

トピックモデルと協調フィルタリングに基づく ユーザ興味を反映した情報推薦システム

Tweet Recommendation System Reflecting User Preference
Based on Latent Dirichlet Allocation and Collaborative Filtering

渡邊 恵太*¹
Keita Watanabe

加藤 昇平*¹
Shohei Kato

*¹名古屋工業大学大学院工学研究科情報工学専攻

Department of Computer Science and Engineering Graduate School of Engineering Nagoya Institute of Technology

Recently, Twitter becomes widely spread rapidly and attracts much attention as large information sources. One of the feature of Twitter is that many information is contributed in real time. However, conventional searching methods like keyword search are not enough to acquire information about user's interest from Twitter immediately, because user must type a keyword every time when they want to search. Consequently, the system which recommends information based on user's interest in real time is required. It is important for such a system to analyze user's contribution and extract user's preference from Twitter. In this paper, we propose a method of extracting user's preference based on collaborative filtering and latent dirichlet allocation. And, we build tweet recommendation system reflecting user's preference based this method. We confirmed the validity of the system by user evaluation experiment.

1. はじめに

近年, Twitter をはじめ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, 投稿が短文に制限されており, ユーザは従来のブログに比べ容易に投稿することができる。そのため, リアルタイム性の高い情報が即座に得られる情報源として注目されており, マイクロブログを対象とした研究も盛んに行われている。一方で, マイクロブログに投稿される情報量は膨大であり, ユーザがその全てを監視し, その中から必要な情報だけを抽出することは困難である。そこで, マイクロブログでは膨大な情報の中からユーザにとって価値のある情報を選別する機能が必要であると考えられる。現在, マイクロブログから情報を得る方法として, 予め着目した投稿者の投稿を監視する方法が一般的に用いられている。また, この方法に従い, ユーザにとって価値ある投稿を多くするユーザを推薦する研究がなされている。しかしながら, 通常投稿者は様々な内容の情報を投稿しており, その全てがユーザにとって価値があるとは考えにくい。したがって, この方法ではユーザにとって価値の無い情報も推薦されてしまうという問題が存在する。このことから, 情報を誰が投稿したのかによって選別するのではなく, 情報の内容がユーザにとって価値があるか否かによって選別すべきである。

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を用いてユーザのトピック分布を推定し, そのトピック分布に基づいて協調フィルタリングによりユーザ興味の推定手法を提案する。また, 推定した興味をもとに, マイクロブログ上の投稿を選別し, ユーザに推薦する情報推薦システムを構築した。興味の推定では, 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単位ではなくトピック単位で推定することができ, 単語間の関連性を考慮することができる。また, 単語からトピックという単位に次元が圧縮されることで, 協調フィルタリングを容易に適用することができ, ユーザが投稿していない内容についても興味対象として推定することが可能となる。情報の推薦では, マイクロブログ上の投稿を投稿内

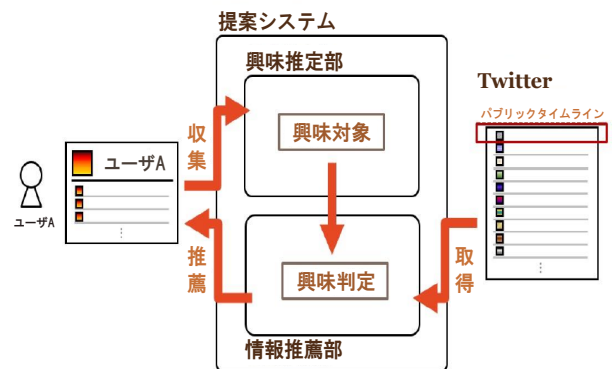


図 1: 提案システム概要

容からユーザの興味に基づき選別することで, リアルタイムにユーザにとって価値のある情報を推薦することができる。

2. 提案システム概要

図 1 に提案システムの概要を示す。提案システムは興味推定部と情報推薦部からなる。提案システムでは, まず Twitter からユーザの投稿を収集し, その内容に基づき興味推定部においてユーザの興味を推定する。Twitter 上の最新の投稿を収集し, 情報推薦部において投稿の内容を判断し, 興味推定部でユーザが興味を持つと推定された内容であれば推薦を行う。3 章において興味推定部について, 4 章において情報推薦部について詳しく説明する。

3. 興味推定部

図 2 に興味推定部の概要を示す。興味推定部では, 潜在的ディリクレ配分法と協調フィルタリングを組み合わせた興味推定手法を提案する。まず Twitter API を用いて Twitter から複数のユーザについて過去の全投稿文を収集しデータベースを作成する。このとき, Twitter API の制約上 3200 件より多

連絡先: 加藤昇平, 名古屋工業大学, 愛知県名古屋市昭和区御器所町, 052-735-5625, shohey@katolab.nitech.ac.jp

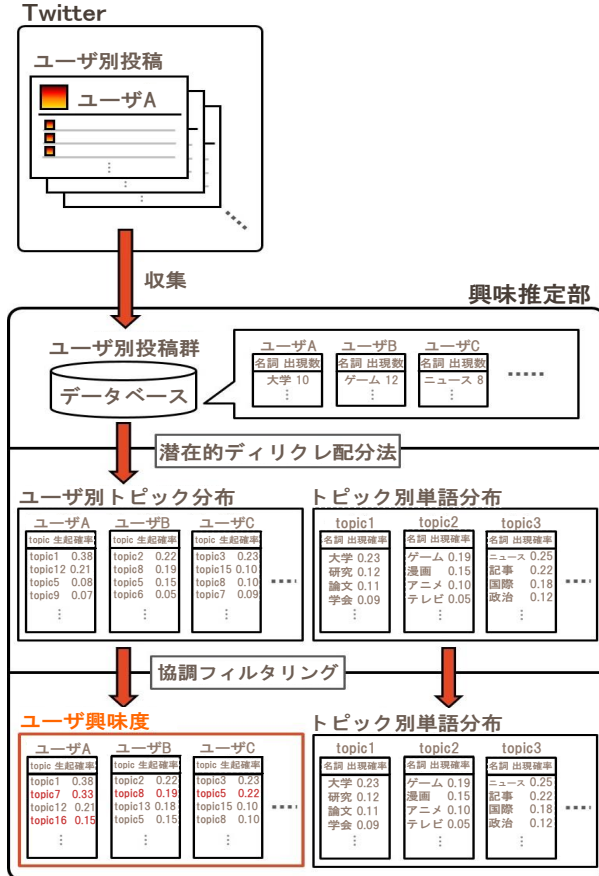


図 2: 興味推定部概要

くの投稿を行っているユーザについては、全投稿を収集できないため、最新の投稿から 3200 件を収集対象とする。このデータベースについて潜在的ディリクレ配分法に基づき、トピックという語の生起確率で表された話題について、トピックごとの単語分布とユーザ毎のトピック分布を推定する。分析対象とする単語を名詞に限定し、投稿文から名詞を抽出して推定を行った。これは、ユーザの興味を表す語は名詞が適切であると考えたためである。名詞の抽出には、形態素解析器 Sen を使用し、抽出の性能を向上させるため宮城ら [1] にない、形態素解析器の辞書に Wikipedia のページタイトル名を名詞として追加している。また、潜在的ディリクレ配分法の適用において Zhao[2] らにならい、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分析の対象から除外した。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。得られたトピック分布において高い生起確率があることは、ユーザが多く投稿しているといえることから興味があるトピックとする。一方低いトピックは、あまり投稿していないことはわかるが、「興味はあるが投稿していない・ユーザにとって未知である」か、「興味が無い」のかどちらかわからない。そこで、生起確率が高いトピックを用いて協調フィルタリングにより、生起確率が低かったトピックの評価値を推定する。ここで、トピックの数を T とすると、生起確率が $\frac{1}{T}$ より高いトピックを興味があるとす。これは、 $\frac{1}{T}$ がトピック分布が一様分布であった時の全トピックの生起確率であり、これより大きければユーザは

そのトピックに偏って投稿を行っているといえると考えたためである。協調フィルタリングによって推定された値をユーザのトピックに対する興味の強さを表す興味度とする。

3.1 潜在的ディリクレ配分法

潜在的ディリクレ配分法 (Latent Dirichlet Allocation, LDA) [3] に基づきユーザがどのような話題を投稿しているか推定する。LDA 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る。LDA によって、ユーザ投稿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。トピック集合の推定には、LDA 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 [4]。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と報告されており [5]、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 サンプリング法では文書が与えられたとき、全ての単語に対してトピックを割り当て、更新式に基づきトピックの更新を繰り返すことで各文書のトピック分布を推定する。Gibbs サンプリング法の更新式は式 (1) で表される。

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + T\alpha} \quad (1)$$

ここで、 $\mathbf{w}$ はデータベース中の語の集合、 $\mathbf{z}_{-i}$ は各語に割り当てられた語の集合から更新対象である i 番目の語を除いたものである。 W は語彙数、 T はトピック数、 α と β はハイパーパラメータである。 $n_{-i,j}^{(w_i)}$ はトピック j に属する語 w_i と同じ語の出現数、 $n_{-i,j}^{(\cdot)}$ はトピック j に属する語の総数を表す。 $n_{-i,j}^{(d_i)}$ は i 番目の単語が属する文書 d_i の語のうち、 w_i を除いてトピック j に属する語の数、 $n_{-i,j}^{(d_i)}$ は i 番目の単語が属する文書 d_i の語の総数を表す。十分な反復回数を行い、更新を行うことでデータベースの語のトピックを推定する。トピック j における語 w の分布 $\phi_j^{(w)}$ の推定量は式 (2)、文書 d におけるトピック j の分布 $\theta_j^{(d)}$ の推定量は式 (3) によって得られる。

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \quad (2)$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n^{(d)} + T\alpha} \quad (3)$$

3.2 協調フィルタリング

LDA 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、興味があっても未投稿な内容は推定できない。また、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 LDA 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。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 [6]。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ユーザの各トピックに対する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

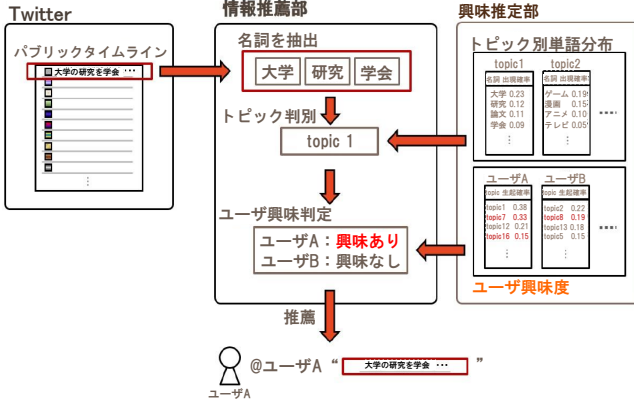


図 3: 情報推薦部概要

できる。まず LDA 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。本研究では、ユーザ間類似度としてコサイン類似度を用いる。ユーザ d を推定対象ユーザとしたとき、ユーザ e との類似度 $w_{d,e}$ は式 (4) によって表される。

$$w_{d,e} = \frac{\sum_{j \in \mathbf{z}_d} r_{d,j} \cdot r_{e,j}}{\sqrt{\sum_{j \in \mathbf{z}_d} r_{d,j}^2} \sqrt{\sum_{j \in \mathbf{z}_e} r_{e,j}^2}} \quad (4)$$

ここで、 $\mathbf{z}_d$ は、トピックの内ユーザ d の生起確率が $\frac{1}{T}$ を超えているトピックの集合である。また、 $r_{d,j}$ はユーザ d のトピック分布におけるトピック j の生起確率である。次にユーザ間類似度をもとに、推定評価値を計算する。潜在的ディリクレ配分法において、生起確率が高いトピックはユーザの投稿に多く現れたことを表し、ユーザが興味を持っているトピックであると仮定する。一方、生起確率が低いトピックは興味の有無がわからないことから、協調フィルタリングによって評価値の推定を行う。ユーザ u のトピック j に対する推定評価値 $P_{u,j}$ は式 (5) によって表される。

$$P_{u,j} = \begin{cases} \theta_j^{(d)} & (\theta_j^{(d)} > \frac{1}{T}) \\ \frac{\sum_{d \in D_{-d}} r_{d,j} \cdot w_{u,d}}{\sum_{d \in D_{-d}} |w_{u,d}|} & (\theta_j^{(d)} \leq \frac{1}{T}) \end{cases} \quad (5)$$

ここで、 D_{-d} は全ユーザからユーザ d を除いた集合を意味する。

4. 情報推薦部

図 3 に情報推薦部の概要を示す。情報推薦部では、まず Twitter のパブリックタイムラインを監視し、新しい投稿を取得する。ただし、取得する投稿はリツイートを一回以上されたものとする。これは、全てのツイートを対象とした場合、ユーザの日常会話のような情報的な価値が低い投稿が多く推薦されてしまうためである。リツイートは投稿者以外が引用したことを表し、情報としての価値が一定以上ある可能性が高いと考え、このような設定とした。次に、その投稿がどのトピックに属するか判別する。この際、トピックは予め興味推定部で生成したものを用いる。ここで、収集した投稿 m が $\mathbf{m} = w_1, w_2, \dots, w_M$ という名詞で構成されているとする。 M は投稿に含まれた名詞の数、 w_i は投稿の i 番目に出現した名詞、 $\mathbf{m}$ は投稿に含ま

表 1: データベースの規模

ユーザ数	投稿総数	単語総数	語彙総数
570	1515804	6505747	141265

れた名詞の集合とする。この投稿が属するトピック k は式 (6) によって決定する。

$$k = \arg \max_{j(j=1,2,\dots,T)} P(z_i = j | \mathbf{m}, \mathbf{z}, \mathbf{w}) \quad (6)$$

ここで、 $P(z_i = j | \mathbf{m}, \mathbf{z}, \mathbf{w})$ は、式 (7) を用いて最大化するように $\mathbf{m}$ の各語の割り当てトピックの更新する。

$$P(z_i = j | \mathbf{m}_{-i}, \mathbf{z}, \mathbf{w}) \propto \frac{n_{-i,j}^{(m_{w_i}, w_i)} + n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(m)} + n_{-i,j}^{(\cdot)} + W\beta} \cdot \frac{n_{-i,j}^{(m)} + \alpha}{n_{-i}^{(m)} + T\alpha} \quad (7)$$

$\mathbf{m}_{-i}$ は $\mathbf{m}$ から i 番目の語を除いた集合である。また、 $n_{-i,j}^{(m_{w_i}, w_i)}$ はトピック j に属する語 w_i の数と投稿 m 中の語 w_i の数のから投稿 m の i 番目の語を除いたものである。 $n_{-i,j}^{(m)}$ は $\mathbf{m}$ から i 番目の語を除き、トピック j に属する語の数である。 n のうち m と関係がないものは、興味推定部で作成したサンプルの値を用いる。よって、この式を用いて $\mathbf{m}$ の各語のトピック割り当てのみを十分回数反復して更新した後、割り当てトピックを決定することで $P(z_i = j | \mathbf{m}, \mathbf{z}, \mathbf{w})$ が求まり、投稿 m が属するトピックが決定できる。取得した投稿の属するトピックが j であったとする。興味推定部で推定した各ユーザの興味度から、トピック j に興味があるユーザを探す。興味推定部では、ユーザ u の興味度 $P_{u,j} > \frac{1}{T}$ であったとき、ユーザ u はトピック j に対して興味があるとした。これより、 $P_{u,j} > \frac{1}{T}$ を満たすユーザ u がいれば、そのユーザに対して投稿の推薦を行う。

5. 性能評価実験

提案システムの性能を検証するため被験者評価実験を行った。興味推定部と情報推薦部のそれぞれについて性能を検証する。Twitter から 800 ユーザの投稿を 2013 年 12 月 16 日から遡り収集し、そのうち投稿数が 300 件に満たないユーザと自動投稿システムであるユーザアカウントを除外した 570 ユーザの投稿をデータベースとした。表 1 にデータベースの規模を示す。興味推定部での潜在的ディリクレ配分法のパラメータ設定として、ハイパパラメータ α と β は Zhao ら [2] に従い、 $\alpha = \frac{50}{T}$ 、 $\beta = 0.01$ とした。また、トピック数 T は、 $T = 10$ から 10 刻みで $T = 200$ について実行し、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 $T = 70$ に設定した。Gibbs サンプルングの試行回数は 1000 回とした。

5.1 興味推定性能

分析対象に含まれる Twitter ユーザ 9 人を被験者として、興味推定結果についてアンケートを行った。アンケートでは生成された 110 トピックについて生起確率上位 10 語を提示し、各トピックに興味がある、どちらでもない、興味がないの 3 つに分類させた。比較手法として LDA の生起確率を用いて興味分析をした場合を示す。表 2 に全被験者の合計性能を示す。興味はアンケートにおいて被験者が興味ありと答えたトピックの総数、提示は各手法が興味ありと推定したトピックの総数、正答

表 2: 興味推定性能

手法	正答	提示	興味	適合率	再現率	F 値
LDA	58	78	317	0.744	0.183	0.294
提案	87	153	317	0.569	0.274	0.370

表 3: 情報推薦性能

手法	提示数	興味	価値	新規性
LDA	244	157	110	100
提案	450	250	159	178
差分	256	93	48	78



図 4: 情報推薦システムの出力例

は推定したうち実際にユーザが興味をもっていたトピックの総数である。各手法について、適合率と再現率、F 値を算出した。提案手法の適合率についてみると、0.569 となっており、潜在的ディリクレ配分法の 0.744 に対して低下している。これは、潜在的ディリクレ配分法は Twitter 上のユーザ自身の投稿に基づき推定を行っているため、出力結果の信頼性が高く適合率が高くなったことが考えられる。これに対して、提案手法では協調フィルタリングを用いているためユーザの興味を推定するのに他のユーザの嗜好を利用しており、これらの情報がユーザ自身の投稿に比べるとノイズが多くなってしまうことが考えられる。また、協調フィルタリングに用いる情報が潜在的ディリクレ配分法によって推定された値であり、実際の興味に完全に一致するわけではないということも原因と考えられる。一方で、再現率については提案手法は 0.274 であり、潜在的ディリクレ配分法の 0.183 に比べ上昇している。提案手法のほうが提示数が多いことから、再現率は一般に高くなりやすくこの結果だけでは性能が向上しているとはいえない。しかし F 値について見ても値が上昇していることから、適合率と再現率を合わせて評価した場合、提案手法は潜在的ディリクレ配分法のものに比べ興味推定性能が向上したといえる。これより、提案手法を用いることで、より多くの興味対象を推定することができ、また、Twitter 上に投稿していない内容についても推定できることが示された。情報推薦において、ユーザの興味推定をする際にその網羅性を高めることは、ユーザにとってより多くの有用な情報を提供できる可能性があり、重要であるといえる。

5.2 情報推薦性能

5.1 節と同じ被験者に対し、投稿の推薦を行いその性能を検証した。図 4 のように Twitter 上に bot を作成し、各被験者に推薦を行った。推薦された投稿に対し、「興味」、「価値」、「新規性」の有無について被推薦者にそれぞれ回答させた。比較として LDA によって推薦された投稿の結果も示す。表 3 に結果を示す。表中の数字は 3 項目について「有」と回答した推薦記事の数を示す。これより、提案手法では LDA のみに比べより多くの興味ある投稿を提示することができたことがわかる。一方で、「興味あり」に比べ「価値あり」と評価された投稿数

が少ないことから、興味ある内容のトピックを推薦することはできているものの、その情報としての価値が十分に判断出来なかったことが考えられる。

6. まとめと今後の展望

本研究では、リアルタイムな情報源として着目されている Twitter において、ユーザの興味に基づき情報を推薦するシステムの提案を行った。潜在的ディリクレ配分法と協調フィルタリングを組み合わせた興味推定手法により、ユーザが Twitter に投稿していない内容についても興味の対象として推定することを可能とした。被験者実験により、興味推定と情報推薦の性能が LDA のみの場合に比べ向上したことを確認した。今後は、情報の流行度や新規性を表す指標の検討を行い、ユーザの興味と情報の価値という二つの指標からより価値ある情報推薦を目指したい。

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] 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).
- [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. M. Blei, A. Y. Ng, M. I. Jordan and J. Lafferty: “Latent dirichlet allocation”, Journal of Machine Learning Research, **3**, p. 2003 (2003).
- [4] T. L. Griffiths and M. Steyvers: “Finding scientific topics”, Proceedings of the National Academy of Sciences, **101**, Suppl. 1, pp. 5228–5235 (2004).
- [5] 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).
- [6] G. Linden, B. Smith and J. York: “Amazon.com recommendations: item-to-item collaborative filtering”, Internet Computing, IEEE, **7**, 1, pp. 76–80 (2003).