Wikipedia を知識源とするニュース・ブログ間のトピック対応付け*

佐藤 由紀† 中崎 寛之‡ 川場 真理子‡ 宇津呂 武仁‡
吉岡 真治§ 福原 知宏¶ 中川 裕志‖ 神門 典子**
筑波大学 第三学群工学システム学類†，筑波大学大学院 システム情報工学研究科‡，
北海道大学大学院 情報科学研究科§，東京大学 人工物工学研究センター¶，
東京大学 情報基盤センター‖，国立情報学研究所**

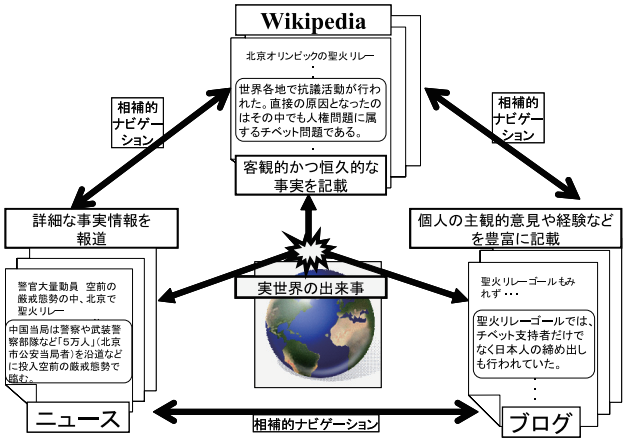

図 1: Wikipedia，ニュース，ブログ間のトピック対応付けの枠組み

1 はじめに

本論文では，Wikipedia，ニュース，ブログの三種類の情報源の間で，密接に関連する項目や記述部分のトピックを相互に対応付ける機能を実現する [佐藤 09](図 1)．そこで，まず，あるトピックについて，Wikipedia のエントリから関連する用語を抽出し，これらの用語を知識源として，ニュース，ブログから関連するニュース記事，ブログサイト，ブログ記事を検索する．この検索のうち，特にブログサイトおよびブログ記事の検索においては，我々はすでに，[川場 08] において，Wikipedia エントリの記述内容をトピックとする有用なブログサイトおよびブログ記事を検索する方式を確立している．この方式においては，Wikipedia エントリ名を表すキーワードを用いて商用検索エンジン API により上位のブログサイトを収集し，これを，当該キーワード，および Wikipedia エントリから抽出した関連語の出現数順に順位付けするという要素技術を用いている．

一方，本論文では，同様の検索手法をニュース記事の順位付けにも用いることにより，Wikipedia エントリの記述内容をトピックとするニュース記事を選別する方式の有効性を示す．さらに，ニュース記事の順位付けにおいては，Wikipedia エントリから関連語を抽出するだけでなく，エントリ本文テキストから名詞句を抽出し，この名詞句と関連語を併用することにより，ニュース記事を順位付けする方式の有効性を示す[1]．

また，上記の，Wikipedia を知識源としてニュース，ブログから関連する項目や記述部分を検索する方式を一般化すると，Wikipedia を知識源としなくても，ニュース，ブログ間で関連する項目や記述を相補的に検索する，あるいは，ニュース，ブログを情報源として，関連する Wikipedia エントリを検索する，という方向でのナビゲーションの実現が可能となる．本論文の研究においては，今後，そのような柔軟な方向性を持った，Wikipedia，ニュース，ブログ間で相補的にトピックを対応付ける方式の研究を進める．

2 Wikipedia エントリからの関連語抽出

ニュース記事およびブログ記事の検索において，Wikipedia エントリを知識源として用いるために，エントリ本文から当該トピックの関連語を抽出する．本論文においては，当該エントリのリダイレクトのタイトル，エントリ本文中の太字，エントリ本文中においてリンク

*Linking Topics of News and Blogs with Wikipedia as Fundamental Knowledge Source

†Yuki Sato, College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba
‡Hiroyuki Nakasaki, Mariko Kawaba, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
§Masaharu Yoshioka, Graduate School of Information Science and Technology, Hokkaido University
¶Tomohiro Fukuhara, Research into Artifacts, Center for Engineering, University of Tokyo
‖Hiroshi Nakagawa, Information Technology Center, University of Tokyo
**Noriko Kando, National Institute of Informatics

[1]本論文の執筆段階においては，開発途中のため，ニュース記事の順位付けには全ての種類の用語を用いているが，ブログサイト・ブログ記事の順位付けにおいては，Wikipedia エントリのタイトル，関連語のみを用いている．また，今後，ニュース記事，および，ブログサイト・ブログ記事の順位付けにおいて，Wikipedia エントリのタイトル，関連語 (リダイレクトおよび強調文字，リンク)，エントリ本文テキスト中の名詞句等の数種類の用語について，個別に効果の評価を行う予定である．

されている他エントリのタイトル，本文中の各段落のタイトル，および，本文テキスト中の全名詞句を関連語として抽出する [川場 08].

3 ニュース記事・ブログ記事の検索と順位付け

3.1 ニュース記事検索

Wikipedia エントリをトピックとするニュース記事の検索においては，Wikipedia エントリのエントリ名を検索クエリとして，検索クエリを含む記事を全て収集した．

3.2 ブログ記事検索

3.2.1 ブログサイトの収集

Wikipedia エントリをトピックとするブログサイトの収集においては，Yahoo!Japan 検索 API を利用し，大手 11 社[2]のブログホストに限って検索を行った．検索の際には，Wikipedia エントリのエントリ名を検索クエリとして，複数のブログホストを一度に指定して検索し，1000 件の記事を取得する．しかし API の検索ではブログ記事単位の検索になるので，同一著者のブログ記事は一つのブログサイトにまとめるという作業を行った．その結果，一トピックあたり約 200 前後のブログサイトを取得することができた．その後，各ブログサイトにおいて，Wikipedia エントリのエントリ名のヒット数を求め，ヒット数が下限未満 (本論文では，10) のブログサイトを削除した．

3.2.2 ブログ記事の選別

次に，収集されたブログサイト中のブログ記事のうち，検索トピックに関連のある記事のみを選別するために，2 節の手順により Wikipedia エントリから抽出した関連語が出現する記事のみを選別する．具体的には，当該 Wikipedia エントリのリダイレクトのタイトル，エントリ本文中の太字，および，エントリ本文中においてリンクされている他エントリのタイトルを関連語として抽出し，それらの関連語のいずれかが出現する記事のみを選別する．

3.3 ニュース記事・ブログ記事の順位付け

検索されたニュース記事およびブログ記事の順位付けにおいては，2 節の手順により Wikipedia エントリから抽出した関連語を用いる．記事の順位付けの際には，2 節において抽出された関連語 t の種類 $type(t)$ ごとに重み $w(type(t))$ を決めておき，記事中に出現する全関連語を用いて以下のスコアを計算し，その降順に記事を順位付けする．

$$\sum_t w(type(t)) \times freq(t)$$

ただし，$freq(t)$ は，記事中における関連語 t の出現頻度である．

ここで，関連語 t の種類 $type(t)$ ごとの重み $w(type(t))$ は，ニュース記事の順位付けにおいては,「他エントリ・リンク」は用いず，その他の重みを全て 1 とした．また，ブログ記事の順位付けにおいては,「リダイレクト」を 3,「太字」を 2,「他エントリ・リンク」を 0.5 とし，その他は用いなかった (今回の評価実験の範囲では，関連語の中にエントリタイトルも含まれている．)

[2]FC2.com，yahoo.co.jp，rakuten.ne.jp，ameblo.jp，goo.ne.jp，livedoor.jp，Seesaa.net，jugem.jp，yaplog.jp，webry.info.jp，hatena.ne.jp

表 1: サンプルトピックおよびサブトピック一覧

トピック	サブトピック
京都議定書	議決内容，締約状況，京都メカニズム，日本の削減量の内訳と現状，京都議定書に関する議論
ミサイル	ミサイルの種類，新たな技術，イージス艦衝突事故，技術漏洩，安全保障，ゲーム (ゲーム内の武器として)，MD，テポドン
ヒラリー・クリントン	大統領選以前，スーパーチューズデイ，女性大統領誕生への期待，ヒラリー凋落，選挙資金，対日本のヒラリーのスタンス，大統領選以降
バンク･オブ･アメリカ	合併前のバンクオブアメリカ社の歩み，合併後のバンクオブアメリカ社の歩み，株価レポート，米金融不安，世界金融市場と日中金融機関，メリルリンチとの合併，米国株式市場，韓国金融市場

表 2: Wikipedia エントリから抽出した用語数

トピック	リダイレクト	太字	他エントリ･リンク	段落タイトル	本文名詞句
京都議定書	0	14	404	9	720
ミサイル	0	1	149	14	988
ヒラリー･クリントン	0	3	247	11	622
バンク･オブ･アメリカ	0	8	81	4	333

4 サンプルトピック・ニュース記事・ブログ記事

本論文において分析対象としたトピックの一覧を表 1 に示す．また，本論文において検索の対象としたニュース

表 3: 新聞社ごとのニュース記事数

新聞社	ニュース記事数
朝日新聞	33039
読売新聞	26657
日経新聞	39164
CNN 日本語版 (アメリカ)	5344
朝鮮日報日本語版 (韓国)	13882
中央日報日本語版 (韓国)	10488
人民網日本語版 (中国)	11697
合計	140271

表 4: 検索されたニュース記事数およびブログサイト数・ブログ記事数 (各トピックごと)

トピック	ニュース記事数	ブログサイト数	ブログ記事数
京都議定書	413	97	1258
ミサイル	739	78	2158
ヒラリー・クリントン	663	68	470
バンク・オブ・アメリカ	264	38	506

表 5: ブログ記事の検索性能の評価 (%)

トピック	関連語	トピック名	
	無関係記事率	無関係記事率	スプログ率
京都議定書	0	10	0
ミサイル	0	0	0
ヒラリー・クリントン	5	0	45
バンク・オブ・アメリカ	0	55	20

記事について新聞社ごとの記事数 (ニュース記事の収集時期は，2008 年 1 月 1 日 ~9 月 29 日[3]) を表 3 に示す．さらに，2 節の手順により Wikipedia エントリから抽出した用語 (表 2 に数を示す) を用いて，前節で述べた手法によりニュース記事およびブログ記事の検索・順位付けを行い，上位 20 記事ずつを収集した．検索されたニュース記事数およびブログサイト数・ブログ記事数を表 4 に示す．さらに，各トピックについて，Wikipedia の段落，および，検索されたニュース記事およびブログ記事中の記述内容を人手で分析し，表 1 のサブトピック一覧を作成した．

5 ニュース記事・ブログ記事の検索性能の評価

前節で述べた手法の有効性を検証するために，トピック名の出現頻度の降順に順位付けしたものとの比較を行った．

ニュース記事の検索性能については，関連語を用いた順位付け，トピック名を用いた順位付けともに，表 1 のサブトピックと無関係な記事は，3 記事のみであり，大きな違いは見られなかった．

ブログ記事の検索性能については，表 5 にブログ記事の検索性能の評価を示す．この結果からわかるように，関連語を用いた順位付けでは，スプログ [佐藤 08] が混入せず，無関係記事もわずかであった．一方，トピック名のみを用いた順位付けでは，無関係記事・スプログが多く混入していた．この結果より，関連語を用いた順位付けの有効性が確認できた．

[3]ニュース記事については，すでに，2006 年分以降を収集済であり，今後，この全体を対象として評価実験を行う予定である．

表 6: 情報源間のサブトピック共有率 (各トピックごと)

トピック	Wikipedia-ニュース間	Wikipedia-ブログ間	ニュース-ブログ間
京都議定書	0.75	0.88	0.82
ミサイル	0.68	0.33	0.73
ヒラリー・クリントン	0.49	0.39	0.62
バンク・オブ・アメリカ	0.28	0.06	0.62

6 Wikipedia・ニュース・ブログ中の記述の比較分析

6.1 情報源間のサブトピック共有率

ここでは，表 1 に挙げた各トピックについて，Wikipedia エントリ中の段落，ニュース記事，ブログ記事の記述内容を人手で比較して，表 1 中のサブトピックがどの程度共有されているかを測定する．具体的には，まず，Wikipedia の各一段落，ニュース記事一記事，ブログ記事一記事を，それぞれ，サブトピックに関する記述の有無を次元とするベクトル d_w，d_n，d_b (ただし，各ベクトルの大きさは，$\mid d_w \mid = \mid d_n \mid = \mid d_b \mid = 1$ に正規化する) によって表現する．次に，一つのトピックについて，Wikipedia エントリ中の全段落，検索結果の全ニュース記事 (本論文では，20 記事)，および，検索結果の全ブログ記事 (本論文では 20 記事) から，各ベクトルの総和 $\sum d_w$，$\sum d_n$，$\sum d_b$ を求める．そして，これらの和ベクトルの間の余弦によって，二つの情報源間のサブトピック共有率を定義する．

$$\text{情報源 } ij \text{ 間のサブトピック共有率} \quad = \quad \frac{\left(\sum d_i\right) \cdot \left(\sum d_j\right)}{\left|\sum d_i\right| \left|\sum d_j\right|}$$

このようにして求めたサブトピック共有率を表 6 に示す．

6.2 各情報源中の記述の比較分析

表 7 に，トピック「京都議定書」のサブトピックについて，Wikipedia，ニュース，ブログに共通する記述内容，および各情報源特有の記述内容の抜粋を示す．この結果から分かるように，Wikipedia で解説されている事項のうちの多くは，ニュースもしくはブログでも取り上げられている．一部では，ニュースでは検索されなかったサブトピックも存在するが，その原因の大半は，ニュース記事の収集時期が短いため，および，調査対象としたニュース記事数が 20 件と少ないためであると考えられる．また，Wikipedia，ニュースともに，情報源特有の記述が観測されている．このうち，Wikipedia 特有の記述内容については，ニュース記事の収集時期を拡大することにより，同一事項に関する記述の有無が正確に判断できると考えている．今後，以上の分析結果をふまえて，Wikipedia，ニュース，ブログにおける個々の記述内容を検索クエリとして，関連記事検索を行う予定であり，この研究の一環において，同一事項の記述の有無の自動

表 7: Wikipedia，ニュース，ブログ中の記述の比較分析 (「京都議定書」)

サブトピック	共通事項	Wikipedia 特有	ニュース特有	ブログ特有
議決内容	参加各国の削減目標および，議決内容に違反した場合の処置について	(特有の記述なし)	(記事なし)	「実はこの地球上で唯一日本だけが削減義務を課せられているのです.」「民間企業に対する二酸化炭素の強制的な排出削減義務化は，多くの企業を強制倒産させ、必ず大不況を招く愚策だと思います.」
締約状況	締結から発行までの流れ，および，参加国の参加体制について	(特有の記述なし)	京都議定書に参加しないアメリカの洞爺湖サミットにおける温暖化対策関連の取り組みについて	「EU では，マイナス 6%などとちんけなことは言わずに，マイナス 20%を目標にすると，1/10 に発表したばかり．すごいですねー．えらいですねー．先進国ですねー．· · · 各国とも排出削減にふんばってる時に，ニホンとアメリカは逆行しちゃってるのだ.」
京都メカニズム	植林活動，国外での活動，削減量の国家間取引など，温室効果ガスの削減をより容易にするための規定について.	既存の森林を CO_2 吸収分として算出できるように規約を拡張した.	日本の世界最先端環境技術について，世界にアピールしていくことにより地球温暖化を防ごう.	「人間は利益がなければ地球環境を守れないということか。」「これからは，地球温暖化防止のためには，国も，企業も、ひいては国民も自分の努力で CO_2 削減が出来ない人はその分お金を出さなくてはいけなくなるよ · · ·」
日本の削減量の内訳と現状	日本の各分野における削減量の具体値および現状		(記事なし)	「電力会社に余剰電力を売電しても，温暖化防止にならないことはこれまで知らされることがありませんでした.」
京都議定書に関する議論	日本に架せられた削減目標の重さ，京都議定書の内容での温暖化防止効果の程度，地球温暖化現象の原因，京都議定書後の取り組み，等の議論の概略	どれほどの効果があるのかという議論や，二酸化炭素の排出量を規制するのは本当に温暖化防止に役立つのかという議論の紹介	中国政府の温室効果ガス削減に関する取り組みと声明	「日本国民にとって一番の問題は，この条約で一番不利なのがどこからどうかんがえても日本だ，ということです.」

判定に取り組む予定である．

7 関連研究

ニュースとブログとの間の相補的な利用については，[小原 05, 池田 05, 石崎 08] などの研究がある．[小原 05] では，ブログ記事中で参照しているウェブサイトやニュース記事をそのユーザの興味の対象として，ブロガーの嗜好を利用したウェブ情報推薦システムを提案している．具体的にはニュースサイトとブログの対応付けを行い，ユーザの嗜好にあったニュース記事を推薦するということを行っている．ただし，ニュース記事とブログの対応付けにおいては，ブログからニュース記事への引用の有無を利用しており，ブログ記事に対するテキスト検索は行われていない．[池田 05] では，ニュース記事とブログ記事との間の文書類似度において，語の出現頻度の推移を考慮した重み付けを用いることにより，ニュース記事に関連したブログ記事を対応付ける手法を提案している．この手法は我々の研究においても有用と考えられるので，今後，本論文の，Wikipedia エントリを知識源とする手法との併用を進める．また，[石崎 08] では，ブログ記事からニュース記事へのアンカーリンクを用いて，ニュース記事に関連するブログ記事の収集を行っている．本論文でも，ブログ記事からニュース記事へのアンカーリンクが抽出できる場合には，その情報を利用する予定であるが，アンカーリンクが抽出できないブログ記事の場合には，主として，Wikipedia エントリを知識源とする本論文の手法による対応付けを用いる．

一方，[吉岡 07] では，同じ事象について，複数の情報源の情報の伝え方の異なりかたを分析することを目的として，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，各々の国の世論がどのように事象を分析しているのかを把握する方式を提案している．

8 おわりに

本論文では，検索エンジン等を用いた検索行動のうちでも，特に，客観的かつ恒久的な事実を記載した Wikipedia，詳細な事実情報を報道するニュース，および，個人の主観的意見や経験などを豊富に記載したブログの三種類の情報源の間で，密接に関連する項目や記述部分の間を相互にナビゲートする方式について提案し，三種類の情報源の記述内容の比較分析を行った結果について述べた．今後は，複数情報源間における同一記述の有無の自動判定について取り組む．

参考文献

[池田 05] 池田大介, 藤木稔明, 奥村学：blog とニュース記事の自動対応付け, 言語処理学会第 11 回年次大会論文集, pp. 1030–1033 (2005).

[石崎 08] 石崎諒, 青野雅樹：Web ニュースに対するブログ意見の分析ツール, 電子情報通信学会技術研究報告, WI2-2008-52, pp. 11–12 (2008).

[川場 08] 川場真理子, 中崎寛之, 宇津呂武仁, 福原知宏：Wikipedia エントリとブログサイトの対応付けによる日本語ブログ空間のトピック分布推定, 情報処理学会研究報告, Vol. 2008, No. (2008–NL–187), pp. 83–90 (2008).

[小原 05] 小原恭介, 山田剛一, 絹川博之, 中川裕志：Blogger の嗜好を利用した強調フィルタリングによる Web 情報推薦システム, 第 19 回人工知能学会全国大会発表論文集 (2005).

[佐藤 08] 佐藤有記, 宇津呂武仁, 福原知宏, 河田容英, 村上嘉陽, 中川裕志, 神門典子：キーワードの時系列特性を利用したスパムブログの収集・類型化・データセット作成, DEWS2008 論文集 (2008).

[佐藤 09] 佐藤由紀, 中崎寛之, 川場真理子, 宇津呂武仁, 福原知宏：Wikipedia を知識源とするニュース・ブログ間の相補的ナビゲーション, DEIM フォーラム論文集 (2009).

[吉岡 07] 吉岡真治：複数のニュース源の差異を考慮したニュース分析の研究, 言語処理学会第 13 回年次大会「大規模 Web 研究基盤上での自然言語処理・情報検索研究」 ワークショップ論文集, pp. 27–20 (2007).