2J3-NFC2-3in

Wikipediaを知識源とするブログ記事の観点分類

Categorizing Blog Posts into Facets by utilizing Wikipedia as a Knowledge Source

横本 大輔[*1] 牧田 健作[*1] 宇津呂 武仁[*1] 河田 容英,[*2] 福原 知宏[*3]
Daisuke Yokomoto Kensaku Makita Takehito Utsuro Yasuhide Kawada Tomohiro Fukuhara

[*1]筑波大学大学院システム情報工学研究科 [*2](株) ナビックス
Graduate School of Systems and Information Engineering, University of Tsukuba Navix Co., Ltd.

[*2]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
Center for Service Research, National Institute of Advanced Industrial Science and Technology

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. 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.

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話題が混在している．例えば「東日本大震災」に関する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．

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「観点に基づくブログ記事集合の分類」の枠組を図 1 に示す．

Wikipedia においては，2011 年 5 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」や「自然災害」などがあり，それらの下位には「原子力事故」や「原子力発電」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」,「チェルノブイリ原発事故」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．例えば「もし福島原発の 1 号機原子炉が爆発したら，チェルノブイリ原子力発電事故と同じ事態になるかもしれない」と述べているブログ記事は,「福島第一原子力発電所」,「チェルノブイリ原子力発電所事故」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 [Tunkelang 09] 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[*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

連絡先: 横本 大輔，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 1-1-1，029-853-5427

[*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．

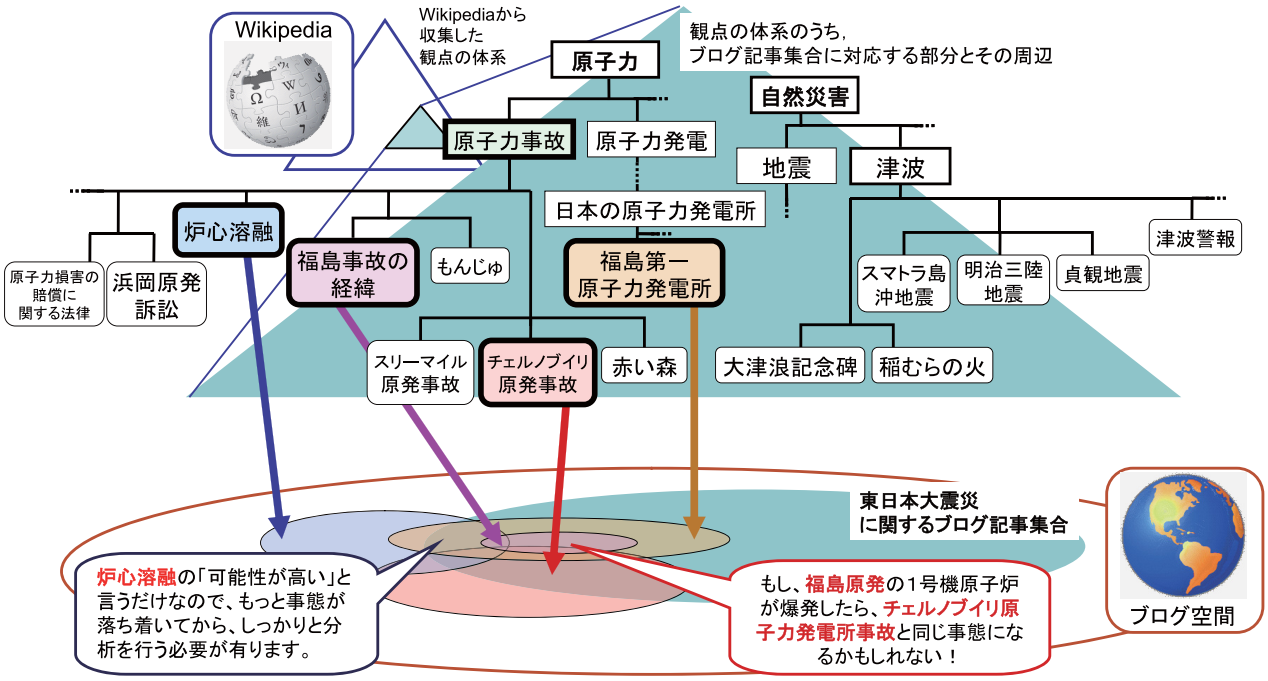

図 1: 「観点の体系」に基づくブログ記事集合の分類および閲覧

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．

2. 特定のトピックに関するブログ記事の収集

本節では，評価の対象としたブログ記事集合の収集方法について述べる．本論文の評価では，日本語ブログホスト大手 8 社[*2]を対象として，2011 年 7~9 月の時期に，[横本 11a] の手法により，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．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．そして，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，観点集合 $F(t_0)$ を構成する[*3].

4. ブログ記事への観点の付与

さらに，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．

*2 fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, hatena.ne.jp

*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．

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，[横本 11a] に基づいて次のように定義する．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)[*4] を重みとして，エントリ e の関連語 idf ベクトル $\vec{I}$ を定義する．

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

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

続いて，類似度に基づいて付与する観点を決定する手順を説明する．まず観点を付与する際の条件として，観点 f を付与

*4 $\mathrm{idf}(r) = \log\left(\frac{\text{Wikipedia の総エントリ数}}{\text{関連語 } r \text{ が出現したエントリ数}}\right)$ として定義する．

表 1: ブログ記事・観点組の評価結果 (4 トピック分のみ抜粋)

初期トピック	観点数	観点	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
喫煙	20	受動喫煙	87.5	7 / 8
		禁煙	100	5 / 5
		禁煙ファシズム	50.0	3 / 6
		その他	63.6	21 / 33
		合計	69.2	36 / 52
臓器移植	18	臓器の移植に関する法律	100	9 / 9
		免疫抑制剤	46.2	6 / 13
		宇和島徳洲会病院	100	4 / 4
		その他	68.4	26 / 38
		合計	68.8	45 / 64
地球温暖化	26	京都議定書	75.0	9 / 12
		再生可能エネルギー	62.5	5 / 8
		環境税	100	4 / 4
		その他	64.8	35 / 54
		合計	67.9	53 / 78
医療事故	14	医師	56.2	9 / 16
		医療訴訟	54.5	6 / 11
		日本医療機能評価機構	100	4 / 4
		その他	48.6	18 / 37
		合計	54.4	37 / 68

するブログ記事の候補は，ブログ記事集合 $D(t_0)$ のうち f のタイトルが文字列として出現するブログ記事の集合 $D_f(t_0)$ に限定する．

次に，各観点 f ($\in F(t_0)$) に対して，集合 $D_f(t_0)$ 中のブログ記事 d のうち，d と f の間の類似度の上位 20 位までのブログ記事を収集する．そして，各ブログ記事 d に対して，観点集合 $F(t_0)$ 中で類似度最大となる観点が f 自身である場合のみ，ブログ記事 d に観点 f を付与することとする[*5]．

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. 評価

本節では，2. で収集したブログ記事集合に対して，3. で作成した観点集合中の観点を付与した結果を評価する．

5.1 評価方法

前節で観点が付与されたブログ記事 d に対して，ブログ記事・観点組 $\langle d, f\rangle$ の評価を行った．評価を行う際には，ブログ記事 d 中の記述内容が初期トピック t_0 および観点 f の双方に関連している場合に，ブログ記事・観点組 $\langle d, f\rangle$ は正解とし，初期トピック t_0 もしくは観点 f の少なくとも一方には関連していない場合には不正解とした．評価尺度としては，以下の正解率を用いた．

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

5.2 評価結果

初期トピックとして,「喫煙」,「臓器移植」,「医療事故」,「高齢化社会」,「アルコール依存症」,「リストラ」,「地球温暖化」,「スマートフォン」,「プリウス」の 9 トピックを対象として評価を行った．そのうち，正解率の高かった 4 トピックについての結果を表 1 に示す[*6]．

初期トピック「喫煙」，観点「受動喫煙」について，正解と判定されたブログ記事では,「受動喫煙症」や「健康増進法」など,「受動喫煙」と密接に関わる語が出現し，Wikipedia エントリ「受動喫煙」との類似度が高くなったため，正しい観点が付与されている．

一方，ブログ記事へ観点を付与した結果における誤りは，以下のように分類できた．

(a) ブログ記事が初期トピックと関連のある場合.

- (a1) 人手でブログ記事に付与した参照用観点が，観点集合 $F(t_0)$ に含まれる．
- (a2) 人手でブログ記事に付与した参照用観点は，観点集合 $F(t_0)$ には含まれないが，Wikipedia に存在
- (a3) 人手でブログ記事に付与した参照用観点が Wikipedia に存在しない
- (a4) ブログ記事は，初期トピックに関連するが，特定の観点の付与は困難

(b) ブログ記事が初期トピックと関連のない場合.

この場合，提案手法によりブログ記事に付与された観点が，ブログ記事に適合している度合いが大きい場合と小さい場合がある．

このうち，まず,「(a) ブログ記事が初期トピックと関連のある場合」の例を示す.

「(a1) 参照用観点が観点集合 $F(t_0)$ に含まれる」場合の例では,「禁煙のために日常生活をどのように改善すればよいか」

*5 本論文の評価では，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．

*6 全ての初期トピックでの評価結果については [横本 11a,Yokomoto 11b] を参照のこと．

について書かれたブログ記事に対して,「ニコチン依存症」という観点が付与されていた．このブログ記事では,「ニコチン」,「喫煙」,「依存症」等，初期トピック「喫煙」に関する他の観点との間で共有される関連語が多く出現し，観点「ニコチン依存症」との類似度が高くなっていた．この問題に対する対策の一つとして，観点集合 $F(t_0)$ 中の各エントリの本文の集合を文書集合とみなして逆文書頻度を測定し，この値を各関連語の重みとする，という手法が考えられる．

「(a2) 参照用観点は観点集合 $F(t_0)$ には含まれないが，Wikipedia に存在」では，ブログ記事中に出現した Wikipedia エントリタイトルを観点候補とすることにより改善が可能である．今後は，この方式の精緻化が必要である.「(a3) 参照用観点が Wikipedia に存在しない」では，ブログ記事中の文字列の中から，観点名として適切な用語を抽出する必要がある．類似の観点を共有するブログ記事が一定数以上存在する場合には，外部知識を用いず，主として，クラスタリング対象の文書集合の情報のみを用いる先行手法 [戸田 05,馬場 09] との併用が効果的であると考えられる．

「(a4) 初期トピックに関連するが，特定の観点の付与は困難」の例では，ブログ記事中に,「喫煙についてのブログ著者の意見」が書かれているが，特定の観点を付与することは困難であった．今後は，このような「特定の観点の付与が困難」なブログ記事の同定に特化した方式を導入する必要がある．

一方,「(b) ブログ記事が初期トピックと関連のない場合」の例では，いずれの場合も，ブログ記事の内容は，初期トピックと関連しない内容であった．ただし，初期トピックが「地球温暖化」の場合の例では，提案手法によって，観点集合 $F(t_0)$ 中の観点「風力発電」がブログ記事に付与されており，この観点は当該ブログ記事に最も適合する観点であった．一方，初期トピックが「臓器移植」の場合の例では，提案手法によって付与された観点「鳩山由起夫」は，観点集合 $F(t_0)$ 中では，ブログ記事の内容にやや近いと言えるが，当該ブログ記事にとってより適切な観点は,「地球温暖化」とは無関係な，より政治色の強い観点であった．しかし，観点集合 $F(t_0)$ 中には，そのような政治色の強い観点が含まれていなかったため，結果的に，観点「鳩山由起夫」が付与された．なお，これらのいずれの例においても，初期トピックとブログ記事との間の類似度や，初期トピックと観点との間の類似度に対して下限を設けることにより，観点付与の性能を改善できる可能性がある．今後の課題として，それらの方式に取り組む．

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 [Macdonald 09] 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．

[原島 10] 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，[戸田 05,馬場 09] 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 [藤村 06] や，Wikipedia の検索に観点を利用するもの [Li 10] などがある．

また，本研究の発展として，[牧田 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．[Lim 11] では，韓国語のブログ記事を対象として本論文の手法を適用し，言語に依存せず，ブログ記事への観点付与が可能であることを示している．

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．提案手法の評価実験を行い，実際に，観点に密接に関連するブログ記事を選定した結果を示した．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．

参考文献

[馬場 09] 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)

[藤村 06] 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)

[原島 10] 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)

[Li 10] Li, C., Yan, N., Roy, S. B., Lisham, L., and Das, G.: Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia, in *Proc. 19th WWW*, pp. 651–660 (2010)

[Lim 11] Lim, D., Yokomoto, D., Makita, K., Utsuro, T., and Fukuhara, T.: 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)

[Macdonald 09] Macdonald, C., Ounis, I., and Soboroff, I.: Overview of the TREC-2009 Blog Track, in *Proc. TREC-2009* (2009)

[牧田 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)

[戸田 05] 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)

[Tunkelang 09] Tunkelang, D.: *Faceted Search*, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2009)

[横本 11a] 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)

[Yokomoto 11b] Yokomoto, D., Makita, K., Kawada, Y., Utsuro, T., and Fukuhara, T.: Utilizing Wikipedia in Categorizing Topic related Blogs into Facets, in *Proc. 12th PACLING* (2011)