Wikipediaを知識源とするトピック対応付け —ニュースに関連するブログ記事の収集—*

佐藤 由紀† 横本 大輔‡ 中崎 寛之† 宇津呂 武仁† 福原 知宏§
筑波大学大学院 システム情報工学研究科†，
筑波大学 第三学群工学システム学類‡，東京大学 人工物工学研究センター§，

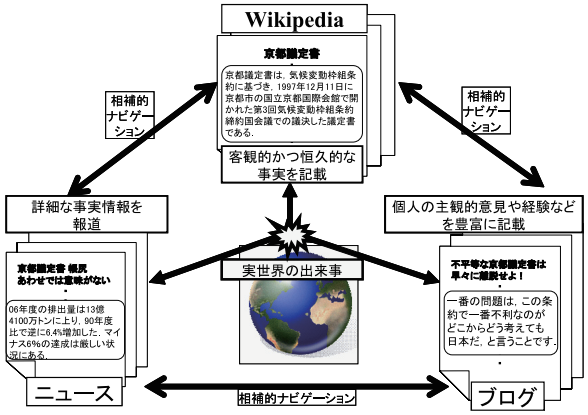


図 1: Wikipedia，ニュース，ブログ間の相補的ナビゲーションの枠組み

1 はじめに

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．日本語では，約 62 万のエントリ (2010 年 1 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．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．

我々はこれまでに，[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た．この手法をふまえて，[佐藤 09a, 佐藤 09b] においては，Wikipedia，ニュース，および，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機能を実現し，利用者の検索行動を支援する枠組みを提案した (図 1)．しかし，これまでに実現した方式では，ニュース記事に関連するブログ記事の順位付けを行う際に，各ブログ記事のブログサイトそのものと密接に関連した Wikipedia エントリのみを知識源として，各ブログ記事のスコアを付けを行っており，この条件が過剰に強い制約となっていた．そのため，特定の Wikipedia エントリとの関連が大きいブログ記事のみが上位に順位付けされて，ニュース記事中の多様な話題のいずれにも密接に関連し，本来上位に順位付けられるべきブログ記事が下位に順位付けられるという弊害を起こしていた．

これに対して，本稿では，ニュース記事との関連が大きい上位 10 個の Wikipedia エントリのすべての関連語を統合してブログ記事の順位付けを行う方式を提案する．実際に，評価実験を通して，本稿で提案する方式の性能が [佐藤 09b] の方式の性能を上回ることを示す．

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 [川場 08]．

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

*Linking Topics through Wikipedia: Collecting Blog Posts related to News

†Yuki Sato, Hiroyuki Nakasaki, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Daisuke Yokomoto, College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba
§Tomohiro Fukuhara, Research into Artifacts, Center for Engineering, University of Tokyo

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．

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

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．また，$R(E)$ は，Wikipedia エントリ E から抽出された関連語集合である．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，

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

とする．

4 Wikipedia エントリ・ニュース記事・ブログ記事の検索・順位付け

4.1 Wikipedia エントリからのニュース記事検索・順位付け

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．ニュース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．

4.2 Wikipedia エントリからのブログ記事検索・順位付け

4.2.1 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 10 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．

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

4.2.2 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．

4.2.3 ブログ記事の順位付け

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．

4.3 ニュース記事・ブログ記事からの Wikipedia エントリの検索・順位付け

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 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 ニュース記事に関連するブログ記事の検索

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，4.3 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，4 節の手順によって行う．検索質問となるニュース記事を D_N として，ブログ記事の検索に知識源として使用する Wikipedia エントリ集合を $ET(D_N)$ として定義する．ただし，以下の定義中の $ET_{10}(D_N)$ は，ニュース記事 D_N との関連性が最も高い Wikipedia エントリ上位 10 個である．また，ニュース記事との関連の大きい Wikipedia エントリを手動で選定する場合，および，選定しない場合の二通りに分けて，$ET(D_N)$ を

定義する．

$$ET(D_N) = \begin{cases} ET_{10}(D_N) & \text{(手動によるエントリ選定なし)} \\ ET_{10}(D_N) \text{ から適切なエントリを手動選定} & \text{(手動によるエントリ選定あり)} \end{cases}$$

さらに，検索対象となるブログ記事を D_B として，両者の間の類似度を以下の式 $Sim_{n,w,b}(D_N, D_B)$ で定義する．ただし，$ET(D_N)$ の各 Wikipedia エントリの関連語集合のすべてを統合してブログ記事の順位付けを行う場合，および，[佐藤 09b] の方式のまま，$ET(D_N)$ の各 Wikipedia エントリの関連語集合の統合を行わずにブログ記事の順位付けを行う場合の二通りを定式化する．

$$Sim_{n,w,b}(D_N, D_B) = \sum_{E \in EE(D_N)} (Sim_{w,nb}(E, D_N) + Sim_{w,nb}(E, D_B))$$

$$EE(D_N) = \begin{cases} ET(D_N) & \text{(複数エントリの関連語集合を統合しない)} \\ \{EU\}, \text{ ただし } R(EU) = \bigcup_{E \in ET(D_N)} R(E) & \text{(複数エントリの関連語集合を統合する)} \end{cases}$$

ただし，EU は，Wikipedia エントリ集合 $ET(D_N)$ において，エントリ毎の関連語集合 $R(E)$ を統合した，関連語集合 $R(EU)$ を持つ仮想統合エントリである．

6 評価

6.1 評価手順

ニュース記事 10 件を入力として，それぞれ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．用いたニュース記事は，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,「喫煙」「京都議定書」「サブプライムローン」「振り込め詐欺」「年金」「臓器移植」「医療事故」をトピックとする記事を選定したものの一部である．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．

表 2: 評価対象ニュース記事タイトル，および，関連 Wikipedia エントリ上位 10 エントリ

ニュース記事 ID	ニュース記事タイトル	関連 Wikipedia エントリ上位 10 エントリ
1	年金記録改ざん問題 社保庁、受給者に直接説明へ	厚生年金，社会保険，社保庁，社会保険庁，拠出，改ざん，改竄，社会保険事務所，標準化，給与
2	日銀、金利を据え置き サブプライム問題の影響見極め	アメリカ合衆国，サブプライムローン，サブプライム問題，金融政策，欧州中央銀行，日本銀行，連邦準備制度理事会，連邦準備制度，中央銀行，金融市場
3	定形小包郵便で現金詐取 振り込め詐欺に新手口	おれおれ詐欺，振り込め詐欺，架空請求詐欺，架空請求，コンビニエンスストア，預金，融資詐欺，融資保証金詐欺，現金書留，書留郵便
4	文書で同意は３件、病気で摘出１１件中 宇和島臓器移植	万波誠，徳洲会，宇和島徳洲会病院，腎臓，器官，泌尿器，愛媛研，フロー，宇和島市，レシピエント
5	病状悪化で売買を決意 臓器移植事件で山下容疑者ら	宇和島徳洲会病院，被疑者，徳洲会，臓器移植法，臓器の移植に関する法律，愛媛県，人工透析，ドナー，貨幣，捜査本部

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．

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

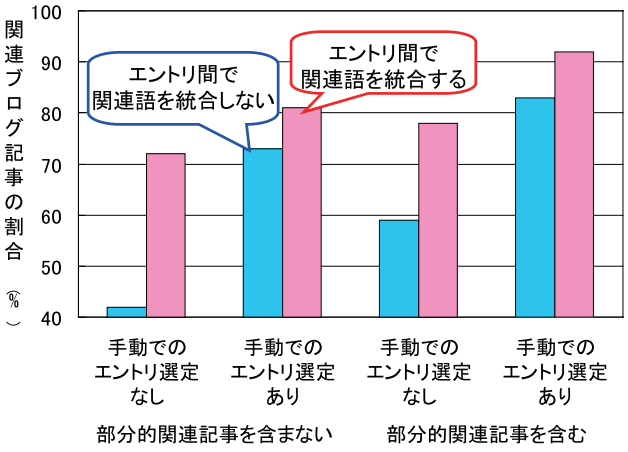
なお，本稿の評価においては，$N = 10$ とした．

6.2 評価結果

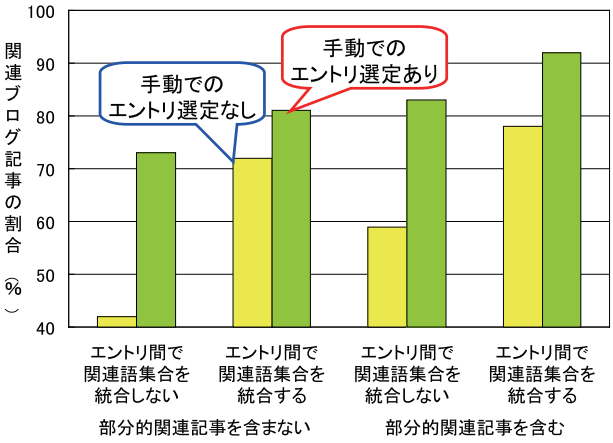
5 節で定義した $Sim_{n,w,b}(D_N, D_B)$ を用いてニュース記事とブログ記事の間の関連性の強さを測定するにあたって，複数エントリの関連語集合の統合が有効に機能しているかどうかの評価を行った．図 2-(i) に，複数エントリの関連語集合の統合の有無を比較した結果を示す．この結果からわかるように，部分的関連記事を含まない場合，部分的関連記事を含む場合のいずれにおいても，複数エントリの関連語集合を統合した方が高い性能となった．手動でのエントリ選定を行わない場合においては，20〜30%性能が改善した．また，手動でのエントリ選定を行う場合においては，10%性能が改善した．

表 1 に，複数 Wikipedia エントリの関連語集合の統合の有無による，関連ブログ記事の順位変動の抜粋を示す．この結果から，複数エントリの関連語集合の統合を行わない場合において，1~10 位の圏外であった関連ブログ記事の多くが，複数エントリの関連語集合の統合を行うことによって，10 位以内に順位が上がっていることがわかる．同様に，複数エントリの関連語集合の統合を行うことによって，ニュース記事に関連しないブログ記事の多くについて順位を下げることができた．

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 $Sim_{n,w,b}(D_N, D_B)$ を用いてニ



(i) 複数エントリの関連語集合の統合の有無の比較



(ii) 手動でのエントリ選定の有無の比較

図 2: ニュース記事に関連するブログ記事の評価結果

表 1: 複数エントリの関連語集合の統合の有無の比較 (関連ブログ記事上位 5 位以内の順位の変動)

ニュース記事 ID		統合ありでの順位 (← 統合なしでの順位, 関連するエントリ数)
1	関連あり	**1 位** (← 2 位, 3 エントリ), **2 位** (← 8 位, 5 エントリ), **3 位** (← 圏外, 2 エントリ), **4 位** (← 圏外, 2 エントリ), **5 位** (← 圏外, 3 エントリ)
2	関連あり	**3 位** (← 圏外, 3 エントリ)
	部分的関連	1 位 (1 位, 3 エントリ), 2 位 (2 位, 3 エントリ), **5 位** (← 7 位, 3 エントリ)
	関連なし	4 位 (← 6 位, 3 エントリ)
3	関連あり	**2 位** (← 圏外, 4 エントリ), **3 位** (← 圏外, 3 エントリ), **4 位** (← 圏外, 1 エントリ), 5 位 (← 2 位, 1 エントリ)
	関連なし	1 位 (1 位, 1 エントリ)
4	関連あり	1 位 (1 位, 3 エントリ), 2 位 (2 位, 5 エントリ), **3 位** (← 6 位, 3 エントリ), **4 位** (← 圏外, 3 エントリ), **5 位** (← 圏外, 3 エントリ)
5	関連あり	**4 位** (← 圏外, 2 エントリ), **5 位** (← 圏外, 3 エントリ)
	関連なし	1 位 (← 圏外, 2 エントリ), 2 位 (← 圏外, 1 エントリ), 3 位 (← 7 位, 1 エントリ)

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，複数エントリの関連語集合の統合が有効であることが分かった．

7 おわりに

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の評価を行った．上位に順位付けされた Wikipedia エントリにおいて，エントリ間で関連語集合を統合することによって，ブログ記事検索結果における関連ブログ記事の割合を改善することができた．今後は，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．具体的には，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，という逆方向でのナビゲーションを実現していきたい．

参考文献

[川場 08] 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).

[佐藤 09a] 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).

[佐藤 09b] 佐藤由紀, 横本大輔, 中崎寛之, 宇津呂武仁, 吉岡真治, 福原知宏, 神門典子, 中川裕志, 清田陽司：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 —, 情報処理学会研究報告, Vol. 2009, No. (2009–NL–194) (2009).