社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Wikipediaを多言語知識源とするブログ集合の話題分析

牧田　健作† 　横本　大輔† 　鈴木　浩子† 　宇津呂武仁† 　河田　容英††

福原　知宏†††

† 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
†† (株) ナビックス　〒 141-0031 東京都品川区西五反田 8-3-6
††† 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　本論文では，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提案する．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．提案手法の適用例として,「地球温暖化」を話題として，日本語ブログ集合，および，韓国語ブログ集合を収集・分類し，観点の分布を比較した結果を分析する．また,「東日本大震災」(そのうち，特に,「原子力発電所」，および,「放射能」) を話題として，日本語 Wikipedia 中の観点を収集し，さらに，それらの観点を用いて英語ブログ集合を収集・分類した結果の分析を行う．以上の分析を通して，提案手法により，特定の検索クエリについて収集されたブログ記事における観点の分布を，素早く俯瞰することが容易になることを示す．

キーワード　ブログ分析，トピック，Wikipedia，観点分類，ファセット

Analyzing Topics of Blogs based on Wikipedia as a Multilingual Knowledge Source

Kensaku MAKITA†, Daisuke YOKOMOTO†, Hiroko SUZUKI†, Takehito UTSURO†, Yasuhide KAWADA††, and Tomohiro FUKUHARA†††

† Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba, 305-8573, Japan
†† Navix Co., Ltd.　8-3-6 Nishi-Gotanda, Shinagawa-Ku Tokyo 141-0031, Japan
††† 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 paper, 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. This paper especially presents examples of applying the proposed framework to Japanese / Korean / English blogospheres. Through those examples, we show that 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, Wikipedia, sub-topic categorization, facets

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．例えば，多言語のブログ記事に対して，ある特定のトピックについて検索を行うことで，そのトピックが世界の各地域でどのように関心を持たれているのかを知ることができる．

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話題が混在している．例えば「東日本大震災」に関する日本語のブログ記事を見てみると,「原子力事故」や「津波」,「電力不足」など，様々な話題が含まれていることがわかる．より細かい粒度で見てみると,「原子力事故」についてのブログ記事の中にも,「福島第一原発」についてのみ書かれているものもあれば,「チェルノブイリ原発事故」など過去の事例にも言及しているものもある．同様に,「東日本大震災」に関する英語のブログ記事においては,「原子力事故」や「津波」に対する関心は高いが,「電力不足」に対する関心は低い，などの傾向が観測される．さらに,「原子力事故」に関するブログ記事においては,「放射能」に対して強い関心が持たれている，などの特徴が見られる．

このように，検索結果にはブログ記事ごとに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把握することは難しい．そこで本研究では，話題とブログ記事空間の対応付けを提案する．このとき，いかにして索引となる話題の体系を構築するか，という問題があるが，これに対して本研究では，Wikipedia を知識源として話題の体系を構築し，ブログ記事空間に対する索引として用いる，というアプローチをとる．

ここで，本研究で提案する「「観点の体系」に基づくブログ記事集合の分類および閲覧」の枠組を図 1 に示す．ここでは，日本語の Wikipedia に基づいた話題の体系の枠組みを用いて，英語ブログ集合を分類・閲覧する例を示す．

日本語の Wikipedia においては，2011 年 6 月の時点において，約 75 万のエントリが含まれている．例えば,「東日本大震災」に関連するカテゴリとしては,「原子力」などがあり，それらの下位には「原子力事故」や「原子力発電所」などのカテゴリが存在する．さらに，それらのカテゴリにおいて,「炉心溶融」,「福島第一原子力発電所」,「スリーマイル原子力発電所」といったエントリが登録されている．このような Wikipedia カテゴリ，および，エントリの体系が,「東日本大震災」に関する話題の体系として提示される．そして，これらの話題の体系を用いることによって,「東日本大震災」に関連する内容が記述された膨大なブログ記事集合中の話題のまとまりに対して，きめ細かな索引付けを行うことができる．また，この体系は多言語のブログ記事に対しても用いることができる．例えば「原子炉がロシアの RBMK 型でなかったことを神に感謝.」と述べている英語のブログ記事は，“Reactor vessel”(原子炉圧力容器),“Fukushima I Nuclear Power Plant” (福島第一原子力発電所) といった Wikipedia エントリが話題として索引付けされる．

このような話題の体系，および，ブログ記事集合中の話題に対する索引体系を用いることにより，ブログ記事検索結果の閲覧者は，話題の体系の中を自由自在に探索し，関心のある話題のブログ記事の存在の有無や，多言語間の差異を容易に把握することができ，該当する話題のブログ記事が存在する場合には，効率よくアクセスすることが可能となる．

ここで，様々な観点からデータにラベルを付与し，検索を行う際には観点ごとにラベルを指定することでデータを絞り込みながら検索を行う，という考え方はファセット検索 [1] と呼ばれており，いわゆるキーワード検索とは異なる考え方である．図 1 に示した話題の体系を用いて,「東日本大震災」に関連するブログ記事集合中の話題を自由自在に探索・閲覧する枠組みは，広範囲の話題に相当する索引からより細かい話題に相当する索引へと閲覧していくという，話題という単位のファセットを扱ったファセット検索とみなすことができる．本研究においては，話題として索引付けされた Wikipedia カテゴリ，および，エントリの体系が，ファセット検索におけるファセットの体系であると考える(注1)．さらに，本研究では，個々のファセットに相当する Wikipedia カテゴリ，および，エントリを「観点」と呼び，図 1 に示した話題の体系を「観点の体系」と呼ぶ．

なお，以下の各節においては，特定の話題に関連するブログ記事集合の収集方法，および，その話題に関連する観点の体系の収集方法としては，現状で実装済みの手法についてのみ説明を行う．ただし，本研究の枠組みにおいては，それらの手法としては多様なものを考えることができる．各手法の実装および比較評価については，今後の課題とする．

以下に本論文の構成を述べる．2.，3.，4. の各節においては，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，観点の収集方法，ブログ記事への観点付与の方法について述べる．そして，これらの各手法に基づいて，5.1 節においては,「地球温暖化」を話題として，日本語ブログ集合 [2,3]，および，韓国語ブログ集合 [4] を収集・分類し，観点の分布を比較した結果を分析する．一方，5.2 節においては,「東日本大震災」(そのうち，特に,「原子力発電所」，および,「放射能」) を話題として，日本語 Wikipedia 中の観点を収集し，さらに，それらの観点を用いて英語ブログ集合を収集・分類した結果の分析を行う．6. では関連研究と本研究の比較を行い，最後にまとめを行う．

2. 特定のトピックに関するブログ記事の収集

ブログ記事の収集においては，日本語，韓国語，英語の各言語における検索エンジン API を用いて，大手ブログホストを対象として，特定の話題を表すキーワード (本論文では，このキーワードを，初期トピック t_0 と呼ぶ) を含むブログ記事を収

(注1)：ファセット検索の枠組みにより，閲覧者が自由自在にブログ記事集合を探索・閲覧する目的において，Wikipedia のカテゴリおよびエントリの体系が，必ずしも最適なファセットの体系であるという保証はない．今後は，ブログ記事集合，および，ファセットの集合に対して，自由自在な閲覧を実現するための階層的な話題の体系を自動構築する方式を確立することが重要である．

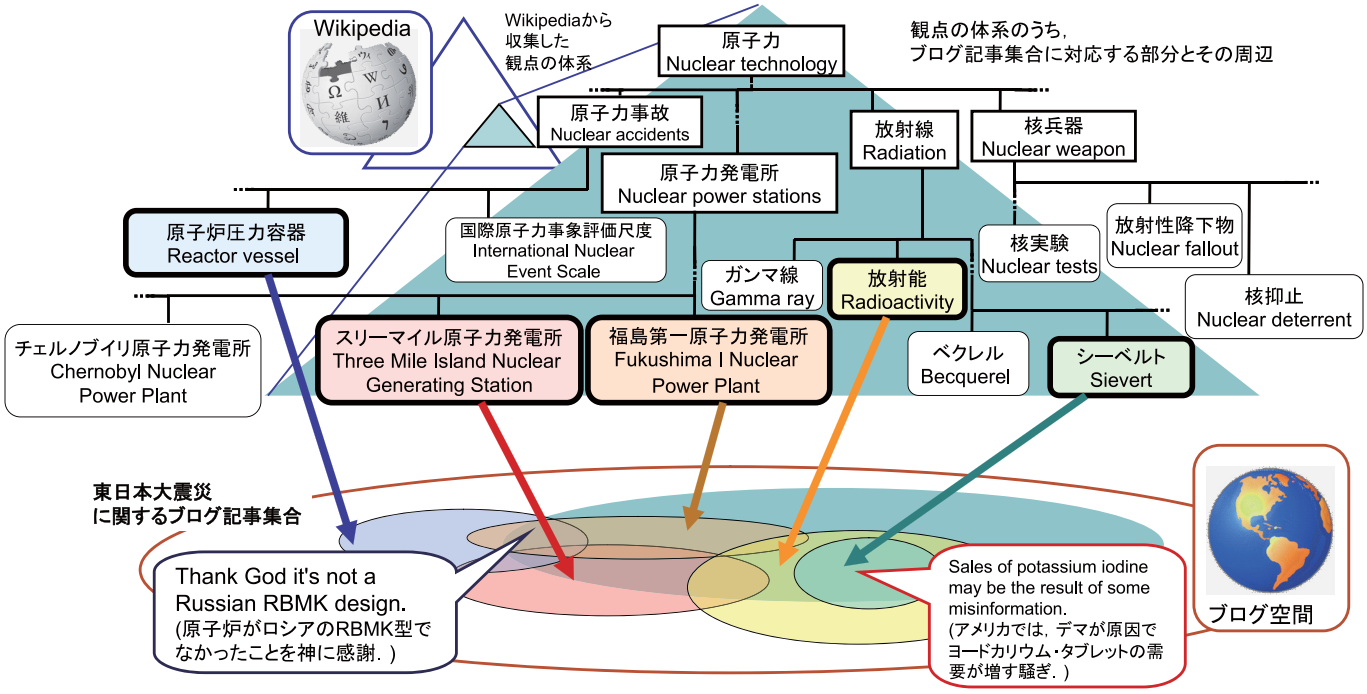

図 1　「観点の体系」に基づくブログ記事集合の分類および閲覧
(例: 日本語の観点体系，閲覧対象: 英語ブログ)

集した集合 $D(t_0)$ を用いた．

3. 観点候補の収集

次に，初期トピック t_0 に対して，Wikipedia から観点の集合 $F(t_0)$ を作成する(注2)．まず，初期トピック t_0 が本文テキスト中に出現する Wikipedia エントリを f_0 とする．ここで，日本語が対象の場合には，f_0 のうち，ブログ記事集合 $D(t_0)$ において，エントリタイトル $u(f_0)$ の文書頻度が 30 以上となるものを選定し(注3)，観点集合 $F(t_0)$ を構成する [2,3]．一方，韓国語が対象の場合には，日本語と比較して相対的にブログ記事の数が少ないため，f_0 のうち，ブログ記事集合 $D(t_0)$ において，エントリタイトル $u(f_0)$ およびそのリダイレクトの文書頻度の総和が 5 以上となるものを選定し，観点集合 $F(t_0)$ を構成する [4].

4. ブログ記事への観点の付与

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 $D(t_0)$ 中の各ブログ記事に対して，$F(t_0)$ 中の観点を付与する．

4.1 Wikipedia エントリとブログ記事の類似度

観点を付与する際には，観点とブログ記事の類似度を計算し，計算された類似度に基づいて付与する観点を決定する．類似度の計算においては，まず Wikipedia エントリ e の本文中に含まれる重要な語を関連語として抽出し，Wikipedia エントリ e を関連語の集合 $R(e)$ として表現する．そして，観点となる Wikipedia エントリ e の関連語 $r(\in R(e))$ がブログ記事 d の本文により多く出現しているほど類似度が高いとする．

具体的には，[2–4] に基づいて，まず，Wikipedia エントリ e から，エントリタイトルや，本文中に出現する太字，他エントリーへのリンクのアンカーテキストなどを関連語 r として収集する．そして，抽出した関連語 r の逆文書頻度 (inverse document frequency, idf) を重み $w(r)$ として(注4)，エントリ e の関連語 idf ベクトル $\vec{I}$ を定義する．

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

一方，ブログ記事についても，Wikipedia エントリ e の関連語 r のブログ記事 d における出現頻度 $freq(d,r)$ を重みとして d のターム頻度ベクトル $\vec{G}(d,e)$ を次のように定義する．

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

そして，Wikipedia エントリ e とブログ記事 d の類似度 $Sim(e,d)$ は，2 つのベクトルの内積として次のように定義する．

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

4.2 ブログ記事への観点の付与手順

最後に，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．ここでは，各ブログ記事 d

(注2)：本論文の執筆時点においては，観点集合を自動作成する手順の実装およびその評価は，日本語および韓国語のみが対象となっており，英語を対象とした実装および評価については，今後別の機会に発表する予定である．

(注3)：一つの Wikipedia エントリ中の記述量が多い場合には，各 Wikipedia エントリの本文中の段落を観点の単位とすることが有効であると考えられ，今後の課題として取り組む予定である．

(注4)：一般には，$w(r) = \mathrm{idf}(r) = \log\left(\frac{\text{Wikipedia の総エントリ数}}{\text{関連語 } r \text{ が出現したエントリ数}}\right)$ として定義するが，韓国語の場合には，経験的に，変則的な重み [4] を用いる．

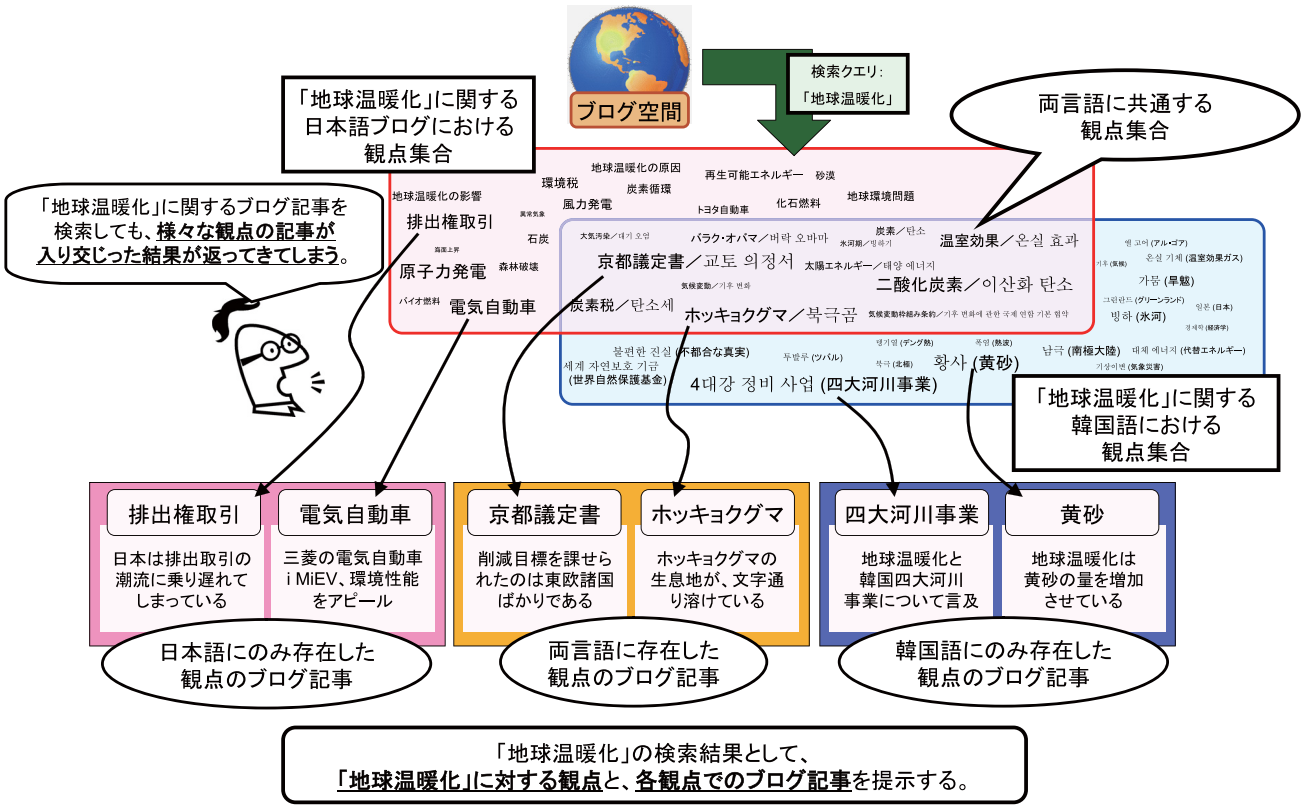

図 2 「地球温暖化」に関する日韓ブログ集合の話題分類の例

に対して，観点集合 $F(t_0)$ 中で類似度最大となる観点 f を付与する(注5)．

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

これにより，ブログ記事 d および付与された観点 f の組 $\langle d, f \rangle$ を作成し，評価の対象とする．

5. 適 用 例

5.1 「地球温暖化」に関する日韓ブログ集合の話題分類

本節においては，「地球温暖化」を初期トピックとして，日本語ブログ集合 [2,3]，および，韓国語ブログ集合 [4] を収集・分類し，観点の分布を比較した結果を分析する．なお，文献 [2,3] および文献 [4] においては，十種類弱の初期トピックを対象として，ブログ記事を収集した結果について，観点付与を行った結果の詳細な評価を行っている．

初期トピック t_0 を含む日本語ブログの収集においては，Yahoo!Japan 検索 API(注6)を利用し，2010 年 7~9 月に，初期トピック t_0 をクエリとして日本語ブログホスト大手 8 社(注7)のドメインを対象としてブログ記事の収集を行った (手順の詳細は文献 [2,3] を参照)．一方，初期トピック t_0 を含む韓国語ブログの収集においては，Naver Open API(注8)を利用し，2010 年 10~11 月に，初期トピック t_0 をクエリとして韓国語ブログホスト大手 4 社(注9)のドメインを対象としてブログ記事の収集を行った (手順の詳細は文献 [4] を参照)．

図 2 に，初期トピック「地球温暖化」について，日本語・韓国語のブログ集合を収集し，各観点に分類した結果の抜粋を示す．この例では，両言語に共通して観測された観点が 12 個，日本語のみで観測された観点が 18 個，韓国語のみで観測された観点が 19 個となった．

両言語に共通して観測された観点としては，「京都議定書」，「ホッキョクグマ」，「二酸化炭素」などが挙げられる．観点として「京都議定書」を付与されたブログ記事の例として，「東欧諸国ばかりが削減目標を課せられている」と批判している日本語ブログ記事が挙げられる．一方，観点として「ホッキョクグマ」を付与されたブログ記事の例として，「ホッキョクグマの生息地が減少している」ことを述べている韓国語ブログ記事が挙げられる．

また，日本語ブログのみで観測された観点としては，「排出権取引」，「電気自動車」，「原子力発電」などが挙げられる．観点として「排出権取引」を付与されたブログ記事の例として，「日本は排出権取引の潮流に乗り遅れている」と述べているブログ記事が挙げられる．一方，韓国語ブログのみで観測された観点としては，「四大河川事業」，「黄砂」，「旱魃」などが挙げられる．

(注5)：本論文の説明では，一つのブログ記事に付与する観点を一つのみとしているが，一つのブログ記事に複数の観点を付与し評価を行うことも可能であり，今後の課題として取り組む．

(注6)：http://www.yahoo.co.jp/

(注7)：fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, hatena.ne.jp

(注8)：http://dev.naver.com/openapi/

(注9)：blog.naver.com, blog.daum.net, blog.cyworld.com, blog.paran.com

表 1 「東日本大震災」に関する英語ブログ集合の話題分析の例

初期トピック	観点	特徴	ブログ記事の日付・内容の要約
原子力発電所 (Nuclear power plant)	福島第一 原子力 発電所 (Fukushima I Nuclear Power Plant)	英語ブログ特有	2011 年 3 月 20 日．米国原子力学会が原発で働く労働者のための基金を設立したことを紹介．関係組織による事故対応に対して肯定的なブロガー．
		英語ブログ特有	2011 年 3 月 12 日．海外の報道は，原子炉のメルトダウンを指摘していることを紹介．関係組織による事故対応に対して否定的なブロガー．
	炉心溶融 (Nuclear meltdown)	英語ブログ特有	2011 年 3 月 13 日．大規模な原子炉爆発を心配する意見を引用．にもかかわらず，オバマ大統領は，原子力のために 360 億ドルの予算を組んでいると批判．関係組織による事故対応に対して否定的なブロガー．
		英語ブログ特有	2011 年 3 月 14 日.「核に関する知識が欠如したジャーナリストが多くの間違った情報を流している」という意見を引用.
	原子炉圧力容器 (Reactor vessel)	英語ブログ特有	2011 年 3 月 17 日．米国 G 社のマーク 1 型原子炉には問題があるにも関わらず，世界の 32ヶ所でまだ使われている，と指摘．関係組織による事故対応に対して否定的なブロガー．
		英語ブログ特有	2011 年 3 月 12 日．ブロガーは元原子力分野の学生．原子炉がロシアの RBMK 型でなかったことを神に感謝.
放射能 (Radioactivity)	シーベルト (Sievert)	英語ブログ特有	2011 年 3 月 17 日．アメリカでは，チェーンメールと右翼系ラジオ局が広めたデマが原因で，ヨードカリウム・タブレットの需要が増す騒ぎがあった.
		日英ブログ共通	2011 年 4 月 19 日．日本政府からの情報伝達の悪さを批判．関係組織による事故対応に対して否定的なブロガー.
	ベクレル (Becquerel)	日本在住者の英語ブログ	2011 年 4 月 13 日．ベクレルの数値において基準を超えた野菜や果物を避けるようにすれば問題ない，として，母国へ帰国した外国人を批判.
		日英ブログ共通	2011 年 4 月 21 日．日本の首相の対応のまずさを強く批判．関係組織による事故対応に対して否定的なブロガー.
	核実験 (Nuclear weapons testing)	日英ブログ共通	2011 年 3 月 29 日．原子力産業界のブログ．原発敷地内のプルトニウム検出量は，冷戦時代の核実験後検出された程度の量だ．関係組織による事故対応に対して肯定的なブロガー.
		英語ブログ特有	2011 年 4 月 30 日．大気圏核実験，チェルノブイリ事故，福島事故の前後では，大気中の放射線量は大きく異なっており，少量の放射線は健康によい，という議論の前提がそもそも成り立っていない．関係組織による事故対応に対して否定的なブロガー.

これらの観点が付与されたブログ記事の例として,「四大河川事業」と「地球温暖化」について言及しているブログ記事や,「地球温暖化によって黄砂が増加している」と主張しているブログ記事が挙げられる.

5.2 「東日本大震災」に関する英語ブログ集合の話題分析

本節においては,「東日本大震災」(そのうち，特に,「原子力発電所」，および,「放射能」) を話題として，日本語 Wikipedia 中の観点を収集し，さらに，それらの観点を用いて英語ブログ集合を収集・分類した結果の分析を行う.

ここでは，まず,「東日本大震災」に関する話題の中でも，海外 (特に，英語圏) における関心が圧倒的に高い「原子力発電所」および「放射能」を初期トピックとして，日本語 Wikipedia において観点候補を収集する．図 1 には,「原子力発電所」および「放射能」を初期トピックとして収集した観点候補の Wikipedia エントリ，および，それらのエントリ周辺の Wikipedia カテゴリの体系を示す.

次に，これらの観点候補の Wikipedia エントリのうち，英語ブログ空間においてブログ記事が多数観測することが予想されるエントリを検索クエリとして英語ブログ記事の収集を行う．ここで，検索クエリを含む英語ブログの収集においては，Yahoo! Search BOSS(注10)を利用し，2011 年 5 月下旬に英語ブログホスト大手 4 社(注11)のドメインを対象としてブログ記事の収集を行った．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，検索結果の URL をブログサイト単位にまとめる．そして，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する.

以上の手順によりブログ記事を収集を行った観点候補のうち,「原子力発電所」を初期トピックとする「福島第一原子力発電

(注10)：http://developer.yahoo.com/search/boss/

(注11)：blogspot.com，wordpress.com，typepad.com，multiply.com

所」,「炉心溶融」,「原子力圧力容器」，および,「放射能」を初期トピックとする「シーベルト」,「ベクレル」,「核実験」についてのブログ記事収集結果の抜粋を表 1 に示す．また，図 1 には，これらのブログ記事の内容の抜粋を模式図として示す．表 1 の「特徴」の欄には，収集されたブログ記事の内容の要約を，日本語ブログ空間中のブログ記事と比較して,「英語ブログに特有の記述内容」であるか，それとも,「日本語ブログにも共通する記述内容が存在する」か，についての判断をした結果を記載している．ただし，一部のブログ記事に対しては，ブロガーの立場や素性を「日本在住者の英語ブログ」と記載している．

表 1 の結果から，特に，英語ブログ特有の記述内容として,「米国原子力学会」に関する話題,「オバマ大統領」に対する批判,「米国 G 社のマーク 1 型原子炉」に関する話題,「ロシアの RBMK 型原子炉」に関する話題,「米国でのデマ」に関する話題，等，興味深い記述内容が観測されている．

6. 関 連 研 究

ファセット検索に関連する研究として，TREC-2009 におけるブログ検索タスク [5] においては，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] 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．これに対し，本研究の手法では，分類対象の情報だけではなく 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 [7,8] 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 [9] や，Wikipedia の検索に観点を利用するもの [10] などがある．

また，本研究の発展として，文献 [11] においては，ブログ記事の時系列の分布，および，ブロガーの分布を考慮して，特定のトピックについて収集されたブログ記事集合における観点分布を提示する方式を提案している．

7. お わ り に

本論文では，特定トピックに関するブログ記事集合に対して，Wikipedia エントリを知識源として観点の体系を構築し，観点ごとにブログ記事を分類する枠組みを提案した．提案手法の適用例として,「地球温暖化」を話題として，日本語ブログ集合，および，韓国語ブログ集合を収集・分類し，観点の分布を比較した結果を分析した．また,「東日本大震災」(そのうち，特に,「原子力発電所」，および,「放射能」) を話題として，日本語 Wikipedia 中の観点を収集し，さらに，それらの観点を用いて英語ブログ集合を収集・分類した結果の分析を行った．以上の分析を通して，提案手法により，特定の検索クエリについて収集されたブログ記事における観点の分布を，素早く俯瞰することが容易になることを示した．

文　　献

[1] D. Tunkelang. *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[2] 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.

[3] D. Yokomoto, K. Makita, Y. Kawada, T. Utsuro, and T. Fukuhara. Utilizing Wikipedia in categorizing topic related blogs into facets. In *Proc. 12th PACLING*, 2011.

[4] D. Lim, D. Yokomoto, K. Makita, T. Utsuro, and T. Fukuhara. Utilizing Wikipedia as a knowledge source in categorizing topic related Korean blogs into facets. 言語処理学会第 17 回年次大会論文集, pp. 876–879, 2011.

[5] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2009 blog track. In *Proc. TREC-2009*, 2009.

[6] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 牧田健作, 横本大輔, 宇津呂武仁, 福原知宏. 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. 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.