4D4-5

Twitter 上のコミュニティとウェブ情報ソースの関係性に基づくユーザセグメンテーションに関する研究

User segmentation of Twitter using Information Diffusion on Online Communities

久保田 修平 *1 Shuhei Kubota
榊 剛史 *2 Takeshi Sakaki
森 純一郎 *3 Junichiro Mori

*1東京大学工学部システム創成学科 Faculty of Engineering, The University of Tokyo
*2株式会社ホットリンク Hottolink, Inc.
*3東京大学 The University of Tokyo

The development of the Internet and social media have changed the way that people get information. As a result, companies have been faced with the need for appropriate web strategy. For companies, it is important to spread the information accurately to the people they want to convey the information. In this paper, we propose user segmentation method of social media by modeling information diffusion among online communities. The proposed method employs clustering of bipartite graph, which consist of user communities and Web information sources. We conducted the experiment of user segmentation using Twitter data. The results show that our proposed method successfully identifies latent relationships between interest-based user groups and online information sources.

1. はじめに

近年、インターネットの発達により情報の流通の仕方が大きく変化してきている。従来人々は、世の中の出来事や話題を新聞やテレビに代表されるようなマスメディアを媒介として享受してきた。しかし、近年では人々が日常的にインターネットを用いるようになり、インターネット上に多数のオンラインメディアが誕生して、インターネットから情報を受け取るようになってきている。また、ソーシャルメディアの発達に伴い、そうしたオンラインメディアから情報を受け取るだけではなく、ユーザが情報の発信源となり、ユーザからユーザへ情報が伝播していくということも起きている。

こうした状況を背景とし、企業のウェブ上における広告費も急増している。情報通信白書の資料 [総務省 15] によれば、企業のウェブ上での広告費は、2003 年には 1183 億円であったものが、2014 年には 1 兆 519 億円にまで増加しており 10 年前の約 10 倍にまでなっている。インターネットを情報源として用いる人の増加に伴い、企業側もウェブ戦略に積極的になっていることが伺える。企業にとっては、情報を伝えたい相手に的確に情報拡散を行うことが重要な意味を持つ。その際、どのようなユーザがどのようなメディアから情報を得るかということに基づいたユーザセグメンテーションに大きな価値が存在している。

本論文では、Twitter 上の 400 万ユーザ以上のデータを用いた分析を行い、ソーシャルメディア上における、情報を拡散させるという意味でのユーザセグメンテーションを行った。まず、本論文ではソーシャルメディア上における情報拡散を情報の発信 (メディア) と情報の受信 (受信者) の連鎖反応であると考える (図 1)。ソーシャルメディア上においてはユーザもまた情報の発信源となりうることを考えれば、ユーザもまたメディアであり、自然な考え方と言えるだろう。ウェブ上にはこのような様々メディアを発端とし、様々なユーザを介した情報拡散の “線” が多数存在すると考えられる。本論文は濃淡を持ってウェブ上に存在する、このような情報拡散の “線” の濃い部分を束として取り出し、情報拡散の経路を抽出・分類することが情報を拡散させるという意味において効果的なユーザセグメンテーションにつながると考え、分析を行った。

ただ、このように情報の経路が抽出でき、あるユーザセグメントがあるメディアに関心があるとわかっても、当然のことながらすべての記事に関心がある訳ではない。メディアの記事の中にも関心がある記事とそうでないものが存在しているからだ。あるユーザセグメントがどのようなコンテンツに対して関心を持っているのかということは、そのセグメントを特徴づけるものであり、広告制作などにおけるクリエイティブの面でも多くの示唆を与えてくれるものであろう。そこで、本研究ではあるメディアから記事を抽出し、各コミュニティがどのような内容のコンテンツに対して反応しているのかを分析することにする。このようにして、本研究では企業のウェブ戦略に有効であるユーザセグメンテーションを実現する。

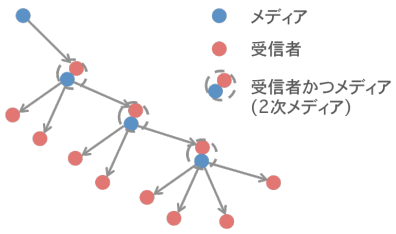


図 1: メディア-受信者の連鎖としての情報拡散

2. 関連研究

ウェブ上の行動に基づくユーザセグメンテーションに関する研究は数多く行われてきた。Ozer らは音楽系ウェブサイトに訪れるユーザーの音楽の興味やコンピュータの使用頻度などを素性としてユーザーセグメンテーションを行い [Ozer 01]、ウェブ上の行動に基づいたユーザのセグメントを示した。また、Zhou らの研究ではウェブサーバーへのアクセスログのデータを用いて、ユーザのセグメンテーションとユーザ行動のモデル化を行っている [Zhou 06]。 また、Twitter を用いたユーザのコミュニティ抽出に関する研究も多く存在する。Java らは Twitter 内のフォローネットワークからコミュニティ分割を行い、コミュニティごとにユーザーの使用目的や使用方法が異

連絡先: 久保田修平, 東京大学大学院工学系研究科, 東京都文京区本郷 7-3-1 工学部 3 号館, shu.kubota78@gmail.com

なっていることを明らかにした [Java 07]。また、Marui らの研究では、ユーザーのメンション関係によって会話ネットワークを構成し、その会話ネットワークをクラスタリングすることによってユーザーコミュニティを抽出する手法を提案した [Marui 14]。しかし、情報の拡散という点に着目しているセグメンテーションに関する研究はあまり見受けられない。

ただし、その中でも本研究と最も関連している研究は鳥海らによる情報拡散に関する研究であろう [鳥海 14]。鳥海らの研究では人工知能の表紙における問題を取り扱い、ユーザーとツイートに含まれる URL の双方をクラスタリングすることによって、情報伝播を可視化している。これにより、どのようなウェブサイトの情報がユーザーコミュニティ間をどのように拡散していったかを明らかにした。しかし、この研究では具体的にツイートしたユーザのみが対象であり、メディアとしてのユーザを介した情報拡散を捉えきれていない。また、これは特定のテーマに関する拡散であるが、どのようなメディアがどのようなセグメントに好まれるかを網羅的に分析した研究はあまり見られない。本研究では 2 次的な拡散を考慮し、特定のテーマ・メディアに限定しない網羅的な分析を行っている。

3. 分析フレームワーク

前述したように、本研究で目指しているのは、メディア-受信者の情報拡散の連鎖の束をすくい上げるようなユーザセグメンテーションである。しかし、ユーザ個人レベルでどういった連鎖の流れがあるかということを追跡するのは粒度が細かすぎて現実的ではない。そこで、本研究ではメディアとしてのユーザが拡散した情報は同様の興味をもつユーザーに対して 2 次的な拡散がなされているという仮定をおき、情報の大本であるメディア (ウェブ情報ソース) から関心を共有するコミュニティに対して情報拡散の “線” が存在するとみなして分析を行うことにする (図 2)。

本研究では、その情報拡散の “線” を取り出すために、内容に URL を含む日本語のツイートデータを用いた。ある URL をツイートしているということは、ツイートを行ったユーザがその URL の表すメディアの情報を 2 次的なメディアとして拡散させていると捉えることができるので、その URL が表すメディアからそのユーザの属するコミュニティへと情報が拡散されたものとみなして、分析を行った。

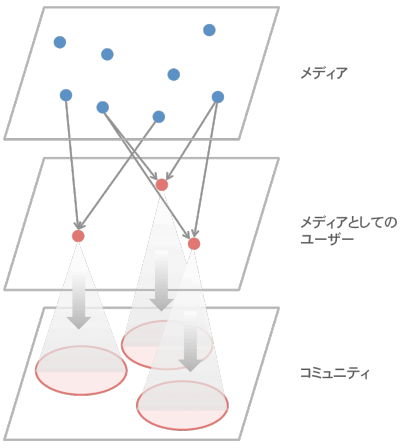


図 2: 今回考える情報拡散の枠組み

3.1 Twitter からのユーザコミュニティ抽出

本節ではユーザをいくつかの興味・関心を共有するコミュニティに分割し、そこから各ユーザーがどういった興味を持つのか、またはどういったグループに所属するのかという特徴を付与することにする。本研究ではコミュニティの抽出に Twitter 上の会話ネットワークを使用することにする。直近でメンションがあったユーザー間にエッジを張ることで会話ネットワークを構成した。多くの分析では、コミュニティ抽出にフォロー関係を用いるが本研究では興味分野の近いコミュニティを抽出したいため、会話ネットワークが適切だと考えた。コミュニティの抽出には Louvain 法を用いることにした [Blondel 08]。

コミュニティの特徴付けには、コミュニティに属するユーザのプロフィール文を用いた。プロフィール文に使用されている語の TF-IDF 値を計算し、値の高い 20 単語をそのコミュニティを表す特徴的な語であるとみなして、各コミュニティを表すキーワードとみなすことにした。

3.2 情報拡散経路の抽出

どういったメディアからどのようなコミュニティに対して情報が拡散しているのかを明らかにしたい。そこで、各コミュニティがどのメディアの URL をツイートしたかという関係データを用いて、コミュニティとメディアの 2 部グラフを作成し、そのグラフに対しネットワーククラスタリングを行うことで、網羅的に関係性の強いコミュニティとメディアを抽出することにした。

具体的には、まず、各ツイートに対して、その中に含まれる URL のドメインを抽出し、そのドメインの表すメディアとツイートを行ったユーザーの所属するコミュニティ間にエッジを張ることでコミュニティとメディアの 2 部グラフを構成していく。

そして、作成したコミュニティとメディアの 2 部グラフに対して、Louvain 法を適用することで、1 度に関係性の強いコミュニティとメディアの混合したグループを抽出した。このようにして得られた、コミュニティとメディアの混合したコミュニティを以後、混合コミュニティと呼ぶことにする。

さらに、2 部グラフのクラスタリングにより出てきた混合コミュニティに所属するユーザの性別・年代的特徴を見るために、分析の対象となる全ユーザの性別・年代を池田らの研究 [池田 11] と同様の手法を用いることで推定し、各混合コミュニティごとに集計することを行った。

3.3 コンテンツに対するコミュニティの反応分析

各コミュニティのコンテンツに対する反応を見るため、今回は朝日新聞デジタルというウェブメディアの記事に着目してコンテンツに関する分析を行った。まず、メディア内の各記事がどのようなコンテンツであるかを特定するために、分析対象となる各記事に対して、LDA を用いてトピック分布を推定した。そして、各コミュニティがどういったトピックに反応しているかを見るために、前節の分析で抽出された混合コミュニティごとに、ツイートした記事のトピック分布の平均値を算出した。そうすることで、各コミュニティごとに反応するトピックの差を明らかにすることができる。式で表現すると、混合コミュニティ i 内のユーザが t 番目にツイートした記事のトピック分布を θ_t とすると、この混合コミュニティのユーザの平均反応トピック I_i は以下のように表すことができる。ただし、N_i は全ツイート数である。

$$I_i = \frac{1}{N_i} \sum_t \theta_t$$

4. 実験と結果

4.1 データセット

全ユーザの 10%をサンプリングし、そのユーザのツイートのうち URL を含むものを分析対象とした。さらに、分析においては、 異常なアカウントを排除するため、対象ツイートで使用されたクライアントのうち、使用回数の多い上位 100 位までに含まれるクライアントによるツイートのみを使用することにした。期間は 2015/9/1 から同年 11/30 の間にツイートされたものを使用した。使用したツイートは 6510 万ツイートで、ユーザの数は 486 万 2609、含まれるメディア数は 2 万 9205 であった。

メディアのコンテンツ分析の際に使用する朝日新聞デジタルの記事としては、対象ツイート内に含まれる朝日新聞デジタルの記事全てを分析の対象とした。記事は URL の情報からスクレイピングすることによって取得した。分析の対象となった記事は 15771 記事であった。

4.2 Twitter からのユーザコミュニティ抽出

会話ネットワークに 1 度だけ Louvain 法を適用したところ 29 万程度のコミュニティが得られた。コミュニティのサイズを適当なものにするため、逐次的にコミュニティのクラスタリングを行い、ユーザ数が 300 以上、10000 以下になるコミュニティに絞ったところ 2717 個のコミュニティが抽出された。本研究ではこの 2717 個のコミュニティを分析の対象とした。

さらに、各コミュニティに対して TF-IDF 値を計算し、キーワード抽出を行った。コミュニティの特徴はそのキーワードから比較的容易につかむことができた。表 1 に抽出したコミュニティとそのキーワードをいくつか示す。多くのコミュニティは (a) 地域の近い高校、(b) 地域の近い大学、(c) 趣味・興味が同じコミュニティの 3 種に大別することができた。

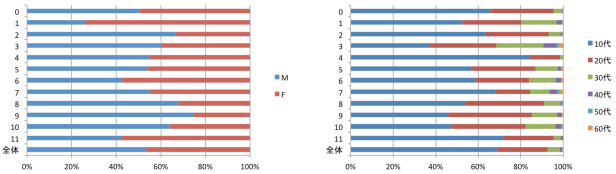
表 1: 会話ネットワークから抽出されたコミュニティ(抜粋)

特徴	キーワード
新潟の高校	新 潟, 長 岡,niigata, 高校,hilcrhyme,instagram
東京の美術系大学	mau, ムサビ,di, 武蔵野美術大学,tzu,tau
トレーダー	トレーダー, 先物, トレード, 投資, スイング, 株式

4.3 情報拡散経路の抽出

コミュニティとメディアの 2 部グラフに対して Louvain 法を適用したところ、12 の混合コミュニティが抽出された。セグメンテーションにおいては比較的解釈のしやすいセグメントに分割され、アニメやキャラクターが好きな層、 俳優や歌手など芸能方面に関心のある層、アニメや鉄道から車やアダルトまで男性的なコンテンツを広く好む層、報道系のメディアを好む層、SNS を強く好む層、バンドやアイドル、お笑いなどに関心のある層などに分かれた。図 3 には各混合コミュニティに関して、所属ユーザの推定された性別・年代割合を載せた。

すべての混合コミュニティについて考察することは、紙面の都合上困難のため、混合コミュニティ4 を例にとる。図 3 から、このコミュニティは年代割合を見ると 10 代の割合が極めて大きいことがわかる。そして、表 2、表 3 には混合コミュニティ 4 に含まれる、ユーザコミュニティとメディアを載せた。それぞれ 2 部グラフ上で同一混合コミュニティのメンバーに対する次数が多い順に表記することとする。メディアの表において、



推定男女割合　　推定年齢割合

図 3: 混合コミュニティの推定された性別・年代割合

タイトルの列にはそのサイトの HTML における TITLE タグを明記している。この結果から、混合コミュニティ4 のユーザには高校生が多く含まれていることがわかり、メディアには Twitter や Instagram といった SNS のサイトが上位に来ていることが確認できる。

表 2: 混合コミュニティ4：コミュニティ

次数	特徴	キーワード
215	ファッション	mer, 自発, マッシュ,sgp, 被写体,target
210	新潟の高校	新 潟, 長 岡,niigata, 高校,hilcrhyme,instagram
202	音楽	乃風,reggae, 湘南, 福井,greeeen, 金沢
160	福岡の高校	genkai,shingu,kashii,fjs, 須恵, 和白丘
155	Youtuber	youtuber,youtube, マホト, カイワレ,mahotonnn, ハンマー
151	熊本の高校	大星, 西原, 長嶺,chr,hyk,kcr

表 3: 混合コミュニティ4：メディア

次数	ドメイン	タイトル
1598	twitter.com	Twitter
1371	instagram.com	Instagram
1354	vine.co	Vine
1328	youtu.be	YouTube
1192	www.monster-strike.com	モンスターストライク (モンスト) 公式サイト
1023	cas.st	モイ株式会社

4.4 コンテンツに対するコミュニティの反応分析

コミュニティごとのコンテンツに対する反応の差を見るため、15771 の朝日新聞デジタルの記事に対してトピック数 50 の LDA を適用し、各記事のトピック分布を推定した。表 4 には得られたトピックの一部を載せ、図 4 には混合コミュニティ 4 の反応トピック分布を載せている。

各コミュニティの反応トピック分布を見ると、全コミュニティが共通して反応しているトピックと反応に差が出るトピックが見られた。「安保」や「憲法」といった当時大きく取り扱われた時事的なニュースにまつわるトピックに対しては、全コミュニティが強く反応していることがわかり、「企業」、「市場」や「ドル」と言った経済的なことに関連するトピックには若者が多い SNS 好きのセグメントはあまり反応しないことなどがわかった。

表 4: 抽出されたトピック（抜粋）

トピックの主要語
法案 国会 安保 反対 安全 憲法 保障 関連 安倍 法
位 大会 戦 選手 試合 リーグ 組 決勝 W杯 サッカー
韓国 大統領 首相 ロシア 会談 首脳 シリア 政府 国連 テロ
消費 税 税率 % 軽減 負担 導入 品 事業 商品
% 経済 中国 化 率 T P P 成長 調査 政府 対策

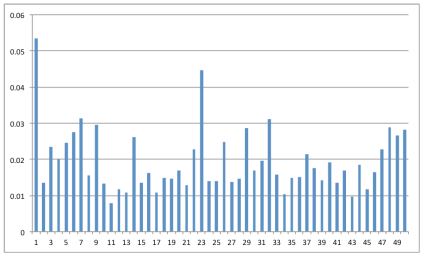


図 4: 混合コミュニティ4 の反応した記事のトピック分布
(横軸はトピック，縦軸はトピックの確率を表す)

4.5 マーケティングへの応用

本論文は、企業が情報を効果的に拡散させる上で適切なユーザセグメンテーションを狙ったものであった。そこで、本節では企業が、これらの知見をマーケティングへどのように応用可能であるかを、前節までの結果を用いて具体的に述べていくこととする。

例えば、SNS を好む若者に対して情報を拡散させたいという状況にあるとする。上で見たように、混合コミュニティ4 がまさに SNS のコンテンツを好み情報拡散を行っていることが確認できる。このことから愚直に考えれば、Twitter や Instagram といった SNS に対して広告を出すなどの施策を行えばいいということになるが、これは自明であり、また、こうしたメディアに広告を出すにはそれなりの資金力が必要であり困難な場合も多い。そのため、他の施策を考えてみることにする。紙面にはスペースの都合上載せていないが、この混合コミュニティ4 は、さらにクラスタリングすることが可能であり、その結果を見ると SNS を好むコミュニティもいくつかに分かれていることが分かる。たとえば、Twitter や Instagram を好む層 (SNS を強く好む層) や、youtube やニコニコ動画など動画メディアを好む層等が抽出されていることがわかる。もし、SNS を強く好む層に対してアクションをしたいとした時には前者のコミュニティに着目すればよく、仮に Twitter や Instagram といった有名メディアへの広告出稿が困難なのであれば、同一混合コミュニティ内に存在しているメディアである Fashionsnap.com であったり、ライブドアブログであるといったサイトに広告を出すということがターゲットに対して効果的に情報を届ける上で有効となることが推察される。また、最初に抽出された混合コミュニティ4 に所属するユーザ全体に広く情報を拡散させたいのであれば、2 度目の分割で得られて混合コミュニティごとに広告を打ち分けていくことも有効な施策となるだろう。

また、こうしたセグメンテーションはこのように広告を効果的に打ち分けることだけに有用なわけではない。これは広告のクリエイティブを考える際にも有用な情報となる。たとえば、先ほどの例で言えば SNS を好んでいる層はファッションに関しても関心を持っている可能性が高いことが分かる。すると、こうした情報は実際の広告を作る際のクリエイティブを考える際にも効果的な視点をもたらしてくれるであろう。

5. まとめ

本論文では、Twitter のデータを用いて、企業のウェブ戦略上有用であると思われる、ソーシャルメディアの普及した現在における情報拡散に適したユーザセグメンテーションを行った。

さて、最後に本論文で抽出してきたものを少し別の角度で考察したい。もちろん、情報拡散の経路ではあるのだが、他にどのような見方ができるのかを考えてみたい。本研究ではメディアとそこから情報拡散していくユーザーを混合コミュニティという形で抽出した。最初に述べたように、情報を拡散させたユーザもまたメディアである。すなわち、混合コミュニティとは似た興味を持つユーザに好まれるメディア群 (情報ソース) と、そのメディア群の情報を好み、拡散させていく力も持つメディアとしてのユーザの集合体ということができる。つまり、混合コミュニティとは似た興味を持つユーザが好む情報を発信する情報源の集合だということができ、混合コミュニティは 1 つの“メディア”であると捉えることが可能である。この意味において本稿で抽出したものは、ソーシャルメディアにおける 12 個の“メディア”と捉えられるだろう。

本研究で示した手法を用いることで、企業がより有効なウェブ戦略を立てられるようになれば幸いである。

参考文献

[Blondel 08] Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E.: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, pp. 10008-10019 (2008).

[Java 07] Java, A., Song, X., Finin, T., and Tseng, B.: Why we twitter: understanding microblogging usage and communities, In WebKDD/SNA-KDD ’07 Proceedings of the 9th WebKDD and 1st SNA-KDD, pp 118-138 (2007).

[Marui 14] Marui, J., Nori, N., Sakaki, T and Mori, J.: Empirical Study of Conversational Community Using Linguistic Expression and Profile Information, In Active Media Technology, Vol. 8610, pp. 286-298 (2014).

[Ozer 01] Ozer, M.: User segmentation of online music services using fuzzy clustering, Omega 29, pp. 193-206 (2001).

[Zhou 06] Zhou, Y. K., and Mobasher, B.: Web user segmentation based on a mixture of factor Analyzers, In Proceedings of the 7th International Conference on E-Commerce and Web Technologies (EC-Web ’06), pp 11-20 (2006).

[鳥海 14] 鳥海不二夫, 榊剛史, 岡崎直観.: 「人工知能」の表紙に関するツイートの分析・続報, 第 4 回 Web インテリジェンスとインタラクション研究会 (2014).

[池田 11] 池田和史, 服部元, 松本一則, 小野智弘, 東野輝夫.: マーケット分析のための Twitter 投稿者プロフィール推定手法, 情報処理学会論文誌コンシューマ・デバイス&システム, pp. 82-93 (2011).

[総務省 15] 総務省.: 平成 27 年版情報通信白書 (2015).