Wikipediaを介した関連ニュース・ブログの対応付け — Wikipediaエントリの分析 —

佐藤由紀[†1] 横本大輔[†2] 中崎寛之[†1]
宇津呂武仁[†1] 吉岡真治[†5] 福原知宏[†3]
神門典子[†4] 中川裕志[†6] 清田陽司[†6]

本研究では，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検索に焦点を当てて，利用者の検索行動を支援する枠組みを提供することを目的とする．本論文では，Wikipediaエントリを介して，関連するニュース・ブログを対応付ける方式において，各Wikipediaエントリの有効性を分析し，ニュース・ブログ間の相補的検索においてWikipediaエントリを選別するためのインタフェースを設計する．特に，本稿では，Wikipediaエントリを介して，ニュース記事に関連するブログ記事を検索する方式を評価した．システムが提示したWikipediaエントリのうち，入力となるニュース記事と密接に関連するエントリを利用者が自在に選択し，選択されたエントリのみを用いて関連ブログ記事の検索を行うための相補的ナビゲーション・インタフェースを作成した．評価実験の結果，利用者が適切なエントリの選択を行うことによって，ブログ記事検索結果における関連ブログ記事の割合を大幅に改善することができた．

Analysis on Wikipedia Entries in Linking Topics of News and Blogs through Wikipedia

Yuki SATO ,[†1] Daisuke YOKOMOTO ,[†2]
Hiroyuki Nakasaki ,[†1] Takehito UTSURO ,[†1]
Masaharu Yoshioka ,[†5] Tomohiro Fukuhara ,[†3]
Noriko KANDO ,[†4] Hiroshi NAKAGAWA [†6]
and Yoji KIYOTA [†6]

We study complementary navigation of news and blog, where *Wikipedia* entries are utilized as fundamental knowledge source for linking news articles and blog feeds/posts. In the proposed framework, given a topic as the title of a Wikipedia entry, its Wikipedia entry body text is analyzed as fundamental knowledge source for the given topic, and terms strongly related to the given topic are extracted. Those terms are then used for ranking news articles and blog posts. In the scenario of complementary navigation from a news article to closely related blog posts, Japanese Wikipedia entries are ranked according to the number of strongly related terms shared by the given news article and each Wikipedia entry. Then, top ranked 10 entries are regarded as indices for further retrieving closely related blog posts. All the retrieved blog posts are finally ranked all together. The retrieved blog posts are then shown to users as blogs of personal opinions and experiences that are closely related to the given news article. In our preliminary evaluation, through an interface for manually selecting relevant Wikipedia entries, the rate of successfully retrieving relevant blog posts improved.

1. はじめに

本研究では，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検索に焦点を当てて，利用者の検索行動を支援する枠組みを提供することを目的とする．本研究では，これら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機能を実現し，利用者の検索行動を支援する (図 1)[8]．

Wikipedia，ニュース，ブログの三者を比較すると，Wikipediaは，インターネット上の最大規模の百科事典として，近年，様々な研究分野において利用されている (例えば，文献 1)，

†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
†2 筑波大学 第三学群工学システム学類
College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba
†3 東京大学 人工物工学研究センター
Research into Artifacts, Center for Engineering, University of Tokyo
†4 国立情報学研究所
National Institute of Informatics
†5 北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
†6 東京大学 情報基盤センター
Information Technology Center, University of Tokyo

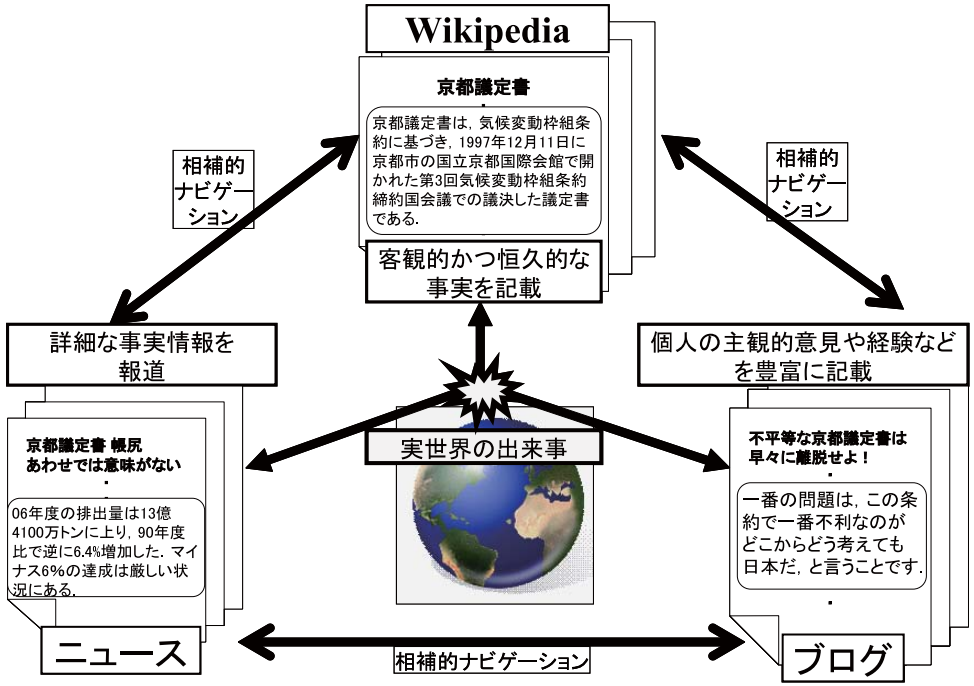


図 1 Wikipedia，ニュース，ブログ間の相補的ナビゲーションの枠組み

9))．日本語では，約 62 万のエントリ (2009 年 10 月時点) が収録されており，しかも，多くの人が自由にエントリを書くことができるため，ニュースやブログで話題となる事項のエントリが，迅速に作成されるという特徴を持っている．Wikipedia を利用した研究事例としては，図書館の分類体系と Wikipedia カテゴリの対応付けを行う研究[9] や，Wikipedia の言語間リンクを利用して多言語対訳辞書を作成するという研究[1] などがある．

ニュースとブログを比較すると，ニュースは，従来より，日々の報道を閲覧するという形で利用されてきた．一方，ブログについても近年，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るのに伴って，様々な情報がブログに記載され，また，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出来るようになった．具体的なサービスの例として，*Technorati*[⋆1]，*BlogPulse*[⋆2]，*kizasi.jp*[⋆3]，*blogWatcher*[⋆4]などが挙げられる．これらの検索サービスは，巨大なブログ空間の索引付けという観点から見ると，キーワードや評判，時系列変化や人手によって作成されたカテゴリ情報などを索引として用いて，利用者の求めるブログ記事やブログサイトを検索する．

本研究において，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を相互にナビゲートする機能を実現するにあたっては，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[8]．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，文献 4)–6) 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 エントリ名をあらわ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．また，文献 8) においては，同様の検索手法をブログ記事・ニュース記事の順位付けに用いることにより，Wikipedia エントリの記述内容をトピックとするブログ記事・ニュース記事を選別する方式の有効性を確認した．また，Wikipedia エントリとニュース記事・ブログ記事間の類似度，および，ニュース記事とブログ記事の類似度について，定式化を行った．

以上の先行研究をふまえて，本稿では，Wikipedia エントリを介して，ニュース記事に関連するブログ記事を検索する方式を評価した．また，上位に順位付けされた Wikipedia エントリのうち，入力となるニュース記事と密接に関連するエントリを利用者が自在に選択し，選択されたエントリのみを用いて関連ブログ記事の検索を行うための相補的ナビゲーション・インタフェースを作成した．評価実験の結果，利用者が適切なエントリの選択を行うことによって，ブログ記事検索結果における関連ブログ記事の割合を大幅に改善することができた[⋆5]．

⋆1 http:/technorati.com/
⋆2 http://www.blogpulse.com/
⋆3 http://kziasi.jp/（日本語のみ）
⋆4 http://blogwatcher.pi.titech.ac.jp （日本語のみ）
⋆5 現時点においては，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の個別の効果を評価できていないが，今後行う予定である．

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[4)–6)]．

3. Wikipedia エントリからのニュース記事・ブログ記事の検索

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 エントリ名を検索クエリとして，検索クエリを含む記事全てを収集した．

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[⋆1]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．

⋆1 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

3.3 Wikipedia エントリとニュース記事・ブログ記事間の類似度

検索されたニュース記事およびブログ記事の Wikipedia エントリとの類似度算出においては，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．具体的には，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，以下の総和によって，Wikipedia エントリ E およびニュース記事・ブログ記事 D の間の類似度 $Sim_{w,nb}(E,D)$ を定義する．

$$Sim_{w,nb}(E,D) = \sum_{t} w(type(t)) \times freq(t)$$

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，

$$w(\text{リダイレクト}) = w(\text{太字}) = w(\text{段落タイトル}) = w(\text{本文名詞句}) = 1,$$
$$w(\text{他エントリ・リンク}) = 0$$

とし，ブログ記事の順位付けにおいては，

$$w(\text{リダイレクト}) = 3, \quad w(\text{太字}) = 2, \quad w(\text{他エントリ・リンク}) = 0.5,$$
$$w(\text{段落タイトル}) = w(\text{本文名詞句}) = 0$$

とする．

3.4 ニュース記事・ブログ記事の順位付け

ニュース記事・ブログ記事の順位付けにおいては，前節で述べた類似度の降順に記事を順位付けする．

4. ニュース記事・ブログ記事からの Wikipedia エントリの検索

ニュース記事・ブログ記事 D からの Wikipedia エントリの検索においては，ニュース記事・ブログ記事中に出現した Wikipedia エントリ名を E_1,...,E_n として，3.3 節で定義した類似度 $Sim_{w,nb}(E_i,D)$ (i=1,...,n) の降順に $E_1,...,E_n$ を順位付けする．

5. Wikipedia エントリを介したニュース記事とブログの対応付け

知識源として Wikipedia エントリを介することにより，ニュース記事もしくはブログ記事を検索質問として，トピックの関連するニュース記事・ブログ記事を対応付けることができる．この際には，ニュース記事もしくはブログ記事を検索質問として，4 節の手順によっ

て検索結果として得られる Wikipedia エントリを知識源として用いる．また，関連するブログ記事もしくはニュース記事の検索は，3 節の手順によって行う．

ここで，検索質問となるニュース記事もしくはブログ記事を D_1，検索対象となるブログ記事もしくはニュース記事を D_2 として，両者の間の類似度を以下の式 $Sim_{n,w,b}(D_1, D_2)$ で定義する．

$$Sim_{n,w,b}(D_1, D_2) = (1 - K_{w,nb})Sim_{n,b}(D_1, D_2) \ + \ K_{w,nb} \sum_{E} (Sim_{w,nb}(E, D_1) + Sim_{w,nb}(E, D_2))$$

ただし，両者の間に介在する Wikipedia エントリを E とする．この類似度 $Sim_{n,w,b}(D_1, D_2)$ においては，$Sim_{n,w,b}(D_1, D_2)$ は，D_1 と D_2 の間の直接の文書間類似度 $Sim_{n,b}(D_1, D_2)$，および，Wikipedia エントリ E を介する際の類似度 $Sim_{w,nb}(E, D_1)$，$Sim_{w,nb}(E, D_2)$ の総和の重み付き和として定義される．D_1 と D_2 の間の直接の文書間類似度 $Sim_{n,b}(D_1, D_2)$ としては，余弦類似度を用いる．ここで，重み $K_{w,nb}$ の値を調整することにより，通常用いられる文書間類似度 $Sim_{n,b}(D_1, D_2)$ と，Wikipedia エントリを介した類似度との間の比率を調整する．

6. ニュース記事に関連するブログ記事の検索

6.1 検索の流れ

本節では，5 節で導入したニュース記事とブログ記事の間の類似度を利用する方式の一つとして，Wikipedia エントリを介して，ニュース記事に関連するブログ記事を検索する枠組みについて述べる．

本研究では，この枠組みのもとで，ニュース記事に関連するブログ記事を利用者が効率的に検索する過程を支援することを目的として，図 2 に示す相補的ナビゲーション・インタフェースを作成した．このインタフェースにおいては，ニュース記事に関連するブログ記事を検索する中間過程において上位に順位付けされた Wikipedia エントリのうち，入力となるニュース記事と密接に関連するエントリを利用者が自在に選択し，選択されたエントリのみを用いて関連ブログ記事の検索を効率よく行うことを支援する．

図 2 のインタフェースにおいては，まず，左上の「ニュース記事選択画面」において，検索の入力となるニュース記事の一覧が示される．この画面上でニュース記事を一つ選択すると，図 2 右上の「Wikipedia エントリ集合選択画面」において，そのニュース記事との関連

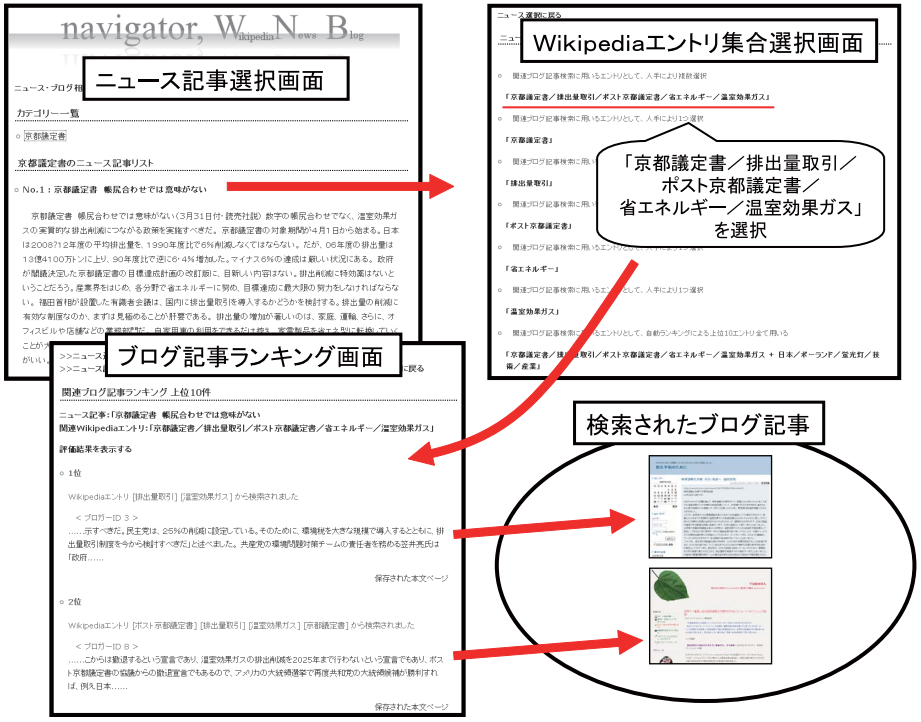


図 2 Wikipedia・ニュース・ブログ間の相補的ナビゲーションのためのインタフェース

性が最も高い Wikipedia エントリ上位 10 個が表示される[*1]．これらのエントリは，3.3 節で導入した Wikipedia エントリおよびニュース記事の間の類似度 $Sim_{w,nb}$ によって順位付けされたうちの上位 10 個である．ここで，実際に，次節で述べる評価実験において用いたニュース記事一覧，および，各ニュース記事に対して，上位に順位付けられた Wikipedia エントリ各 10 個を表 1 に示す．なお，これらのニュース記事は，2008 年 1 月 1 日 ~9 月 29 日の期間に収集した記事集合のうち，特に「京都議定書」をトピックとする記事を選定

*1 図 2 右上の「Wikipedia エントリ集合選択画面」のインタフェースは，プロトタイプ版のため，ニュース記事との関連性が最も高い Wikipedia エントリ上位 10 個のうちの任意の要素を選択するという形式になっていない．正式版のインタフェースにおいては，インタフェースの提示する上位 10 エントリの任意の要素を選択する機能が実現される．

表 1 評価対象ニュース記事概要，関連 Wikipedia エントリ，および，関連ブログ記事概要

ニュース記事 ID	ニュース記事概要	検索された Wikipedia エントリ		ブログ記事概要
		自動順位付け上位 10 エントリ	手動選定エントリ	
1	日本の環境キャンペーンの紹介．排出した温室ガスをお金と環境事業などで補う「カーボンオフセット（Carbonn offset）」運動について．電気使用量削減と温暖化防止について．（新聞社：朝鮮中央日報日本語版，日付：2008-01-25）	環境問題，京都議定書，日本，自動車，カーボンオフセット，交通，アメリカ合衆国， ホテル，二酸化炭素，寄与	京都議定書，カーボンオフセット，二酸化炭素	京都議定書には大きな意義があるが，しかし問題も多い (ブロガー A) 新潟県佐渡市の「カーボンオフセット」運動，温暖化対策と同時にトキの森を豊かに (ブロガー B)
2	二酸化炭素排出量取引の導入に関する有識者会議について．排出量の増加が著しいのは家庭，運輸，オフィスビルや店舗などの業務部門．家電製品の省エネ転換が大切．（新聞社：読売新聞，2008-03-31）	京都議定書，排出量取引，日本，ポスト京都議定書，省エネルギー，ポーランド，蛍光灯，技術，温室効果ガス，産業	京都議定書，排出量取引，ポスト京都議定書，省エネルギー，温室効果ガス	「温室効果ガス 家庭と産業で増」などのニュースを紹介 (ブロガー C) 日本は排出量取引などの経済的手法に頼らざるを得ない状況にある (ブロガー A)
3	森林が吸収する二酸化炭素量の見積もり方や，排出量獲得の際の取り決めなどポスト京都議定書の制定に向けた課題検討について．（新聞社：日経新聞，日付：2008-08-28）	ポスト京都議定書，国際連合，議定書，二酸化炭素，アメリカ合衆国，ディベート，京都，温室効果ガス，国務大臣，ポーランド	ポスト京都議定書，二酸化炭素，温室効果ガス	「ポスト京都議定書も国別目標． ダボス会議で福田首相が表明」などのニュースを紹介 (ブロガー C)
4	地球温暖化問題についての討論．途上国の地球温暖化対策への参加について．日本の省エネルギー技術を世界にもっとアピールすべきだ．（新聞社：読売新聞，日付：2008-06-29）	日本，地球温暖化，環境問題，アメリカ合衆国，政治，資源，第 34 回主要国首脳会議，インド，化石燃料，社会	地球温暖化，第 34 回主要国首脳会議，化石燃料	間違った温暖化対策しかしない政治家たちと違って，日本の電力会社の技術陣は研究開発を着実に進めている (ブロガー D) 日本の気象学者根本氏は「温暖化の原因は化石燃料の燃焼ではない」としている (ブロガー D)

したものの一部である．このうち，図 2 においては，表 1 中において ID=2 のニュース記事を選択した場合の流れを示している．

次に，図 2 右上の「Wikipedia エントリ集合選択画面」において，入力されたニュース記事と密接に関連するエントリのみを利用者が選択し，選択されたエントリのみを用いてブログ記事の順位付けを行った結果を，左下の「ブログ記事ランキング画面」に示す．このブログ記事の順位付けにおいては，5 節で導入した類似度 $Sim_{n,w,b}$ を入力ニュース記事と検索対象ブログ記事の間に適用し，類似度の降順にブログ記事を順位付けする．ここで，$Sim_{n,w,b}$ の計算における Wikipedia エントリ E としては，図 2 右上の「Wikipedia エントリ集合選択画面」において手動で選択されたエントリのみを用いる．表 1 には，入力として用いた各ニュース記事に対して，上位に順位付けられた Wikipedia エントリ各 10 個のうちで，入力されたニュース記事と密接に関連するエントリのみを人手で選定した結果，および，最終的に上位に検索されたブログ記事のうち主要なものの概要を示す．

最後に，図 2 左下の「ブログ記事ランキング画面」においてブログ記事を選択することにより，ブログ記事を閲覧することができる．

6.2 評　　価

6.2.1 評 価 手 順

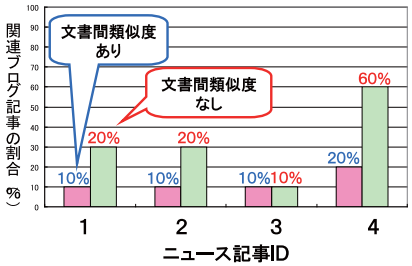
表 1 に挙げた各ニュース記事を入力として，関連ブログ記事の順位付けを行い，その結果を人手で評価した．順位付けされたブログ記事に対して，入力として用いたニュース記事との間の関連性の強さを以下の三段階で判定した．

(a) ニュース記事の内容に密接に関連するブログ記事である．
(b) ニュース記事の内容に部分的に関連するブログ記事である．
(c) ニュース記事の内容に関連しないブログ記事である．

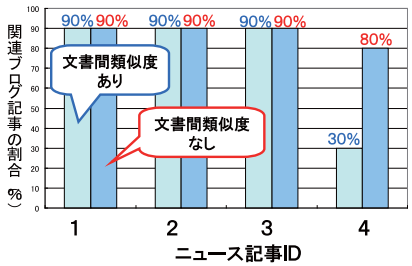
そして，評価対象とするブログ記事数を N とし,「関連するブログ記事」として，上記の (a) のみを対象とする場合，および，(a) と (b) の両方を対象とする場合の二通りについて，以下の「関連ブログ記事の割合」を測定した．

$$\text{関連ブログ記事の割合} = \frac{\text{関連するブログ記事数}}{N}$$

なお，本稿の評価においては，$N = 10$ とした．

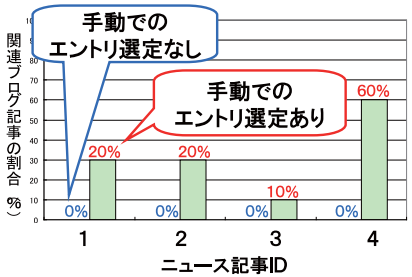


(i) 部分的に関連するブログ記事を含まない

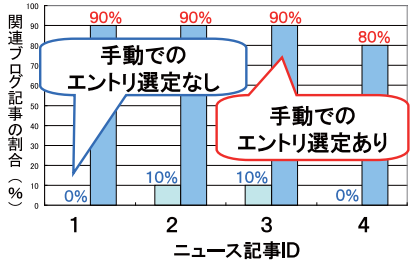


(ii) 部分的に関連するブログ記事を含む

図 3 ニュース記事・ブログ記事の文書間類似度の有無の比較 (手動での Wikipedia エントリ選定後)



(i) 部分的に関連するブログ記事を含まない



(ii) 部分的に関連するブログ記事を含む

図 4 手動での Wikipedia エントリ選定の有無の比較

6.2.2 ニュース記事・ブログ記事の文書間類似度の有無の比較

まず，5 節で導入した類似度 $Sim_{n,w,b}$ を用いてニュース記事とブログ記事の間の関連性の強さを測定するにあたって，ニュース記事とブログ記事の間の直接の文書間類似度 $Sim_{n,b}$ が有効に機能しているかどうかの評価を行った．

まず，この評価を行うにあたっては，図 2 右上の「Wikipedia エントリ集合選択画面」において，手動で Wikipedia エントリの選定を行う場合，および，手動での Wikipedia エントリの選定を行わず，提示された Wikipedia エントリの上位 10 個をそのまま用いる場合，の二通りの設定が考えられる．ここで，現時点の実装では，提示された Wikipedia エントリの上位 10 個をそのまま用いる場合では，「関連ブログ記事の割合」が 10%以下と著しく低かったため，以下では，手動で Wikipedia エントリの選定を行った場合の結果を述べる．

以上の設定において，(i) 部分的に関連するブログ記事 (b) を含まない場合，および，(ii) 部分的に関連するブログ記事 (b) を含む場合，の各々について，ニュース記事とブログ記事の間の直接の文書間類似度 $Sim_{n,b}$ の有無を比較した結果を図 3 に示す．この結果から分かるように，(i)，(ii) のいずれにおいても，文書間類似度を用いない方が高い性能となった．(i) の場合においては，4 ニュース記事のうち 3 記事において 10~40%性能が改善した．また，(ii) の場合においては，4 ニュース記事のうち 1 記事において 50%性能が改善した．

以上の結果から，本稿の評価実験の範囲においては，5 節で導入した類似度 $Sim_{n,w,b}$ を用いてニュース記事とブログ記事の間の関連性の強さを測定するにあたって，ニュース記事とブログ記事の間の直接の文書間類似度 $Sim_{n,b}$ は不要であることが判明した．

6.2.3 手動での Wikipedia エントリ選定の有無の比較

次に，図 2 右上の「Wikipedia エントリ集合選択画面」において，手動で Wikipedia エントリの選定を行う場合，および，手動での Wikipedia エントリの選定を行わず，提示された Wikipedia エントリの上位 10 個をそのまま用いる場合，の二通りの設定に対して，「関連ブログ記事の割合」の比較を行った．なお，前節の結果から，本節の評価においては，5 節で導入した類似度 $Sim_{n,w,b}$ において，ニュース記事とブログ記事の間の直接の文書間類似度 $Sim_{n,b}$ を加算せずに評価を行った．

以上の設定において，(i) 部分的に関連するブログ記事 (b) を含まない場合，および，(ii) 部分的に関連するブログ記事 (b) を含む場合，の各々について，手動での Wikipedia エントリ選定の有無を比較した結果を図 4 に示す．この結果から分かるように，手動での Wikipedia エントリ選定を行わない場合の「関連ブログ記事の割合」は，(i) では 0%，(ii) では 0~10%であった．一方，手動での Wikipedia エントリ選定を行うことにより，(i) での性能は 10~60%に，(ii) での性能は 80~90%に，それぞれ改善した．以上の結果から，現時点の実装では，手動での Wikipedia エントリ選定を行わなければ，上位 10 以内に関連ブログ記事が含まれる割合はごくわずかであるが，手動での Wikipedia エントリ選定を行うことにより，関連ブログ記事の検索が可能であることが分かった．したがって，適切な Wikipedia エントリの自動選定を実現することができれば，本論文の方式を用いた関連ブログ記事検索において一定の性能が達成できることがわかった．

ここで，表 1 において，「自動順位付け上位 10 エントリ」と「手動選定エントリ」を比較すると分かるように，関連ブログ記事検索において「関連ブログ記事の割合」を下げる要因となっている Wikipedia エントリの多くは，一般語や国名といった専門性の低い用語で

ある．これらの用語の多くは，その Wikipedia エントリ本文のテキストが比較的長く，また，ウェブやブログでのヒット数も大きい，といった特徴を持つ．また，一定数のニュース記事を収集した記事集合においても，文書頻度が大きい傾向があることが予測される．したがって，今後は，これらの特徴を用いて，関連 Wikipedia エントリの順位付け結果において，一般語や国名等をより適切に順位付けする方式を実現する．

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，文献 2)，3)，7) などの研究がある．文献 7) 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．文献 2) 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，文献 3) 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，文献 10) 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

本研究では，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検索に焦点を当てて，利用者の検索行動を支援する枠組みを提供することを目的とした．本稿では特に，Wikipedia エントリを介して，ニュース記事に関連するブログ記事を検索する方式の評価を行った．また，上位に順位付けされた Wikipedia エントリのうち，入力となるニュース記事と密接に関連するエントリを利用者が自在に選択し，選択されたエントリのみを用いて関連ブログ記事の検索を行うための相補的ナビゲーション・インタフェースを作成した．評価実験の結果，利用者が適切なエントリの選択を行うことによって，ブログ記事検索結果における関連ブログ記事の割合を大幅に改善することができた．

本研究の，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 を知識源として，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だけでなく，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，という逆方向でのナビゲーションの実現も可能となる．本研究においては，今後，そのような柔軟な方向性を持った，Wikipedia，ニュース，ブログ間の相補的ナビゲーションの研究を進める．

参 考 文 献

1) 新井嘉章，福原知宏，増田英孝，中川裕志：Wikipedia を用いた多言語ブログ検索のための訳語抽出，情報処理学会第 70 回全国大会講演論文集，Vol.5, 情報処理学会，pp. 55–56 (2008).
2) 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).
3) 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).
4) 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).
5) 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のための特定トピックのブログサイト検索，電子情報通信学会第 19 回データ工学ワークショップ，第 6 回日本データベース学会年次大会 (DEWS2008) 論文集 (2008).
6) 川場真理子，中崎寛之，宇津呂武仁，福原知宏：多言語 Wikipedia エントリを用いた特定トピックブログサイト検索と日英対照ブログ分析，第 22 回人工知能学会全国大会論文集 (2008).
7) 小原恭介，山田剛一，絹川博之，中川裕志：Blogger の嗜好を利用した協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).
8) 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，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 (2009).
9) 田村悟之，清田陽司，増田英孝，中川裕志：図書館における自動レファレンスサービスシステムの実現Web 上の二次情報と図書館の一次情報の統合，情報処理学会研究報告， Vol.2007, No.(2007–FI–179), pp.1–8 (2007).
10) 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).