4E1-7

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[*1] 飯塚 修平[*1] 大澤 昇平[*1] 中山 浩太朗[*1] 高須 正和[*2] 嶋田 絵理子[*3]
Jun Hozumi Shuhei Iitsuka Shohei Ohsawa Kotaro Nakayama Masakazu Takasu Eriko Shimada

須賀 千鶴[*3] 西山 圭太[*3] 松尾 豊[*1]
Chizuru Suga Keita Nishiyama Yutaka Matsuo

[*1]東京大学 The University of Tokyo
[*2]チームラボ株式会社 Team Lab Inc.
[*3]経済産業省 Ministry of Economy, Trade and Industry

The Japanese government adopted a national strategy, *Cool Japan*, which assists national culture industries in exporting Japanese media contents (e.g., manga) to Asian countries. The strategy has been not succeed yet in the point of business view due to the lack of measures for international consumer consumption trends. In this paper, we propose a consumption trend predicting system, *ASIA TREND MAP* (ATM). ATM employs Wikipedia as input to predict weekly manga sales data in several countries. We show ATM can predict consumption trend in 6 months with 96.5% accuracy. Our prediction method in ATM can be appried to wide countries and contents.

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していくことを目的とした国家戦略であるクール・ジャパン戦略が推進されている。また、経済産業省は、クール・ジャパン戦略を行う上で、アジアを重要なターゲットとし、海外進出を促進するための推進事業を行なっている。また、世界で日本のコンテンツは評判を得ており、クール・ジャパン戦略は一定の成果を上げているといえる。しかし、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おらず、コンテンツの海外展開は必ずしも成功しているとはいえない [杉山 06]。

コンテンツを海外に輸出してビジネスとして成功を収めるには、適切なターゲット市場や購買層を選定し、そこに向けた内容の翻訳や、現地の文化に合わせた内容の修整を行うことが重要である。しかし現状では、消費トレンド[*1]を正確に把握するためには現地での市場分析やアンケート調査といった手法を取らざるを得ず、それらを行うには多大な手間と時間を要する。その一方、先進国では、Web 上のデータのマイニングを用いて、リアルタイムで消費者性向の分析を行うサービスが存在する。例えば、我が国においては、株式会社ホットリンク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。このサービスのような方法で外国の消費トレンドを予測できれば、従来のトレンド把握方法のような手間と時間を要さずに海外の消費トレンドの把握が行えるので、便利であると考えられる。

しかし、従来の消費者性向分析サービスは単一の言語を対象としており、他言語圏へ適用することはあまり考えられていなかった。その理由として、各言語ごとにデータを収集する際のクエリを適切に選定する際に手間を要し、また、他言語の中には、形態素解析や構文解析といった高度な自然言語処理を実行できるプログラムが開発されていないものが存在することが挙げられる。

以上の背景を踏まえ、本研究では Web 上の情報をもとに、世界各国における日本製コンテンツの消費トレンドを統一的に予測するシステム **ASIA TREND MAP** を開発する。

本研究では消費トレンドを日本国内のマンガの売上部数をもとに設定し、その予測モデルを設計する。消費トレンドの種類をマンガに限定した理由は、他のコンテンツに比べて安価で購入しやすいことから、販売部数が消費トレンドをより直接的に反映していると考えられ、また、個別の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部数情報を一括して収集できるという点にある。まず、検索エンジンや Twitter、Wikipedia から情報を収集する。これらの Web サービスは世界各国で共通に使用されており、API や公式に提供されるデータを使用することで、Web サイトやブログのクローリングによる情報収集に比べて、より短時間に国や言語を指定した情報の収集が行うことができる。次に、収集した情報をもとに、消費トレンドの予測に用いる素性を作成し、それらを用いて、現在から 6 ヶ月後までの消費トレンドを、サポートベクトル回帰 (SVR) を用いて予測するモデルを複数設計する (図 1)。ただし、この方法で設計した予測モデルが、日本以外の国の、マンガ以外の他種のコンテンツのトレンドの予測にも応用できるようにするため、素性を作成する際には、形態素解析や構文解析などの自然言語処理は用いず、また、マンガでいう刊行巻数のような、特定のコンテンツに特有の性質は用いないこととする。そして、それぞれのモデルの精度を比較評価し、消費トレンドの予測にどの素性が有効であるかを調べる。本研究の新規性と有用性は、以下のとおりである。

- 消費トレンド指標を予測するモデルを設計する際に、検索エンジンと Twitter、Wikipedia という 3 つの Web サービス[*4]からなる素性を組み合わせて用いている。実験の結果、これらの素性を組み合わせることで予測モデルの精度が向

連絡先: 保住 純, 東京大学大学院工学系研究科, 東京都文京区弥生 2-11-16 工学部 9 号館 hozumi@weblab.t.u-tokyo.ac.jp

∗1 本論文では消費トレンドを、与えられた国と期間内におけるコンテンツの人気獲得の程度を数値化したもの、と定義する。

∗2 http://www.hottolink.co.jp/kakaricho/

∗3 http://www.ntt.com/marketing/bf/

∗4 本論文では、Twitter だけではなく、検索エンジンや Wikipedia も含めて Web サービスと呼ぶこととする。

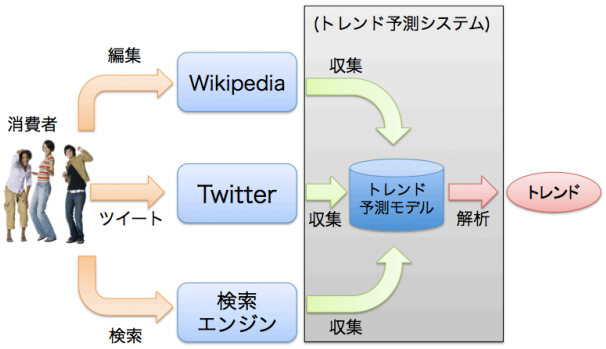

図 1　消費トレンド予測モデルの概念図

上することを示している。

- 本研究で設計するモデルは、複数の言語圏における適用を前提として設計されるので、得られる知見は対象とする言語圏に依らず、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適用可能である。

2. ASIA TREND MAP

前章の内容を踏まえ、アジアにおける消費トレンド予測情報システム ASIA TREND MAP を開発する。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マンガ、アニメ、ゲームであり、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 2 は本サービスのトップページである。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最も高いと予測されるコンテンツの名称と画像が、世界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。

ASIA TREND MAP 内で扱う、検索エンジンや Twitter、Wikipedia のデータは、それぞれ異なるプログラムにより収集、解析され、それぞれ異なるテーブルで管理される。プログラムやテーブルをデータ収集元の Web サービスごとに別々のものを設計することで、Web サービスの仕様変更に伴ってデータの収集方法を変更しなければならなくなった場合や、Web サービスから新たなデータが収集できるようになったときにそのデータを収集する場合、予測精度を高めるためにデータを収集する Web サービスの追加するという場合においても、Web サービス別にテーブルやデータ収集プログラムを変更していけばよいので、仕様の変更に柔軟に対応できる。

また、消費トレンド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ので、ユーザーの興味が高いと思われる一部のコンテンツのみ、各国ごとの Wikipedia 解析データベースからアジア各国ににおけるコンテンツのタイトルを取得し、それをクエリとして検索エンジンや Twitter からデータを収集し、それらから作成される素性を用いて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。

図 2　ASIA TREND MAP トップページ

3. 消費トレンド予測モデル

3.1 提案手法

本研究で設計する予測モデルは、世界各国における様々なコンテンツ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、解析対象の 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がある。そこで本研究ではモデルを設計する際に 1 章で述べ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。
- コンテンツの種類特有の性質は用いない。(例: マンガにおける刊行巻数)

また、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、消費トレンド指標を算出する際に、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ということはできない。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システムと、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システムの、計 2 種類のシステムを設計する。

3.1.1 データセット

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、Web サイトのクローリングに比べて、比較的短い時間で国や言語を指定したデータの収集が行えることである。また消費トレンド指標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。

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検索回数とする。検索回数の集計には、ASIA TREND MAP で実際にそのデータを使用する

表 1　本研究で使用した素性一覧とその説明

素性	記号
月間検索回数 †‡	$S_0, S_1, S_2, S_3, S_4, S_5$
月間つぶやき回数 †‡	$T_0, T_1, T_2, T_3, T_4, T_5$
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}

こととなる、チームラボ株式会社より提供される推定 Web 検索回数データを使用する。

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の数とする。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長期間にわたって収集されている、株式会社ホットリンクのデータベースを使用する。

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 日における Wikipedia フランス語版ダンプデータに解析を行って作成したデータベースを使用する。このデータベースからは、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の、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収集することができる。

消費トレンドを作成する元データには、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推定売上部数表を使用する。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。

3.1.2　素性の作成

3.1.1 項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 † がついているものについては、m ヶ月前の数値を添字 m を付与して X_m と記し、さらに過去 6 ヶ月分の月の平均 $\overline{X}$ 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。また、素性の語尾に ‡ がついているものは、桁数が大きく、そのまま素性として使用すると予測精度が低下する [Goel 10] ので、元の素性に対し自然対数をとった値を素性として使用する。

3.1.3　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。したがって、マンガの週別売上部数を直接消費トレンドとみなすことは適切とは言えない。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 とし、以下の補正を行うことで補正済週別マンガ売上部数 Y_w を算出し、それをもとに消費トレンドを作成する。

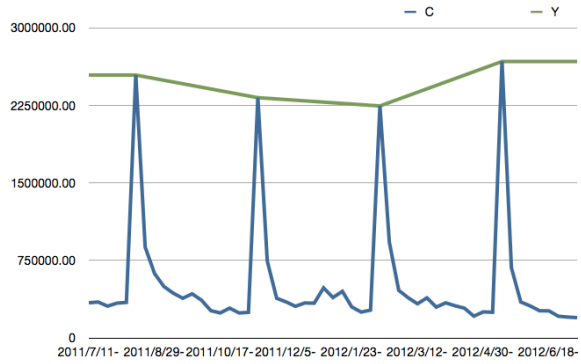

図 3　C_w と Y_w の対応関係例

まず、前週に比べて C_w が 5 倍以上になっている週をピーク週と定義し、ピーク週は $Y_w = C_w$ とする。そしてピーク集でない週については、最初のピーク週以前の週の Y_w は最初のピーク週の C_w を、最後のピーク週以後の週の Y_w は最後のピーク週の C_w を、そしてそれ以外の週の Y_w は、その週の直前と直後のピーク週における C_w を直線で結んだときにその週に該当する Y_w の値とする (図 3)。

この方法で求めた補正済週別マンガ売上部数 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、教師事例における消費トレンドの最大値が 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。

本研究で設計するシステムは、サポートベクトル回帰 (SVR) を用いて、表 1 にある組成から消費トレンド y_m の予測を行う。カーネル関数として RBF カーネルを使用し、パラメータを調整するために $K = 5$ の K-fold 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。そして、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、そのときのパラメータをモデルに採用する。

3.2　実験

本章では、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示す。また、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、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)、2 ヶ月後、4 ヶ月後、6 ヶ月後の消費トレンドを SVR を用いて算出した。

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 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 Wikipedia のデータを 3.1.1 項で述べた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やカテゴリリンク数といったページの構造情報を、過去のある時点での値を遡って収集することができないので、本実験では Wikipedia 日本語版解析データベースに含まれている 2012

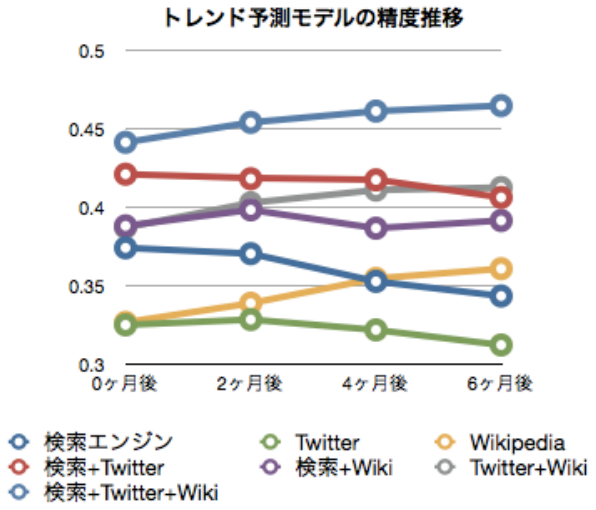

図 4　用いた Web サービスの違いによる予測モデルの精度変化

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については教師事例から取り除き、最終的に合計 1499 個の教師事例を作成した。

3.3　実験結果

3.3.1　複数の Web サービスからなる素性を用いた予測モデル

まず、検索エンジン、Twitter、Wikipedia の記事編集回数のデータからなる素性を使用し、モデルを作成した。単数の Web サービスによる素性を用いた予測モデルでは、2 ヶ月の予測までは検索回数を素性として用いたモデルの予測精度が高く、4 ヶ月以降の予測は Wikipedia ページ編集情報を素性として用いたモデルの予測精度が高かった。そして、検索回数による素性を用いた予測モデルでは消費トレンドを予測する時期が先になるほど精度が低下していくことに対し、Wikipedia の編集情報による素性を用いた予測モデルでは、消費トレンドを予測する時期が先になるほど精度が向上した。また、複数の Web サービスによる素性を用いた予測モデルは、単一の複数の Web サービスによる素性を用いた予測モデルに比べて、より高い予測精度を示した (図 4)。

3.3.2　Wikipedia のみからなる素性を用いた予測モデル

Wikipedia の情報から得られる素性のみを用いた予測モデルは、ページ情報が過去の値を遡って収集することができないことから、6 ヶ月後のみの予測精度を算出し、精度を検証した。

Wikipedia からのみからなる素性を用いた予測モデルは、ページの構造情報を素性として用いることで予測モデルの精度を大きく向上させることができた。特に、ページのリンク、被リンク数が精度の向上に大きく寄与した。Wikipedia から得られるデータによる素性を順次追加していくことで、モデルの精度を大きく向上させることができた (表 2)。

表 2　Wikipedia のみよるモデルの 6 ヶ月後の予測精度

使用した素性	6 ヶ月後の予測精度
WIki 記事編集履歴情報	0.4057
編集履歴 + リンク数	0.8235
編集履歴 + 被リンク数	0.7946
編集履歴 + ページ構造情報	0.9656

4.　関連研究

検索エンジンにおける検索回数から商品の売上データを予測する研究が、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、予測する期間が一ヶ月先になると、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検索回数しか用いておらず、予測する先の範囲が最大 1.5 ヶ月後までにとどまり、より長期間先の範囲におけるデータの予測実験は行われていない。本研究では、現在から最大で 6 ヶ月後までの、より期間の長い予測実験を行った。

5.　結論

本研究の結果、複数の Web サービスによる素性や、Wikipedia のページ構造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モデルが設計できることを示した。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システム ASIA TREND MAP を開発する。

今後は、さらにモデルの精度を高めていくために、検索エンジンや Twitter のデータを用いた予測モデルについて、詳細な精度検証を行う。また、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。

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

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

[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).

[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)