Wikipedia エントリに関連するブログサイトの収集*

横本 大輔† 中崎寛之‡ 宇津呂 武仁‡ 福原 知宏§
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‡，
東京大学 人工物工学研究センター§

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．具体的なサービスの例として，*Technorati*[1]，*BlogPulse*[2]，*kizasi.jp*[3]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．

ここで，本稿では個々のブログ記事ではなく，ある同一のトピックについてまとまった規模の記述が書かれたブログサイトに注目する．そして，そのような専門的内容を含むブログサイトを選択的に検索する手法を提案する．本稿において実現をめざす手法と比較すると，既存の検索エンジン API を用いたブログ検索においては，被リンク数の多い人気ブログサイトの記事から優先的に検索される傾向にある．したがって，既存の検索エンジン API を用いた場合は，被リンク数は多くないが，特定のトピックについて詳細な情報を載せているブログサイトが検索されにくい．この問題に対して，本稿の手法では，特定のトピックについての詳細な情報を含むブログサイトを選択的に検索することを実現するために，各トピックについての Wikipedia エントリ中の記述を知識源として利用する (3 節)．特に，本稿では日英両言語のブログサイトの検索を行うが，その際には，Wikipedia における日英両言語のエントリを知識源とする．そして，トピック名がタイトルである Wikipedia エントリを知識源として，トピックに密接に関連する用語を抽出し，それらの関連語がより多く含まれるブログサイトを検索するという手法を用いる．実際に，本稿の手法を既存の検索エンジン API と比較し，ブログサイトの検索性能において本稿の手法が優れていることを示す (4 節)．

2 評価対象トピック

まず，検索手法の精度比較を日英両言語で行うために用いるトピックとしては，日英両言語のブログ空間において，そのトピックについて詳細な記述を掲載しているブログサイトが十分な数存在する可能性が高いトピックが望ましい．そのための手がかりとして，予備調査として，日本語・英語それぞれのブログ空間におけるトピック名のヒット数の範囲と，詳細な記述を掲載しているブログサイトの有無との相関を分析した結果，ブログ空間におけるヒット数が 10,000 以上であれば，詳細な記述を掲載しているブログサイトが存在する可能性が比較的高いことが分かった [川場 09a]．実際に，日英両言語の間で言語間リンクを介して対訳関係にあるエントリが存在するエントリのうち，日本語 Wikipedia エントリのタイトルが日本語ブログ空間中で 10,000 以上のヒット数を持ち，かつ，英語 Wikipedia エントリのタイトルも英語ブログ空間中で 10,000 以上のヒット数を持つエントリは，約 6,000 個存在した．本稿では，そのうち，特に，社会現象および社会問題に関するトピックに焦点を当てて，約 50 個のトピックを選定し，日英ブログサイト・ブログ記事の検索・自動選定を行った．それらのトピックのうち，本稿では，「アルコール依存症 (alcoholism)」，「リストラ (restructuring)」，「著作権侵害 (copyright infringement)」，計 3 トピックについての評価結果について述べる[4]．

3 Wikipedia を用いた関連語の収集

Wikipedia とは多くの人が自由に書くことができるインターネット上の巨大な辞書のことであり，日本語で約 64 万，英語で約 315 万のエントリ (2010 年 1 月現在) があ

*Automatic Collection of Blog Feeds closely related to Wikipedia

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

[1] http://technorati.com/
[2] http://www.blogpulse.com/
[3] http://kizasi.jp/（日本語のみ）

[4] これまでに，提案手法によって日本語および英語ブログサイト・ブログ記事を自動順位付けした結果を用いて日英各言語のブログ分析を行っているが，5 節 における評価に用いた 3 トピック以外のいずれのトピックにおいても，定性的評価の範囲では，検索エンジン API による順位付けと比較して，提案手法によって一定の改善を達成できている．

表 1: 各トピックごとの Wikipedia 関連語数およびブログサイト・記事数 (日本語/英語)

評価用トピック	Wikipedia 関連語数	ブログサイト数	ブログ記事数
アルコール依存症	60 / 161	183 / 100	2158 / 5185
リストラ	70 / 20	102 / 102	2640 / 6944
著作権侵害	88 / 108	56 / 99	1195 / 4448

表 2: 各トピックの Wikipedia 関連語の抜粋 (リダイレクト / 太字 / 他エントリへのリンクのアンカーテキスト)

日本語トピック名 (英語トピック名)	日本語	英語
アルコール依存症 (alcoholism)	慢性アルコール中毒，酒乱 / アルコール中毒 / 日本酒，飲酒運転，ビール	Alcohol addiction, Alcohol misuse / abuse, problem use /Blood alcohol content
リストラ (restructuring)	事業再構築 / リストラクチャリング / バブル崩壊，レイオフ，ワークシェア，終身雇用	Corporate restructuring // Bankruptcy, Layoff, Spin-out, Voluntary Redundancy
著作権侵害 (copyright infringement)	/ 依拠性，創作的，類似性 / フェアユース，レコード輸入権，告訴，裁判所，実用新案権，著作権，著作物，日本音楽著作権協会	Copyright violation, Illegal copying, Unlawful copying / Bootleg recording , piracy /EU Copyright directive, Peer-to-peer

る．さらに，10 個程度の主要カテゴリ以下にサブカテゴリ，エントリが連なる，巨大なグラフ構造になっている．また，カテゴリがグラフ構造の節にあたり，エントリが節内に列挙されている．

Wikipedia は多くの言語で書かれており，言語間リンクを辿ることで他の言語で書かれたエントリを読むことができる．これまでに，すでに，世界の主要な言語版の Wikipedia が存在するため，十分な種類・数のブログが書かれている言語を対象として本稿の手法を適用することは比較的容易である．また，他の知識源と比較した場合の Wikipedia の最大の利点として，日常的に，新たなエントリの作成と記述の更新が行われており，ブログにおける分析対象となり得る主要なトピックが網羅されている点が挙げられる．

トピック名がタイトルである各言語の Wikipedia エントリを知識源として，トピック名に密接に関連する Wikipedia 関連語を収集する．特に，本稿においては，各エントリのリダイレクト，各エントリ本文中の太字，および，本文中における他エントリへのリンクのアンカーテキストを Wikipedia 関連語として収集する．本稿において評価対象とする各トピックについて，収集した Wikipedia 関連語数を表 1 に，Wikipedia 関連語の抜粋を表 2 に，それぞれ示す．

4 Wikipedia エントリに対応するブログサイトの収集・順位付け

まず，分析対象となるブログサイトの候補を収集するために，本稿では，日本語ブログの検索には，Yahoo!Japan 検索 API[5]を，英語ブログの検索には，米 Yahoo!検索 API[6]をそれぞれ利用する．ただし，日本語ブログホスト大手 10 社[7]，および，英語ブログホスト大手 10 社[8]のブログ会社のドメインに限って検索を行った．検索の際には，複数のドメインを一度に指定して検索し，1000 件の記事を取得する．しかし検索エンジン API を用いた検索ではブログ記事単位の検索になるので，ブログ記事検索後，ブログサイト単位にまとめた．その結果，1 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．

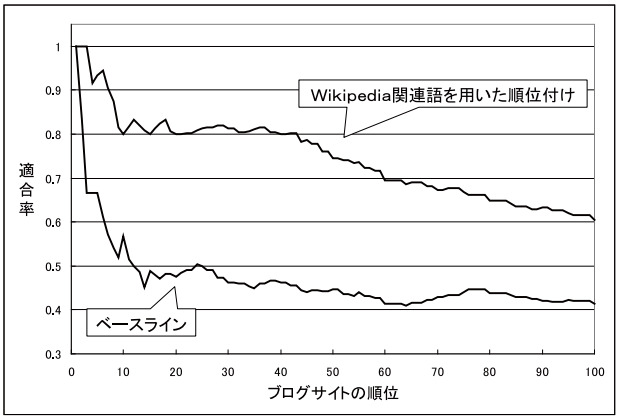
次に，収集した日英ブログサイト集合の中から，トピックについて詳しく書かれたブログ記事を選定する．手法としては，3 節において収集した Wikipedia 関連語のいずれかが出現する各言語のブログ記事を選定する．本稿で評価対象とした各トピックについて，以上の手法により収集したブログサイト数，および，収集したブログサイト中で Wikipedia 関連語のいずれかが出現したブログ記事数を表 1 に示す．

最後に，トピックについて詳細な記述を多く含むブログサイトおよびブログ記事をより上位に順位付けするために，3 節で抽出した Wikipedia 関連語を用いて，ブログ記事およびブログサイトにスコアを付与する．まず，ブログ記事 p のスコアとして，以下を用いる．

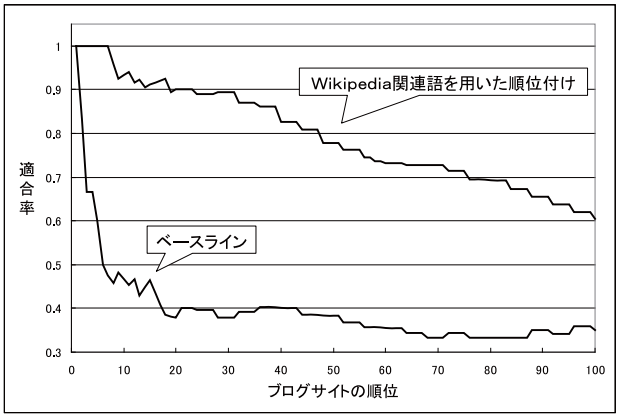
$$PostScore(p) \quad = \quad \sum_{t}(weight(type(t)) \times freq(t))$$

ここで，$weight(type(t))$ は，Wikipedia 関連語 t の種類 $type(t)$ に付与する重みで，$freq(t)$ は，ブログ記事 p 内における Wikipedia 関連語 t の出現頻度である．関連語 t の種類 $type(t)$ としては，Wikipedia エントリタイト

[5] http://www.yahoo.co.jp/
[6] http://www.yahoo.com/
[7] 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
[8] blogspot.com, spaces.live.com, livejournal.com, vox.com, multiply.com, typepad.com, aol.com, blogsome.com, wordpress.com, blogster.com

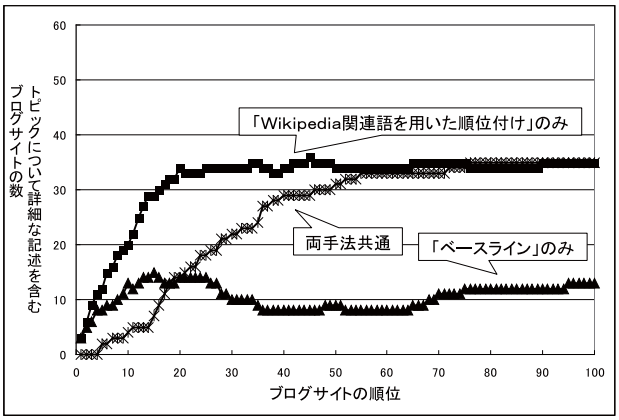


(a) 日本語

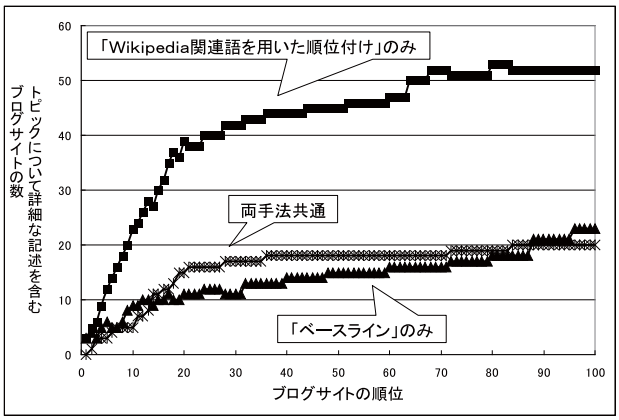


(b) 英語

図 1: 特定トピックのブログサイト検索における適合率の評価



(a) 日本語



(b) 英語

図 2: 特定トピックのブログサイト検索における「トピックについて詳細な記述を含むブログサイトの数」の比較

ル，リダイレクト，エントリ本文中の太字，本文中における他エントリへのリンクのアンカーテキストの 4 種類を考慮し，それぞれの重みは 1 または 0 とする．評価実験を通して最適な重みの組み合わせを求め，全ての重みを 1 とした[9]．さらに，ブログサイト s のスコアとして以下の式を用いる．

$$SiteScore(s) = \sum_{p} PostScore(p)$$

ただし，ブログ記事 p は，ブログサイト s に含まれるブログ記事である．

5 評価

5.1 手順

2 節で挙げた「アルコール依存症 (alcoholism)」,「リストラ (restructuring)」,「著作権侵害 (copyright infringement)」の 3 トピックを対象として，4 節で述べた手法を用いて収集したブログサイトを，4 節で述べた手法で順位付けした結果の人手評価を行った．各トピックを対象として順位付けされた日英のブログサイトのうち，それぞれ，上位 20 ブログサイト，および，100 位までの 20 ブログサイトを等間隔にサンプリングした合計 40 ブログサイトを手動で評価した．評価結果としては,「当該トピックについて詳細な記述を含むか否か」の判定を付与した．また，比較対象として，検索エンジン API に

[9]これらの重みの設定の際には，まず，日本語 Wikipedia エントリに限定して 60 個のエントリを選定し，これを用いて収集したブログサイトを評価対象として，重みの調整を行った．具体的には，まず，各エントリタイトルをトピックとして，前節までの手順により，順位付け対象となるブログサイトを収集した．次に，その中から評価対象として用いるブログサイトをサンプリングして選定し，人手で「当該トピックについて詳細な記述を含むか否か」の判定を付与した．そして,「当該トピックについて詳細な記述を含む」ブログサイトをより上位に順位付けするように，Wikipedia 関連語の重みを設定した．さらに，5 節においてブログサイト検索手法の評価対象として用いた 3 トピック (日本語および英語) に対しても，Wikipedia 関連語の重みが最適な値に設定されていることを確認した．

よるブログ記事の順位付けを変更せずに，ブログサイト単位にまとめた順位付けに対しても，同様に合計 40 ブログサイトを手動で評価した．

5.2 評価結果

評価結果を比較するために，横軸に各ブログサイトの順位をとり，その順位までのブログサイトをすべて「当該トピックについて詳細な記述を含む」と自動判定した場合の適合率を縦軸にとって，その推移をプロットしたものを図 1 の「Wikipedia 関連語を用いた順位付け」(提案手法)，および「ベースライン」に示す．ただし，このプロットには，3 トピック分を平均した結果を示す．この結果から分かるように，日英どちらの言語においても，提案手法によりベースラインの適合率を大幅に改善することが分かる．

また，図 2 には，横軸に各ブログサイトの順位をとり，「トピックについて詳細な記述を含むブログサイトの数」(3 トピック分) の推移を，両手法の間で比較した．具体的には，各順位までで，両手法によって共通に出力されたブログサイト数，片方の手法によってのみ出力されたブログサイト数 (提案手法およびベースラインを区別して 1 プロットずつ) の比較を行った．この結果から明らかなように，提案手法のみによって出力されたブログサイト数は，ベースラインのみによって出力されたブログサイト数よりもはるかに多い．これにより，提案手法は，「トピックについて詳細な記述を含むブログサイト」のうち，既存の検索エンジン API において下位に順位付けされたサイトを上位に押し上げていることが分かる．

6 関連研究

TREC の 2007 年度の Blog Distillation タスク [Macdonald07] においては，ある特定のトピックについて検索したときに，そのトピックについて詳しく書かれていて，繰り返し見たいと思うブログサイトを検索するというタスクを行っている．このタスクにおいて上位の成績を収めた [Elsas07] においては，本稿の手法と同様，Wikipedia エントリ中の他エントリへのリンクを用いた検索質問拡張が採用されている．一方，本稿では，Wikipedia 中のリダイレクトおよび太字といった，エントリとの関連性がより高い知識も合わせて用いてブログサイトの検索を行っている．また，ベースラインとして，既存の検索エンジン API によるブログサイトの順位付けからの改善を実証している．さらに，本稿では，日英二言語において提案手法の有効性を実証した．

その他には，ブロガーの熟知度に基づき，ブログサイトをランキングする手法 [中島 08] などがある．この手法では，マニアの多そうなキーワードを集めたマニア辞書をあらかじめ作成しておき，このマニア辞書に基づいてブログサイトの順位付けを行う．本稿の手法が，Wikipedia を知識源として利用するのに対して，この手法では，あらかじめマニア辞書を作成する必要がある点が大きく異なっている．

また，筆者らによる [川場 09b] では，Wikipedia エントリによって指定されたトピックとブログサイトとの間の記述内容の対応を判定するタスクにおいて，機械学習によって対応の有無の判定を行っている．この研究では，記述内容の対応の有無を二値で判定することに焦点が当られているのに対して，本稿の手法では，Wikipedia エントリとブログサイトの間の記述内容の対応の度合いを測定することに焦点が当てられている．実際に，筆者らが，i) 記述内容の対応の有無の二値判定，ii) 記述内容の対応の度合いの測定，の二種類のタスクにおいて両者の手法の性能を比較したところ，タスク i) では [川場 09b] の機械学習を用いた手法がやや上回り，タスク ii) では本稿の手法がやや上回る，という結果であった．

7 まとめ

本稿では，個々のブログ記事ではなく，ある同一のトピックについてまとまった規模の記述が書かれたブログサイトに注目し，そのような専門的内容を含むブログサイトを選択的に検索する手法を提案した．本稿の手法では，特定のトピックについての詳細な情報を含むブログサイトを選択的に検索することを実現するために，各トピックについての Wikipedia エントリ中の記述を知識源として利用した．実際に，本稿の手法を既存の検索エンジン API と比較し，ブログサイトの検索性能において本稿の手法が優れていることを示した．

参考文献

[Elsas07] Elsas, J., Arguello, J., Callan, J. and Carbonell, J.: Retrieval and Feedback Models for Blog Distillation, *Proc. TREC-2007 (Notebook)*, pp. 170–175 (2007).

[川場 09a] 川場真理子，中崎寛之，宇津呂武仁，福原知宏：Wikipedia 概念体系を用いた日本語ブログ空間のトピック分布推定, 人工知能学会研究会資料, SIG-SWO (2009).

[川場 09b] 川場真理子, 中崎寛之, 横本大輔, 宇津呂武仁, 福原知宏：Wikipedia 概念体系とブログ空間の間のトピック対応の推定, 日本データベース学会論文誌, Vol. 8, No. 1, pp. 17–22 (2009).

[Macdonald07] Macdonald, C., Ounis, I. and Soboroff, I.: Overview of the TREC-2007 Blog Track, *Proc. TREC-2007 (Notebook)*, pp. 31–43 (2007).

[中島 08] 中島伸介, 稲垣陽一, 草野奉章：ブロガーの熟知度に基づいたブログランキング方式の提案, 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).