4K2-3

Wikipediaの編集履歴による各国のユーザ嗜好に基づいたコンテンツの分析

Analysis of Contents Based On User Preferences of Each Country by Editing History of Wikipedia

野中　尚輝 *1　　中山　浩太郎 *1　　松尾　豊 *1
Naoki Nonaka　　Kotaro Nakayama　　Yutaka Matsuo

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

The contents industry is one of the important industries in Japan, and its popularity is increasing also in overseas in recent years. In developing content works overseas, it is important to investigate the market in each country/region. In recent years, with the spread of social networks, it has become easy to gather information on overseas users, and it is attempted to predict the popularity of contents in each country/region using information obtained from the web there. Under such circumstances, capturing multifaceted information of contents based on consumers and users' preferences of each country/region is an important issue for companies considering secondary use of content holders and contents. In this research, we focused on the multilinguality and completeness of Wikipedia and attempted to obtain multifaceted information including content genres based on Wikipedia user's preferences.

1. はじめに

コンテンツ産業は，日本の重要な産業の一つであり，近年海外でもその人気が高まっている．代表的なコンテンツであるアニメ産業の市場規模は，2015 年には前年比 12.0%増となっており，著しい成長を見せている [aja]．また，海外におけるコンテンツ市場の規模も拡大を続けており，海外の市場を開拓することにより，さらなるコンテンツ産業の発展が期待される [met]．

海外へのコンテンツ作品の展開を行う上で，それぞれの国・地域における市場を調査することは重要である．国・地域ごとに文化を始めとする条件が異なり，流行するコンテンツは異なると考えられるため，それぞれの国・地域ごとの市場調査が重要となる．近年，ソーシャルネットワークが普及したことにより，海外のユーザの情報を収集することが容易になっており，ウェブ上から得られる情報を用いて各国・地域におけるコンテンツの人気を予測することが試みられている [hoz]．

このような背景の中で，各国・地域の消費者・ユーザの嗜好に基づくコンテンツの多面的な情報を捉えることは，コンテンツホルダーおよびコンテンツの二次利用を考える企業にとって重要な課題である．コンテンツについてのジャンルをはじめとする多面的情報を比較することで，類似するコンテンツの特定やコンテンツ市場でのポジショニングを知ることができると考えられる．

そこで本研究では，Wikipedia の持つ多言語性と網羅性に着目し，Wikipedia のユーザの嗜好に基づくコンテンツのジャンルを始めとする多面的な情報を得ることを試みた．Wikipedia は，幅広い内容と多くの言語を網羅しており，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である．Wikipedia の各ページは，そのページに対する興味関心の高いユーザにより編集されるため，ユーザの嗜好を反映していると考えられる．また各言語のページは，その言語に習熟した人が行うことが多いと考えられるため，ユーザによる編集系列は，言語ごとの特性を反映していると考えられる．したがって，各言語ごとの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，コンテンツ作品についてのジャンルをはじめとする多面的な情報を得られると考えられる．

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

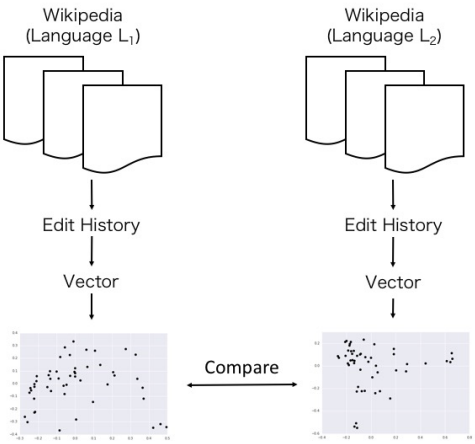


図 1: 提案手法の全体像

本研究では図 1 に示すように，各言語の Wikipedia におけるコンテンツ作品に関するユーザの編集行動履歴の系列データから，ユーザの嗜好に基づくコンテンツの多面的な情報を学習する．その後，学習された結果について各言語と対応する国・地域間での傾向の違いを分析する．コンテンツ作品に関する編集回数の多かった，英語，日本語，フランス語，ドイツ語，スペイン語，ロシア語，中国語，イタリア語のデータを用いた．

本論文は以下のように構成される．2 章にて，関連研究について触れ，3 章にて提案手法について説明する．4 章において，取り扱ったデータとその前処理について述べ，5 章に実験結果を記す．最後に 6 章にて，実験結果に対する考察と結論を述べる．

2. 関連研究

この章では，本研究に関連する研究について述べる．

2.1 コンテンツの人気を分析した研究

商品やコンテンツの人気予測はマーケティング戦略を決定する上で重要な課題であり [Kuo 98]，多くの研究が行われている．近年インターネットやスマートフォンの普及により，一般のユーザが情報を発信することが容易になった．その結果，ウェブ上には多くの情報が存在するようになり，それらを分析することで商品の流行や人気を予測することが先行研究にて行われている．特に，コンテンツ作品についての人気予測として，各種ソーシャルメディアから得られた素性を組み合わせて用いる研究 [hoz] が存在している．

商品の販売や人気を予測する上で，消費者の嗜好をもとにした商品の多面的な情報を考慮することが重要であると考えられる．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]．また，商品の人気には自己強化的な側面があり，人気は消費者の意思決定に影響を与えることが知られている [Salganik 06], [Chen 11]．このことから，消費者の嗜好と商品の人気の間には関連性が存在すると考えられる．

加えて，海外における商品の展開を考えた場合，その国・地域ごとに消費者の嗜好を考慮することが重要となる．しかしながら，これまでの研究では異なる国・地域間の消費者についてその嗜好の違いを分析した研究は存在しなかった．

本研究では，消費者の嗜好に基づいたコンテンツの多面的な情報を，Wikipedia の各言語について学習し，対応する国・地域について分析している点が新しい．

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 [hoz] 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．

3. 編集履歴からの多面的情報の取得

本章では，Wikipedia 上でのユーザの編集履歴をもとにコンテンツ作品の多面的な情報をベクトル表現として取得する提案手法について説明する．学習される多面的な情報はベクトル表現として得られるため，以下ではコンテンツベクトルと記す．

コンテンツベクトルの学習と国・地域間での比較を行う上で，Wikipedia は望ましい性質を有する．Wikipedia は幅広い内容を網羅し，コンテンツ作品についての情報も豊富である．また多くのユーザにより利用され，頻繁に編集が行われる．加えて，多言語の同一内容ページとの間にリンクが存在するため，異なる言語間において対応する項目を紐づけることも容易である．このような性質を有する Wikipedia におけるページの編集履歴に着目し，コンテンツベクトルの学習を行う．

Wikipedia において，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，これを用いることでコンテンツベクトルを学習できる．ユーザの興味関心についての系列的な情報と捉えることができるユーザごとの編集履歴を用いることで，ユーザの嗜好に基づくジャンルをはじめとする多面的な情報を学習できると考えれる．また，ユーザの編集するページは自身の得意とする言語であることが多いと考えられるため，各言語ページの編集者はその言語を母語とする可能性が高い．したがって，各言語における編集履歴から学習を行うことで言語ごとのユーザの嗜好を反映したコンテンツベクトルが得られることが期待される．

提案手法では，入力としてある言語 L についての Wikipedia におけるコンテンツ作品 c^L についての編集系列を与え，出力としてその言語におけるコンテンツベクトル C^L を得る．ユーザごとの編集履歴を時系列に並べ，あるコンテンツの前後に出現するコンテンツが与えられた場合に元のコンテンツを予測するように学習を行う．より厳密には，あるコンテンツの系列 $c_1, c_2, c_3, ..., c_T$ 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 $c_{t-k}, ..., c_{t+k}$ により予測できるように各コンテンツ c のベクトル表現を学習する．ここで，各コンテンツ c を単一のベクトルにマッピングする行列を C とする．コンテンツの系列が与えられた時，以下の平均対数確率を最大化することでコンテンツベクトルを学習する．

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(c_t|c_{t-k}, ..., c_{t+k}) \tag{1}$$

予測タスクは通常，ソフトマックス関数に代表される多クラス分類によって行う．

$$p(c_t|c_{t-k}, ..., c_{t+k}) = \frac{\exp(y_{ct})}{\sum_i \exp(y_i)} \tag{2}$$

ここで y_i は，各コンテンツ c_i についての正規化されていない対数確率であり，

$$y = b + Uh(c_{t-k}, ..., c_{t+k}; C) \tag{3}$$

で算出される．U および b はソフトマックス関数のパラメータであり，h は C から得られるコンテンツのベクトル表現の平均値である．

このように学習された言語 L のコンテンツベクトル C^L を比較・分析する．

4. データ

本章では，実験において用いたデータとその前処理について述べた後，Wikipedia の各言語版におけるユーザの嗜好に基づくコンテンツの多面的な情報を獲得した結果について述べる．

4.1 データソース

取得したデータから対象とするコンテンツ作品に関するページの特定を行った．日本語版の Wikipedia のカテゴリー「アニメ」,「マンガ」,「ゲーム」の下に存在するページのうち,「○○の一覧」といったページを除いたものを対象のコンテンツ作品とした．その後，日本語版のページに存在する言語リンクを元に日本語以外の言語における各作品のページを特定した．な

お，言語リンクが存在しない場合は，その言語における対象のコンテンツ作品のページは存在しないものとして扱った．

続いて，対象としたコンテンツ作品ページに関わる編集系列を各言語ごとに取得した．WikiMedia から取得したデータに含まれる各ページの編集履歴の中から，コンテンツ作品に関わるページについて，編集者と編集日時の情報を取得した．これを編集者ごとに時系列順に並べて，編集系列として，実験に用いた．

4.2 関連するコンテンツの統合処理

コンテンツのベクトル表現を学習する前に，得られた編集系列に対して，コンテンツの統合処理を施した．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．これらのページは，同一コンテンツに関するページであるので，編集系列上においては同一のコンテンツとして扱う．これを実現するため，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 を利用する．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する．

4.3 各言語における編集系列

コンテンツの統合処理を経て得られた編集系列から，コンテンツのベクトル表現を学習するモデルに入力する編集系列を選択する方法について述べる．コンテンツのベクトル表現を学習する際に，一定以上の系列長が必要となる．そこで，本研究ではいずれの言語についてもコンテンツ作品に関する編集系列が長い上位 2,000 名の編集者の編集系列を用いてベクトル表現の学習を行った．対象とした言語は，上位 2,000 名の編集系列の長さの最低値が 10 以上であった英語，日本語，スペイン語，中国語，ドイツ語，イタリア語，フランス語，ロシア語の 8 つとした．

5. 実験結果と考察

本章では，提案手法により学習されたベクトル表現について述べる．

5.1 ベクトル表現の学習と可視化

ベクトル表現はコンテンツ作品に関する編集系列が十分に取得できた，英語，日本語，スペイン語，中国語，ドイツ語，イタリア語，フランス語，ロシア語の 8 言語について行った．学習の際のウィンドウサイズを 10，出力されるベクトルを 50 次元とし，Continuous Bag-of-Words(CBOW) モデル [Mikolov 13] により学習を行った．なお，実装は python ライブラリ Gensim を用いて行った．

学習されたベクトル表現は，RBF カーネルを用いたカーネル主成分分析により，2 次元に射影し可視化した．可視化は，対象とした 8 言語のうち 7 言語以上に存在するコンテンツの中で，日本語版 Wikipedia での被リンク数上位 50 件のコンテンツを対象として行った．これは，被リンク数の多いコンテンツは知名度が高い傾向にあり，得られた可視化結果の解釈が行いやすくなると考えたためである．また可視化の際に一部のコンテンツに対してラベルを付与した．

本手法で得られたベクトル表現にて，近接するコンテンツ作品同士はユーザの嗜好に基づいて何らかの形で似ていると考えられる．これは二次元に射影し，可視化した結果についても同様である．一方，同一のコンテンツ作品であっても，学習されるベクトル表現は言語ごとに異なる．そのため言語が異なる場合，同一の次元であっても同じ意味表現であるとは限らないため，得られたベクトル表現について異なる言語の間でのベクトルの計算を行うことはできない．

5.2 可視化結果の分析

まず，得られた可視化結果を分析し，学習されたベクトル表現に含まれる情報について調べた．可視化結果を図 2 に示す．日本語版の Wikipedia から得られた可視化結果（図 2(a)）では，図の上側に女性向け漫画のジャンルに含まれる「ハチミツとクローバー」や「NANA」が位置し，左下には「機動戦士ガンダム」や「マジンガー Z」といったやや古い男性向けの作品が現れた．また,「俺の妹がこんなに可愛いわけがない」や「魔法少女まどか☆マギカ」といった今回扱った編集系列の期間の中では比較的新しい作品が右下に出現した．この可視化結果から，学習されたベクトル表現において近接するコンテンツ作品はジャンルや作品が登場した年代といった情報を内包していることが示唆された．

続いて，日本語以外の言語にて得られた可視化結果について分析を行った．英語（図 2(b)）および中国語（図 2(d)）での可視化結果では，ガンダム系の作品である「機動戦士ガンダム」,「新機動戦記ガンダム W」,「機動戦士ガンダム SEED」の三作品が非常に近くに存在していた．日本語で得られた結果においては,「機動戦士ガンダム SEED」はその他の二作品よりも「進撃の巨人」といった作品に近い位置に存在した．これは「機動戦士ガンダム SEED」の捉えられ方が，日本と英語圏および中国語圏で異なることを示唆している．

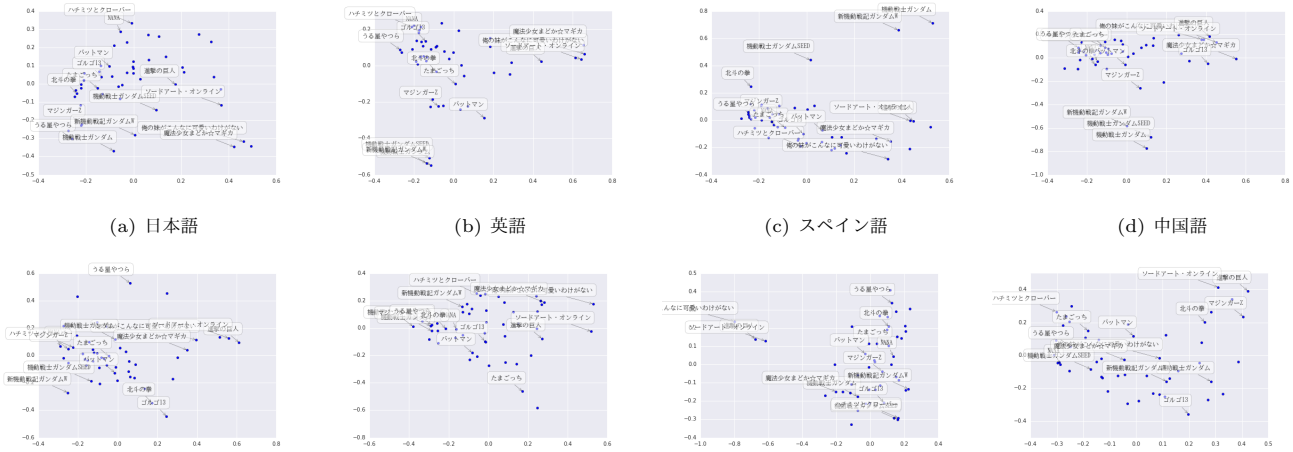
ドイツ語（図 2(e)）では,「うる星やつら」,「エースをねらえ！」,「らんま 1/2」といった作品が図の上側に位置し，下側には「北斗の拳」や「ゴルゴ 13」といった作品が位置していた．また，イタリア語（図 2(f)）では,「たまごっち」や「逆転裁判」といったゲームに関連する作品が下側に出現する傾向にあった．

また,「ソードアートオンライン」や「進撃の巨人」といった作品は，今回対象としたいずれの言語においても近い距離で存在していた．この二作品は，編集系列を取得した期間において新しい作品であったため，いずれの言語においても近い距離に出現したと考えられる．

以上の結果から，編集系列から学習されたコンテンツのベクトル表現は，作品のジャンルや作品が登場した年代といった多面的な情報を含んでいることが示唆されることがわかった．また他の言語における結果から，各言語に対応する国・地域におけるユーザの嗜好が反映されたと考えられるベクトル表現が得られることがわかった．

6. 結論

本研究では，各言語の Wikipedia におけるコンテンツ作品に関するユーザの編集系列データから，ユーザの嗜好に基づくコンテンツの多面的な情報を学習し，各言語と対応する国・地域間での傾向の違いを分析した．日本語版の Wikipedia から得られたベクトル表現の可視化結果から，日本のユーザの嗜好を一定程度反映したものとなることが示唆された．可視化された結果からは，得られたベクトル表現にはコンテンツ作品のジャンルや登場した年代といった情報が含まれていると考えられた．また他の言語から得られた結果から，各言語に対応する国・地域におけるユーザの嗜好が反映されたと考えられるベクトル表現が得られていることがわかった．今後は，得られた各言語ごとのベクトル表現を組み合わせて用いることでコンテンツ作品の人気予測を行うことを考えている．



(a) 日本語　(b) 英語　(c) スペイン語　(d) 中国語

(e) ドイツ語　(f) イタリア語　(g) フランス語　(h) ロシア語

図 2: 学習されたベクトル表現の可視化結果

謝辞

本研究は JSPS 科研費 JP25700032, JP15H05327, JP16H06562 の助成を受けたものです。

参考文献

[aja] アニメ産業レポート 2016 サマリー（日本語版）1.1

[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)

[Chen 11] Chen, Y., Wang, Q., and Xie, J.: Online social interactions: A natural experiment on word of mouth versus observational learning, *Journal of marketing research*, Vol. 48, No. 2, pp. 238–254 (2011)

[hoz] 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

[Kuo 98] Kuo, R. J. and Xue, K.: A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights, *Decision Support Systems*, Vol. 24, No. 2, pp. 105–126 (1998)

[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)

[met] コンテンツ産業の現状と今後の展開の方向性（経済産業省）

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (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)

[Potthast 08] Potthast, M., Stein, B., and Anderka, M.: A Wikipedia-based multilingual retrieval model, in *European Conference on Information Retrieval*, pp. 522–530Springer (2008)

[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. 175–186ACM (1994)

[Salganik 06] Salganik, M. J., Dodds, P. S., and Watts, D. J.: Experimental study of inequality and unpredictability in an artificial cultural market, *science*, Vol. 311, No. 5762, pp. 854–856 (2006)

[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)

[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)