, B. A. and Chi, E. H.: He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.453-462 (2007)
- [Loizou 10] Loizou, A. and Dasmahapatra, S.: Using Wikipedia to alleviate data sparsity issues in Recommender Systems. *Semantic Media Adaptation and Personalization 5th International Workshop*, pp.104-111 (2010)
- [McGray 02] McGray D.: Japan's Gross National Cool. *Foreign Research*(2002)
- [Nastase 10] Nastase, V., Strube, M., Börschinger, B., Zirn, C. and Elghafari, A.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. In *Proceedings of The International Conference on Language Resources and Evaluation*, Vol.10, pp.1015-1022 (2010)
- [Thiesing 97] Thiesing, F. M., Vornberger, O.: Sales forecasting using neural networks. *International Conference on Neural Networks*, Vol.4, pp.2125-2128 (1997)
- [佐々木 10] 佐々木隆: クール・ジャパン マンガアニメの現状と展望について. 多正堂 (2010)
- [杉山 06] 杉山知之: クール・ジャパン 世界が買いたがる日本. 祥伝社 (2006)

〔担当委員: 高間 康史〕

2014 年 1 月 27 日 受理

—— 著 者 紹 介 ——

保住 純

2013 年 3 月 東京大学工学部システム創成学科知能社会システムコース卒業。現在、同大学院工学系研究科技術経営戦略学専攻修士課程在籍中。

飯塚 修平

2012 年 3 月 東京大学工学部システム創成学科知能社会システムコース卒業, 2014 年 3 月 同大学院工学系研究科技術経営戦略学専攻修士課程修了。2014 年 Google 株式会社ウェブデベロッパー。

中山 浩太郎

2007 年 大阪大学大学院情報科学研究科博士後期課程修了(情報科学), 同年大阪大学大学院情報科学研究科特任研究員。2008 年 東京大学知の構造化センター助教, その後講師に就任, 現在に至る。大規模解析, 人工知能, Web マイニングに関する研究に従事。

高須 正和

チームラボ株式会社テクニカルエバンジェリスト・ニコニコ学会 幹事/実行委員・ニコニコ技術部。

嶋田 絵理子

2005 年 3 月 慶應義塾大学文学部卒業。2006 年 4 月 経済産業省入省, 商務流通グループ流通政策課配属。2011 年 8 月 商務情報政策局文化情報関連産業課係長。2014 年 4 月 貿易経済協力局安全保障貿易審査課 審査官。

須賀 千鶴

経済産業省 商務情報政策局メディア・コンテンツ課 総括補佐。2003 年 東京大学法学部卒業。2009 年 米国ペンシルベニア大学ウォートン校経営学修士課程修了(MBA, 医療経営専攻)。2003 年に経済産業省に入省し, 途上国支援, 資源外交, 気候変動交渉, 税制改革等に携わり, 現在はメディア・コンテンツ課にて, 映画・アニメ・音楽・出版等のコンテンツを通じて日本の強みを世界に発信するクール・ジャパン戦略の一端を担う。

西山 圭太

1985 年 東京大学法学部卒業と同時に通商産業省(現 経済産業省)入省。1992 年 オックスフォード大学にて哲学・政治学・経済学 MA 修了。2002 年 内閣府産業再生機構準備室企画官, 通商政策局情報調査課長, アジア太平洋州課長, 経済産業政策局産業構造課長を歴任し, 2005 年 中央大学大学院公共政策研究科客員教授, 2009 年 産業革新機構発足後は同機構執行役員, 専務執行役員に就任。2011 年 内閣官房東京電力経営・財務調査タスクフォース事務局長, 2012 年 7 月より経済産業省大臣官房審議官(経済社会政策担当/経済産業政策局担当)を経て, 現職に至る。

松尾 豊(正会員)

1997 年 東京大学工学部卒業。2002 年 同大学院博士課程修了。博士(工学)。同年より, 産業技術総合研究所研究員。2005 年 10 月よりスタンフォード大学客員研究員。2007 年 10 月より, 東京大学大学院工学系研究科 技術経営戦略学専攻 准教授。2007 年より国際 WWW 会議プログラム委員, 2013 年より Web Mining 部門のチェア。2012 年より人工知能学会 理事・編集委員長。人工知能学会論文賞, 情報処理学会長尾真記念特別賞, ドコモモバイルサイエンス賞など受賞。専門は, Web 工学, 人工知能。