ネットワーク構造に基づく新聞記事の分類

○園田 亜斗夢　鳥海　不二夫（東京大学）中島 寛人 郷治 雅（日本経済新聞社）

Classification of Articles in a News Service from Network Structures

* Atom Sonoda and Fujio Toriumi (The University of Tokyo)

Hiroto Nakajima　Miyabi Gouji (Nikkei Inc.)

概要－　情報が電子媒体で発信されるようになり，情報の即時性や情報量の増加が進んでいる．その中で，読者が情報を選択する際の労力は増加しており，そのような負担を減らす推薦サービスの導入も進んでいる．一方で，過度な推薦により，ユーザに偏った情報のみを提供するフィルターバブルが発生しているとの指摘もある．本研究では，推薦に先立つ記事の分類について，記事を読んだユーザによるネットワーク構造から記事を分類し，テキスト情報や既存のタグから分類するより，より興味に基づいた分類ができることを示した．

キーワード: 記事，トピック分類，ネットワーククラスタリング

1　はじめに

新聞社の発信するニュースの主な媒体が新聞からwebサイトに拡大するに従い，記事の速報性の向上や記事数の増加が進んでいる． これに伴い，ユーザが記事を選択する際に必要な時間と労力が増大していると推測される．このような負担を減らし，満足度を向上させるため，ユーザの嗜好に応じた推薦サービスは多くの分野で導入されている。一方で，過度の推薦によってユーザに偏った情報のみを提供するフィルターバブルが発生しているとの指摘もある[1)]

記事の推薦のためには，記事の適切な分類が必要である．本研究では，新聞社の1つである日本経済新聞社のwebサービス日経電子版に注目し，記事の分類手法を提案する．記事の分類には，既に日経電子版にはジャンルや産業といった分類がなされている．一方で，本研究では鳥海らがtwitter投稿を分類する際に提案した手法[2)]に基づき，言語情報を利用することなしに，ユーザの行動履歴，つまり記事を閲覧したか否かのみに注目し，日経電子版の記事の分類手法を提案する．すなわち，ユーザの興味を反映した記事の選択という行動に注目することで，記事をネットワーク化し，内容の類似性がある記事のクラスタリングを行う．実際の推薦システムでは，ユーザの行動履歴を閲覧した記事のクラスタの割合を分析し，それに基づき記事を推薦し，結果，読む記事のクラスタの割合が変化するかどうかを観察することを目指す．そのために本研究では，将来的なフィルターバブルの発生を抑制し，多様な意見に触れる機会を提供できるような推薦システムの開発に向け記事のクラスタリングを行う．

2　分析に用いるデータ

本研究では，日本経済新聞社のwebサービス日経電子版のデータを用いる．まず，2017年5月21日から８月20日までの3ヶ月間に日経電子版の全会員がweb上で読んだ記事のデータを収集した． この間に読まれた記事は全部で約60万記事であり，会員数は約200万人であった．この期間の記事ごとの閲覧数と，ユーザごとの閲覧数をFig. 1， ２に示す．

この中で，一定程度以上読まれている記事のみを扱うため，今回は読者が100人以下の記事は対象から除いた．これは，将来的な推薦システムの開発には，ニュースという即時生の高いコンテンツであるという性質上，24時間以内の記事を優先して推薦するため，定期的な更新が必要となり，計算時間の観点からデータを減らす必要があることと，推薦ではフィルターバブルを抑制するという目的から全体のユーザ数に対し極端に読者の少ない特殊な記事を推薦することは考えていないためである．分析対象となるデータはこれらの処理を行った約7万記事である．

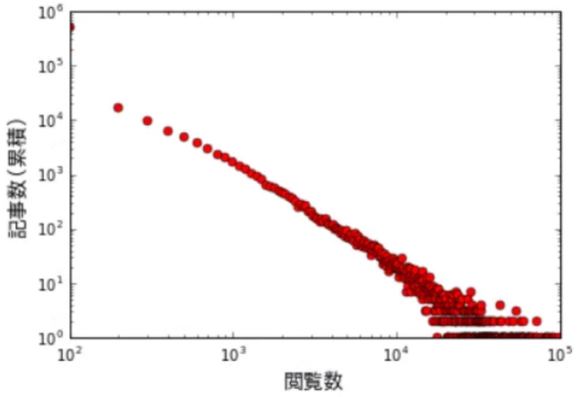

Fig. 1:記事ごとの閲覧数

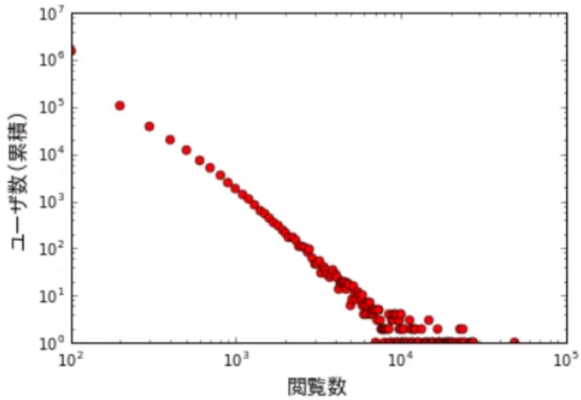

Fig. 2:ユーザごとの閲覧数

3　関連記事ネットワークの構築

ある２つの記事を読んだユーザが複数人いた場合，２つの記事は共通した興味を引く内容を有していると

第15回社会システム部会研究会，第59回システム工学部会研究会（2018年3月14日-16日・沖縄）

考えられる．つまり，記事を読んだユーザの重複度から記事の類似性を求めることができる．そこで，ユーザの重複度の高い記事同士をリンクで結ぶことで，記事ネットワークの構築を行う．

記事間類似度の算出

２つの記事a_i，a_jのユーザ群U_i，U_jの重複率はSimpson係数を用いて次のように求められる．

$$Sim(a_i, a_j) = \frac{|U_i \cdot U_j|}{\min(|U_i|, |U_j|)}$$

なお，このような類似度を測る指標としては，Simpson係数のほかにJaccard係数，Dice係数などがあるが，共起を用いた関係性の強さを表現するための指標としてはSimpson係数が適切であるとされている[3]．

ネットワークの構築

ここでは，前述の類似度$Sim(a_i, a_j)$が閾値th以上の記事の間にリンクを張ることで，重みあり無向ネットワークを構築した．実際には，まず，類似度$Sim(a_i, a_j)$をすべての記事ペアについて算出し，得られたペアのうち類似度が閾値th以上の記事ペアを抽出する．そして，得られた記事ペアの間にリンクを張ることで構築される．

4　記事のクラスタリング

得られたネットワークについてコミュニティ抽出を行い，記事の類似性に基づいたクラスタを獲得する．前章で得られた重みつきネットワークに対し，ネットワーククラスタリングの手法を適用する．クラスタリング手法としては，モジュラリティを基準とするLouvain法[4]を用いた．モジュラリティとは各クラスタの結合度合いを表す指標であり，コミュニティ間のリンクが疎であるほど高い値を出す指標であり，モジュラリティを最大化することで，結合度の高いコミュニティを抽出する．

記事間類似度の設定

ここでは，関係性の少ない２つの記事，すなわちSimpson係数の小さいリンクの影響を排除するため，前章の類似度Sim(a_i,a_j)が閾値th以上のリンクのみを利用することとした．また，ほかの記事とリンクで結ばれていない記事すなわち独立したノードは，今回の分析対象から除外した．ここでは，0.31≦th≦0.97で変化させた．

このとき，クラスタリング結果を評価するために，モジュラリティを用いた．これは，適切なパラメータを選んだ場合，モジュラリティが大きくなりすなわち結合度の高いコミュニティが抽出できることで，記事の分類が適切に行われると考えるためである．

クラスタリング結果

クラスタリング結果の評価をFig. 3に示す．閾値th

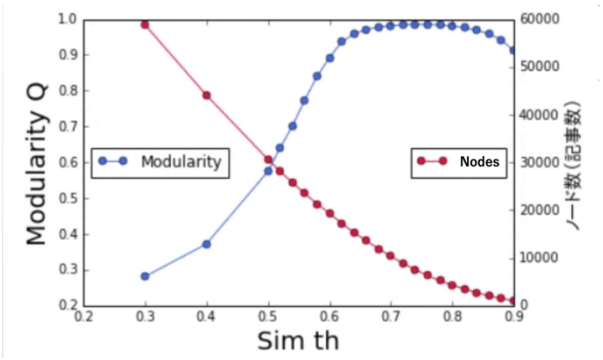

Fig. 3:類似度の閾値 th によるモジュラリティ Q の変化

が上がるに従い，モジュラリティは増大するが，一定の th 以降は減少することがわかる．また，一方で，閾値 th が上がるに従いネットワークに含まれるノードである記事数（Node News）が減少することもわかる．これより，th=0.7464 のとき，モジュラリティ Q=0.98562 と最大値となったが，．このとき，リンクは 87201842 件，ノードは 75071259 件となり，元の約 7 万記事から大幅に減少し，また，ネットワークの平均次数も 1.16 となるため，本研究の目的である推薦で使うには不適切である．そこで，モジュラリティと記事数の関係から，モジュラリティが 0.9 を超えているなかから，th=0.62 を選び分析した．このとき，モジュラリティ Q=0.936，リンクは 38002 件，ノードは 17233 件であった．

このとき，2470 件のクラスタが得られた．得られたクラスタ毎の内容を記事のタイトルとカテゴリによって確認した．これは，記事をクリックして初めて本文を読めるため，ユーザが興味に従い閲覧するかどうか決める際にはタイトルで判断していると推測され，タイトルがより興味を反映していると考えられるためである．結果は，クラスタ毎にある程度記事の属性が限定されており，コミュニティ抽出がおおむね成功したと言える．しかしながら，トップ記事に表示されている記事はクリック率が高く，トップ記事に表示されていたその時間帯に日経電子版を利用したユーザの多くが読んでいることから，ユーザの興味によらない部分があり，トップ記事をまとめたクラスタのほかに，選挙記事などのトップ記事になりやすいものはすべてのクラスタに分散していた．得られたクラスタのうち，ノード数の多い方から 10 個のクラスタについて Table 1 に示す．これらのクラスタには，日経電子版に元からあるジャンルや産業は異なっているが，例えば東アジアなど同じトピックで繋がっていると判断できる記事などが多数含まれていた．このことから，記事のジャンルを超えたユーザの興味に基づいた分類ができたことが確認できた．

ノード数の下位のクラスタについては，より内容に類似性が認められた．例えば，

Table 1:抽出されたクラスタ

記事数	主な内容
673	トップ記事
646	外交・経済
497	IT 関連
447	有名人コラム＋犯罪
380	国際・産油国
363	東アジア
320	国際・アメリカ
304	景気
289	東アジア・経済
265	経済・マーケット

Table 2:LDA によるトピック分類された単語の抜粋

ラベル	トピックを構成する単語
東京市場	円，株，反発，東証，反落
市場情報	益，６月，%，期，純利益
石油	%，増，ドバイ，５月，原油
新興市場	米，新興株，北朝鮮，ジャスダック
プロ野球	プロ野球，連敗，目，連勝，巨人
テニス	テニス，強，審査，ウィンブルドン
情報通信	情報，市場，人事，通信
ゴルフ	ゴルフ，松山，男子ゴルフ，位
築地問題	エレクトロニクス，豊洲，移転
ベンチャー	VB, 仮想通貨，米，京セラ

・人事情報
・ゴルフコラム
・マーケット情報
・書評

といった内容を含むクラスタが存在した. このように，コラムや書評といった通常のニュースとは違った情報をまとめたクラスタも存在した．これは，日経電子版に人事情報をまとめて見る機能などがあることなどから, 人事情報を多く見る層といったユーザ集合があり，その興味に従いクラスタリングされた結果だと考えられる．また，コラムなどに関しては，日経電子版の機能でタグというものが存在し，同様のコラムがリンクされていることに由来すると考えられる．

5　他のクラスタリング手法との比較

ここでは, ネットワーククラスタリング比較対象として採用したクラスタリング手法として，言語情報を用いてクラスタリングを行う LDA[5]と，記事ごとの読者ベクトルを基にした K-means 法 [6]を用いてクラスタリングを行い，ネットワーククラスタリングの結果と比較した．

LDA によるクラスタリング

コーパスの作成には, ネットワーククラスタリングの対象となった記事と同じ約 7 万記事のタイトルを用いた．それを基に LDA モデルを生成し，その後，記事ごとにタイトルを基にその記事のトピックを LDA モデルで推定した．ここで，トピック数は 50 とした．これは，トピック数 50 の時, 各トピックに含まれる記事が 782 から 5019 と比較的偏りなく分類できたためである．

Table 2 に，LDA により得られたトピックに含まれる記事数が多い順から 10 のトピックについて, トピックとそれを構成する単語の抜粋を示す．ただし，トピックは上位の単語より人間の手によって付与した．これらのトピックに含まれる記事に関して確認したところ，それぞれラベルに代表されるような記事が含まれていた．また，ジャンルやカテゴリなども似たものが集まっており，例えば東京市場というトピックでは反落や反発という言葉でつながりのある海外市場の為替や先物取引などが含まれており，市場情報では個々の企業に関する決算情報などが多くなっていた．石油に関しては，テキスト情報を基にクラスタリングするという LDA の特徴から，タイトルで確認した際にはネットワーククラスタリングよりよくまとまっていることが確認された．

記事のタイトルは文章としては短いが, 記事の内容が端的にまとまっており，また，同じジャンルの記事は新聞独自の文法により同じような形式と単語で記されているため，テキスト情報に基づき分類した場合，極端に同じ内容のものがまとまりやすい傾向があると推測される．例えば，ラベル東京市場に含まれる記事であれば，「ＮＹ円、続落　１ドル＝112 円 30～40 銭、１カ月半ぶり安値 対ユーロの売り波及」のように「市場」「通貨」「市場動向を示す単語」「為替レート」「考察原因」などを書くというフォーマットが決まっており，特徴となる単語が普通の文章より現れる確率が高いと言える．

Table 3 に東京市場トピックと石油トピックに含まれる記事の抜粋を示す．ここで選んだトピック以外に関しても，含まれる記事にはトピックごとの決まったフォーマットが確認された．本研究の分析対象は日経新聞の記事に限られるため, この傾向が顕著であった．このことは LDA によるクラスタリングの利点であり，かつ欠点でもある．つまり，クラスタリング結果は日経電子版に元からあるジャンルや産業と非常に近いものとなり，タイトルで判断した場合は非常によくまとまっているという印象を得るが, 本研究の目的である，多様な意見に触れる機会を提供できるような推薦システムの開発には寄与しないと考えられるからである．興味に基づいた記事の分類という観点からは，ジャンルを超えて分類が可能なネットワーククラスタリングの方が優れていると言える．

K-means によるクラスタリング

K-means によるクラスタリングにも, ネットワーククク

Table 3:トピックごとの記事の抜粋

東京市場
外為 17 時　円続伸、109 円台前半　一時 109 円 22 銭、１カ月半ぶり高値
ＮＹ債券、続落　10 年債利回り 2.20%、２年債利回りは１カ月ぶり高水準
ＮＹ債券、長期債小反発　10 年債利回り 2.15%、低調な米経済指標受け買い
外為 17 時　円、３日続落し 111 円台後半　一時約１カ月ぶり 112 円台
ＮＹ円、続落　１ドル＝112 円 30～40 銭、１カ月半ぶり安値 対ユーロの売り波及
石油
４月の建機出荷額 23%増　中国など堅調で６カ月連続プラス
４月の自動車生産、前年比 16.3%増の 74 万 9087 台　６カ月連続増
５月の軽含む新車販売、前年比 12.4%増　７カ月連続プラス
５月のマネーストック、「Ｍ３」は前年比 3.4%増 「Ｍ２」3.9%増
５月の中国新車販売、トヨタ 9.6%増　日産は 5.7%増

ラスタリングの対象となった記事と同じ約 7 万記事の読者の情報を用いた．ここで，K-means における重心の更新には，消費メモリと計算量を削減する目的からクラスタに属する記事の読者について，クラスタ内の半数以上の記事を読んでいる読者の集合を重心とした．以下に，今回用いた K-means のアルゴリズムを示す．

1. クラスタ数 k の分だけ初期値となる値を決める．この際，初期値は約 7 万記事の中からランダムに選ぶ．
2. 重心との距離 d に基づき，約 7 万記事の属するクラスタを更新する．距離 d は２つの記事a_i, a_jのユーザ群U_i, U_jにより以下のように求められる．

$$d(a_i, a_j) = \frac{U_i \cdot U_j}{|U_i||U_j|}$$

3. クラスタに属する記事の読者について，クラスタ内の半数以上の記事を読んでいる読者の集合を重心として，クラスタの重心を更新する．

以上のような手順でクラスタリングを行なった．ここで，クラスタ数は LDA の時と同じ 50 とした．この時，各クラスタに含まれる記事数は 243 から 6585 となった．

得られたクラスタのうち，ノード数の多い方から 10 個のクラスタについて Table 4 に示す．この表から，トップ記事が多いことがわかる．これはうまく分類できていないことから，人目でラベルをつけた際にトップ記事としたことによる．しかし，ノード数の少ないクラスタの中には人事やスポーツなど限られたジャンルでよく分類できていたものもあった．

Table 4:抽出されたクラスタ

記事数	クラスタの内容
6585	トップ記事
5317	不明
5237	企業業績
4684	アパレル・家具
3739	市場
3204	東京市場
2643	企業業績
2583	スポーツ
2536	研究開発
2355	地方

そこで，各クラスタの重心を分析することで，どのような特徴があるかを確認したところ，どの重心にも 0 から 103 ユーザしか含まれていなかった．特に，重心に含まれるユーザが 0 から 3 の場合は偶然，特定のユーザが読んだ記事が単に集まっただけで，内容のつながりが強い記事ではないことから，内容のまとまりが少ないクラスタであった．また，103 ユーザが含まれたクラスタは含まれる記事数も 6585 と最も多いトップ記事の集まったクラスタであった．内容のまとまりが認められたクラスタに関しても，このクラスタリングでは非常に限られた数十ユーザの興味に基づいて分類されていたということになる．この傾向はクラスタ数を変化させても変わらなかった．これは，重心の更新をクラスタ内の半数以上の記事を読んでいる読者の集合としたことに起因すると考えられる．また，各ユーザの読む記事数は第二章で見た通り限られており，また記事ごとの読者数も限られていることから，同じ興味を惹く記事であってもそのうち半分以上を読んでいる読者はほとんどいないと考えられる．

このことから，単純に読者と記事の組み合わせから k-means により記事を分類することは困難であり，たとえうまく分けられているように見えるクラスタであっても限られたユーザの興味にしか基づいていないため，多様な意見に触れる機会を提供する推薦システムの開発といった観点からは，不適切であると言えよう．

6　不適切なクラスタの特定

前章までで，ネットワーククラスタリングによる分類が有効であることが明らかになったことを受けて，本章ではネットワーククラスタリングによって得られたクラスタの特徴について分析する．

本研究では，ユーザの類似度によってネットワーク

を構築し，それを基にクラスタリングを行ったため，クラスタ毎の記事を読んだユーザの分布はクラスタ毎に異なっていると考えられる．また，興味が同じユーザは同じカテゴリを見ると期待されるため，クラスタ毎の記事の属性もある程度一致していると予測される．そこで，クラスタ毎の記事の特徴，ユーザの特徴をそれぞれ集計し，クラスタ毎のユーザの特徴を明らかにした．それにより，偏りが大きく記事の推薦には不向きな記事の含まれるクラスタを特定した．

クラスタ毎の読者の属性の特徴

クラスタ毎の記事の読者の分布を，性別，年齢，職業によって集計した．まず，クラスタリングの対象となるネットワーク含まれる記事を読んでいたユーザは約200万人であった．それ以外の記事は，類似度の閾値によりネットワークとのリンクが切れたノードでクラスタリングの対象から外れた．

なお，ユーザの性別，年齢，職業毎の割合の事前分布は，日経新聞社による公開情報[7]によると， Table 5のとおりである．性別に関しては男性が80%と8割を占め，年齢に関しては40代が最も多く，職業に関してはお勤めが71%となっていることがわかる．これは，日本経済新聞特有の特性によるものだと考えられる．

これらの属性の分布に関して，ユーザ属性の集中度，すなわちユーザ属性エントロピーを計算した．ここでは，属性は性別，年齢，職業の交差属性を考えた．ユーザ属性エントロピーが高いクラスタとは，様々な属性の読者が読んでいるクラスタであり，対応する興味が幅広いと考えられる．p_iを各クラスタにおけるユーザ属性iの存在確率として，ユーザ属性エントロピーは以下のように表される．

$$H(A) = -\sum_i p_i \log p_i$$

各クラスタ(クラスタに含まれる記事を読んだ固有のユーザN人)におけるエントロピーはp_i=(属性iのユーザ数)/Nとして計算した．得られたクラスタに含まれる記事17233件全体でのユーザ属性エントロピーは1.209であった．クラスタに含まれる記事が30以上のクラスタ65件に関して分析したところ，Fig. 4のようになった．平均が1.11で全体でのユーザ属性エントロピーと比べ減少していることがわかる．その中でもユーザ属性エントロピーが0.934，1.023と小さくなったクラスタの記事を確認したところ，それぞれゴルフのコラムと私の履歴書と呼ばれる著名人の自伝風連載であった．これより，フィルターバブルを避けて速報性の高い記事を推薦するといった目的に使うにはユーザ属性エントロピーが低いものは避けた方が良いと考えられる．

クラスタ毎の読者の集中度

次に，各クラスタに含まれる記事について，記事間のユーザの集中度を分析した．そのために，各クラスタ（クラスタ内全閲覧数N）におけるエントロピーをp_i=(ユーザiの閲覧記事数)/Nとして上記の式に従いユーザエントロピーを計算した．ユーザエントロピーが

Table 5:ユーザ属性の割合

男性	80%	70代以上	5%
女性	20%	60代	13%
お勤め（会社員、公務員など）	71%	50代	23%
自営・自由業	10%	40代	25%
無職	7%	30代	19%
学生	9%	20代以下	15%
主婦（パート含む）	3%		

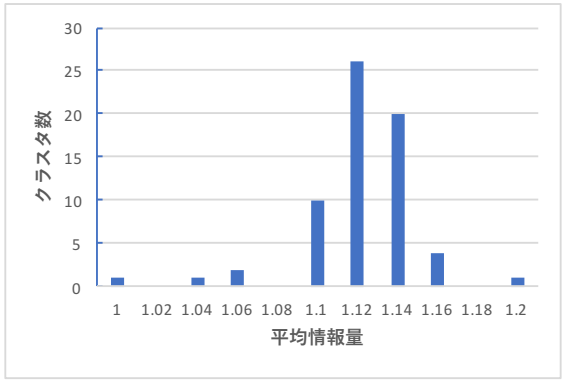

Fig. 4:ユーザ属性のエントロピー

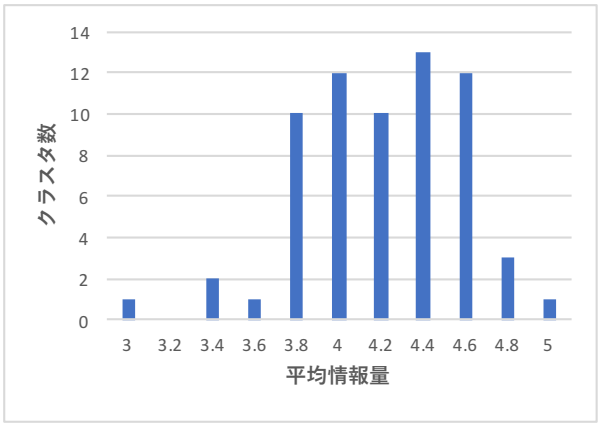

Fig. 5:ユーザのエントロピー

高いクラスタは，クラスタ内の記事を読んだユーザがクラスタ内の他の記事もよく読んでいるというクラスタであり，この値が低ければクラスタ内の記事を読んでいるユーザの内の一部のユーザによりネットワークのリンクが貼られていたということを表す．つまり，5章のK-meansによるクラスタリングと同様に，限られたユーザの興味に基づいてまとめられたクラスタであると考えられる．

得られたクラスタに含まれる記事17233件全体でのユーザエントロピーは5.435であった．クラスタに含まれる記事が30以上のクラスタ65件に関して分析したところ，Fig. 5のようになった．

ユーザエントロピーが2.52，3.56，3.67と小さくなったクラスタの記事を確認したところ，それぞれ私の履歴書と春秋と呼ばれるコラムと決算短信や選挙の速報記事など数行しかない記事によって構成されるクラス

Table 6:抽出されたクラスタ

ユーザ属性エントロピー	ユーザエントロピー	クラスタの内容
1.183	3.67	春秋
1.158	4.01	アメリカ関連
1.147	4.25	経済・景気
1.141	3.92	テクノロジー
1.131	4.60	国際政治
1.124	4.62	アジア・中東
1.124	4.93	中国＋医療
1.097	4.66	西日本＋スポーツ

タであった.

エントロピーの大きいクラスタ

ユーザエントロピー，ユーザ属性エントロピーの高いクラスタの代表をTable 6に示した．ユーザエントロピー，ユーザ属性エントロピーがともに高いクラスタに関しては内容も良くまとまっており，また，記事の内容もコラム等に偏っておらず，推薦に適したクラスタであると考えられるものであった．表に載せたクラスタのうち唯一，春秋というコラムがまとまった推薦に不適切なクラスタが存在したが，これは，ユーザ属性エントロピーは高いが，ユーザエントロピーが低いと判断でき，両方のエントロピーのいずれか一方でも低いと,推薦に不適切なクラスタであることがわかる．

考察

ユーザエントロピー，ユーザ属性エントロピーいずれかが低いクラスタに関しては，一般的な記事が少なく，数行の記事やコラムが多く含まれることが確認された.

4章で確認したように各クラスタには属性が類似した記事がまとまっており，コミュニティの抽出はある程度成功していることを確認した．それに加え，本章ではユーザの集中度という観点からクラスタの特徴を分析することで，フィルターバブルを抑制する推薦といった目的には適さない偏りの大きいクラスタを特定することが可能であることを確認した.

7　結論

本研究では，日経電子版の web 上の記事について，ユーザの閲覧記録に基づいて，類似度ネットワークを構築し，ユーザの興味に基づいた記事のクラスタリングを行なった．それぞれのクラスタは経済や政治，金融など内容ごとに情報がまとめられていることを確認した．また，クラスタ毎のユーザの集中度，ユーザ属性の集中度を測定することで，クラスタの特徴を明らかにし，推薦に適切なクラスタの選定が可能であることも示した.

今後の課題は，クラスタリングの妥当性の分析，これらの結果を利用した推薦システムの開発などが挙げられる．本研究結果から，ユーザの興味に基づいた記事の分類ができるということが確認されたため，推薦システムの開発においては，読者が読んでいる記事の属するクラスタにより読者の興味を推定し，それから読者の持つ興味とは異なった記事を推薦することで，フィルターバブルの発生を抑制し，多様な意見に触れる機会を提供できるような推薦システムの開発が可能になると期待される.

参考文献

1) Pariser, Eli. *The filter bubble: What the Internet is hiding from you*. Penguin UK,(2011).
2) 鳥海不二夫, and 榊剛史. "バースト現象におけるトピック分析." 情報処理学会論文誌 58.6 : 1287/1299(2017).
3) 松尾豊, et al. "Web 上の情報からの人間関係ネットワークの抽出." 人工知能学会論文誌 20.1 : 46/56 (2005).
4) Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 : P10008 (2008).
5) Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan: 993/1022(2003).
6) MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14.(1967).
7) http://ps.nikkei.co.jp/adweb/download/index.html 日経 Web 広告ガイド 日経電子版 メディアリポート 2017 年 8 月 (2017)