

# 日中ブロガー・コミュニティの収集・俯瞰・対照分析

鄭立儀<sup>1</sup> 小池大地<sup>1</sup> 宇津呂武仁<sup>2</sup> 河田容英<sup>3</sup> 神門典子<sup>4</sup>

**概要:** 本論文では、日本語および中国語を対象として、ブログ記事の書き手であるブロガーに注目し、数百人規模のブロガー集合のブログ記事に対してトピックモデルを適用することにより、ブロガー集合をコミュニティへと分類する。そして、トピックモデルによって推定されたトピックを利用することにより、ブロガー・コミュニティの俯瞰を行う。最後に、特定の分野を対象として、日中の間で、コミュニティの分布の違い、および、コミュニティ内の関心の動向の違いの比較対照分析を行う方式を提案する。

**キーワード:** ブロガー, ブログ, コミュニティ, 俯瞰, トピックモデル, 日中文化間差異

## Collecting / Overviewing / Comparative Analysis of Japanese and Chinese Bloggers' Communities

LIYI ZHENG<sup>1</sup> DAICHI KOIKE<sup>1</sup> TAKEHITO UTSURO<sup>2</sup> YASUhide KAWADA<sup>3</sup> NORIKO KANDO<sup>4</sup>

**Abstract:** This paper first studies how to apply a topic model to Chinese and Japanese blog posts collected from a few hundred Chinese and Japanese bloggers and then to classify bloggers into communities. The estimated topics are exploited in the task of overviewing the Chinese and Japanese bloggers' communities and comparing them between Chinese and Japanese in order to discover differences in the communities of the two languages. In the evaluation, we collect a few hundred bloggers from a well-known "Sina Blog" bloggers categories' directory in China, and an also well-known blogger community service "Nihon Blog Mura" in Japan. As a case study, we focus on the "health" category in the services of both languages, and generate communities based on a topic model, and then overview and compare them between Chinese and Japanese. We actually discover certain differences in bloggers' communities regarding "health" between Chinese and Japanese.

**Keywords:** blogger, blog, community, overview, topic model, cultural difference between Japan and China

### 1. はじめに

現代の情報社会においては、情報の氾濫、すなわち、いわゆる情報爆発が起こっている。そして、そのように爆発する情報の集約や、俯瞰をするための技術の開発が強く望まれている。中でも、情報爆発が最も顕著に現れているのはウェブである。ウェブ上の情報の一例として、近年、一

般個人が自由に情報を発信するツールであるブログが世界中で普及し、各地域の人々がそれぞれインターネット上で個人の意見や評判を発信することが可能になった。それに伴い、様々な情報がブログに記載され、様々な人々の意見や評判がウェブ上に氾濫するようになった。

ここで、本論文では、ウェブ上における個人の意見や関心に関する情報を対象として、効率的な俯瞰を行うために、ブログの書き手であるブロガーに注目し、同一の興味を持つブロガーのコミュニティを作成し、俯瞰を行う。ブログ記事を直接俯瞰するのではなく、ブロガーという情報発信者の単位で俯瞰を行うことで、人々の話題への関心の度合いを考慮した、効率的な俯瞰が実現できると考えられる。特に、本論文では、日本語から見れば外国語の一つである中国語のウェブ空間における個人の意見や関心の情報を俯

<sup>1</sup> 筑波大学大学院システム情報工学研究科  
Graduate School of Systems and Information Engineering,  
University of Tsukuba, Tsukuba, 305-8573, Japan

<sup>2</sup> 筑波大学 システム情報系  
Faculty of Engineering, Information and Systems, University  
of Tsukuba, Tsukuba, 305-8573, Japan

<sup>3</sup> (株) ログワークス  
LOG WORKS Co., Ltd., Tokyo 141-0031, Japan

<sup>4</sup> 国立情報学研究所  
National Institute of Informatics, Tokyo 101-8430, Japan

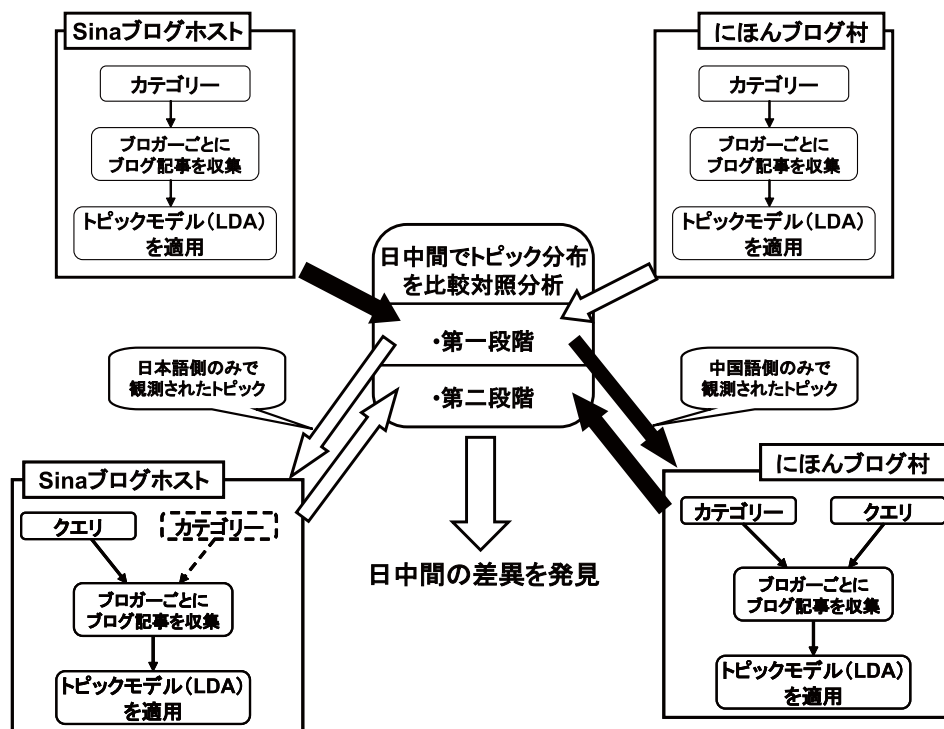


図 1 日中プロガー・コミュニティの比較対照分析

眺めることを目的として、文献 [4] の手法により、日本語および中国語プロガー・コミュニティの収集および俯瞰を行なう。さらに、図 1 に示すように、日中二言語間で、プロガー・コミュニティの比較対照分析を行う方法を提案し、その適用事例について報告する。

まず、日本において人手で作成されたプロガー・コミュニティの代表的なものとして、「にほんブログ村」\*1 (登録プロガー数 68 万人、カテゴリー数 121、サブカテゴリー数 約 5,500) が挙げられる。一方、中国のブログサービスを対象としては、代表的なブログホストとして、「Sina ブログホスト」\*2 (登録プロガー数 2.5 億人、カテゴリー数 23) が知られているが、個々のプロガー自身にはカテゴリーへの登録を手動で行う権限はなく、記事数・閲覧回数・ヒット数に基づいて、少数の人気プロガーのみがカテゴリーに登録されるという方式を採用している。また、「Sina ブログホスト」と「にほんブログ村」を比較すると、「Sina ブログホスト」にはプロガーコミュニティ作成の際にコミュニティ候補の手がかりとなるようなカテゴリー情報が十分には含まれていない。本論文では、これらの日中プロガー・カテゴリー/ディレクトリを対象として、図 1 に示す手順により研究を進める。

## 2. 中国語プロガー及びブログ記事の収集

「Sina」は、中国最大手メディア運営会社であると同時に同国最大の広告会社であり、「Sina ブログホスト」は、中国においても人気の高いブログホストの一つである。表 1 に

表 1 「Sina ブログホスト」及び「にほんブログ村」のカテゴリー数・プロガー数

	カテゴリー数	サブカテゴリー数	プロガー数
Sina ブログホスト	23	—	12,570
にほんブログ村	121	約 5,500	681,041

表 2 分析対象プロガー数およびブログ記事数

	カテゴリー	プロガー数	ブログ記事数
中国語	「健康」	268	7,708
日本語		300	9,380

示すように、「Sina ブログホスト」は、23 個のカテゴリーを持つが、全 2.5 億人の登録プロガーのうち、カテゴリーに登録され、閲覧者が閲覧可能となっている少数の人気プロガーの数は、12,570 人のみである。

本論文では、「Sina ブログホスト」のカテゴリーのうち、「健康」のカテゴリーに着目し、「健康」カテゴリーに属するプロガーを対象として、各プロガーごとに、最新の記事を最大 50 記事収集した。その結果、記事が正しく収集され、分析対象となったプロガー数は、表 2 の「中国語」の欄に示す 268 プロガーとなり、収集されたブログ記事数は 7,708 記事となった。

## 3. 日本語プロガー及びブログ記事の収集

「にほんブログ村」とは、日本最大級のプロガー・コミュニティであり、表 1 に示すように、約 68 万人の登録プロガーが 121 のカテゴリー、および、約 5,500 のサブカテゴリーに登録されている。

\*1 <http://www.blogmura.com/>

\*2 <http://blog.sina.com.cn/>

本論文では、「にほんブログ村」のカテゴリーのうち、「健康」のカテゴリーに着目し、カテゴリーに属するブロガーを収集した。まず、人気ランキングの上位ブロガーから、日本語ブログホスト大手6社<sup>\*3</sup>のドメインを対象として、「健康」のカテゴリーに所属する300ブロガーを対象として、各ブロガーごとに、最新の記事を最大50記事収集した。表2の「日本語」の欄に示すように、記事が正しく収集され、分析対象となったブロガーを、人気ランキングの上位より300ブロガー選定した。収集されたブログ記事数は9,380記事となった。

## 4. トピックモデルを用いたブロガー・コミュニティの俯瞰

### 4.1 概要

本節では、本論文における「ブロガー・コミュニティ」の定義について述べる。

本論文においては、2節、および、3節において収集したブロガーが書いたブログ記事集合に対して、トピックモデルを推定することによりブロガー・コミュニティを生成する。具体的には、ブロガーのブログ記事に対して、トピックモデルを適用することによってトピックを推定し、トピックに対してブログ記事を分類する。そして、あるブロガーが書いたブログ記事が一定数以上、同一トピックに分類された場合に、そのブロガーを当該トピックに分類する。

以上の手順によって、各トピックに分類されるブロガーの集合を、「ブロガー・コミュニティ」と定義する。この際、各トピックに分類されたブログ記事の話題を手手で分析することにより、各ブロガー・コミュニティが一定の話題に対応することとする。

### 4.2 トピックモデル

本論文では、トピックモデルとして潜在的ディリクレ配分法(LDA; Latent Dirichlet Allocation) [2]を用いる。LDAを用いたトピックモデルの推定においては、語 $w$ の列によって表現された文書の集合と、トピック数 $K$ を入力として、各トピック $z_n$  ( $n = 1, \dots, K$ )における語 $w$ の確率分布 $P(w|z_n)$  ( $w \in V$ )、及び、各文書 $b$ におけるトピック $z_n$ の確率分布 $P(z_n|b)$  ( $n = 1, \dots, K$ )を推定する。これらを推定するためのツールとしては、GibbsLDA++<sup>\*4</sup>を用いた。LDAのハイパーパラメータである $\alpha$ 、 $\beta$ には、GibbsLDA++の基本設定値である $\alpha = 50/K$ 、 $\beta = 0.1$ を用いた。LDAではトピック数 $K$ を手手で与える必要があるが、今回はもっともトピックにおける記事のまとまりが良かった50を採用した。

### 4.3 文書に対するトピックの割り当て

本研究では、ブロガーの書いた各ブログ記事に対してトピックを一意的に割り当てることで、ブログ記事を分類することとした。ブログ記事集合を $D$ 、トピック数を $K$ 、1つの文書を $d$  ( $d \in D$ ) とすると、トピック $z_n$  ( $n = 1, \dots, K$ )のブログ記事集合 $D(z_n)$ は以下の式で表される。

$$D(z_n) = \left\{ d \in D \mid z_n = \underset{z_u (u=1, \dots, K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

これはつまり、文書 $d$ におけるトピックの分布において、確率が最大のトピックに、文書 $d$ を割り当てていることになる。

### 4.4 ブロガー・コミュニティへのブロガーの割り当て

本節では、ブロガーを割り当てることで、ブロガー・コミュニティを作成する手法について述べる。

まず、評価対象のブロガー集合 $B$ について、 $B$ 中のブロガーを $b$  ( $b \in B$ ) とする。そして、トピック $z_n$ におけるブロガー $b$ の記事集合を $D(z_n, b)$ とする。ここで、 $D(z_n, b)$ が5記事以上となるような、トピック $z_n$ に対して、ブロガー $b$ を割り当てる。そして、 $B$ 中のすべての $b$ について、トピックへブロガーの割り当てを行い、トピックごとのブロガー集合 $B(z_n)$ を、以下のように定義する。

$$B(z_n) = \left\{ b \in B \mid |D(z_n, b)| \geq 5 \right\}$$

## 5. 日中ブロガー・コミュニティの比較対照分析

### 5.1 分析手順

本論文においては、図1に示すように、二段階の過程を経て日中間のブロガー・コミュニティの比較対照分析を行う。

第一段階においては、「にほんブログ村」と「Sina ブログホスト」から、分析対象とするブロガー及びブログ記事を収集し、前節の手法を用いて、それぞれブロガー・コミュニティを生成する。そして、生成した日中のブロガー・コミュニティを手動で比較し、中国語側のみに存在する話題を持ったブロガー・コミュニティ、および、日本側のみに存在する話題を持ったブロガー・コミュニティをそれぞれ選定する。

第二段階においては、片言語のみにおいて観測されたブロガー・コミュニティを対象として、相手言語側での存在の有無を検証する。

具体的に、中国語側でのみ観測されたブロガー・コミュニティを対象としては、「にほんブログ村」のカテゴリー・サブカテゴリー情報を参照した検証、および、検索エンジンAPIとしてYahoo! Search BOSS API<sup>\*5</sup>を用いた検証の二通りの検証を行う。まず、中国語側でのみ観測された

<sup>\*3</sup> fc2.com, yahoo.co.jp, ameblo.jp, goo.ne.jp, livedoor.jp, hatena.ne.jp

<sup>\*4</sup> <http://gibbslda.sourceforge.net/>

<sup>\*5</sup> <http://developer.yahoo.com/search/boss/>

表 3 日中共通に観測された話題のコミュニティ (「健康」カテゴリー、図 1 の「第一段階」の日中比較対照分析後)

話題の大分類	「にほんブログ村」		「Sinaブログホスト」	
	ブロガー・コミュニティの話題	ブロガー数	ブロガー・コミュニティの話題 (日本語訳)	ブロガー数
病気全般	病気全般、種々の病気について	5	各种疾病(病気全般)	6
美容商品	美容商品、アロマに関する話題	9	美容产品(美容商品)	13
こころの健康	臨床心理士のカウンセラー、 レイキヒーリングについて	4	心理健康(心の健康) (2トピック)	18
	ヒーリング能力伝授について	6		
	心理カウンセリングの話題、 カウンセリングの講座	2	心理学的理论知识 (心理学の知識)	5
	心の健康	22		
飲食と健康	健康な食品に関する紹介、 食品と栄養について	12	健康的饮食(飲食健康) (2トピック)	25
	糖砂栄養素の健康効果	2		
	冷え性対策としての食事療法、 食育、レシピ紹介	12	饮食和营养(饮食と栄養)	21
	薬膳、薬膳茶の紹介	2	健康的食谱 (健康な飲食レシピ)	15
歯の健康	歯科矯正	2	牙齿的健康(歯の健康)	6
筋力 トレーニング	体操、ハビトレなどの 筋力トレーニング (2トピック)	20	运动和健康(運動と健康)	3
整体	整体院のお知らせ	6	鼻などの整容(整容)	6
			整形(整形)	4
健康について の研究	免疫力、糖鎖、サプリメント、 アンチエイジングについての 健康に関する研究	9	关于健康的研究 (健康に関する研究)	8
	睡眠、運動、血液などの 健康情報	13		

ブロガー・コミュニティの話題を、「にほんブログ村」におけるカテゴリーおよびサブカテゴリーと比較し、対応するカテゴリーあるいはサブカテゴリーが存在する場合には、3節の手順によってブロガーおよびブログ記事を収集した後、トピックモデルを適用して、新たにブロガー・コミュニティの生成を行う。一方、適切なカテゴリーもしくはサブカテゴリーが存在しない場合は、検索エンジン API として Yahoo! Search BOSS API を用いて、手動で適切なクエリを与えることによりまずブログ記事集合を収集する。次に、収集したブログ記事に対してブロガーのドメイン情報を抽出し、各ブロガーをドメインとして指定して、クエリを再度与え、検索エンジン API として Yahoo! Search BOSS API を用いてブログ記事集合を適用する。最後に、以上の手順により収集したブログ記事集合を対象としてトピックモデルを適用し、新たにブロガー・コミュニティの生成を行う。

同様に、日本語側でのみ観測されたブロガー・コミュニティを対象としては、「Sina ブログ」において有用なカテゴリーが新たに利用できる可能性が低いことを考慮して、検索エンジン API として Yahoo! Search BOSS API を用いた検証のみを行う。この場合も、手動で適切なクエリを与えることによりブログ記事集合を収集し、収集したブログ記事集合を対象としてトピックモデルを適用し、新たにブロガー・コミュニティの生成を行う。

## 5.2 「健康」カテゴリーのブロガーから生成されたコミュニティにおける日中比較対照分析

前節の手順により、「健康」カテゴリーのブロガーを情報源として、日中ブロガー・コミュニティを生成し、まず、図 1 の「第一段階」の日中比較対照分析を行った結果の各コミュニティごとの話題の一覧を表 3 (日中共通に観測された話題)、表 4 (日本語側でのみ観測された話題)、表 5 (中国語側でのみ観測された話題) にそれぞれ示す。次に、図 1 の「第二段階」の日中比較対照分析を行った結果、片言語側でのみ出現したブロガー・コミュニティの日中比較対照分析を行った結果の例を表 6 に示す。

中国語側でのみ観測したブロガー・コミュニティとして、「二十四節気と健康」があるが、このコミュニティに対しては、「二十四節気」および「健康」をクエリとして Yahoo! Search BOSS API を用いて AND 検索を行った結果では、ブログ記事を収集することができなかった。参考情報として、同様の検索を Google<sup>\*6</sup> を検索エンジンとして行った結果では、一般のウェブページにおいて「二十四節気と健康」を話題とするものを見つけることはできたが、ブログ空間において「二十四節気と健康」を話題とするブログ記事はほとんどみつけれなかった。この結果から、日本においては、「二十四節気と健康の間には何らかの関係がある」との認識が一部で観測されるものの、一般個人の間にも広く浸透しているとは言えないことがわかった。

一方、日本語側でのみ観測したブロガー・コミュニティ

\*6 <http://www.google.jp>

表 4 日本語側でのみ観測された話題のコミュニティ（「健康」カテゴリ、図 1 の「第一段階」の日中比較対照分析後）

話題の大分類	「にほんブログ村」	
	ブロガー・コミュニティの話題	ブロガー数
色彩心理テスト	カラーセラピストによる色に関する話題	12
	オーラソーマに関する知識と体験	6
	フラワーセッションセラピー	4
育毛	育毛のためのシャンプー、育毛剤、植毛	12
眼科	レーシックなどの角膜手術	6
冷えとりの治療	冷えとりのための衣類と靴下	8
足の健康	インソールと足の健康	4
健康講座	健康に関する各種講座	4
禁煙	禁煙に関する話題、体験	5
旅行体験	旅行の体験記、観光日記	7
ヨガ	ヨガインストラクター、ヨガレッスンについて	9
鍼灸、指圧	鍼灸に関する話題	10
	指圧講座・指圧治療の紹介	3
神秘的な治療法	レイキに関する交流会、セミナー	5
	病脈靈感法などの触手療法	3
	真氣光というスピリチュアル系の内容	4
	気功に関する内容、神秘系、スピリチュアルな内容の気功	2
日常記録	日常の出来事 (2トピック)	20

表 5 中国語側でのみ観測された話題のコミュニティ（「健康」カテゴリ、図 1 の「第一段階」の日中比較対照分析後）

話題の大分類	「Sinaブログホスト」	
	ブロガー・コミュニティの話題（日本語訳）	ブロガー数
性教育	性教育（性教育）	14
各種病気	男性病的予防と治療（男性病の予防と治療法）	9
	女性病的予防と治療（女性病の予防と治療法）	20
	关于糖尿病（糖尿病に関して）	8
	高血压，心脏病的注意事项（高血压、心脏病の注意事項）	2
	癌症的预防与早期癌症的发现（ガン予防と早期発見）	5
二十四節気と健康	关于肝炎（肝炎について）	3
	二十四节气相对应的健康注意事项 (二十四節気の各節気での健康に関する注意事項) (2トピック)	4
製薬企業	社会和制药业（社会と製薬企業）	9
医者と患者	对于医疗制度的批判和探讨（医療制度に対する批判と検討）	9
	关于医疗政策的探讨（医療政策についての検討）	2
経済	医患关系（医患関係）	2
	房地产，股票等（不動産、株など）	6
中華風の健康	道德健康学（道德健康学）	3
	中医与民间土方（漢方医学と中国民間療法）	9
	经络和太极拳的健康价值（ツボ、太極拳の健康価値）	5
子供の健康	儿童健康（子供の健康）	5

として、「育毛」に関するもの、および、「レイキ療法」に関するものが確認できた。「育毛」コミュニティに関しては、図 1 の手順の第一段階において、日本語側でのみブロガー・コミュニティが観測された。そこで、第二段階において、「育毛」、「育毛剤」、「ハゲ」の中国語訳をクエリとしてブログ記事を収集し、ブロガー・コミュニティを生成した。ブロガー・コミュニティにおける話題を日中間で比較対照分析したところ、両言語に共通の話題として、「育毛剤を用いて治療」というものが観測されたが、「植毛」については日本語側でのみ観測され、「生姜を頭皮に塗る等の自然療法」については中国語側でのみ観測される、という結果となり、日中間の差異が発見された。「レイキ療法」に関しても、図 1 の手順の第二段階において、中国語側でのブログ記事収集を試みたが、結果的に、ブロガー・コミュニ

ティの生成にまでは至らず、日本に固有の民間療法であるという事前知識を裏付ける結果となった。

## 6. 関連研究

本論文の先行研究として、我々は、文献 [5] において、特定の話題について、日本語ブログ記事、および、英語ブログ記事を収集し、関心事項や賛否に関する文化間差異発見過程を支援する方式を提案した。この方式の成果として、「捕鯨」や「臓器移植」など、日本と欧米圏との間で社会制度上の違いや食文化の差異が大きい話題について、ブログ空間における関心の違いを容易に観測することができた。一方、[8] においては、特定の話題に関するブログ記事集合において、日本語・英語二言語での観点を分類・比較・対照分析する手法が提案されている。また、[6] においては、

表 6 「健康」カテゴリーのプロガーから生成されたコミュニティにおける日中間差異の例  
 (図 1 の「第二段階」の日中比較対照分析後)

	話題	日本語コミュニティ		中国語コミュニティ	
		コミュニティ収集手順	話題	コミュニティ収集手順	話題
中国語側のみで観測	「二十四节气与健康」 (二十四節気与健康)	「二十四節気 and 健康」をクエリとして、ブログ記事を検索したが、収集されなかった		「健康」カテゴリーのプロガーから生成されたコミュニティの一つとして存在	1年間における各節気での健康に関する注意事項
日本語側のみで観測	「育毛」	「健康」カテゴリーのプロガーから生成されたコミュニティの一つとして存在	・日中共通の治療方法(育毛剤) ・日本でのみ観測された治療方法(植毛)	「生发」(育毛)、「生发剂」(育毛剤)、「秃头」(ハゲ)をクエリとして、ブログ記事を収集して、トピックモデルを適用し、「育毛」に関するコミュニティを生成	・中国でのみ観測された治療方法:自然療法(例:生姜を頭肉に塗る)
	「レイキ療法」	「健康」カテゴリーのプロガーから生成されたコミュニティの一つとして存在	代替医療であり、一種の手当て療法である(日本の民間療法)	「灵气疗法」(レイキ療法)、「灵气」(レイキ)、「灵能力」(ヒーリング)、「苏摩」(ソーマ)をクエリとして、ブログ記事を収集して、トピックモデルを適用し、「レイキ療法」に関するコミュニティが生成されなかった	

日中質問回答サイトを対象として、トラブル情報の比較対照分析を行い、文化間差異発見支援を行う方式を提案している。ただし、これらのブログおよび質問回答サイトを対象とした研究においては、トピックモデルによって話題のまとまりを同定する過程が欠如しており、比較的小規模な文書集合を対象とした人手による分析に重点が置かれている点が、本研究とは大きく異なる。

一方、複数情報源からのニュースの多言語間差異分析を行っている研究として、文献 [1,7,9,10] が挙げられる。文献 [9] は、32 言語における 1,000 以上の情報源を分析し伝染病に関するレポートをまとめあげる研究を行っている。文献 [7] では、32 言語におけるニュース記事群から特定の人物名を収集し、その人物の人間関係やその人物について言及している各国のニュース記事を継続的に分析する研究を行っている。文献 [10] は、複数の国の代表的なメディアが発信するニュースを情報源として、同一事象に対する各国のニュースの伝え方の差異分析方式を提案している。文献 [1] では、9 言語間における同一事象に対する主観情報の差異分析の研究を行っている。これらの研究は主にニュース記事を対象に分析を行っている点で本論文とは異なる。

また、本論文に関連して、文献 [3] においては、日中の時系列ニュースに対して時系列トピックモデルを適用し、日中単言語のトピックの間の言語間対応をとることにより、同一の話題に関するニュース記事の集合を持つ日中各言語のトピックを同定する方式を提案している。

## 7. おわりに

本論文では、日中二言語間で、プロガー・コミュニティの比較対照分析を行う方式を提案し、その適用事例について報告した。今後は、Wikimedia 等を情報源とする日中対訳知識を利用することにより、日中間のプロガー・コミュニティの対応付けを自動的に行う手法を確立するとともに、

既存のプロガー・コミュニティ・サービスにおける全カテゴリを対象として、日中間差異の発見を網羅的に行う予定である。

## 参考文献

- [1] Bautin, M., Vijayarenu, L. and Skiena, S.: International Sentiment Analysis for News and Blogs, *Proc. ICWSM*, pp. 19–26 (2008).
- [2] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).
- [3] 胡 碩, 高橋佑介, 鄭 立儀, 宇津呂武仁, 吉岡真治, 神門典子: 日中時系列ニュースにおけるバースト・トピックの推定と二言語間対応付け, 言語処理学会第 19 回年次大会論文集, pp. 204–207 (2013).
- [4] 牧田健作, 鈴木浩子, 小池大地, 鄭 立儀, 宇津呂武仁, 河田容英, 神門典子: トピックモデルを用いたプロガー・コミュニティの収集と俯瞰, 第 5 回 DEIM フォーラム論文集 (2013).
- [5] 中崎寛之, 川場真理子, 横本大輔, 宇津呂武仁, 福原知宏: 多言語 Wikipedia エントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析, 人工知能学会論文誌, Vol. 25, No. 5, pp. 613–622 (2010).
- [6] 轟 添, 新井翔太, 宇津呂武仁, 河田容英: 日中質問回答サイトの比較対照分析および文化間差異発見支援, 第 27 回人工知能学会全国大会論文集 (2013).
- [7] Poulliquen, B., Steinberger, R. and Belyaeva, J.: Multilingual Multi-document Continuously-updated Social Networks, *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 25–32 (2007).
- [8] 鈴木浩子, 横本大輔, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏: Wikipedia を知識源とする日英ブログ記事集合の観点分類と言語間対照分析, 情報処理学会研究報告, Vol. 2011–DBS–153 (2011).
- [9] Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D. and Steinberger, R.: Combining Information about Epidemic Threats from Multiple Sources, *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 41–48 (2007).
- [10] Yoshioka, M.: IR Interface for Contrasting Multiple News Sites, *Prof. 4th AIRS*, pp. 516–521 (2008).