F-002

潜在的ディリクレ配分法に基づく協調フィルタリングを用いたマイクロブログユーザの興味対象分析

Analyzing User Preference Using Collaborative Filtering with Latent Dirichlet Allocation

渡邊 恵太 †　　加藤 昇平 †
Keita Watanabe　Shohei Kato

1　はじめに

近年，Twitter に代表されるマイクロブログが急速に普及している．マイクロブログでは，投稿できる文が短く制限されているためユーザは手軽に投稿することができるとされている．そのため，マイクロブログには現実社会で起きているイベントの情報がリアルタイムに投稿されており，今社会で何が起きているかの情報を即座に入手することができる．一方で，次々と投稿される膨大な情報の中からユーザが興味を持つ情報を手作業でリアルタイムに抽出することは困難であり，マイクロブログに存在する有用な情報が十分に活用できない問題が存在する．したがって，マイクロブログに投稿されたリアルタイムな情報の中からユーザが興味を持つ情報を選別・抽出しユーザに提示するシステムが必要である．このようなシステムの実現には，ユーザの興味対象を明らかにすることが重要である．

ユーザの興味対象を明らかにする方法には，ユーザ自身に興味対象を表す語を提示させることがある．しかし，通常ユーザは複数の興味を持つことからその全てを挙げることはユーザにとって負担が大きい．また，単語一つを挙げるだけでは，語の言い換えや略語，同義語などに対応できない．そこで，ユーザの投稿を tf·idf 法を用いてユーザの興味対象を表す語を分析する手法がある．また，ユーザ自身の投稿だけでなくユーザの友人の投稿も関連付けて tf·idf 法によりユーザの嗜好分析を試みた研究がある [1]．この手法でも，単語単位での分析のため単語間の関係を考慮できない．また，Twitter の投稿を潜在的ディリクレ配分法などのトピックモデルを用いてトピックという単位でその構成を推定する研究がある [2, 3]．ユーザが興味対象に関する内容を多く投稿すると仮定すれば，ユーザの投稿から潜在的ディリクレ配分法を用いて推定されたトピックの分布はユーザのトピックに対する興味の強さを表すと考えられる．一方で，潜在的ディリクレ配分法に基づき推定されたトピックの分布は，あくまで文書中に存在する単語の構成によって推定されたものである．したがって，トピックの生起確率が低いことは，文書にほとんどそのトピックが現れないことを意味するが，ユーザの興味対象分析においては，必ずしもユーザがそのトピックに興味が無いことを意味しない．つまり，ユーザがそのトピックに興味が無いから投稿していないのか，マイクロブログ上では投稿していないだけなのか判断できない．しかし，情報推薦を行う上ではユーザの投稿内容だけに限らず興味対象を推定できることは，推薦の範囲を広げることができ，重要である．

そこで，本研究では Twitter を対象とし潜在的ディリクレ配分法に基づきユーザのトピックの分布を推定し，そのトピック分布を用いて協調フィルタリングによりユーザの興味を推定する手法を提案する．潜在的ディリクレ配分法を用いることで，ユーザの興味対象を単語ではなくトピックという単位で推定することができる．また，協調フィルタリングによりユーザが投稿していなくても興味対象としてトピックを推定することができる．事例では，実際に Twitter ユーザを対象として提案手法による興味対象の分析を行った結果を示す．

† 名古屋工業大学, Nagoya Institute of Technology

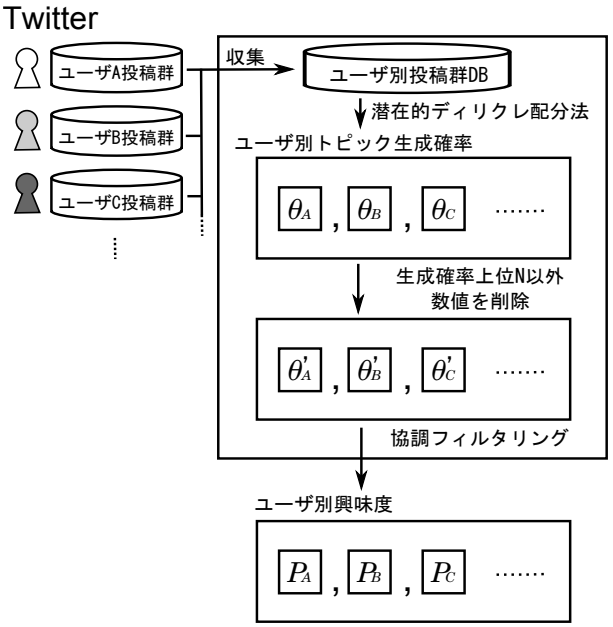

図 1　概要図

2　提案手法概要

本研究で提案する手法は，Twitter ユーザの投稿からトピックという語の集合単位でユーザの興味対象を推定する手法である．トピック単位という単語よりも話題としてわかりやすい単位で推定が可能であり，また協調フィルタリングを用いることで，ユーザが投稿していない内容であっても興味対象として推定することができる．

提案手法の概要を図 1 に示す．以下に提案手法の流れを説明する．提案手法では，まず Twitter API を用いて Twitter からユーザの投稿を収集しユーザ毎の投稿群を集めてデータベースを作成する．収集対象とする投稿は，ユーザの過去の全投稿であるが，Twitter API の制約上，最大 3200 投稿までしか収集できないため，それ以上の投稿を行なっているユーザについては最新の投稿から 3200 投稿を収集対象とする．次に，収集した各投稿を形態素解析する．ここで，本稿では興味を表す語の品詞として名詞が適切であると考え，分析対象語を名詞に限定した．よって，形態素解析の結果品詞が名詞である語だけを抽出する．形態素解析には，MeCab の Java 移植版である Sen を用いた．また，宮城ら [4] にならい，形態素解析器の辞書に Wikipedia のページ名として登録されている語を名詞として追加した．各ユーザに対して，収集した投稿に含まれる名詞を抽出し，投稿群としてまとめる．このとき，Zhao[2] らにならい全文書の 70 %以上に出現する名詞，つまり収集した全ユーザの 70 %以上が使用していた名詞については，作成した投稿群

から除外する．これは，多くのユーザに用いられる語は一般的な名詞であり，興味対象の分析において重要ではないと考えられるためである．一人のユーザの投稿群を一つの文書として，潜在的ディリクレ配分法により各ユーザのトピック分布 θ_u を推定する．ここで，θ_u はユーザ u のトピック分布を表す．ここで得られたユーザごとのトピック分布は，ユーザの投稿内容に基いており，ユーザが多く投稿しているトピックはユーザが興味を持つトピックであると仮定のもと，ユーザの興味の強さを表していると考えられる．潜在的ディリクレ配分法については 3 章において詳しく説明する．次に，トピック分布を用いて，協調フィルタリングを適用するが，その際，ユーザ毎にトピックの生起確率上位 N トピックの生起確率を用いて協調フィルタリングを行う．ここで，ユーザ u のトピック分布 θ_u から生起確率上位 N 以外のトピックの生起確率を削除したものを θ'_u とする．潜在的ディリクレ配分法をによって表現されたトピック分布では，全てのトピックに対して生起確率が与えられており，生起確率を協調フィルタリングによって推定する必要はない．しかし，ユーザの興味対象となるトピックを推定する場合においては，トピック分布があくまでユーザの投稿内容に基づいているため，トピックの生起確率が低いことはユーザの投稿にそのトピックがあまり出現しない，ことを表すが，ユーザの興味が無いのかあるいは単に Twitter 上で投稿していないだけかわからない．したがって，生起確率が低いトピックに対しては協調フィルタリングによって興味対象となり得るかを判定する．θ'_u をもとに協調フィルタリングによってユーザのトピックに対する興味の強さである興味度 P_u を推定する．ここで，P_u はユーザ u の各トピックに対する興味度の配列である．協調フィルタリングによって，ユーザと似通ったトピック分布を持つ，つまりユーザと同じトピックについて多く投稿をしているユーザが好んでいる他のトピックについてもユーザの評価値の推定値が与えられることになる．したがって，ユーザが投稿をしていないトピックであっても興味対象として推定することが可能となる．協調フィルタリングについては 5 章において詳しく説明する．

3 潜在的ディリクレ配分法

本研究では，ユーザの投稿上の潜在的なトピックの推定に潜在的ディリクレ配分法（Latent Dirichlet Allocation）[5] を用いる．潜在的ディリクレ配分法は，文書の生成過程を確率的にモデル化したトピックモデルの一つであり，一つの文書中に複数のトピックが存在することを表現できる潜在的意味解析手法である．潜在的ディリクレ配分法によって，文書を構成するトピックの多項分布，各トピックを構成する単語の多項分布を表現することができる．本稿では，文書を Twitter の一ユーザの過去の投稿をまとめた投稿群としてトピック分布，単語分布を推定する．これにより，ユーザがどのようなトピックをどのような分布によって投稿しているのか得ることができる

図 2 に潜在的ディリクレ配分法のグラフィカルモデルを示す．ここで，θ は文書におけるトピックの多項分布，ϕ はトピックにおける単語の多項分布を表し，z は語 w に割り当てられたトピックを表す．また，K はトピック数を，D は文書数を，N は単語数を表している．α と β はそれぞれトピック分布 θ が従うディリクレ分布のハイパーパラメータ，単語分布 ϕ が従うディリクレ分布のハイパーパラメータを表している．潜在的ディリクレ配分法において，各文書はそれぞれトピック分布 θ を持ち，文書中の各単語について，トピック分布 θ に従ってトピック z が選択され，トピック z に対応する単語分布 ϕ に従い単語 w が生成される．また，同一の単語であっても必ずしも同一のトピックとはならず，トピック z は単語 w が出現する文書の位置によって生成される．なお，文書が与えられたとき，観測変数は w で，それ以外は潜在変数または未知パラ

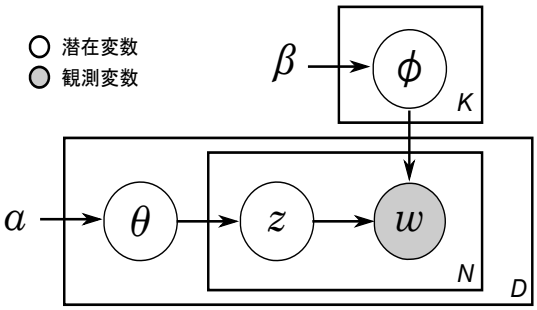

図 2　潜在的ディリクレ配分法のグラフィカルモデル

メータである．

本稿において，文書 d は投稿者毎の投稿をまとめて一文書としたものであり，文書数 D はユーザ数にあたる．よって，投稿者ごとに投稿する際のトピックを選択するトピック分布を持っており，各トピックに対応する単語分布が存在すると考えることができる．つまり，Twitter の投稿はまずユーザのトピック分布に従い投稿するトピックが選択され，選択されたトピックに対応した単語分布に従い単語が選択されるという流れで生成される．

潜在的ディリクレ配分法において文書は以下のように生成される．

1　各トピック $k = 1, \cdots, K$ について：
　(a)　単語分布 ϕ_k を生成
　　$\phi_k \sim Dir(\beta)$
2　各文書 $d = 1, \cdots, D$ について：
　(a)　トピック分布 θ_d を生成
　　$\theta_d \sim Dir(\alpha)$
　(b)　各単語 $n = 1, \cdots, N_d$ について：
　　(i)　トピックを生成
　　$z_{dn} \sim Multi(\theta_d)$
　　(ii)　単語を生成
　　$w_{dn} \sim Multi(\phi_{z_{dn}})$

ここで，ϕ_k はトピック k の単語分布，θ_d は文書 d のトピック分布，z_{dn} は文書 d の n 番目の語の潜在トピック，w_{dn} は文書 d の n 番目の単語を表す．また，$Dir(\cdot)$ はディリクレ分布を，$Multi(\cdot)$ は多項分布を表す．

トピック集合 Z と文書集合 W の完全尤度は，式（1）の通りである．なお，V は語彙数，$\Gamma(\cdot)$ はガンマ関数を表す．

$$P(Z, W|\alpha, \beta) = P(W|Z, \beta)P(Z|\alpha) \tag{1}$$

$$P(W|Z, \beta) = \left(\frac{\Gamma(\beta V)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{V} \Gamma(N_{kw} + \beta)}{\Gamma(N_k + \beta V)} \tag{2}$$

$$P(Z|\alpha) = \left(\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\right)^D \prod_{d=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(N_{kd} + \alpha)}{\Gamma(N_d + \alpha K)} \tag{3}$$

トピック集合の推定には，潜在的ディリクレ配分法の提案論文では変分ベイズ法が用いられ，また，Gibbs サンプリング法を用いる方法も提案されている [6]．Gibbs サンプリング法は，十分な反復回数を行えば，高い精度が得られる [7] と報告されており，本稿では，トピック集合の推定には Gibbs サンプリング法を用いることとした．Gibbs サンプリング法での更新式は式（4）で表される．

$$P(z_j = k|Z_{\setminus j}, W) \propto \frac{N_{dk\setminus j} + \alpha}{N_{d\setminus j} + \alpha K} \cdot \frac{N_{kw_j\setminus j} + \beta}{N_{k\setminus j} + \beta V} \tag{4}$$

ここで，N_{dk} は文書 d におけるトピック k が割り当てられた単語数，N_{kw} はトピック k における単語 w の出現回数，$N_k = \sum_{k=1}^{K}$，$\backslash j$ は文書 d の n 番目の単語を除いたときの回数もしくは変数を表す．

十分な反復回数を行い，更新を行うことで文書ごとのトピック分布，トピックごとの単語分布を得られる．文書 d のトピック分布 θ_d の推定量は式（5），トピック k の単語分布 ϕ_k の推定量は式（6）によって得られる．

$$\hat{\theta_d} = \frac{N_{dk} + \alpha}{N_d + \alpha K} \tag{5}$$

$$\hat{\phi_k} = \frac{N_{kw} + \beta}{N_k + \beta V} \tag{6}$$

4 潜在的ディリクレ配分法を用いた Twitter 分析の事例

潜在的ディリクレ配分法を用いて，Twitter ユーザを対象に各ユーザのトピック分布推定を行った．推定対象として Twitter から 309 アカウントの投稿を 2013/6/22 日から遡り収集した．また，ハイパパラメータ α と β は Zhao ら [2] に従い，$\alpha = K/50$，$\beta = 0.01$ とした．トピック数 K については，$K = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110$ についてそれぞれ実行し，人手で調べた結果最も適切にトピックが構成されたと判断した $K = 70$ について事例を示す．Gibbs サンプリングの試行回数は 2000 回とした．

推定されたトピックについて，一部を抜粋しその代表語を表 1 に示す．なお，トピック id は各トピックを区別するだけであり，数字に意味はない．また各トピックの代表語は，各トピックの単語生起確率上位五語を示している．

5 協調フィルタリング

潜在的ディリクレ配分法によって推定されるユーザごとのトピック分布は，あくまでユーザの投稿中の単語から推定されたものであり，興味があっても投稿していない内容は当然推定できない．また，情報推薦を目的とした興味対象の推定の場合，投稿内容にのみ依存した興味対象分析では，意外性のある情報を推薦することは難しい．そこで本研究では，ユーザの興味対象となるトピックを，投稿内容だけに限らず推定するために，複数のユーザのトピック分布を用いてユーザベース協調フィルタリングを行う．

協調フィルタリングは，商品の推薦などにおいて用いられる手法であり，オンラインショップの Amazon においても用いられている [8]．ユーザベース協調フィルタリングでは，あるアイテム群に対して評価が似通っているユーザは，あるアイテムに一方が高い評価をすればもう一方も高い評価をする可能性が高いという仮定に基づいている．そこで，本研究では潜在的ディリクレ配分法のトピック分布を各トピックに対するユーザの評価と捉えることで協調フィルタリングを行う．これにより，投稿トピックが似通っているユーザのトピック分布によって，興味分析対象ユーザのトピックに対する評価を推定することができる．

本稿における協調フィルタリングの流れは，まず潜在的ディリクレ配分法により表現されたユーザ毎のトピック分布をもとにユーザ間の類似度を計算する．この時，各ユーザのトピック分布の行列は θ と表される．次に，θ から各ユーザについて生起確率上位 N 以外のトピックを削除した θ' を作成する．この θ' を用いて協調フィルタリングを行うことは，分析対象ユーザの投稿にあまり含まれないトピックについてユーザの評価が無いものとして扱い，他ユーザのトピック分布から推定することを意味する．ここで，協調フィルタリングにより推定された値を，ユーザのトピックに対する興味を表す興味度とする．ユーザ u のトピック k に対する興味度は $P_{u,k}$ で表される．

5.1 ユーザ間類似度

協調フィルタリングにおいて，ユーザ間の類似度の尺度には Pearson 相関係数を用いる手法 [9] やコサイン類似度を用いる手法がある．本稿では，潜在的ディリクレ配分法により表現された各ユーザのトピック分布をユーザの評価値と捉え協調フィルタリングを行う．つまり，各ユーザの評価値として用いる値はトピックの生起確率であり，ユーザ間で評価基準に差がない．したがって，各ユーザの評価値平均を考慮する必要がないことから，本稿では協調フィルタリングにおけるユーザ間類似度の計算にコサイン類似度を用いる．ユーザ u とユーザ v の類似度 $w_{u,v}$ は式（7）によって表される．

$$w_{u,v} = \frac{\sum_{t \in T_{u,v}} r_{u,t} \cdot r_{v,t}}{\sqrt{\sum_{t \in T_{u,v}} r_{u,t}^2} \sqrt{\sum_{t \in T_{u,v}} r_{v,t}^2}} \tag{7}$$

ここで，$T_{u,v}$ は，ユーザ u と v がともに生起確率上位 N 以内であるトピックの集合であり，$r_{u,t}$ はユーザ u におけるトピック t の生起確率を表す．この $w_{u,v}$ は，$T_{u,v}$ のトピック数の違いを考慮しないため，$|T_{u,v}|$ が小さくても類似性が高くでる問題がある．したがって本稿では，トピック数を考慮したユーザ間類似度 $w'_{u,v}$ を用いる．ユーザ u とユーザ v のトピック数を考慮したユーザ間類似度は式 (8) で表される．

$$w'_{u,v} = w_{u,v} \cdot \frac{T_{u,v}}{N} \tag{8}$$

ここで，N は協調フィルタリングで推定に用いるトピック生起確率の数である．

5.2 トピックに対する興味度の推定

ユーザ間類似度を用いて，トピックに対する興味の強さである興味度を推定する．各ユーザについて興味度を推定する対象トピックは，潜在的ディリクレ配分法により推定されたトピックの生起確率が上位 N 位以内に入らなかった全トピックである．ユーザ u のトピック k に対する興味度 $P_{u,k}$ は式（8）によって表される．

$$P_{u,k} = \begin{cases} \theta_{u,k} & (k \text{ の生起確率が上位 } N \text{ 以内}) \\ \frac{\sum_{v \in U_{k \backslash u}} r_{v,k} \cdot w_{u,v}}{\sum_{v \in U_{k \backslash u}} |w_{u,v}|} & (\text{その他の場合}) \end{cases} \tag{9}$$

ここで，$U_{k \backslash u}$ は全ユーザ集合 U のうちトピック k に対する評価値を持つユーザ集合からユーザ u を除いた集合を意味する．トピック k がユーザ u のトピック分布において，生起確率上位 N 以内であれば興味度 $P_{u,k}$ は潜在的ディリクレ配分法により表現された生起確率をそのまま用いる．生起確率上位 N 以内でなければ，協調フィルタリングによって興味度を推定する．これにより，ユーザが投稿していないトピックに対しても興味の強さを推定することができる．

6 まとめと今後の展望

本研究では，Twitter を対象にユーザの投稿から潜在的ディリクレ配分法により表現されたトピック分布をもとに協調フィルタリングを用いることでユーザの興味対象を推定する手法を提案した．潜在的ディリクレ配分法を用いることで，単語ではなくトピック単位で興味対象を推定でき，また，協調フィルタリングを用いることでユーザが投稿していない内容であっても興味対象として推定することができる．本稿では，潜在的ディリクレ配分法により Twitter ユーザのトピック分布を推定した．今後は，提案手法のユーザ評価実験を行うことで，提案手法の有効性を確認する．また，ユーザの投稿には，投稿を構

表 1　トピック代表語の抜粋

トピック id	代表語
3	稽古，公演，劇場，芝居，初日
5	新刊，単行本，コミケ，イラスト，表紙
7	研究，バイト，大学，ゼミ，授業
8	デッキ，デュエル，対戦，遊戯王，コンボ
10	アイマス，アイドルマスター，765，公式サイト，日本一
26	クレオパトラ，ホール，単独，前売，劇団
33	ホテル，新幹線，渋谷，横浜，業務
37	脚本，ウルトラマン，監督，特撮，プリキュア
39	作曲，アレンジ，レコーディング，演奏，作詞
50	シュタインズ・ゲート，シュタゲ，STEINS;GATE，ロボティクス・ノーツ，劇場版
56	研究，論文，言語，Haskell，プログラミング
61	政治，国民，批判，責任，国

成するトピックに対しポジティブな内容の場合，ネガティブな内容の場合が存在する．投稿内容をポジティブと　ネガティブに判別，分類を行い提案手法に適用することで，より正確なユーザ興味対象の推定が実現できる可能性がある．さらに，提案手法によって推定されたユーザの興味対象をもとに，リアルタイムな情報推薦システムの構築を目指したい．

謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 25280100，および，25540146）の助成により行われた．

参考文献

[1] 早川, 岡部, 尾内：“Twitter を利用したソーシャルニュース記事推薦システム”, 情報処理学会研究報告. データベース・システム研究会報告, **2011**, 16, pp. 1–4 (2011).

[2] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan and X. Li: “Comparing twitter and traditional media using topic models”, Proceedings of the 33rd European conference on Advances in information retrieval, ECIR’11, Berlin, Heidelberg, Springer-Verlag, pp. 338–349 (2011).

[3] D. Ramage, S. T. Dumais and D. J. Liebling: “Characterizing microblogs with topic models.”, ICWSM (Eds. by W. W. Cohen and S. Gosling), The AAAI Press (2010).

[4] 宮城, 當間, 遠藤：“日本語オントロジー辞書システム ontolopedia の構築と興味抽出手法への応用検討”, 知能と情報：日本知能情報ファジィ学会誌：journal of Japan Society for Fuzzy Theory and Intelligent Informatics, **21**, 5, pp. 815–826 (2009).

[5] D. M. Blei, A. Y. Ng, M. I. Jordan and J. Lafferty: “Latent dirichlet allocation”, Journal of Machine Learning Research, **3**, p. 2003 (2003).

[6] T. L. Griffiths and M. Steyvers: “Finding scientific topics”, Proceedings of the National Academy of Sciences, **101**, Suppl. 1, pp. 5228–5235 (2004).

[7] Y. W. Teh, D. Newman and M. Welling: “A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation”, Advances in Neural Information Processing Systems, Vol. 19 (2007).

[8] G. Linden, B. Smith and J. York: “Amazon.com recommendations: item-to-item collaborative filtering”, Internet Computing, IEEE, **7**, 1, pp. 76–80 (2003).

[9] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl: “Grouplens: an open architecture for collaborative filtering of netnews”, Proceedings of the 1994 ACM conference on Computer supported cooperative work, CSCW ’94, New York, NY, USA, ACM, pp. 175–186 (1994).