DEIM Forum 2011 B6-4

興味分析に基づくウェブログ筆者検索

渡辺　一史† 　岡部　誠†,†† 　尾内理紀夫†

† 電気通信大学　〒 182–8585 東京都調布市調布ヶ丘 1–5–1
†† 独立行政法人 科学技術振興機構
E-mail: †{k_watanabe,okabe,onai}@onailab.com

あらまし　我々は，ウェブログの内容が時間的に変化する点に着目し，筆者を分類するための手法を提案する．本研究では，トピックモデルに基づいてブログ文書の潜在意味（トピック）を分析し，各トピックに対する言及量をもとに筆者を興味分野という軸でモデリングする．さらに，構築した筆者のモデルをインタラクティブに検索するための直感的なインターフェースを開発した．ユーザは約千人の Twitter ユーザから自分と同じ興味を持つ筆者や，ある特定の興味の移り変わりを経験した筆者を効率良く検索することができる．
キーワード　興味分析，情報推薦

Blogger Search based on Interest Analysis

Kazufumi WATANABE†, Makoto OKABE†,††, and Rikio ONAI†

† University of Electro-Communications, 1–5–1 Chofugaoka, Chofu, Tokyo, 182–8585, Japan.
†† JST PRESTO
E-mail: †{k_watanabe,okabe,onai}@onailab.com

Abstract　Since blog contents are rapidly increasing on the web, it's becoming difficult for the readers to find blogs that they want to read most. To efficiently search for interesting blogs, we propose a method to find a blogger whose blogs are interesting to the reader. We model each blogger focusing on the fact that blogger's interest is always changing. We first analyze blogs using latent dirichlet allocation, and extract tens of topics from them. We manually classify the extracted topics into five categories that correspond to bloggers' interests, "IT", "life", "politics and economics", "entertainment", and "game and animation". We visualize time-varying interests of each blogger in the categories. Our user interface allows the user to efficiently create a query by editing a value of each category and its temporal variation. We demonstrate our method enables the user to effectively find a blogger who has a specific interest or whose interest has been changing in a specific way.
Key words　Interest analysis, Recommendation

1. はじめに

近年，ブログによる情報発信が活発になったことで，ユーザは様々な情報を入手することができるようになった．しかし，ブログの数が膨大になったことで，自分の興味に合ったブログを見つけることは難しくなっている．

このような問題に対して，推薦システムの研究が広く行われている．推薦システムとは，ユーザに適した情報や物（以下，アイテムと呼ぶ）を選出して提示するシステムであり，これを実現するための手法として内容ベースフィルタリングと協調フィルタリングがある．内容ベースフィルタリングは，ユーザの嗜好をモデリングしたユーザプロファイルと，各アイテムの特徴に従ってモデリングしたコンテンツモデルとを比較することで推薦するアイテムを決定する手法である．協調フィルタリングでは，アイテムに対する嗜好パターンが似ている別のユーザを見つけ，そのユーザが好むものを推薦する．

しかし最近の動向として，推薦精度自体は向上したものの，ユーザの好みではあるが似たようなアイテムばかり推薦されるという問題が発見されている．これを踏まえて，単純に推薦を行うだけでなく，推薦理由を提示したり，ユーザプロファイルを自由に編集できるようにしたりすることで，ユーザの自発的な探索を促し，全体的な推薦の質を向上させる方法が提案されている．また，特に内容ベースフィルタリングにおいて，アイテムの特徴を生かしたモデリングが重要となっている．文書に対しては，tf-idf を用いた単語の重み付けを行うのが一般的である．これをブログに対して適用した場合，ブログには筆者の

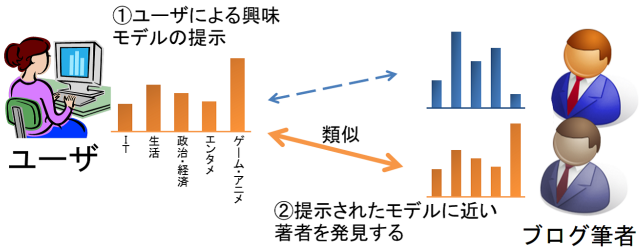

図 1　興味モデルを用いたブログ筆者の発見方法

興味が反映されやすいことから，筆者の興味の度合いによるモデリングが行われる．しかしこの手法には，単語の意味的な類似性や，筆者の興味が時間的に変化することを考慮していないという問題がある．

そこで本研究では，人間が直感的に理解できる形で興味モデルを構築し，人の興味がどのように，またどのくらいの時間で変化するかをユーザが自由に編集しながら条件に適するブログ筆者を発見できるインタラクティブな検索システムを提案する．このシステムの最も簡単な使用例は，ある特定の興味を持った筆者を探すことである．図 1 に示すような 5 次元のグラフが表示されるので，例えば，ゲーム・アニメのバーを高くして検索を行うと，それについて強い興味を持っている筆者が推薦される．もう 1 つの重要な使用例は，ある特定の興味の移り変わりを経験した筆者を探すことである．例えば，1ヶ月前はスポーツのバーが低かったが，ここ数日だけスポーツのバーが著しく高くなっているというようなクエリを与えることで，昨日の日本代表のサッカーの試合に熱狂しているミーハーなサッカーファンだけを見つけることが可能となる．他にも，ユーザが自分の現在の興味とは別の興味を伸ばしたいという時に，実際にそのような変化を経験した筆者のブログが推薦されれば，そこから有益な情報が得られると期待できる．

このようなシステムを実現するためには，ブログ文書から特徴的な興味の強さを抽出し，さらにそれをユーザが理解・編集しやすい形に抽象化する必要がある．そこで我々は，代表的なトピックモデルの一つである Latent Dirichlet Allocation(LDA) [1] を利用して，ブログ文書に含まれるトピックを抽出する．これらトピックに対する言及の度合いを各筆者の特徴として利用するが，このとき興味のカテゴリとして「IT」,「生活」,「政治・経済」,「エンタメ」,「ゲーム・アニメ」の 5 つを設定し，このいずれかに分類されるトピックのみを用いて興味の特徴ベクトルを構築する．ユーザは，特徴ベクトルを直感的に編集できるインターフェースを用いることで，柔軟性の高いクエリを与えることができ，またそれに対応した筆者を発見することができる．

2. 関 連 研 究

推薦システムにユーザによるプロファイル編集機能を取り入れたものとして Hijikata ら [2] の研究がある．これは音楽データに対するユーザの嗜好モデルを決定木により表現し，可視化するというもので，ユーザは後からそのパラメータを編集することで好みの音楽を探索することができる．また，音楽的特徴量を理解していないユーザのために，特徴空間へ評価済みの音楽データをマッピングすることで，視覚的な編集支援を行っている．本研究ではブログから抽出した筆者の興味の度合いを特徴として用いており，特徴量そのものが理解しやすいため，より直感的な探索が可能となっている．

ブログからのユーザモデリングを行っているものとしては，桑原ら [3] の研究がある．これはマイクロブログにおける個々の記事について，トピック分類とそれに対する感情抽出を行うというもので，各トピックのスコアを筆者のモデリングに使用する．肯定的な感情を持っているトピックについて，スコアが高い順に 10 件選出したものを特徴ベクトルとし，最終的に興味の類似したユーザの推薦を行う．

野田ら [4] は，個々のブログ記事を Wikipedia のカテゴリにマッピングすることで各カテゴリを得意分野とするブログ筆者の発見を行っている．記事のマッピングはカテゴリ名によるキーワードマッチングによって行われ，ブログ記事の中に「iPhone」や「Android」といったカテゴリ名が含まれていれば該当するカテゴリにマッピングされる．また，Wikipedia のカテゴリには WordNet にあるような意味的な上位カテゴリ，下位カテゴリの構造があり，例えば上記のカテゴリには「スマートフォン」という共通の上位カテゴリが存在する．この構造を用いて,「スマートフォン」について専門性の高い筆者の発見を行っている．

Ramage ら [5] は，Twitter ユーザの行動調査に基づいて substance, social, status, style の 4 つのカテゴリを設定し，Twitter 上の文書の分類を行う．文書集合にトピックモデルを適用し，得られたトピックに単語を割り当てることで，上記のカテゴリへの分類が行われる．ユーザはこのカテゴリをフィルタリングに利用することで，読みたいツイートを簡単に読むことができる．

このように，ブログ記事の内容から筆者の特徴を抽出する研究はいくつか行われているが，ユーザに編集可能な形で提示することを考えると，全ての筆者を共通の指標によって表現する必要がある．[5] の 4 カテゴリによる分類がそれに近いが，これは Twitter の利用目的や対象による分類であるため，筆者の興味を直接表現しているとは言えない．そのため，我々は興味分野をモデリングに用いる特徴とし，特徴量抽出のために LDA を用いることにした．

また，ブログの時間情報に着目したものとしては，話題抽出，コミュニティ抽出といった研究がある．これは，ある期間において様々なブログで使用回数が増加している語句から話題を抽出したり，抽出した語句をもとにコミュニティを発見するというものである．関口ら [6] は，同じ興味を持つ筆者の間で特徴的に出現する語句を話題語として抽出する手法を提案している．このように，特定語句の使用回数の変化から筆者の類似性を求める研究はされているが，特定の興味変化を経験した筆者の発見ということはあまり行われていない．

3. システム概要

提案システムは，興味モデルの構築と検索によって構成される．まず，興味モデルの構築では次の 3 段階の処理が行われる．

（1） ブログ文書集合からのトピック抽出
（2） 興味カテゴリの定義
（3） 各文書のトピック推定および興味カテゴリへの集約
この処理の結果，ブログ文書が興味の特徴ベクトルに変換される．

また検索の際には，ユーザに対して，筆者検索を行うためのインターフェースを提示する．ここでは作成した特徴ベクトルをヒストグラムとして提示する．ユーザは自分や他人のヒストグラムが時間変化する様子を閲覧・編集しながら好みの筆者を探す．ここでは検索結果として，特徴ベクトルの生成に用いたブログ記事の提示も行う．

4. 興味モデルの構築

ここで用いる LDA は，単語の共起関係を利用して文書集合中に存在する話題を学習する．これはトピックと呼ばれ，トピックごとに存在確率の高い単語集合によって表現される．本研究では，ギブスサンプリング [7] を用いて実装した LDA を用いて，文書集合からのトピック抽出および新規文書に対するトピック推定を行う．

4.1 Twitter からのトピック抽出

ブログ文書集合を用いて，その文書内でよく語られている話題の分析を行う．今回は，データ収集の容易さ，ユーザ数が十分であるといった理由から，Twitter のデータを利用した．Twitter の提供する API に Streaming API があり，筆者が非公開にしていない文書の一部を収集することができる．これを利用して 1 週間分のデータを収集し，そこから日本語と思われるもの約 500 万件を抽出した．さらに前処理として本文に含まれるユーザ名やハッシュタグなどを除去し，形態素解析によって得られた名詞のみをトピック抽出に用いた．

このとき，LDA によって抽出するトピック数は 100 とし，サンプリング回数は 1000 とした．その結果の一部を，表 1 に示す．この表では，各トピックにおいて出現確率の高い単語と，それらの単語から主観的に決定したトピック名を便宜的に表示している．

4.2 興味カテゴリの定義

ここでは，興味の対象になりやすいと思われる「IT」,「生活」,「政治・経済」,「エンタメ」,「ゲーム・アニメ」という 5 つの興味カテゴリを設定し，抽出したトピックをこの興味カテゴリに主観的に分類する．このいずれのカテゴリにも属さないトピックは，興味モデルの構築において不要であると考え，使用しないことにする．そのため，挨拶やアスキーアート，学校や仕事といった人の属性を表すトピックなど，表 2 に示されるような興味とは関係なく発信されると思われる情報はモデル構築の際にはフィルタリングされることになる．結果として，100 トピック中約 4 割を興味モデルに利用することになり，その構成は表 3 のようになった．

興味カテゴリを設定する理由として，LDA によって筆者の各トピックへの興味の度合いをその言及量で表すことはできるが，抽出したトピックをそのまま用いると 100 トピックによるトピック分布となるため，筆者がどういう方向に興味を持っているのかをモデルの形状から推測するのは難しくなってしまうということがあった．ユーザが筆者の特徴を理解し，好みの筆者を発見するための編集を行えるようにするには，モデルの形状を単純に，かつ人の興味を反映したものにする必要がある．

表 1 Twitter から抽出されるトピック例

出現確率の高い単語	推定されるトピック名
用 パソコン 携帯 mac Air iPhone 型	デジタル機器
試合 応援 戦 野球 チーム 選手 中日	野球
風邪 今年 秋 度 体調 冬 大事 季節 喉 気	風邪
風 月 心 綺麗 海 空 花 水 星 光 夜 きれい	自然
猫 犬 山 散歩 うち 川 匹 クマ 鳥 森 虫	動物・ペット

表 2 興味モデルの要素として不適切なトピック例

出現確率の高い単語	推定されるトピック名
時 分 度 時間 午後 秒 朝 時半 お知らせ	時間
(´ Д `) ‘) ‘ 艸 (; ´ (;´ (° * ´ °	アスキーアート
今日 雨 今日は 朝 天気 昨日 今朝 日 傘	天気
そう 無理 絶対 ダメ だめ 嫌 駄目 口 これ	感情
金 ⇒ 簡単 方法 お金 パック 完全 成功	広告

表 3 興味カテゴリの構成

興味カテゴリ	構成トピック (一部)
IT	デジタル機器，モバイル
生活	ファッション，交際，飲食，運動，部屋，育児
政治・経済	政治，経済，国際
エンタメ	野球，芸能，自然，ライブ音楽
ゲーム・アニメ	ゲーム，アニメ，本，配信

4.3 筆者の興味モデルの構築

LDA を用いて抽出したトピックを，実際にブログ文書の各単語に割り当て，特徴ベクトルを生成していく．ここで用いる文書数は，1 人の筆者に対して最大 1000 件とした．これについても Twitter API によって収集し，トピック抽出と同様の前処理を行う．そして各単語に対し，ギブスサンプリングを用いてトピックを割り当てる．各文書に対するサンプリング回数は 50 とし，そこから 3 回の推定結果を用いてその文書のトピック分布を生成する．最後に，設定した興味カテゴリごとにトピックの値を集約することで興味モデルを構築した．

5. 興味モデルの編集・検索

システムのインターフェースを，図 2 に示す．ユーザが Twitter のユーザ ID とモデル構築に利用する記事の期間を入力すると，その筆者のトピック分布が 5 つの興味カテゴリの値に変換され，ヒストグラムとして左上に表示される．期間というのは，例えば 1ヶ月と指定すれば，1ヶ月分の記事によって興味モデルが生成される．

次に，画面右上のヒストグラムについて説明する．このヒストグラムは編集することができ，図 3 に示すようにヒストグラムの上部を上下に動かすことで，カテゴリの比率が変化するようになっている．このヒストグラムは，左側のヒストグラムが指定した期間を経て変化した後の興味モデルを想定したものとなっている．具体的な使用例としては，左側のモデルを基準と

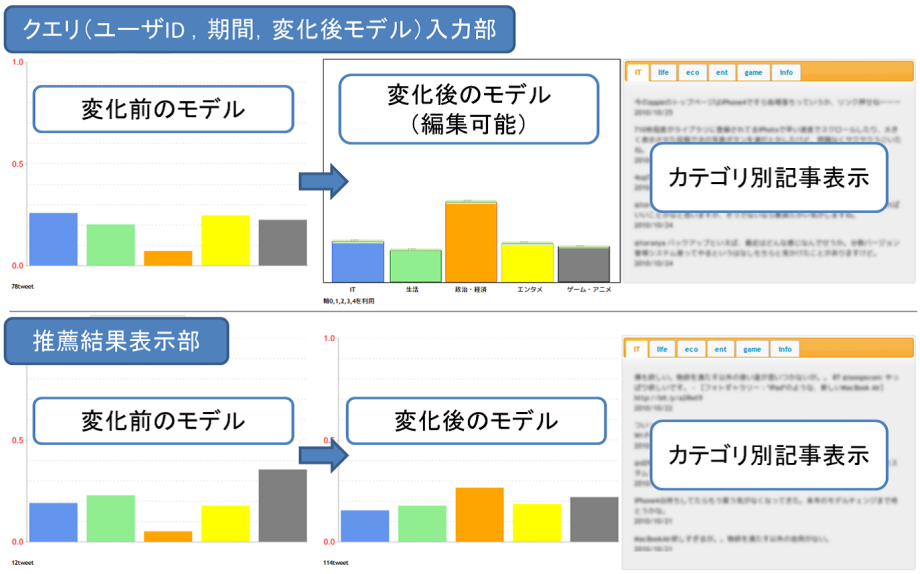

図 2　筆者検索画面．この画面では，政治・経済に対する興味が上昇した著者が推薦されている．

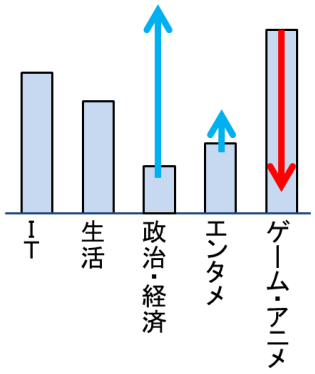

図 3　クエリの入力方法

して，今後伸ばしていきたい興味を引き上げるなど，単純な類似検索では満足できなかったときの検索手法として用いることが考えられる．このように，ユーザはある期間における興味の移り変わりをシステムに提示することができる．これら 2 つのヒストグラムが筆者検索におけるクエリとなり，このクエリと距離の近い筆者が検索される．距離尺度には，マンハッタン距離

$$d(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k|$$

を用いている．ただし $\mathbf{x}$ はクエリ側の興味モデルベクトル，$\mathbf{y}$ は推薦対象である筆者の興味モデルベクトルである．n は興味カテゴリ数で，この場合は 5 である．システムは，変化前，変化後のヒストグラム同士で距離を計算し，合計した値の小さい筆者を推薦する．

推薦の結果，提示された筆者の情報が画面下部分に提示される．この例でいえば，過去 2ヶ月目から過去 1ヶ月目までの記事によって作られたモデルが左下に，過去 1ヶ月間の記事によって作られたモデルが右下にヒストグラムとして表示される．また，興味モデルと合わせて，そのモデル構築に使われた文書が画面右側に表示される．これらにより，ユーザは提示された筆者の情報や文書を見て，購読すべきかどうかといった判断を行うことができる．

6. 評　　価

本システムは，ユーザの指定する興味変化を経験した筆者を推薦する．このため，例えば IT の興味カテゴリの値がある期間で上昇したという筆者のブログからは，IT 関係の変化が起こった原因が推定できると考えられる．そこで，任意の興味カテゴリの値が大幅に上昇するクエリによって検索を行った上で，提示された筆者 10 人の投稿内容を読み，その上昇の理由が判断できるかを評価した．ここで，変化の期間は 1ヶ月とし，自動投稿 Bot などは予め評価対象から除外することにした．

表 4　筆者の変化原因の推定結果

上昇した興味カテゴリ	原因推定率	主な判断理由
IT	50 %	デジタル機器の調査，購入
生活	60 %	運動，出産，住居の移転
政治・経済	20 %	海外出張，経済の勉強
エンタメ	30 %	音楽趣味の再燃
ゲーム・アニメ	40 %	ゲームの購入，制作
全体	40 %	

結果と判断理由は表 4 のようになった．大まかな特徴としては，IT カテゴリが上昇した筆者は新製品，特にモバイル機器やパソコンの発売時期に応じて，それに対する興味の上昇が見られるケースが多かった．生活カテゴリでは，運動を始めるといった生活スタイルの変化や，出産，入院，あるいは住居の移転といった生活環境の変化が起こった筆者が推薦された．政治・経済カテゴリとエンタメカテゴリは，変化を判断できる筆者は比較的少なかったものの，経済の勉強を始めた筆者や，落語や音楽のイベントに参加した筆者などを発見することができた．ゲーム・アニメカテゴリでは，ゲームやシナリオの制作を始めた筆者や，IT カテゴリと同様のケースで，新作ゲームの発売日に向けて発言が増加したというのがあった．

逆に，変化の原因が推定できなかった例としては，推薦結果の中で上昇させたいカテゴリがあまり上昇していない筆者が推薦されてしまうことがあった．これは，上昇させたいカテゴリ以外の類似度の高さによる影響によるものと考えられ，あるカテゴリの類似度だけを見るように推薦アルゴリズムを変えたり，あるいは推薦に利用する筆者の数を増加させたりすることで改善できると思われる．また，評価に使う文書の量が極端に少ない場合，一つの単語が持つ影響力が強くなってしまい，少しの発言の増加が急激な変化として捉えられてしまうという問題があった．これに関しては，推薦に利用する際の文書量の閾値を

設けたり，あるいは発言の量に依存しない興味尺度を検討する必要があると思われる．

以上より，本システムでは特定の興味カテゴリの変化による推薦を行ったとき，筆者自身の変化を捉えた結果を４割含む推薦結果が得られることが分かった．

7. 考　　察

7.1 興味モデル

本研究では，筆者を 5 つの興味カテゴリに対する興味の度合いによって表現することで，人間が直感的に理解でき，また筆者同士の比較や類似検索なども容易に行えるモデルを作成することができた．本システムを利用するユーザが Twitter を利用していれば，自分自身の特徴ベクトルを見ることができ，自分がどの分野に興味を持っているか，あるいは持っていないかという客観的な評価を得ることができる．また，自分以外のモデルを見ることにより，知り合いの筆者のことを理解する手がかりや，自分のモデルとの差異など新たな知見を得るといったことも期待できるだろう．実際，5 つのカテゴリの割合と筆者の記事はうまく対応しており，大まかな区分であれば自分の求めるユーザ像を再現することができるようになっている．

ただし，ユーザ間でヒストグラムが類似していても話題の傾向が違うということはよく起こった．これは興味カテゴリを構成している個々のトピックの割合の違いによるものと考えられる．例えば，現状のシステムでは野球好きもサッカー好きも同じようにスポーツへの興味が強いユーザとして表現されてしまう．これは筆者検索をより実用的にする際には考慮しなければならない問題である．

一つの解決策としては，スポーツや音楽などの個々のトピックに関しても，筆者検索に利用できるようにするということが考えられる．推薦に利用するトピックをユーザがカスタマイズできるようにすることで,「スポーツへの興味が強くて，音楽も少し興味を持っている」というように，より具体的な指定をして筆者を探すことができる．しかしそういった機能を追加していく場合には，直感的な操作性が失われないよう，インターフェースを工夫していく必要がある．

7.2 情報推薦・スパムフィルタへの応用

現在は，検索によって発見された筆者の情報やブログ文書を抽出して提示しているだけであるが，扱っているのが人の興味であることから，様々な応用が考えられる．ブログ本文に含まれる URL 先の情報や，同じ URL を閲覧している他の筆者の情報を利用することでさらに各個人に適した情報推薦を行ったり，強い興味を持っている分野に関する商品推薦を行うシステムも簡単に構築することができる．

また，今回は Twitter から筆者モデルの構築を行ったが，Web ページの閲覧履歴に対して適用することで，Twitter やブログをやっていなくても自分のモデルを構築，閲覧することができるといった拡張が考えられる．

モデルの編集機能の応用例としては，あるカテゴリ（例えば IT）だけが高いようなモデルを作ることもできる．このとき，検索される筆者を見ると大抵は自動的にニュースなどを投稿する Bot となっているという特徴が見られた．これは普通の筆者（人間）というのは極端に偏った話題はしないという性質に合致しており，興味モデルの妥当性を示しているといえる．この性質を利用すると，ある特定のトピックの割合が一定値を超えたり，あるいは他の Bot とヒストグラムの形状が近いといった指標によって，本システムをスパムフィルタに応用することも考えられる．

8. おわりに

本研究では，各ユーザが自分に適した筆者を発見することを目的として，文書による書き手のモデリングおよび，直感的なモデルの編集インターフェース，筆者検索システムを開発した．現在ネット上ではあらゆる人が情報発信を行っているが，そこへ実際にアクセスする手段はキーワード検索・タグ検索などが未だに中心となっており，自分が考えているような興味を持った人にはなかなか出会うことができない．本システムは，筆者の興味を客観的な指標によって表現することにより，人間が直感的に理解でき，また筆者同士の比較や類似検索などが容易に行えるモデルを構築することができた．このため，モデルを編集し，検索を行うことによって目的とする筆者を発見することができるようになった．今後は筆者の内面の興味変化をよりうまく捉えることができる筆者検索を目指していきたい．

なお本研究は，楽天技術研究所の支援を受けた．ここに記して深謝する．

文　　献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, March 2003.

[2] Yoshinori Hijikata, Kazuhiro Iwahama, and Shogo Nishida. Content-based music filtering system with editable user profile. In *Proceedings of the 2006 ACM symposium on Applied computing*, SAC '06, pp. 1050–1057, New York, NY, USA, 2006. ACM.

[3] 桑原雄, 稲垣陽一, 草野奉章, 中島伸介, 張建偉. マイクロブログを対象としたユーザ特性分析に基づく類似ユーザの発見および推薦方式. 研究報告データベースシステム (DBS), Vol. 2009-DBS-149, No. 18, November 2009.

[4] 野田陽平, 清田陽司, 中川裕志. Wikipedia カテゴリを用いたブログ著者の得意分野プロファイリング. NLP 若手の会 第 3 回シンポジウム, September 2008.

[5] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[6] 関口裕一郎, 川島晴美, 奥田英範. ブログ発信者の特徴を利用した話題抽出手法. *DBSJ letters*, Vol. 5, No. 1, pp. 9–12, June 2006.

[7] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5228–5235, April 2004.