3K3-3

言語モデルを用いたウェブメディアの特徴抽出

Using Language Model for Web Media Analysis

関 喜史 *1　松尾 豊 *2
Yoshifumi Seki　Yutaka Matsuo

*1株式会社 Gunosy　*1*2東京大学
Gunosy Inc　University of Tokyo

We attempt to extract features of media using author topic model. What kind of pick up the topic is different by the media. The different of media is subjective. Our purpose is to quantitatively interpret the difference using language model. We construct author topic model by 275 media and 28,942 articles. Analyzing author topic distribution classify media to 9 category and compare same category media.

1. はじめに

スマートフォンの普及，ネイティブ広告やネットワーク広告などによる収益化などを背景に新しいウェブメディアが数多く誕生し，急成長している．また新聞社などの既存のメディアも需要の頭打ちによる危機感から，ウェブメディアに力をいれつつある．

ウェブ以前では，新聞を購読したり雑誌を買ったりテレビのチャンネルを選択したり情報を消費する際には必ずメディアの選択を行っていた．しかしウェブにおいては，検索やソーシャルメディア，ポータルやキュレーションサービスなどにおいてその記事がどのメディアかを意識することなく記事に触れ，情報を消費することができる．

ウェブメディアが収益を高めるためにはは Page View(PV) 数を高めることが必要となる．そのためにメディアは検索流入の最適化を行ったり，ソーシャル上で拡散を促進するような仕組みを導入したり，キュレーションサービスやポータルへ掲載されるような取り組みを行ったりしている．これらの取り組みはすべて記事単位で流入を増大させるためのものである．このようにウェブにおいてメディア自体の特性というのは意識されないことが多く，PV 数を伸ばすためにメディアが路線を変更することも多い．

本研究ではウェブメディアの記事に Author Topic Model[Michal 04] を適用しメディアごとの特徴を抽出し分析することを試みる．各メディアがどのようなメディアかということはこれまで主観的に述べられてきた．言語モデルを用いることでメディアごとの特性を客観的に分析することにより，現在の市場環境の可視化や，想定されるユーザ層の分析，競合分析などが行えると期待される．

2. 手法

本研究ではある一定期間にウェブメディアから配信された記事データを用いる．記事数は 28,942 件であり，メディア数は 275 件であった．記事にはタイトルと本文情報があり，それらのテキストデータを形態素解析し Bag of Words 化して用いる．形態素解析には mecab[*1] を, 辞書には IPA 辞書を用いた．分析には形態素のうち名詞であり，”数”, ”代名詞”, ”接尾”, ”地域”, ”非自立” ではないものを用いた．この際未知語処理のルールを変更し記号は名詞にならないようにしている．この結果単語数は 138,656 個であった．

各メディアの特徴抽出のために Author Topic Model を用いる [Michal 04]．Author Topic Model は LDA[Blei 02] に著者要素を加えて拡張したモデルである．Author Topic Model のグラフィカルモデルを図 1 に示す．

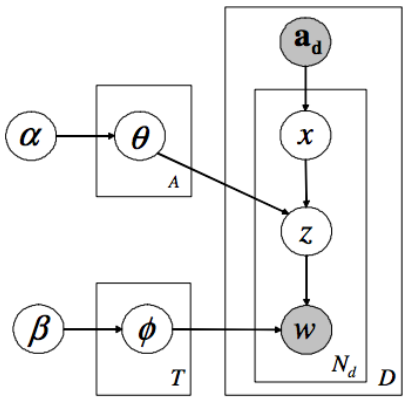

図 1: Authot Topic Model のグラフィカルモデル

Author Topic Model ではトピックの確率分布が Author ごとに決まる．この Author をメディアと考えることでメディアごとのトピックの確率分布を得て，それぞれのメディアがどのような記事を配信しているのかを明らかにする．今回トピック数は 300, $\alpha = 0.1$, $\beta = 0.01$ とした．

このようにして得られたメディアごとのトピッ分布を元にメディアの特徴分析を行う．

3. クラスタ分析

メディアごとのトピック分布を用いてメディア間の類似度を求めメディアをノードとする重み付きグラフを構築しクラスタ分析を行う．メディア間の類似度はトピック分布をベクトルと考えて正規化しコサイン類似度を求める．コサイン類似度 0.3 以下のメディア間についてはエッジを生成しないことした．そ

連絡先: 関 喜史，株式会社 Gunosy，東京都港区六本木 6-10-1 六本木ヒルズ森タワー， 03-6455-4560, yoshifumi.seki@gunosy.com

*1 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

の結果, 結合グラフがノード数 259, エッジ数 2,963 のものと，ノード数 2, エッジ数 1 の 2 つの結合グラフが得られた．分析対象のメディア数は 275 件であったため，14 件のメディアは他のメディアとのエッジが得られなかった．

後者のノード数 2 のグラフはこれ以上分割できないため，前者のグラフのみクラスタ分析を行う．ノード数 2 のグラフは両メディアともテニス情報の専業メディアであった．クラスタリングには louvian method[Blondel 08] を用いる．louvian method はネットワークにおける modularity という指標を最適化するようにクラスタリングをを来なう手法であり，高速で優れた手法として知られている [Fortunato 10]．クラスタリングの結果グラフは 8 つのクラスタに別れた．各クラスタの特徴とメディア数を表 1 に示す．

表 1: クラスタの概要とメディア数

	クラスタの概要	ユーザ数
(a)	新聞社などの大手メディア，経済誌	39
(b)	観光・食・ファッション	41
(c)	芸能ニュース・ゴシップ	49
(d)	恋愛・健康・美容などの女性向けコラム	60
(e)	スポーツニュース	19
(f)	アプリ・ガジェットなど	28
(g)	車	3
(h)	2ch まとめサイト	20

このようにメディアがその特性に合わせて分離されていることが見て取れることから，Author Topic Model を用いたメディアの特徴抽出はある程度有効であるといえる．

4. 各メディアのトピック分布

本章では実際に各メディアでどのようなトピック分布が得られどのような知見が得られたかについて述べる．まず大手新聞社 2 社のトピック分布を図 2, 3 に示す．

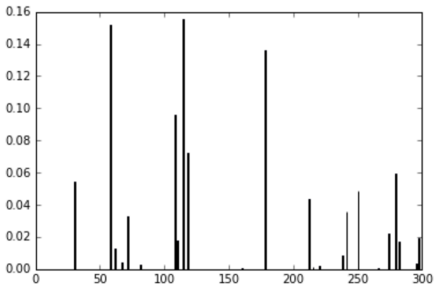

図 2: 大手新聞社 A 社のトピック分布

実際の値については微差があるが，両メディアの上位 3 つのトピックは共通している．共通しているトピックは「国際問題を中心とした政治」「経済」「社会」の 3 つであった．A 社と B 社を比較した際には B 社がより「社会」を重きを置いて発信しており，B 社がより「政治・経済」に重きを置いて発信していると捉えることができる．

次に女性向けメディアの特徴について述べる．表 1 におけるクラスタ (d) の各メディアのトピック分布を平均化したトピック分布を図 4 に示す．

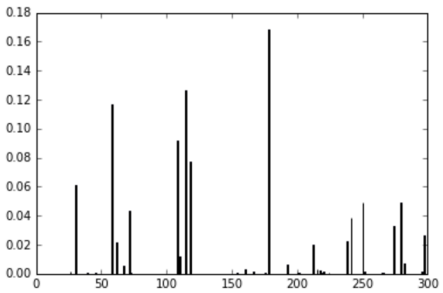

図 3: 大手新聞社 B 社のトピック分布

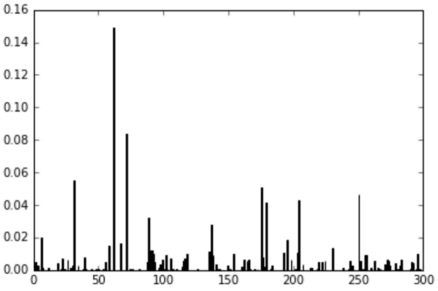

図 4: クラスタ (d) の平均トピック分布

女性向けメディアで最も高い値を出したトピックは「恋愛」に関するものであった．続いて「ハウツー」「健康」「美容」「キャリア」と続いた．続いていくつかのメディアのトピック分布を見る．

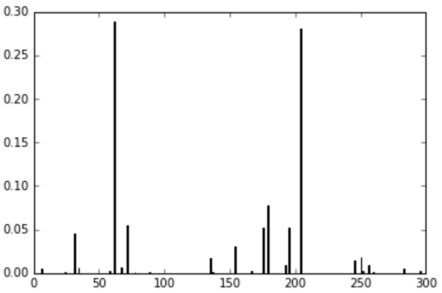

図 5: 女性向けメディア C のトピック分布

図 5 はクラスタの特徴である恋愛トピックと同様に大きな特徴となっているトピックがある．そのトピックの内容も恋愛に関するものであった．この 2 つのトピックの違いとしては，平均トピックで強く現れたトピックの特徴語には「自分」「気持ち」「人生」といった個人に関する語が現れたのに対し，メディア C で強く現れたトピックでは「男性」「女性」といった一般論について扱っていることを示唆する特徴語や,「デート」「結婚」など具体的な事案について扱っていることを示唆する特徴語が現れており，恋愛に関するトピックであっても大きな違いが見て取れた．このことからメディア C は女性向けメディアの中でもより恋愛に特化したメディアであると考えられる．

図 6 では平均トピックで現れた「美容」に関するトピックが 3 番目に高いトピックであった．最も高いトピックには「睡

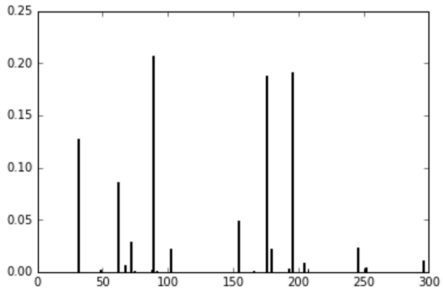

図 6: 女性向けメディア D のトピック分布

眠」「顔」「シワ」など美容についてより具体的な内容に言及していることが伺える特徴語が見られた．また 2 番目に高いトピックでは「春」「ファッション」「肌」「メイク」など，美容とファッションについての深い記述があるトピックが現れた．このようにクラスタ特徴ベクトルでは平均的な内容が現れ，各メディアのトピック分布ではより踏み込んだそのメディアの特性・強みなどを知ることが可能であることが示唆された．

5. まとめ

本研究では Author Topic Model を用いたメディアの特徴抽出を試みた．その結果得られたメディアのトピック分布を用いてメディアの特徴を得ることができ，それぞれのメディアがどのような記事を配信しているのか，同じようなメディアはどれかなどについて論じることが可能であることを示した．今後は word2vec[Mikolov 13] に代表される語の分散表現を用いた手法や，ユーザの行動などを用いてさらにさまざまな側面からメディアの特徴について分析していきたい．

参考文献

[Blei 02] Blei, D., Ng, A., and Jordan, M.: Latent Dirichlet Allocation, in Dietterich, T., Becker, S., and Ghahramani, Z. eds., *Advances in Neural Information Processing Systems 14*, Cambridge, MA (2002), MIT Press

[Blondel 08] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008 (2008)

[Fortunato 10] Fortunato, S.: Community detection in graphs, *Physics Reports*, Vol. 486, No. 3, pp. 75–174 (2010)

[Michal 04] Michal, R.-Z., Thomas, G., Mark, S., and Padhraic, S.: The author-topic model for authors and documents, in *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (2004)

[Mikolov 13] 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)