DEIM Forum 2011 A4-4

トピックに関する話題の時系列分布に着目したブログ分析

牧田　健作[†1]　　横本　大輔[†2]　宇津呂武仁[†2]　福原　知宏[†3]

†1 筑波大学理工学群工学システム学類　〒 305-8573 茨城県つくば市天王台 1-1-1
†2 筑波大学大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
†3 独立行政法人 産業技術総合研究所 サービス工学研究センター　〒 135-0064 東京都江東区青梅 2-3-26

あらまし　本論文においては，あるトピックについて収集したブログ記事集合に対して，記事内容における詳細な話題の違いを考慮して，詳細な話題ごとにブログ記事を分類して俯瞰的に提示する方式について述べる．この方式においては，特定トピックに関して詳細な記述を含むブログ記事集合に対して，Wikipedia エントリを知識源として，特定トピックにおける観点ごとにブログ記事を分類する枠組みを利用する．この枠組みにおいては，Wikipedia 中において特定トピックのキーワードが出現するエントリを収集し，特定トピックにおける観点の候補とする．さらに，Wikipedia エントリ中の関連語の情報を利用して，ブログ記事を各観点に分類する．本論文では，さらに，ブログ記事の日付情報，および，ブロガーの異なりを考慮して，時期の違いによる詳細な話題の分布，および，ブロガーの違いによる詳細な話題の分布を俯瞰的に提示する枠組みを提案する．
キーワード　ブログ分析，トピック，時系列分布，Wikipedia

Analyzing Temporal Distribution of Sub-topics in Blogs related to a Topic

Kensaku MAKITA[†1], Daisuke YOKOMOTO[†2], Takehito UTSURO[†2], and Tomohiro FUKUHARA[†3]

†1 College of Eng. Sys., School of Science and Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
†2 Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
†3 Center for Service Research, National Institute of Advanced Industrial Science and Technology, Tokyo, 135-0064, Japan

Abstract　Given a search query, most existing search engines simply return a ranked list of search results. However, it is often the case that those search result documents consist of a mixture of documents that are closely related to various sub-topics. This is also true for the case of our previously developed framework of retrieving blog posts which are closely related to a certain topic. In this article, we propose a framework of categorizing blog posts according to their sub-topics, where, given a search query, those blog posts are automatically collected from the blogosphere. In our framework, the sub-topic of each blog post is identified by utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a sub-topic label. Furthermore, we propose to consider temporal distribution of sub-topics as well as distribution of sub-topics that are dependent on identity of bloggers. We show that, with this framework, it becomes much easier to quickly overview the distribution of sub-topics over the whole blog posts collected with a certain search query.
Key words　blog analysis, topic, temporal distribution, Wikipedia

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

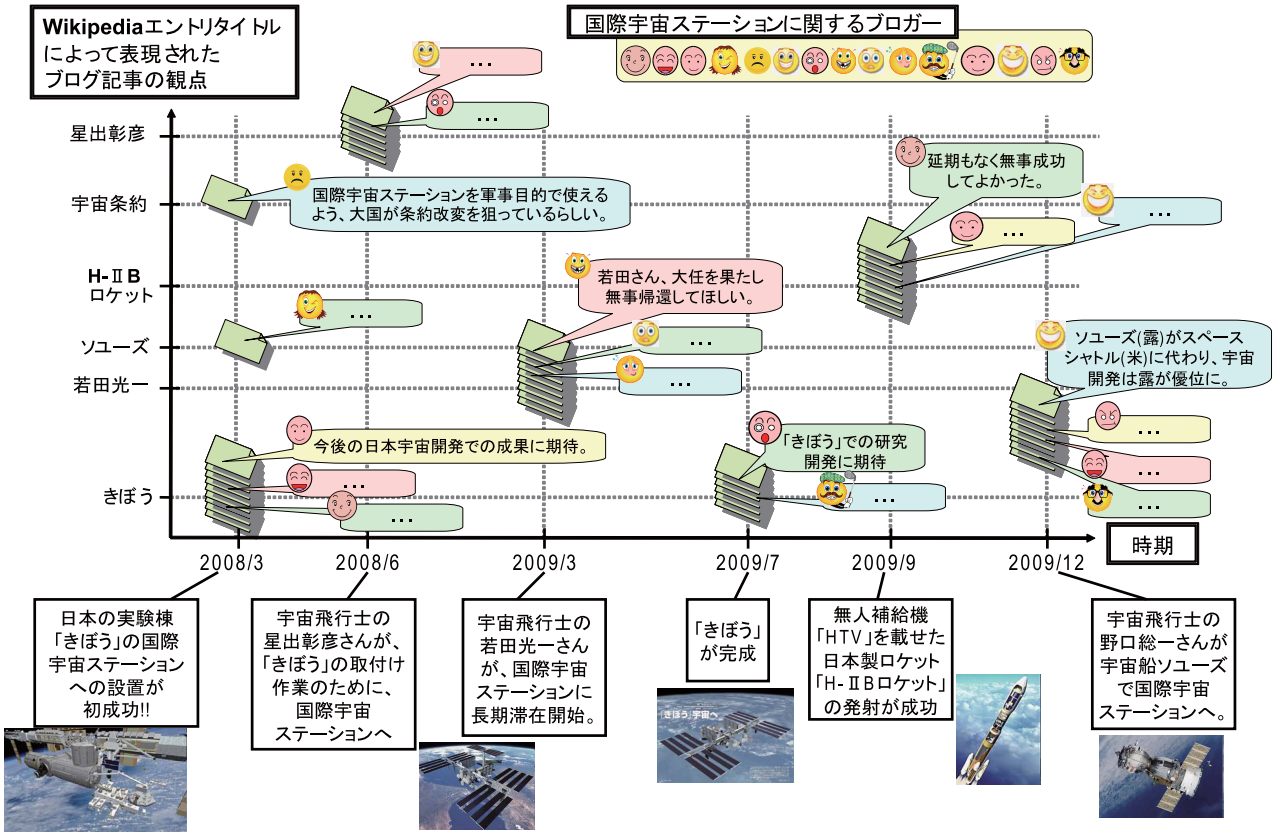

図 1　「国際宇宙ステーション」を初期トピックとするブログ記事集合における観点，および，ブロガーの時系列分布

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なった場合でも，その検索結果には様々な話題について書かれたブログ記事が存在している．例えば，トピック「国際宇宙ステーション」を話題とするブログ記事では，a) 日本の国際宇宙ステーションの実験棟である「きぼう」の設置，b) 国際宇宙ステーションに補給物資を届ける HTV を搭載した「H-IIB ロケット」の打ち上げ成功，c) 野口宇宙飛行士がロシアの宇宙船「ソユーズ」に搭乗し，国際宇宙ステーションに向かうこと，等が，それぞれ，異なるブログ記事において話題となっている．また，トピック「臓器移植」を話題とするブログ記事では，d) 「臓器の移植に関する法律」の改正，e) 疾患を持つドナーから腎臓移植を行なう「病気腎移植」の是非，f) 臓器移植を望む子供への募金活動団体である「○○ちゃんを救う会」の活動，等が，それぞれ，異なるブログ記事において話題となっている。このように，検索結果には様々な話題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話題が含まれているのかを手軽に知ることは容易ではない。

また，上述のような話題の多様性の分析においては，さらに踏み込んで，実際にブログ記事が投稿された時期や，ブログ記事を投稿したブロガーの特性を考慮して分析を行うことが不可欠な場合もある。例えば，トピック「国際宇宙ステーション」を話題とするブログ記事の場合を例にとると，それぞれの話題のブログ記事が投稿された時期は，実世界においてイベントが発生した時期と密接に関係している．具体的には，日本の国際宇宙ステーションの実験棟である「きぼう」の設置された時期は 2008 年 3 月であるが，このことを話題とするブログ記事が最も多く投稿された時期は，実世界におけるイベント発生時期とほぼ同時期である．上述の b)「H-IIB ロケット」の打ち上げ成功，および，c) 野口宇宙飛行士の「ソユーズ」搭乗，についても，同様のことが言える．

一方，トピック「臓器移植」を話題とするブログ記事の場合，個々のブロガーごとに，話題の傾向の特性が大きく異なっている．具体的には，「臓器の移植に関する法律」の改正を話題とするブロガーは，もっぱら，法律関係者や，政治に詳しい人物が多く，ブログ記事の話題の傾向も，法律・政治関連であることが多い．また，「病気腎移植」の是非を話題とするブロガーは，医療関係者や，医療問題に興味を持つ人物であることが多い．

そこで，本論文においては，あるトピックについて収集したブログ記事集合に対して，記事内容における詳細な話題の違いを考慮して，詳細な話題ごとにブログ記事を分類して俯瞰的に提示する方式について述べる．この方式においては，特定トピックに関して詳細な記述を含むブログ記事集合に対して，Wikipedia エントリを知識源として，特定トピックにおける観点ごとにブログ記事を分類する枠組み [8] を利用する．この枠組みにおいては，Wikipedia 中において特定トピックのキーワードが出現するエントリを収集し，特定トピックにおける観点の候補とする．さらに，Wikipedia エントリ中の関連語の情報を利用して，ブログ記事を各観点に分類する．本論文では，さらに，ブログ記事の日付情報，および，ブロガーの異なりを考慮

して，時期の違いによる詳細な話題の分布，および，ブロガーの違いによる詳細な話題の分布を俯瞰的に提示する枠組みを提案する．

図 1 に，本論文の枠組みによって,「国際宇宙ステーション」をトピックとするブログ記事集合における観点，および，ブロガーの時系列分布を模式図化した結果を示す．図 1 には,「国際宇宙ステーション」を話題とするブログ記事集合において，2008 年 3 月から 2009 年 12 月までの期間で，主要な話題のブログ記事が一定数以上観測された月における，各ブログ記事の観点およびブロガー情報が示されている．例えば，2008 年 3 月には，あるブロガーが「きぼう」という観点について記事を投稿しており，2009 年 9 月には,「H-IIB ロケット」という観点について別のブロガーが記事を投稿している．また，2008 年 3 月と 2009 年 7 月という異なる時期でも,「きぼう」という同じ観点について投稿された記事が存在している．一方で，ある一人のブロガーが，2009 年 9 月，および，2009 年 12 月という異なる時期に，それぞれ,「H-IIB ロケット」，および,「ソユーズ」という異なる観点についてのブログ記事を投稿している．また，2008 年 3 月には，異なるブロガーが，それぞれ,「きぼう」，および,「宇宙条約」という別々の観点で記事を投稿している場合もある．このように，ある特定のトピックについて，様々な観点，投稿時期，ブロガー情報が混在したブログ記事集合に対し，本論文では，図 1 に示すように，ブログ記事における観点，投稿時間，ブロガー情報に基づく分類を行なう．

以下に本稿の構成を述べる．2. では，本研究において分類の対象とするブログ記事集合の収集方法について述べ，3. では，Wikipedia エントリとブログ記事の類似度について述べる．4. では，初期トピックに関連するブログ記事集合中の各ブログ記事に対して付与される観点の候補を，Wikipedia 中から収集し，それらの観点候補のうち適切なものをブログ記事に付与する手順を述べる．5. では，特定のトピックに関して収集したブログ記事集合を対象として，観点に基づいて分類を行った結果を分析する．6. では関連研究と本研究の比較を行い，最後にまとめと今後の課題について述べる．

2. 特定のトピックに関するブログ記事の収集

本研究においては，初期トピック t_0 に対して，関連するブログ記事集合を収集した結果に対して，観点の分類を行う．そこで，本節ではまず，初期トピック t_0 を含むブログ記事の収集方法を述べる．

初期トピック t_0 を含むブログ記事の収集においては，Yahoo!Japan 検索 API(注1)を利用し，日本語ブログホスト大手 8 社(注2)のドメインに限って検索を行った．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，初期トピック 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初期トピック t_0 を含むブログ記事を収集し，ブログ記事集合 $P(t_0)$ を作成する．

(注1)：http://www.yahoo.co.jp/

(注2)：fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, hatena.ne.jp

3. Wikipedia エントリとブログ記事の類似度

3.1 Wikipedia エントリの関連語 idf ベクトルの生成

3.1.1 Wikipedia 関連語

トピック名がタイトルである Wikipedia エントリ e を知識源として，トピック名に密接に関連する Wikipedia 関連語を収集する．Wikipedia エントリ「排出取引」の場合について，エントリのスナップショットの抜粋，および，Wikipedia 関連語の例を図 2 に示す．この例の場合は，エントリタイトル「排出取引」の別名であるリダイレクト「排出量取引」で検索を行った結果，エントリタイトル「排出取引」の下に，リダイレクト「排出量取引」が表示され，エントリとしては「排出取引」の本文が提示されている．その他，エントリ中の太字「炭素クレジット」，および，他エントリへのリンクのアンカーテキスト「吸収源活動」,「カーボンオフセット」,「認証排出削減量」が提示されている．本稿においては，各エントリのリダイレクト，各エントリ本文中の太字，および，本文中における他エントリへのリンクのアンカーテキストを Wikipedia 関連語として収集する．Wikipedia エントリ e について収集された関連語集合を $R(e)$ とする．

3.1.2 Wikipedia エントリの関連語 idf ベクトル

Wikipedia エントリ e を表現するベクトルとして，収集されたそれぞれの関連語 $r \in R(e)$ を次元とし，値をその関連語の重み $w(r)$ とするベクトル $\vec{I}(e)$ を定義する．ここで，関連語 $r \in R(e)$ の重み $w(r)$ は，ブログ記事との類似度を測る際の関連語 r の重要度に基づいて設定する．例として,「排出取引」エントリの関連語としては,「排出枠」や「温室効果ガス」など,「排出取引」というエントリとの類似度を測る際に重要となる関連語が収集される．しかし一方で,「社会」,「日本」など,「排出取引」との関連が弱く，類似度を測る際に重要でない関連語も収集されてしまう．これらの重要度の違いを考慮するために，本稿では，逆文書頻度 (inverse document frequency, idf) を用いる．本研究では，Wikipedia の全エントリを文書集合 W として，関連語 r の逆文書頻度 $\mathrm{idf}(W, r)$ を定義する．

$$\mathrm{idf}(W, r) = \log \frac{\left|W\right|}{\left|\left\{e \in W \middle| r \in R(e)\right\}\right|}$$

そして，エントリ e の関連語 r ($\in R(e)$) の重み $w(r)$ として，この逆文書頻度 $\mathrm{idf}(W, r)$ を用いる．

$$w(r) = \mathrm{idf}(W, r)$$

この重み $w(r)$ を用いて，Wikipedia エントリ e の関連語集合 $R(e)$ に対して，関連語 idf ベクトル $\vec{I}$ を以下のように定義する．

$$\vec{I}(e) = \left(w(r_1),\ \ldots,\ w(r_n) \right)$$

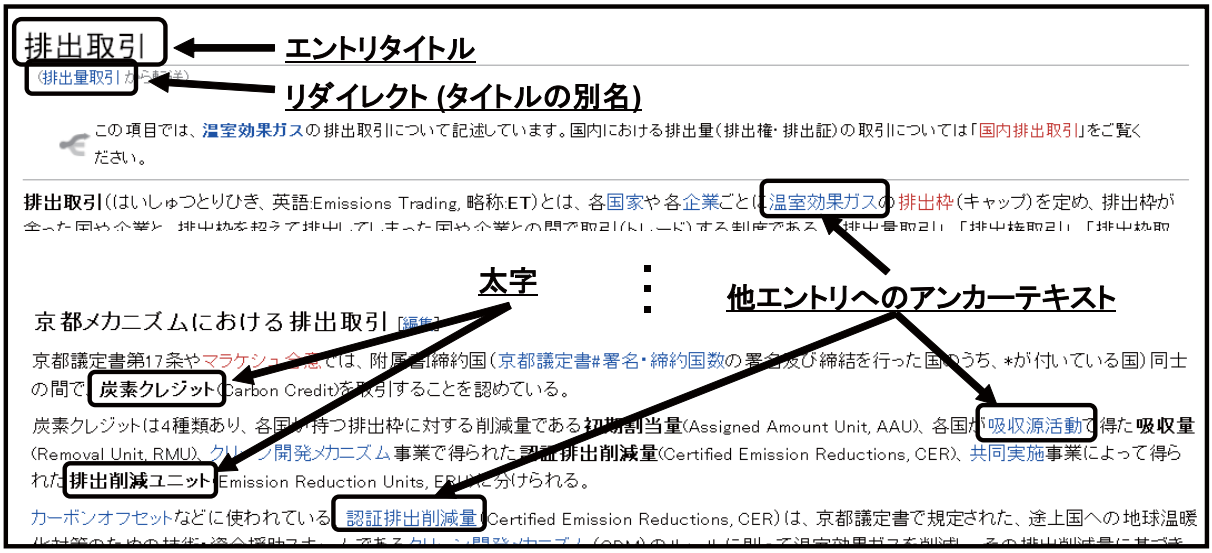

図 2　Wikipedia エントリおよび Wikipedia 関連語の例

3.2　ブログ記事のターム頻度ベクトルの生成

Wikipedia エントリ e，および，2. で収集したブログ記事集合 $P(t_0)$ 中の各ブログ記事 p ($\in P(t_0)$) の組に対して，Wikipedia エントリ e の関連語 r_i ($\in R(e),\ i = 1, \ldots, n$) を次元とする p のターム頻度ベクトル $\vec{G}(p, e)$ を次のように定義する．

$$\vec{G}(p, e) = \left(\ freq(p, r_1),\ \ldots,\ freq(p, r_n) \ \right)$$

ただし，$freq(p, r_i)$ は，Wikipedia エントリ e の関連語 r_i ($\in R(e),\ i = 1, \ldots, n$) のブログ記事 p における出現頻度である．

3.3　Wikipedia エントリとブログ記事の類似度の定式化

Wikipedia エントリ e とブログ記事 p の類似度 $Sim(e, p)$ は，3.1 節で定義した Wikipedia エントリの関連語 idf ベクトル $\vec{I}$ と，3.2 節で定義したブログ記事のターム頻度ベクトル $\vec{G}(p, e)$ の内積として，次のように定義する．

$$Sim(e, p) \ = \ \vec{I}(e) \cdot \vec{G}(p, e) \ = \ \sum_{r \in R(e)} w(r) \times freq(p, r)$$

4.　ブログ記事への観点の付与

4.1　観点候補の収集

初期トピック t_0 に対して，収集したブログ記事に付与する観点の集合 $F(t_0)$ を作成する．具体的には，まず，本文中に，初期トピック t_0 が出現する Wikipedia エントリを f_0 とする．次に，f_0 のうち，ブログ記事集合 $P(t_0)$ において，エントリタイトル $t(f_0)$ の文書頻度が k 以上となるものを選定し，観点集合 $F(t_0)$ を構成する．

$$F(t_0) \ = \ \left\{ \ f \ \middle| \ \mathrm{df}(\ P(t_0),\ t(f)\ ) \ \geq k \right\} \qquad (1)$$

なお，5. で述べる分析においては，初期トピック t_0 に対して 2. において収集したブログ記事集合 $P(t_0)$ 中のブログ記事 p に対して，「初期トピック t_0 をタイトルとする Wikipedia エントリ $e(t_0)$ の関連語の頻度の総和が 10 より大きい」という条件を課し，初期トピック t_0 に対する関連性の高いブログ記事の集合 $P'(t_0)$ ($\subseteqq P(t_0)$) を選定している．具体的には，p のターム頻度ベクトル $\vec{G}(p, e(t_0))$ の次元の総和に対する下限

$$\vec{G}(p, e(t_0)) \ = \ \left(\ freq(p, r_1),\ \ldots,\ freq(p, r_n) \ \right)$$
$$\sum_{i=1}^{n} freq(p, r_i) \ > \ 10 \qquad (2)$$

を課している．そのうえで，$P(t_0)$ のかわりに $P'(t_0)$ に対して，式 (1) の条件 (ただし，$k = 15$) を課して，観点集合 $F(t_0)$ を選定している．

4.2　ブログ記事への観点の付与手順

次に，特定トピックに関連するブログ記事集合中の各ブログ記事に対して，観点を付与する手順の詳細を以下で述べる．以下では，2. において，初期トピック t_0 を含むブログ記事を収集して作成した集合 $P(t_0)$ 中の各ブログ記事に対して観点を付与する．

まず，各ブログ記事 p に対して，観点集合 $F(e)$ 中で類似度最大となる観点 f を付与する．

$$f \ = \ \underset{f' \in F(t_0)}{\mathrm{argmax}}\, Sim(f', p)$$

次に，ブログ記事 p および付与された観点 f の組 $\langle p, f \rangle$ を作成する．なお，次節の分析においては，各観点 f ($\in F(t_0)$) に対して，集合 $P(t_0)$ 中のブログ記事 p と観点 f との組 $\langle p, f \rangle$ のうち，p と f の間の類似度の上位 10 位までを分析対象としている．

$$\langle p_1, f \rangle,\ \langle p_2, f \rangle,\ \ldots,\ \langle p_9, f \rangle,\ \langle p_{10}, f \rangle$$
$$(\ \text{ただし} \quad i < j \quad \text{ならば} \quad Sim(f, p_i) \geq Sim(f, p_j)\ )$$

5.　分　析　例

本節では，「国際宇宙ステーション」を初期トピック t_0 として収集したブログ記事集合を対象として，観点の付与を行った結果について分析する．

表 1　ブログ記事・観点組の評価結果 (初期トピック:「国際宇宙ステーション」. 観点ごと. 観点との類似度の上位 10 記事までが評価対象)

観点	各ブログ記事に対して類似度最大の観点のみを付与した場合		各ブログ記事に対して任意の観点を付与した場合	
	正解率 (%)	正解と判定されたブログ記事・観点組数 / 評価対象のブログ記事・観点組数	正解率 (%)	正解と判定されたブログ記事・観点組数 / 評価対象のブログ記事・観点組数
H-IIB ロケット	100	5 / 5	87.5	7 / 8
ソユーズ	100	8 / 8	66.7	4 / 6
宇宙ステーション補給機	100	8 / 8	100	4 / 4
スペースシャトル	77.8	7 / 9	87.5	7 / 8
宇宙食	66.7	2 / 3	66.7	2 / 3
きぼう	62.5	5 / 8	50.0	4 / 8
宇宙航空研究開発機構	37.5	3 / 8	42.9	3 / 7
アメリカ航空宇宙局	20.0	1 / 5	50.0	3 / 6
宇宙服	—	0 / 0	100	3 / 3
宇宙旅行	—	0 / 0	66.7	2 / 3
山崎直子	—	0 / 0	50.0	1 / 2
土井隆雄	—	0 / 0	50.0	2 / 4
野口聡一	—	0 / 0	40.0	2 / 5
若田光一	—	0 / 0	14.2	1 / 7
欧州宇宙機関	—	0 / 0	—	0 / 0
合計	72.2	39 / 54	60.8	45 / 74

5.1　ブログ記事への観点の付与結果の分析

5.1.1　分析対象

まず，2010 年 7 月に,「国際宇宙ステーション」を初期トピック t_0 として，2. の手順により，t_0 を含むブログ記事を収集し，ブログ記事集合 $P(t_0)$ を作成した．次に，ブログ記事集合 $P(t_0)$ に対して，4.1 節の式 (2) の条件を満たすブログ記事のみを選定したところ，ブログ記事投稿時期が 2004 年 2 月から 2010 年 7 月にわたる合計 814 記事が得られた．この 814 記事を用いて，4.1 節の手順により観点集合 $F(t_0)$ を作成し，合計 15 個の観点が得られた(注3)．

次に，4.2 節の手順により，15 個の各観点に対して，観点との類似度の上位 10 記事を収集した．ここで，814 記事中の投稿記事数を各月ごとに集計するとともに，これらの上位 10 記事の投稿時期の分布を調べたところ，その大部分は，月ごとの投稿記事数順の上位の 19 月 (2008 年 3 月から 2010 年 6 月のうちの 19 月) に集中していた．そこで，以降の分析においては，この 19 月に投稿されたブログ記事 538 記事を対象とする．

そして，この 538 記事を対象として，再度，4.2 節の手順により，15 個の各観点に対して，観点との類似度の上位 10 記事を選定した．さらに，その後，各観点の上位 10 記事に対して，

ブログ記事本文中に，当該観点名，もしくは，その観点名のリダイレクトが出現する．

という条件を課し，この条件を満たさないブログ記事・観点組を除外した結果，54 組のブログ記事・観点組，異なりで 8 個の観点が残った．これらのブログ記事・観点組を対象として，観点ごとに以下の正解率を算出した結果を表 1「各ブログ記事に対して類似度最大の観点のみを付与した場合」の欄に示す．

$$\text{正解率} = \frac{\text{正解と判定されたブログ記事・観点組数}}{\text{評価対象のブログ記事・観点組数}}$$

また，この評価とは別に，4.2 節においてブログ記事に観点を付与する段階において，

> 各ブログ記事 p に対して任意の観点 f を付与し，ブログ記事 p および付与された観点 f の組 $\langle p, f\rangle$ を作成する．そのうえで，各観点 f $(\in F(t_0))$ に対して，集合 $P(t_0)$ 中のブログ記事 p と観点 f との組 $\langle p, f\rangle$ のうち，p と f の間の類似度の上位 10 位までを分析対象とする．
>
> $$\langle p_1, f\rangle, \langle p_2, f\rangle, \ldots, \langle p_9, f\rangle, \langle p_{10}, f\rangle$$
> $$(\text{ ただし}\quad i < j\ \text{ ならば }\ Sim(f, p_i) \geqq Sim(f, p_j)\ )$$

という手順にしたがったところ，74 組のブログ記事・観点組，異なりで 14 個の観点が評価対象として残った．これらのブログ記事・観点組を対象として観点ごとの正解率を算出した結果を表 1「各ブログ記事に対して任意の観点を付与した場合」の欄に示す．

(注3)：実際には，合計 28 個の観点が得られたが，分析の都合上，初期トピックの上位語，あるいは，重要な観点の上位語に相当する語，一般性の高い語，等，13 個の観点を除去した．これらの除去対象となった観点のうちのいくつかは，ブログ記事に対する観点付与の性能評価において正解率を下げることがわかっている．これらの不要な観点を自動で同定することが今後の重要な課題の一つである．

表 2　ブログ記事・観点組の評価結果 (初期トピック:「国際宇宙ステーション」. 月ごと. 観点との類似度の上位 10 記事までが評価対象.) (2008 年 3 月 ~2009 年 9 月)

時期	その月の主要な出来事	(評価対象以外のブログ記事を含む) 主要なブログ記事の傾向	評価対象のブログ記事の分析	
			正解率 (%) $\left(\frac{\text{正解と判定されたブログ記事・観点組数}}{\text{評価対象のブログ記事・観点組数}}\right)$	主要な観点の正解率および ブログ記事の分析結果
2008 年 3 月	国際宇宙ステーションに「きぼう」取り付け成功	主要な観点:「きぼう」	50 (2/4)	「きぼう」: 100% (2/2) 「宇宙航空研究開発機構」: 0% (0/2) 「ISS へのきぼうの取り付け」が話題 「宇宙航空研究開発機構」へも言及しているが, 主たる観点は「きぼう」
2008 年 6 月	星出さん「きぼう」取り付け作業のため ISS へ	主要な観点:「きぼう」	100 (2/2)	「きぼう」: 100% (1/1) 「ISS へのきぼうの取り付け」が話題 「アメリカ航空宇宙局」: 100% (1/1) 「NASA についての本」のレビュー
2008 年 12 月	HTV が JAXA で公開, など, いくつかの話題を観測	主要な観点なし	50 (1/2)	「スペースシャトル」: 100% (1/1) 「スペースシャトルを輸送機で運搬」が話題
2009 年 7 月	「きぼう」完成	主要な観点:「きぼう」	100 (5/5)	「きぼう」: 100% (2/2) 「スペースシャトル」: 100% (1/1) 「ソユーズ」: 100% (1/1) 「「きぼう」完成」が話題の記事など
2009 年 9 月	HTV を搭載した H-IIB ロケット発射成功	主要な観点:「H-IIB ロケット」「宇宙ステーション補給機」	81.8 (9/11)	「H-IIB ロケット」: 100% (5/5) 「宇宙ステーション補給機」: 100% (3/3) 「H-IIB ロケットによる HTV の打ち上げ」が話題

5.1.2　分 析 結 果

表 1 の結果から分かるように,「各ブログ記事に対して類似度最大の観点のみを付与した場合」の方が約 12%高い正解率となった. このことから, ブログ記事と観点との類似度の値には, 一定の信頼性があることが分かる.

不正解であったブログ記事・観点組は 15 組であったが, これらの中には,「国際宇宙ステーション」との関連性が低いブログ記事が 6 記事含まれていた. 一方, 残りの 9 記事については, 人手で付与した参照用観点は, いずれも, 観点集合 $F(t_0)$ 中の 15 観点に含まれていた. また, 誤り例のいずれの場合においても, 類似度最大の観点と参照用観点との間で多くの関連語が共有されていることが, 誤りの主たる原因となっていた.

一方, 表 1 の「各ブログ記事に対して任意の観点を付与した場合」には, 74 組のブログ記事・観点組中の異なり記事数は 36 記事であり, 一記事あたりの観点数は 2.1 であった. また, 正解と判定されたブログ記事・観点組 45 組中の異なり記事数は 32 記事であり, 一記事あたりの正解観点数は 1.4 であった.

5.2　月ごとの分析

次に, 前節において,「各ブログ記事に対して類似度最大の観点のみを付与した場合」として分析対象とした 19 月中の 54 組のブログ記事・観点組のうち, 一月あたり 2 記事以上のブログ記事を含む月は 11 月, 一月あたり 1 記事のみ含む月は 4 月, ブログ記事を含まない月は 4 月であった. このうち, 本節では, 一月あたり 2 記事以上のブログ記事を含む 11 月を対象に, ブログ記事への観点付与の性能を月ごとに分析する. まず, これらの 11 月の各月を対象として, 以下について調査した結果を表 2, および, 表 3 に示す.

- 「国際宇宙ステーション」に関連する主要な出来事
- 評価対象以外のブログ記事を含む主要なブログ記事においてどのような話題の傾向があるか, を観測した結果
- 評価対象のブログ記事の分析
  - ブログ記事に対する観点付与の正解率
  - 主要な観点の正解率, および, ブログ記事の分析結果

この結果から分かるように, 全体としては, 評価対象以外のブログ記事を含む主要なブログ記事において大きな話題の傾向がある場合には, その話題を中心とする観点が付与されることにより, 観点付与の正解率が高くなっている. ここで, 誤り例のいくつかについては, このような月ごとの話題のまとまりを考慮することにより, 正解率を改善できると考えられる. この課題については, 今後取り組む予定である.

表 3　ブログ記事・観点組の評価結果 (初期トピック:「国際宇宙ステーション」. 月ごと. 観点との類似度の上位 10 記事までが評価対象.) (2009 年 10 月 ~2010 年 6 月)

時期	その月の主要な出来事	(評価対象以外のブログ記事を含む) 主要なブログ記事の傾向	評価対象のブログ記事の分析	
			正解率 (%) $\left(\frac{\text{正解と判定されたブログ記事・観点組数}}{\text{評価対象のブログ記事・観点組数}}\right)$	主要な観点の正解率およびブログ記事の分析結果
2009 年 10 月	宇宙ステーション補給機が ISS から分離	主要な観点:「宇宙ステーション補給機」	80 (4/5)	宇宙ステーション補給機: 100% (3/3) 「宇宙ステーション補給機が ISS から分離」が話題
2009 年 11 月	スペースシャトル「アトランティス」打ち上げ, など, いくつかの話題を観測	主要な観点なし	83.3 (5/6)	「宇宙食」: 100% (2/2) 「宇宙航空研究開発機構」: 100% (1/1) それぞれの話題のブログ記事が単発で出現
2009 年 12 月	野口さんがソユーズで国際宇宙ステーションへ	主要な観点:「ソユーズ」	100 (5/5)	「ソユーズ」: 100% (5/5) 「野口さんがソユーズで ISS へ向かう」が話題
2010 年 1 月	国際宇宙ステーションが日本上空を通過	主要な観点:「国際宇宙ステーションを肉眼で見る」(Wikipedia エントリなし)	0 (0/4)	「きぼう」: 0% (0/2) 「スペースシャトル」: 0%(0/1) 「NASA の宇宙開発計画」,「Falcon9 ロケット」等が話題 観点「きぼう」,「スペースシャトル」と関連語を共有するため, 誤った観点を付与
2010 年 4 月	山崎直子さんスペースシャトル「ディスカバリー」で ISS へ	主要な観点:「山崎直子」	100 (3/3)	「スペースシャトル」: 100% (2/2) 「山崎直子さんの打ち上げ」が話題
2010 年 6 月	野口飛行士が ISS より帰還	主要な観点なし	33.3 (1/3)	「ソユーズ」: 100% (1/1) 「宇宙航空研究開発機構」: 0% (0/2) 「野口飛行士がソユーズで帰還」が話題 「はやぶさの帰還」が話題のブログ記事は,「国際宇宙ステーション」とは関連なし

6. 関 連 研 究

文献 [3] は, 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を提案している. この手法では, 分類対象の Web ページの情報のみを利用してクラスタリングを行うため, データが十分に存在しない場合, まとまりのよい分類を行うことが難しくなる. Wikipedia を知識源として利用しているため, 分類対象が少ない場合でも分類を行うことができるという利点がある.

また, 文献 [1,7] では, 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を提案している. これらの手法でも, ラベルを付与する対象のページの情報しか用いていない. これに対し, 本研究の手法では, 観点となる Wikipedia エントリのタイトルをラベルとしている. このように, ラベルの付与においても, 付与対象の情報に加えて, Wikipedia の知識も用いることで, より容易にラベルを付与することができていると考えられる.

その他に観点に基づいて検索結果を提示する研究としては, トピック, ブロガー, リンク先, 感想といった観点でブログを閲覧するもの [2] や, Wikipedia の検索に観点を利用するもの [5] などがある.

また, Web 上のニュースサイトにおける時系列分析の研究として, 文献 [6] においては, Kleinberg のバースト解析手法 [4] を用いて時系列のニュース記事からバーストキーワードを抽出し, これを話題ごとに集約する手法を提案している. この手法と本論文の手法を併用することにより, 話題の分布およびキーワードのバースト特性の双方を考慮して, ブログ空間における時系列トピック分布の集約的把握をより高性能に実現すること

が可能になると考えられる。

7. おわりに

本論文においては，あるトピックについて収集したブログ記事集合に対して，記事内容における詳細な話題の違いを考慮して，詳細な話題ごとにブログ記事を分類して俯瞰的に提示する方式について述べた．この方式においては，特定トピックに関して詳細な記述を含むブログ記事集合に対して，Wikipedia エントリを知識源として，特定トピックにおける観点ごとにブログ記事を分類する枠組みを利用した．この枠組みにおいては，Wikipedia 中において特定トピックのキーワードが出現するエントリを収集し，特定トピックにおける観点の候補とする．さらに，Wikipedia エントリ中の関連語の情報を利用して，ブログ記事を各観点に分類する．本論文では，さらに，ブログ記事の日付情報，および，ブロガーの異なりを考慮して，時期の違いによる詳細な話題の分布，および，ブロガーの違いによる詳細な話題の分布を俯瞰的に提示する枠組みを提案した．

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] 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.

[3] 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.

[4] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th SIGKDD*, pp. 91–101, 2002.

[5] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das. Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In *Proc. 19th WWW*, pp. 651–660, 2010.

[6] 高橋佑介, 宇津呂武仁, 吉岡真治. ニュースにおけるバーストキーワードの話題への集約. 第 3 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2011.

[7] 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.

[8] 横本大輔, 林東權, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏, 神門典子, 吉岡真治, 中川裕志, 清田陽司. 特定トピックに関するブログ記事集合の観点分類における Wikipedia の利用. 第 3 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2011.