Vol.2012-DBS-155 No.11
2012/11/19

Wikipediaを知識源とする分野トピックモデルの推定と分析

牧田 健作[1]　鈴木 浩子[1]　小池 大地[1]　宇津呂 武仁[2]　河田 容英[3]

概要：本論文では，特定のキーワードをクエリとして収集したブログ記事集合を対象として，ブログ記事集合中の話題の広がりを俯瞰することを目的として，Wikipedia を知識源とする分野トピックモデルを提案し，その推定法，および，ブログ記事集合への適用結果について述べる．具体的には，ブログ記事集合から抽出した Wikipedia エントリタイトルに対して，「地球温暖化」における「気象学・天文学・生物学・エネルギー・工業」といった分野に対応するトピックモデルを推定し，その特性を分析する．特に，この Wikipedia を知識源とする分野トピックモデルを，ブログ記事集合から推定した通常のトピックモデルと比較して，両者の特性の違いを分析し，ブログ記事集合中の話題の広がりを俯瞰する目的において両者が相補的な関係にあることを示す．

キーワード：ブログ，Wikipedia，トピックモデル，LDA，トピック分析

Estimating and Analyzing a Domain Topic Model of Wikipedia Entries

Kensaku Makita[1]　Hiroko Suzuki[1]　Daichi Koike[1]　Takehito Utsuro[2]　Yasuhide Kawada[3]

Abstract: In order to address the issue of quickly overviewing the distribution of the contents of the collection of blog posts, this paper proposes a framework of estimating a topic model, namely "a domain topic model", which is a topic model estimated with the texts of Wikipedia entries extracted from the collection of blog posts. In this "domain topic model" of Wikipedia entries, each topic represents domains such as meteorology, astronomy, biology, energy, and industry, that are closely related to a query term, e.g., "global warming". We compare the proposed approach of topic modeling *with* Wikipedia knowledge source and the standard topic modeling *without* Wikipedia knowledge source. Both topic modeling results have quite different nature and contribute to quickly overviewing the search result of blog posts in a quite complementary fashion.

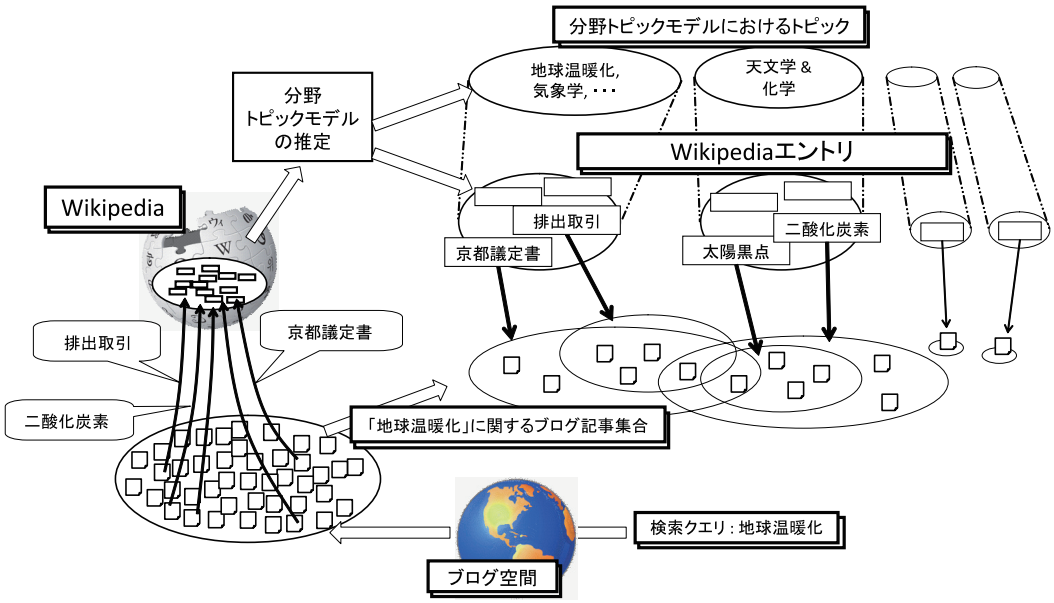
Keywords: blog, Wikipedia, topic model, LDA, topic analysis

1. はじめに

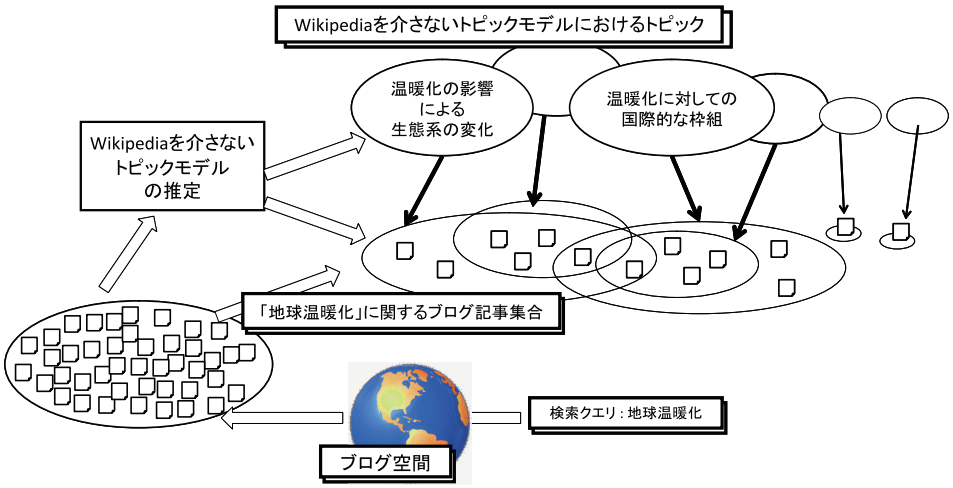
近年，世界中でブログサービスやブログツールが普及し，各地域の人々がそれぞれインターネット上で個人の意見や評判を発信することが可能になった．それに伴い，様々な情報がブログに記載され，商用ブログ検索サービスを利用することでそれらの情報を取得することができるようになった．

しかし，特定のトピックについて検索を行った場合でも，その検索結果には様々な観点が混在している．例えば「地球温暖化」というトピックを検索クエリとしてブログ記事を収集した結果においては，生物学的観点から，生態系への影響を話題にしているブログ記事や，政治学的観点から，温暖化対策の一つである排出取引について書いているブログ記事，天文学的観点から，地球温暖化の原因は二酸化炭素などではなく太陽活動の変化である，と述べているブログ記事など，「地球温暖化」について様々な観点で書かれたブログ記事が得られる．このように，検索結果には様々な

[1] 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
[2] 筑波大学 システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, 305–8573, Japan
[3] (株) ログワークス
LOG WORKS Co., Ltd., Tokyo 141-0031, Japan



(a) Wikipedia を知識源とする分野トピックモデル



(b) Wikipedia を介さないトピックモデル

図 1 Wikipedia を知識源とする分野トピックモデルおよび Wikipedia を介さないトピックモデル

観点が混在しているため，検索結果を単なるリストとして提示するだけでは，検索結果にどのような観点が含まれているのか知ることができない．

このような課題を解決するための一つのアプローチとして，情報検索分野においては，ファセット検索の考え方 [15] が広く知られている．ここで，一般に，ファセット検索の枠組みにおいては，検索対象の各文書に対して，あらかじめ人手もしくは自動でファセットラベルを付与しておく必要がある．そこで，この問題を解決するために，我々は，これまでに，文献 [17] において，Wikipedia を知識源として，Wikipedia の各エントリ本文と検索対象の文書との間の文書類似度を測定し，類似する Wikipedia エントリごとに文書をクラスタリングすることにより，Wikipedia エントリをファセットラベルとするファセット検索の枠組みを自動構築する方式を提案した．

しかし，この方式においては，収集された文書集合における話題の偏りや分布状況を考慮することができず，各文書ごとに独立にファセットラベルの付与が行われており，この点が根本的な問題となっていた．また，Wikipedia に登録されている専門的な用語と検索対象文書の類似度が大きくなり，そのような専門的用語のみがファセットラベルとして付与される場合も多く観測され，検索された文書集合全体にわたる分野や話題の分布を俯瞰することは容易ではなかった．

ここで，前者の，文書集合における話題の偏りや分布状況を考慮することができないという問題点に対しては，潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [3] をはじめとするトピックモデルを適用することにより，Wikipedia 等の外部知識を用いずに，文書集合中の話題の分布を把握することができる．「地球温暖化」を

検索クエリとして収集したブログ記事集合を対象として，LDA 等の典型的なトピックモデルを適用し，ブログ記事集合中の話題の分布を俯瞰する枠組みの模式図を，図 1(b)「Wikipedia を介さないトピックモデル」に示す．また，表 6 の $BT1 \sim BT16$ に，「地球温暖化」を検索クエリとして収集したブログ記事集合を対象として，実際に LDA を適用して得られたトピックの特徴を人手で記述した結果を示す．この結果から分かるように，収集されたブログ記事集合に対して LDA を直接適用した場合には，Wikipedia 等の外部知識を用いていないため，文書集合における話題の偏りや分布状況を直接反映する形でトピックの分布が推定される点が大きな利点となる．その一方で，推定されたトピックの一つ一つが比較的専門性の高い話題のブログ記事の集まりに対応しており，検索された文書集合全体にわたる分野や話題の分布を俯瞰することは容易ではない．

以上の状況をふまえて，本論文では，図 1(b) に示す「Wikipedia を介さないトピックモデル」のように，収集された文書集合に直接適用することにより推定されたトピックモデルに対して相補的な役割を担うトピックモデルとして，Wikipedia を知識源として用いる**分野トピックモデル**を提案する．本論文で提案する分野トピックモデルの枠組みの模式図を図 1(a) に示す．Wikipedia を知識源とする分野トピックモデルの推定時には，まず，収集された文書集合 (図 1(a) の場合には，ブログ記事集合) から Wikipedia エントリのタイトルが抽出される．そして，各タイトルのエントリ本文を Wikipedia から収集し，収集された Wikipedia エントリ本文の集合を対象としてトピックモデルを推定し，得られたトピックモデルを分野トピックモデルとする．この分野トピックモデルは，表 6 の $WT1 \sim WT8$ の例に示すように，「地球温暖化」における「気象学・天文学・生物学・エネルギー・工業」といった分野に対応している．この分野トピックモデルを，表 6 の $BT1 \sim BT16$ に示す従来型の「Wikipedia を介さないトピックモデル」の各トピックとあわせて相補的に用いることにより，それらの従来型の複数のトピックを包含し，検索された文書集合全体にわたる分野や話題の分布を俯瞰することが可能となり，収集された文書集合の効率的閲覧が促進される．

以下の各節においては，本論文で提案する分野トピックモデルの推定手順，および，推定された各トピックを用いて文書集合における分野や話題の分布を俯瞰する手法について述べる．以下の各節のうち，特に，4.2 節においては，分野トピックモデルの推定手順について述べ，次に，4.3 節において，分野トピックモデルに対して，トピックモデルにより文書を生成するクエリ尤度モデル [13,16] を適用し，Wikipedia エントリからブログ記事を生成する確率をモデル化する．そして，この確率を用いて，ブログ記事との間の適合度合いにしたがって，Wikipedia エントリの順位付けを行う．さらに，4.4 節において，各ブログ記事に対して，分野トピックモデルのトピックを分野ラベルとして付与する方式を提案する．以上の方式について，5 節において評価実験の結果を示し，それぞれのタスクにおいて安定した性能が達成できることを示す．最後に，6 節において，分野トピックモデルと図 1(b) に示す「Wikipedia を介さないトピックモデル」との間の比較対照分析を行う手順およびその結果について述べ，両者のトピックモデルの相補的特性について考察する．

2. 関連研究

本論文に関連して，ファセット検索の研究分野においては，TREC-2009 におけるブログ検索タスク [11] において，ファセット検索によるブログサイト検索タスクが導入され，「意見の有無」，「個人的情報・公的情報の別」，「トピックについて専門的あるいは詳細な情報を含むか否か」の 3 種類のファセットをブログサイトに付与するタスクが行われた．

文献 [6] においては，検索対象の文書に対して自動的にファセットラベルを付与し，ファセット検索を行う枠組みとして，トピック，ブログ記事の書き手 (ブロガー)，ブログ記事のリンク先，ブログ記事中の主観表現といったファセットラベルを付与し，ファセット検索の枠組みによりブログ記事集合を閲覧する枠組みを提案している．一方，文献 [9] においては，Wikipedia 中の記事を閲覧対象として，ファセットラベルそのものも Wikipedia 中から自動収集し付与することにより，Wikipedia 中の記事集合を俯瞰する枠組みを提案している．本論文の研究とこの方式との間の最も重要な違いとして，本論文の方式においては，Wikipedia 中の記事集合にとどまらず，任意の文書集合を閲覧対象とできる点が大きな利点である．また，ファセットラベルの体系に相当する分野トピックモデルの推定において LDA を用いているため，閲覧対象の文書集合に応じて，臨機応変にファセットラベルの体系が構築される点が長所である．

また，その他に，Web ページの検索結果を分類し，各分類に対して適切な要約文を付与するという手法 [7]，および，検索された個々の Web ページに対してラベルの付与を行い，付与されたラベルに基づいて分類を行う手法 [1,5,14]，階層的なトピックの体系を推定する手法 [2] 等が提案されている．これらの手法においては，いずれも，閲覧対象の文書集合のみを用いて，ファセット体系およびファセットラベルに相当する情報を抽出している．一方，本論文の手法において推定される分野トピックモデルにおいては，Wikipedia を知識源として，検索された文書集合全体にわたる分野や話題の粒度にまで抽象化されたトピックをファセット体系とする点が大きく異なる．

さらに，本論文の研究に関連して，文書集合をクラスタリングした結果の各クラスターのラベル付けにおいて，

Wikipedia を知識源として用いる手法 (例えば，文献 [4]) 等も提案されている．しかし，これらの手法においては，本論文の分野トピックモデルのように，複数のクラスターを包含する分野ラベルに相当する構造を俯瞰することは行っておらず，この点が大きく異なっている．

一方，トピックモデルとして LDA を用いて文書を生成するクエリ尤度モデルの研究 [16] においては，文書モデルのスムージング手法における比較対象として，pLSI を用いる手法 [8]，および，文書クラスターを用いる手法 [10] をとり上げ，LDA を用いることによりそれらの短所を改善できると論じている．

3. 分析対象ブログ記事の収集

本論文においては，初期クエリ t_0 に密接に関連するブログ記事の候補を収集し，これを分析対象とする．具体的には，以下の手順にしたがい，初期クエリ t_0 を含むブログ記事を収集し，これを分析対象ブログ記事集合 $BP(t_0)$ とする．

初期クエリ t_0 を含むブログ記事の収集においては，Yahoo!Japan 検索 API[*1]を利用し，t_0 をクエリとして，日本語ブログホスト大手 8 社[*2]のドメインに限定し，2010 年 7〜9 月の期間に検索を行った．検索の際には，複数のドメインを一度に指定して検索し，1,000 件の記事を取得する．次に，ブログ記事検索後，検索結果の URL をブログサイト単位にまとめる．その結果，一つの検索クエリあたり約 200 前後のブログサイトが取得される．次に，各ブログサイトをドメイン指定し，t_0 を検索クエリとすることにより，各ブログサイト中において t_0 を含むブログ記事を収集し，ブログ記事集合 $BP(t_0)$ を作成する．

4. Wikipedia を知識源とする分野トピックモデル

4.1 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) [3] 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 の集合を V として，語 $w(\in V)$ 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_k $(k = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_k)$ $(w \in V)$，及び，各文書 d におけるトピック z_k の確率分布 $P(z_k|d)$ $(k = 1, \ldots, K)$ を推定する．これらを推定するためのツールとしては，GibbsLDA++[*3] [12] を用いた．LDA のハイパーパラメータである α，β 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 を人手で与える必要があるが，本論文では，トピック数を 50 とした．

4.2 ブログ記事集合に対する分野トピックモデルの推定

ブログ記事集合 $BP(t_0)$ に対して分野トピックモデルを推定するために，まず，$BP(t_0)$ 中に出現する Wikipedia エントリタイトルを収集する．ここでは，予備実験の結果をふまえて，Wikipedia エントリ E のタイトル $t(E)$ に対して，ブログ記事集合 $BP(t_0)$ における文書頻度 df の下限を 10 とし，以下の式にしたがって，ブログ記事集合 $BP(t_0)$ に対する Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ を作成する．

$$\mathbb{E}(BP(t_0)) = \left\{ E \;\middle|\; \mathrm{df}(\ BP(t_0),\ t(E)\ ) \geq 10 \right\}$$

次に，Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ を，各 Wikipedia エントリの本文テキストを要素とする文書集合とみなして，前節の手順にしたがい，LDA を適用しトピックモデルを推定する．ただし，その際，語 w の集合 V としては，日本語 Wikipedia 中のタイトルの集合[*4]を用いる．以上の手順により推定したトピックモデルを，ブログ記事集合 $BP(t_0)$ に対する**分野トピックモデル**と呼ぶ．分野トピックモデルの各トピックを z^e_k $(k = 1, \ldots, K)$ と記述すると，分野トピックモデルの推定結果としては，各トピック z^e_k $(k = 1, \ldots, K)$ における語 w の確率分布 $P(w|z^e_k)$ $(w \in V)$，及び，Wikipedia エントリ E におけるトピック z^e_k の確率分布 $P(z^e_k|E)$ $(k = 1, \ldots, K)$ が得られる．

4.3 分野トピックモデルによりブログ記事を生成するクエリ尤度モデル

通常，クエリ尤度モデル [13] においては，クエリ q に対して，文書 d が適合する確率 $P(d \mid q)$ によって，文書 d の順位付けを行う．ここで，ベイズの定理を用い，また，$P(q)$ は文書 d に依存しないので定数とみなすとともに，文書 d に関しての何らかの事前知識がない限り，$P(d)$ は一様であるとみなすことにより，次式による簡略化を行う．

$$P(d \mid q) = \frac{P(q \mid d)P(d)}{P(q)} \propto P(q \mid d)P(d) \propto P(q \mid d)$$

このように，クエリ尤度モデルにおいては，文書 d からクエリ q が生成される確率 $P(q \mid d)$ をモデル化し，この確率を用いて，クエリ q に対する文書 d の順位付けを行う．

本論文では，このクエリ尤度モデルに基づき，ブログ記事 $B(\in BP(t_0))$ をクエリとして，ブログ記事 B 中に含まれる Wikipedia エントリタイトル $t(E)$ のエントリ本文 E がブログ記事 B に適合する度合いによって，Wikipedia エントリの順位付けを行う．そして，ブログ記事 B に対し

[*1] http://www.yahoo.co.jp/
[*2] fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, hatena.ne.jp
[*3] http://gibbslda.sourceforge.net/
[*4] 日本語 Wikipedia としては，2010 年 2 月にダウンロードした，エントリ数約 65 万 8,000 のものを用いた．

て，順位付けが上位の Wikipedia エントリタイトルをラベル付けする，というアプローチをとる．

具体的には，まず，ブログ記事 B に対して，順位付けの対象となる Wikipedia エントリ E としては，前節で作成した Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ の要素に限定する ($E \in \mathbb{E}(BP(t_0))$)[*5]．そして，ブログ記事 B 中の語 w を，日本語 Wikipedia 中のエントリのタイトルに限定したうえで，ブログ記事 B 中における複数の語の間の独立性を仮定して，Wikipedia エントリ E からブログ記事 B を生成する確率 $P(B \mid E)$ を次式で定義する．

$$P(B \mid E) = \prod_{w \in B} P(w \mid E)$$

次に，文献 [16] にしたがい，確率 $P(w \mid E)$ を，トピックモデルとして LDA を用いた場合の確率 $P_{lda}(w \mid E)$，Wikipedia エントリ E における語 w の最尤推定値 $P_{ML}(w \mid E)$，および，Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ の全体における語 w の最尤推定値 $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ の線形補間として，次式によってモデル化する[*6]．

$$\begin{aligned} P(w \mid E) \;=\; & \lambda \Big\{ \mu P_{ML}(w \mid E) \;+ \\ & (1-\mu) P_{ML}(w \mid \mathbb{E}(BP(t_0))) \Big\} \\ & + (1-\lambda) P_{lda}(w \mid E) \end{aligned} \tag{1}$$

ただし，トピックモデルとして LDA を用いた場合の確率 $P_{lda}(w \mid E)$ は，分野トピックモデルのトピックを z_k^e として，次式によって与えられる．

$$P_{lda}(w \mid E) \;=\; \sum_{k=1}^{K} P(w \mid z_k^e) P(z_k^e \mid E)$$

4.4 ブログ記事への分野ラベルの付与

前節の手順により，Wikipedia エントリ E からブログ記事 B を生成する確率 $P(B \mid E)$ に基づいて，ブログ記事 B に対して，Wikipedia エントリ E を順位付けた．一方，4.2 節の手順によって分野トピックモデルを推定した結果，Wikipedia エントリ E におけるトピック z_k^e の確率分布 $P(z_k^e|E)$ $(k = 1, \ldots, K)$ を得た．そこで，本節では，これらの確率を用いて，次式によって，各ブログ記事 B に対して，分野トピックモデルにおける各トピック z_k^e の重み $score_e(B, z_k^e)$ を求め，

$$score_e(B, z_k^e) = \sum_{E \in \mathbb{E}_{30}(B)} P(B \mid E) P(z_k^e \mid E) \tag{2}$$

この重みが上位のトピックをブログ記事 B に付与する，という考え方を導入する．ただし，本論文では，重み $score_e(B, z_k^e)$ を求める際に参照する Wikipedia エントリ E としては，確率 $P(B \mid E)$ の上位 30 エントリに限定することとし，それらの Wikipedia エントリの集合を $\mathbb{E}_{30}(B)$ と記述する[*7]．

ここで，5.2 節の評価・分析において述べるように，分野トピックモデルの各トピックは，例えば，クエリ「地球温暖化」における「気象学・天文学・生物学・エネルギー・工業」といった分野に対応したトピックとなっている．そこで，本論文では，分野トピックモデルの各トピックが，情報の粒度として，「分野」程度のものを表現すると考えて，ブログ記事に対して，分野トピックモデルのトピックを付与することを，「ブログ記事に対して**分野ラベルを付与**する」と呼ぶ．

5. 分野トピックモデルの評価および分析

本論文で提案する分野トピックモデルに対して，その性能を評価するために，4.3 節で述べた「分野トピックモデルによりブログ記事を生成するクエリ尤度モデル」によって，ブログ記事をクエリとして，Wikipedia エントリを順位付けした結果の評価を行った．さらに，4.4 節で述べた手法により，ブログ記事に対して分野ラベルを付与した結果の評価を行った (ただし，4.1 節より $K = 50$)．

5.1 ブログ記事をクエリとする Wikipedia エントリの順位付け

5.1.1 評価手順

初期クエリとして，表 1 に示す 5 種類のキーワードを対象として，評価および分析を行った．表 1 には，これらの 5 種類の初期クエリを対象として収集したブログ記事集合 $BP(t_0)$ 中のブログ記事数もあわせて示す．本節の評価においては，各初期クエリについて無作為に 60 記事のブログ

表 1 初期クエリおよび評価対象ブログ記事数

初期クエリ t_0	評価対象ブログ記事数 $\lvert BP(t_0) \rvert$
喫煙	8,834
臓器移植	1,402
地球温暖化	7,199
医療事故	1,823
プリウス	4,211

[*5] 実際には，さらに，ブログ記事 B 中に含まれる低頻度語の影響を緩和するために，Wikipedia エントリ E のタイトル $t(E)$ に対して，ブログ記事 B における頻度 $freq(B, t(E))$ が 3 以上であるという下限を設けている．

[*6] Wikipedia エントリ E における語 w の最尤推定値 $P_{ML}(w \mid E)$，および，Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ の全体における語 w の最尤推定値 $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ の補間において，文献 [16] で述べられているディリクレ・スムージングと本節で用いている線形補間との比較を行ったところ，$\lambda = \mu = 0.7$ の場合に，線形補間を用いた場合の性能が，ディリクレ・スムージングを用いた場合の性能を上回ったため，本論文においては，線形補間の方を採用した．

[*7] 5.1 節における「ブログ記事をクエリとする Wikipedia エントリの順位付け」の評価結果から分かるように，Wikipedia エントリの順位付け結果においては，適合率が 70%程度の場合に再現率が 60%程度となっている．ここで，評価対象となった Wikipedia エントリは，上位の 50 エントリであることから，上位の 30 エントリ中に含まれる関連エントリが再現率 60%に対応し，その場合の適合率が約 70%となることから，重み $score_e(B, z_k^e)$ の計算においても，上位の 30 エントリを用いることとした．

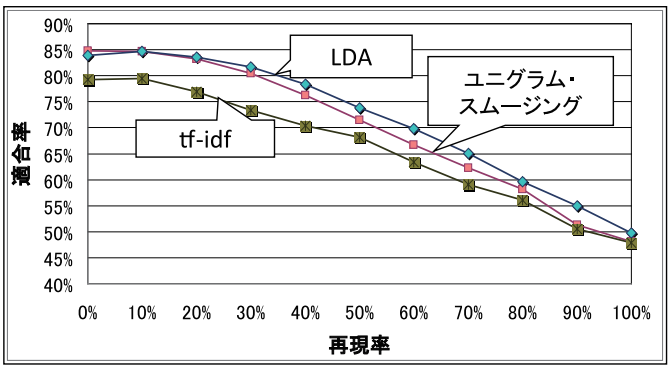


図 2 評価結果: ブログ記事をクエリとする Wikipedia エントリの順位付け

記事を選定し，合計 300 記事のブログ記事を対象として，4.3 節で述べた「分野トピックモデルによりブログ記事を生成するクエリ尤度モデル」によって Wikipedia エントリを順位付けした結果の評価を行った．評価の際には，各ブログ記事に対して，順位付けされた Wikipedia エントリの各々に対して，クエリとして用いたブログ記事との間の関連性の有無を人手で判定した．そして，以下の三種類の手法の間で，判定結果の比較を行った．

(1) 本論文の分野トピックモデルに基づいて，4.3 節で述べた確率 $P(B \mid E)$ を用いて Wikipedia エントリの順位付けを行う (図 2 中の評価結果においては，「LDA」と表記).
(2) (1) の確率 $P(B \mid E)$ の定義において，式 (1) において，トピックモデルによる項 $P_{lda}(w \mid E)$ を用いず，Wikipedia エントリ E における語 w の最尤推定値 $P_{ML}(w \mid E)$，および，Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ の全体における語 w の最尤推定値 $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ の線形補間 ($\mu = 0.7$) のみとする (図 2 中の評価結果においては，「ユニグラム・スムージング」と表記).
(3) Wikipedia エントリ E を，クエリであるブログ記事 B との間の文書類似度の降順に順位付けする．ただし，文献 [17] にしたがい，文書類似度は，クエリであるブログ記事 B 中の語の頻度ベクトルと Wikipedia エントリ E の本文中の語の逆文書頻度ベクトルとの間の内積によって表現される (図 2 中の評価結果においては，「tf-idf」と表記).

5.1.2 評価結果

前節の三種類の手法の各々について，各ブログ記事に対して順位付けられた Wikipedia エントリのうち，最大で上位の 50 エントリを評価対象として，クエリとなったブログ記事と各 Wikipedia エントリとの間の関連性の有無を人手で判定した．評価対象の Wikipedia エントリのうち，クエリとなったブログ記事との間で関連性があり，かつ，順位最下位の Wikipedia エントリ，および，それより上位の Wikipedia エントリを対象として測定した再現率を 100%として，再現率が 0%，10%，...，90%，100%となる 11 点において，評価対象の 300 ブログ記事における適合率・再現率のミクロ平均をプロットしたものを図 2 に示す．

表 2 評価結果: ブログ記事への分野ラベルの付与

初期クエリ t_0	$score_e(B, z_k^e)$ 最大となる分野ラベル (=分野トピックモデルのトピック z_k^e) の正解率 (%)
喫煙	82.5 (429/520)
臓器移植	78.7 (365/464)
地球温暖化	71.8 (348/485)
医療事故	80.1 (370/462)
プリウス	72.2 (372/515)

評価対象の 300 ブログ記事において，平均 16.0 個の Wikipedia エントリが評価対象となっており，そのうち，平均 6.0 個がクエリとなったブログ記事との間で関連性があると判定された．さらに，「LDA」，「ユニグラム・スムージング」，「tf-idf」の三手法のうち，概ね，「LDA」が最も性能がよい，という結果となった．また，「LDA」と「tf-idf」の性能の差について，正規分布に基づく母比率の差の統計的有意差検定を行ったところ，10%，...，90%の 9 点において，有意水準 1%で有意な差となった．「LDA」，「ユニグラム・スムージング」と，「tf-idf」との間の最も大きな違いとして，前者においては，確率 $P(B \mid E)$ の計算において，ブログ記事集合 $BP(t_0)$ から収集した Wikipedia エントリの集合 $\mathbb{E}(BP(t_0))$ 中における語 w の分布が考慮されるのに対して，後者においては，クエリとなるブログ記事 B と Wikipedia エントリ E との間の文書類似度の計算において，B および E における語 w の分布のみが考慮される点が挙げられる．

5.2 ブログ記事への分野ラベルの付与

次に，各初期クエリ t_0 に対して，ブログ記事集合 $BP(t_0)$ 中のブログ記事のうち，評価対象の 500 記事程度を選定し[*8]，ブログ記事に対して分野ラベルを付与した結果の評価を行った．具体的には，評価対象のブログ記事 B に対して，4.4 節の式 (2) で定義した重み $score_e(B, z_k^e)$ が最大となる分野ラベル (分野トピックモデルのトピック z_k^e) が適切であるか否かの判定を人手で行った[*9]．この評価結

[*8] 500 記事程度の選定においては，まず，各ブログ記事 B に対して，重み $score_e(B, z_k^e)$ が最大となる分野ラベル (分野トピックモデルのトピック z_k^e)，および，重み $score_e(B, z_k^e)$ の最大値を付与する．そして，ブログ記事集合 $BP(t_0)$ における分野ラベル，および，重みの最大値の分布を反映するように 500 記事程度を選定する．

[*9] 評価作業の際には，評価作業を円滑に進めるために，分野トピックモデルの各トピック z_k^e に対して，

$$z_k^e = \operatorname*{argmax}_{z'_k} P(z'_k \mid E) \quad (3)$$

となる Wikipedia エントリ E をふまえて，各トピック z_k^e に対して，「生物学」，「地学」，「気象学」，「地球温暖化」といった分

表 3 ブログ記事への分野ラベル (=分野トピックモデルのトピック z_k^e) 付与の例 (初期クエリ: 「地球温暖化」の場合) (1)

分野トピックモデルのトピックの ID		*WT1*: 生物学	*WT2*: 地学，気象学，地球温暖化	*WT3*: 天文学，化学	*WT4*: 政治学
ブログ記事 *B1*	$\mathbb{E}_{30}(B1)$ 中の Wikipedia エントリ (抜粋)	—	地球温暖化の原因，温室効果，地球寒冷化	地球，太陽放射，二酸化炭素，太陽黒点	—
	ブログ記事の概要	—	CO2 は地球温暖化の原因ではない．太陽黒点の影響のため，地球温暖化ではなく地球寒冷化が起こっている．		—
ブログ記事 *B2*	$\mathbb{E}_{30}(B2)$ 中の Wikipedia エントリ (抜粋)	ウミガメ，絶滅，孵化	砂浜，地球温暖化，環境，太平洋	—	—
	ブログ記事の概要	地球温暖化がウミガメの生態に影響を与えている．		—	—
ブログ記事 *B3*	$\mathbb{E}_{30}(B3)$ 中の Wikipedia エントリ (抜粋)	—	京都議定書，地球温暖化，排出取引	—	民主党，マニフェスト
	ブログ記事の概要	—	京都議定書の規定を守ることは日本にとっては不利益である．	—	民主党のマニフェストを実行することは日本にとって不利益である．

表 4 ブログ記事への分野ラベル (=分野トピックモデルのトピック z_k^e) 付与の例 (初期クエリ: 「地球温暖化」の場合) (2)

分野トピックモデルのトピックの ID		*WT5*: 発電，エネルギー	*WT6*: 工業	*WT7*: 農業，林業	*WT8*: 金融，経済
ブログ記事 *B4*	$\mathbb{E}_{30}(B4)$ 中の Wikipedia エントリ (抜粋)	固定価格買い取り制度，代替エネルギー，太陽光発電	—	—	税，価格，政策，会社
	ブログ記事の概要	日本の固定価格買い取り制度を諸外国の制度と比較．	—	—	日本の税制を英国の税制と比較．
ブログ記事 *B5*	$\mathbb{E}_{30}(B5)$ 中の Wikipedia エントリ (抜粋)	—	自動車，石油，燃料	トウモロコシ，小麦	—
	ブログ記事の概要	—	地球温暖化が原因でトウモロコシの収穫高が減少していることを指摘して，バイオエタノールの普及に反対している．		—
ブログ記事 *B6*	$\mathbb{E}_{30}(B6)$ 中の Wikipedia エントリ (抜粋)	—	—	桃，収穫，農家，農業	生産，市場
	ブログ記事の概要	—	—	地球温暖化の影響で，ブランド桃の出荷時期が早まり，本来の商品価値を損ねている．	

果を表 2 に示す．この結果から分かるように，平均的に 70~80%程度の正解率を達成できている[*10]．

ここで，初期クエリ t_0 が「地球温暖化」の場合について，6 種類のブログ記事 $B1 \sim B6$ をとりあげ，各ブログ記事 B ごとに，重み $score_e(B, z_k^e)$ の値の上位 2 トピック

野名の付与を補助的に行い，式 (3) を満たす Wikipedia エントリ E とあわせて，この分野名を補足的に参照して評価作業を行う．ただし，分野ラベルの評価作業は，この分野名の付与作業を行った作業者と同一の作業者が行っているため，分野名の付け方によって評価結果が左右されることはない．

[*10] 分野ラベルの付与のタスクは，Wikipedia のエントリ集合を対象として，分野トピックモデルによって推定されたトピックをブログ記事に付与するタスクであり，分野トピックモデルのみが対象となるタスクである．「Wikipedia を介さないトピックモデル」において，同様の分野知識を付与するタスクを設計することは原理的に困難である．

表 5 分野トピックモデルによって推定された分野ラベルの抜粋

初期クエリ t_0	分野ラベル (=分野トピックモデルのトピック z_k^e)
喫煙	食，司法，政治，疾病，タバコ・薬物，社会保障
臓器移植	医療，事件，生物，法，社会保障，病気，社会問題
地球温暖化	生物学，地学，気象学，天文学，化学，政治学
医療事故	政治，解剖，法，社会保障，出産・育児，病気
プリウス	工業製品，交通，トヨタ自動車，エネルギー，電気

の例を表 3 および表 4 に示す．ただし，これらの二つのトピックはいずれも，各ブログ記事に対して適切な分野ラベルであると判定されたものとなっている．これらの表中で示した分野ラベル (本論文における説明の都合上，人手で付与したもの) は，いずれも，「生物学」，「地学，気象学，地球温暖化」のように，一定の分野に対応するものとなっている[*11]．また，これらの分野トピックモデルの各トピック z_k^e に対して，式 (3) を満たす Wikipedia エントリ E のうち，各ブログ記事 B との間で関連性があると判定されたエントリの抜粋を，「$\mathbb{E}_{30}(Bi)$ 中の Wikipedia エントリ (抜粋)」$(i = 1, \ldots, 6)$ の欄に示す．

また，表中には，各ブログ記事の概要もあわせて示す．ただし，ブログ記事 $B3$，$B4$ の場合には，各分野ラベルの観点を考慮して，各分野ラベルごとに個別に概要を記載している．具体的には，ブログ記事 $B3$ の場合には，分野ラベル $WT2$「地学，気象学，地球温暖化」の観点からは，「京都議定書の規定の遵守の必要性」が論じられているのに対して，分野ラベル $WT4$「政治学」の観点からは，「民主党のマニフェストの実行の必要性」が論じられている．一方，ブログ記事 $B4$ の場合には，分野ラベル $WT5$「発電，エネルギー」の観点からは，「日本の固定価格買い取り制度」について論じられているのに対して，分野ラベル $WT8$「金融，経済」の観点からは，「日本の税制」について論じられている．また，他のブログ記事 $B1$，$B2$，$B5$，$B6$ においても，それぞれ，分野ラベルに密接に関連する内容の概要が記載されていることが分かる．このことから，分野トピックモデルのトピックによって表現された分野ラベルをファセットラベルとみなして，ブログ記事集合 $BP(t_0)$ を閲覧することによって，ブログ記事集合を効率よく俯瞰できることが分かる．

6. Wikipedia を介さないトピックモデルとの比較対照分析

最後に本節では，本論文の分野トピックモデルと図 1(b) に示す「Wikipedia を介さないトピックモデル」との間の比較対照分析を行う．

[*11] 評価・分析対象とした 5 種類の初期クエリについて，分野トピックモデルによって推定された分野ラベルの抜粋を表 5 に示す．

6.1 トピック間の対応関係

まず，表 1 に示す 5 種類の初期クエリの各々について，表 1 に示した数のブログ記事を対象として，4.1 節で述べた設定のもとで LDA のツールキットを適用し，図 1(b) に示す「Wikipedia を介さないトピックモデル」の推定をおこなった．ただし，語 w の集合 V としては，4.2 節において分野トピックモデルを適用した場合と同様に，日本語 Wikipedia 中のタイトルの集合を用いた．以上の手順により推定したトピックモデルの各トピックを z_k^b $(k = 1, \ldots, K)$ と記述する (ただし，4.1 節より $K = 50$)．

次に，各トピック z_k^b に対して，次式にしたがい，確率 $P(z'_k|B)$ を最大化するトピック z'_k が z_k^b となるブログ記事 B を収集し，集合 $\mathbb{B}_b(z_k^b)$ を構成する．

$$\mathbb{B}_b(z_k^b) = \left\{ B \middle| z_k^b = \operatorname*{argmax}_{z'_k} P(z'_k|B) \right\}$$

同様に，分野トピックモデルの各トピック z_k^e に対しても，次式にしたがい，重み $score_e(B, z'_k)$ を最大化するトピック z'_k が z_k^e となるブログ記事 B を収集し，集合 $\mathbb{B}_e(z_k^e)$ を構成する．

$$\mathbb{B}_e(z_k^e) = \left\{ B \middle| z_k^e = \operatorname*{argmax}_{z'_k} score_e(B, z'_k) \right\}$$

そして，分野トピックモデルのトピック z_i^e と，「Wikipedia を介さないトピックモデル」のトピック z_j^b のあらゆる組に対して，以下の Dice 係数を算出し，Dice 係数が大きく，相関の強いトピックの組について分析を行った．

$$Dice\big(\mathbb{B}_e(z_i^e), \mathbb{B}_b(z_j^b)\big) = \frac{2 \times \left|\mathbb{B}_e(z_i^e) \bigcap \mathbb{B}_b(z_j^b)\right|}{\left|\mathbb{B}_e(z_i^e)\right| + \left|\mathbb{B}_b(z_j^b)\right|}$$

表 6 に，初期クエリ t_0 が「地球温暖化」の場合について，分野トピックモデルのトピックのうち，表 3 および表 4 において分析対象とした $WT1 \sim WT8$ と，「Wikipedia を介さないトピックモデル」のトピック z_j^b の組のうち，Dice 係数の値が 0.05 以上となるものの抜粋，および，共有するブログ記事の数 $\left|\mathbb{B}_e(z_i^e) \bigcap \mathbb{B}_b(z_j^b)\right|$ を示す．表中に示した「Wikipedia を介さないトピックモデル」のトピック z_j^b の ID は，$BT1 \sim BT16$ であり，ブログ記事集合 $\mathbb{B}_b(z_k^b)$ 中のブログ記事の内容をふまえて，各トピック z_j^b に対して説明のためのラベルを人手で付与した．

この例から分かるように，分野トピックモデルの各トピック z_i^e は，生物学，地学，気象学，地球温暖化，天文学，化学，政治学，発電，エネルギー，工業，農業，林業，金融，経済といった分野に対応している．一方，「Wikipedia を介さないトピックモデル」の各トピック z_j^b は，より粒度の小さい話題に対応しており，「地球温暖化の影響による生態系の変化」，「太陽活動・宇宙線による地球温暖化」，「日本政府による地球温暖化対策」，といった，「地球温暖化」に関する詳細な話題であることが分かる．以上の結果より，

表 6 Wikipedia を介さないトピックモデルとの比較対照分析の例 (初期クエリ: 「地球温暖化」の場合)

分野トピックモデルのトピック z_i^e の ID, ブログ記事数 $\left\|\mathbb{B}_e(z_i^e)\right\|$	Wikipedia を介さないトピックモデルのトピック z_j^b の ID	共有ブログ記事数 $\left\|\mathbb{B}_e(z_i^e)\bigcap\mathbb{B}_b(z_j^b)\right\|$	Dice 係数 $Dice\big(\mathbb{B}_e(z_i^e),\mathbb{B}_b(z_j^b)\big)$
WT1: 生物学 $\left\|\mathbb{B}_e(z_i^e)\right\|=172$	*BT1*: 地球温暖化の影響による生態系の変化	66	0.37
WT2: 地学, 気象学, 地球温暖化 $\left\|\mathbb{B}_e(z_i^e)\right\|=1,974$	*BT1*: 地球温暖化の影響による生態系の変化	71	0.07
	BT2: 地球温暖化の影響による異常気象	114	0.11
	BT3: 地球温暖化懐疑論	51	0.05
	BT4: 地球温暖化の影響による海面上昇	319	0.27
	BT5: 日本政府による地球温暖化対策	188	0.16
	BT6: 地球温暖化に対する国際的枠組み	157	0.14
	BT7: 温室効果ガスと地球温暖化	139	0.13
	BT8: 太陽活動・宇宙線による地球温暖化	72	0.07
WT3: 天文学, 化学 $\left\|\mathbb{B}_e(z_i^e)\right\|=192$	*BT7*: 温室効果ガスと地球温暖化	58	0.29
	BT8: 太陽活動・宇宙線による地球温暖化	37	0.24
WT4: 政治学 $\left\|\mathbb{B}_e(z_i^e)\right\|=255$	*BT5*: 日本政府による地球温暖化対策	19	0.06
	BT9: 政局分析	39	0.23
	BT10: 政党政治	110	0.54
WT5: 発電, エネルギー $\left\|\mathbb{B}_e(z_i^e)\right\|=219$	*BT11*: エネルギーと環境	169	0.69
WT6: 工業 $\left\|\mathbb{B}_e(z_i^e)\right\|=588$	*BT12*: CO2 対策製品	47	0.13
	BT13: 環境対策住宅	131	0.31
	BT14: 新エネルギー開発	106	0.28
	BT15: ごみ問題	82	0.23
WT7: 農業, 林業 $\left\|\mathbb{B}_e(z_i^e)\right\|=128$	*BT1*: 地球温暖化の影響による生態系の変化	18	0.11
	BT16: 食糧問題	31	0.28
WT8: 金融, 経済 $\left\|\mathbb{B}_e(z_i^e)\right\|=396$	*BT5*: 日本政府による地球温暖化対策	56	0.15
	BT6: 地球温暖化に対する国際的枠組み	21	0.06
	BT9: 政局分析	14	0.06
	BT12: CO2 対策製品	12	0.05

分野トピックモデルと「Wikipedia を介さないトピックモデル」は，それぞれの特性が大きく異なっており，これら二種類のトピックモデルを相補的に参照することにより，検索対象の文書集合の俯瞰および効率的閲覧がより容易になると言える．

6.2 初期クエリとの関連性の分析

本研究の枠組みにおいては，分野トピックモデルにおいても，Wikipedia を介さないトピックモデルにおいても，初期クエリとは関連性が低いトピックが一定数含まれる．まず，初期クエリを用いたブログ記事集合を収集する段階において，ブログ記事中に初期クエリが含まれているが，ブログ記事の主題は初期クエリとは無関係である場合が一定数含まれる．このため，Wikipedia を介さないトピックモデルにおいて，初期クエリとは無関係なトピックが一定数含まれることになる．また，分野トピックモデルにおいても，ブログ記事集合中の Wikipedia エントリのうち，初期クエリとの関連性がほとんどない Wikipedia エントリタイトルが一定数含まれ，それらのエントリが集まって分野の集まりが形成される場合がある．

そこで，本節では，分野トピックモデルの 50 トピック，および，Wikipedia を介さないトピックモデルの 50 トピックに対して，表 7 に示すように，初期クエリとの関連性を分析し，集計を行った．まず，表 7(a)「分野トピックモデル」においては，初期クエリ「地球温暖化」の場合のトピック「日本の歴史」,「戦争，軍事」,「デジタル機器」のように，初期クエリとは関連性がほとんどないトピック (表 7(a) の「関連性無」の欄) が多く含まれており，関連性が大きいトピックの 1.5∼2.5 倍程度含まれる．また，「日付」(Wikipedia に登録されている実際の日付についてのエントリの集まりによって構成される),「年号」(Wikipedia に登録されている実際の年号についてのエントリの集まりによって構成される),「数字」(Wikipedia に登録されている実際の数字についてのエントリの集まりによって構成さ

表 7 初期クエリとの関連性の分析結果 (全 50 トピックのうちのトピック数)

(a) 分野トピックモデル

初期クエリ t_0	関連性大	関連性無	分野としての有用性低	分野としてまとまっていない
喫煙	15	20	9	6
臓器移植	15	21	5	9
地球温暖化	15	25	4	6
医療事故	10	23	7	10
プリウス	10	27	4	9

(b) Wikipedia を介さないトピックモデル

初期クエリ t_0	関連性大	関連性はあるが主題が異なる	関連性無	話題がまとまっていない
喫煙	36	5	4	5
臓器移植	29	5	5	11
地球温暖化	23	12	8	7
医療事故	30	8	5	7
プリウス	18	19	11	2

れる) のように，分野としてはまとまっているが，分野としての有用性が低いトピック (表 7(a) の「分野としての有用性低」の欄)，および，分野としてまとまっていないトピック (表 7(a) の「分野としてまとまっていない」の欄)，も一定数含まれる．一方，表 7(b)「Wikipedia を介さないトピックモデル」においても，初期クエリ「地球温暖化」の場合のトピック「政局分析」,「高速道路無料化」のように，初期クエリとの関連性は少しはあるが，ブログ記事の主題が初期クエリからずれているトピック (表 7(b) の「関連性はあるが主題が異なる」の欄) が一定数含まれる．また，「世界金融問題」,「税と社会保障」のように，話題としてはまとまっているが，初期クエリとの関連性が無いトピック (表 7(b) の「関連性無」の欄)，および，話題としてまとまっていないトピック (表 7(b) の「話題がまとまっていない」の欄)，も一定数含まれる．ただし，分野トピックモデルと比べると，初期クエリとの関連性の大きいトピックが二倍程度含まれることが分かる．このことは，表 6 に示す，分野トピックモデルと「Wikipedia を介さないトピックモデル」との間のトピックの対応関係の例からも妥当な結果であると言える．つまり，分野トピックモデルの一つのトピックが「Wikipedia を介さないトピックモデル」の二つ以上のトピックと対応しているため，初期クエリとの関連性の大きいトピックの数は半分程度でも十分な数あると言える．

7. おわりに

本論文では，特定のキーワードをクエリとして収集したブログ記事集合を対象として，ブログ記事集合中の話題の広がりを俯瞰することを目的として，Wikipedia を知識源とする分野トピックモデルを提案し，その推定法，および，ブログ記事集合への適用結果について述べた．特に，ブログ記事集合から抽出した Wikipedia エントリタイトルに対して,「地球温暖化」における「気象学・天文学・生物学・エネルギー・工業」といった分野に対応するトピックモデルを推定し，その特性を分析した．さらに，ブログ記事に対して Wikipedia エントリが適合する度合いにしたがって，Wikipedia エントリの順位付けを行う方式，および，各ブログ記事に対して，分野トピックモデルのトピックを分野ラベルとして付与する方式を提案し，評価実験において安定した性能が達成できることを示した．最後に，従来型の「Wikipedia を介さないトピックモデル」の各トピックとあわせて，分野トピックモデルを相補的に用いることにより，検索された文書集合全体にわたる分野や話題の分布を俯瞰することが可能となり，収集された文書集合の効率的閲覧が促進されることを示した．具体的には，表 3 および表 4 の例に示すように，分野トピックモデルの各トピックを共有するブログ記事の集合を容易に俯瞰することができる．また，表 6 の例に示すように,「Wikipedia を介さないトピックモデル」によって，分野トピックモデルの各トピックとブログ記事集合との間の中間的な俯瞰を促進することができる．今後の課題として，本論文では，分野トピックモデル，および,「Wikipedia を介さないトピックモデル」の双方において，各トピックの内容を表すラベルを人手で付与した上で評価実験を行ったが，今後は，このラベル付け過程の自動化手法を確立する必要がある．

参考文献

[1] 馬場康夫，黒橋禎夫：キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰，情報処理学会論文誌，Vol. 50, No. 4, pp. 1399–1409 (2009).
[2] Blei, D. M., Griffiths, T. L., Jordan, M. I. and Tenenbaum, J. B.: Hierarchical Topic Models and the Nested Chinese Restaurant Process, *NIPS'03* (2003).
[3] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).
[4] Carmel, D., Roitman, H. and Zwerdling, N.: Enhancing Cluster Labeling Using Wikipedia, *Proc. 32nd SIGIR*, pp. 139–146 (2009).
[5] de Winter, W. and de Rijke, M.: Identifying Facets in Query-Biased Sets of Blog Posts, *Proc. ICWSM*, pp. 251–254 (2007).
[6] 藤村　考，戸田浩之，井上孝史，廣嶋伸章，片岡良治，杉崎正之：マルチファセット型ブログ検索システム BLOGRANGER の開発，電子情報通信学会技術研究報告，OIS2005-92，pp. 19–24 (2006).
[7] 原島　純，黒橋禎夫：PLSI を用いたウェブ検索結果の要約，言語処理学会第 16 回年次大会論文集，pp. 118–121 (2010).
[8] Hoffman, T.: Probabilistic Latent Semantic Indexing, *Proc. 22nd SIGIR*, pp. 50–57 (1999).

[9] Li, C., Yan, N., Roy, S. B., Lisham, L. and Das, G.: Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia, *Proc. 19th WWW*, pp. 651–660 (2010).
[10] Liu, X. and Croft, W. B.: Cluster-based Retrieval using Language Models, *Proc. 27th SIGIR*, pp. 186–193 (2004).
[11] Macdonald, C., Ounis, I. and Soboroff, I.: Overview of the TREC-2009 Blog Track, *Proc. TREC-2009* (2009).
[12] Phan, X.-H. and Nguyen, C.-T.: GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) (2007).
[13] Ponte, J. M. and Croft, W. B.: A Language Modeling Approach to Information Retrieval, *Proc. 21st SIGIR*, pp. 275–281 (1998).
[14] 戸田浩之，中渡瀬秀一，片岡良治：特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案，情報処理学会論文誌：データベース， Vol. 46, No. SIG 13(TOD 27), pp. 40–52 (2005).
[15] Tunkelang, D.: *Faceted Search*, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2009).
[16] Wei, X. and Croft, W. B.: LDA-Based Document Models for Ad-hoc Retrieval, *Proc. 29th SIGIR*, pp. 178–185 (2006).
[17] Yokomoto, D., Makita, K., Utsuro, T., Kawada, Y. and Fukuhara, T.: Utilizing Wikipedia in Categorizing Topic related Blogs into Facets, *Procedia - Social and Behavioral Sciences*, Vol. 27, pp. 169–177 (2011).