Wikipedia エントリに対応するトピックのブログサイト検索*

川場 真理子† 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†，東京大学 人工物工学研究センター‡

1 はじめに

近年，ブログの爆発的普及により，多くの人が個人の関心や評判などをウェブ上で発信するようになった．それに伴い，多くの情報がブログを通じてウェブ上から取得できるようになった．ブログからの情報収集の方法としては，既に多くのサービスがあり，様々な研究もなされている．特定のキーワードに対する評判情報や時系列分布をブログから取得するサービスには Kizasi.jp[1]などがあり，また，キーワードでブログを検索するサービスには Yahoo!ブログ検索[2] や Google ブログ検索[3]がある．これらの検索サービスは，巨大なブログ空間に対する索引付けという観点から見ると，キーワードや評判，時系列変化などによる索引付けを行い，それらの索引を用いて利用者の検索要求を満たすブログ記事やブログサイトを検索する，と位置付けることができる．また，テクノラティ[4]のようなカテゴリ式のブログ検索サービスもよく知られている．この場合，ブログ空間に対する索引付けという観点から見ると，主として人手により付与されたカテゴリ情報が，ブログ空間に対する索引であると位置付けることができる．

ここで，これらの既存のブログ検索サービスは，ブログ空間に対する索引付けの粒度と体系化の二点において不十分であると言える．まず，カテゴリ式のブログ検索サービスにおいては，人手により設定されたカテゴリの体系が十分な網羅性を持つとは言えず，また，実際の検索要求に比べて，カテゴリの粒度が粗すぎる傾向がある．一方，キーワードや評判，時系列変化などによるブログ検索サービスの場合は，個々の索引の粒度が細かく，また，それらの索引全体を体系化してとらえることが困難である．したがって，利用者が，検索要求に対して適切な索引を想起することができなければ，巨大なブログ空間に対して容易にはアクセスできない．

このような現状をふまえて，本研究では，巨大なブログ空間へのアクセスを実現するにあたって，より適切な粒度で，しかも，十分に体系化された索引付けの一つの方式として，

あらゆる事柄が詳細に体系化された知識体系である Wikipedia とブログサイトを対応付ける

アプローチを提案する．Wikipedia は誰でも自由に情報を書き込むことのできる巨大なウェブ百科事典として知られており，あらゆる分野に関する詳細な情報が書き込まれている．Wikipedia を使用してブログ空間に索引付けを行うということが達成されると，検索要求に対し，的確なブログサイトを得ることができる．また，キーワードに対するブログの有無などを知ることによって，現存するブログ空間における話題の分布の傾向を把握することが容易に実現できる．さらに，検索対象のブログの単位を，特定のトピックに対するブログの記事ではなく，ブログサイトとすることによって，ブログ空間において，個々の記事よりもより大きい，ブログ著者の単位での索引を付けるアプローチをとる．

2 商用ブログ検索サービスの現状

2.1 キーワード入力式ブログ検索サービス

ユーザが自由に選んだキーワードを入力して検索する検索サービスには代表的なものとして Google ブログ検索や Yahoo!ブログ検索などがあげられる．これらの検索サービスでは，ユーザが好きなキーワードを自由に選んで検索することができるという利点がある．しかし，これらの検索サービスは人気の高いブログ優先的に検索する．そのため，マニアックなために多くの人に知られていないが，面白い情報を載せているブログが上位に検索されにくくなっている．本稿の目的を達成するためには，現在の検索サービスでは不十分であると言える．

2.2 カテゴリ式ブログ検索サービス

あらかじめ人手で用意したカテゴリを使用してブログの検索を行うサービスには代表的なものとしてテクノラティなどがあげられる．このような検索エンジンは，検索したいトピックがカテゴリに無い場合に，検索したいトピックと近いトピックで検索しなければならない．より詳細なカテゴリが必要だといえる．

*Blog Distillation from Wikipedia Entries

†Mariko Kawaba, 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://kizasi.jp/
[2] http://blog-search.yahoo.co.jp/
[3] http://blogsearch.google.co.jp/
[4] http://www.technorati.jp/

図 1: Wikipedia の構造

3 Wikipedia

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 45 万，英語で約 220 万のエントリ (2008 年 1 月現在) がある．大きな特徴として，Wiki を利用して作られており，だれでも自由に情報を書き込むことができる．さらに，11 のメインカテゴリ以下にサブカテゴリ，エントリが連なる，巨大な木構造になっている．また，カテゴリが木構造のノードにあたり，エントリが木構造の葉に相当する．図 1 に示すように，日本の電気通信事業者カテゴリというノードの下にさらにサブカテゴリがノードとしてつながっており，さらにそのカテゴリの下の NTT グループサブカテゴリの下には日本電信電話エントリが葉となってつながっている．

また，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．本稿の実験に用いた日本語キーワードに対応する英語キーワードは Wikipedia の言語間リンクの情報を使用した．

4 Wikipedia エントリに対応するブログサイトの検索

4.1 TREC 2007 Blog Distillation タスク

TREC-2007 のブログ検索のタスクの一つである Blog Distillation タスク [Macdonald07] は

ある特定のトピック X について検索したときに，そのト ピック X 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

というものである．特定のトピック X を与えると，システムは X について長期的に詳しく書かれていて，そのトピック X について興味のある人に RSS リーダなどに登録して定期的に読むことを勧めることができるようなブログサイトを返す．TREC の検索トピックは番号，タイトル，説明，補足で構成されているが，[Macdonald07] の報告によると，大半の参加者がタイトルのみを索引語として使用することで，各参加者の最高の性能を達成している．そこでこの結果を元に，本稿では，Wikipedia のエントリのタイトルのみでの検索実験を行った．

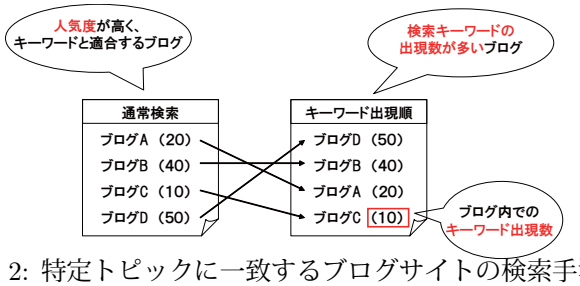

図 2: 特定トピックに一致するブログサイトの検索手法

4.2 本研究の枠組み

本研究の目的は，**Wikipedia の中のある特定のトピックから，そのトピックについての意見や評判などの情報が書かれているブログサイトを探し，対応づける**ということである．

そこで，本稿ではそのトピックについてどれだけ多く述べられているかで，検索トピックについて述べられているブログサイトかどうかを判断した．つまり，**検索トピックの出現数が多いブログサイト**を検索するというアプローチをとる．具体的には図 2 に示すように，

通常の検索方法でブログサイトを検索し，検索されたブログ集合を検索トピックの出現数が多い順にソートする．

また，本研究の発展として，外国語 Wikipedia エントリを利用した多言語間でのブログサイトの対照分析があげられる [中崎 08]．そのため，本稿の実験では日本語のブログサイトと英語のブログサイトの検索を行った．

4.3 評価手順

ブログサイトを検索するために，本実験では日本語ブログサイトの検索には，Yahoo!Japan 検索 API[5]を，英語ブログサイトの検索には米 Yahoo!検索 API[6]を利用し，日本語ブログサイトでは大手 11 社[7]，英語ブログサイト検索では大手 12 社[8]のドメインに限って検索を行った．

[5] http://developer.yahoo.co.jp/search/
[6] http://developer.yahoo.com/search/
[7] FC2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesa.net, jugem.jp, yaplog.jp, webry.info.jp, hatena.ne.jp
[8] blogspot.com, msnblogs.net, spaces.live.com, livejournal.com, vox.com, multiply.com, typepad.com, aol.com, blogsome.com, wordpress.com, blog-king.net, blogster.com

英語ブログ　　日本語ブログ

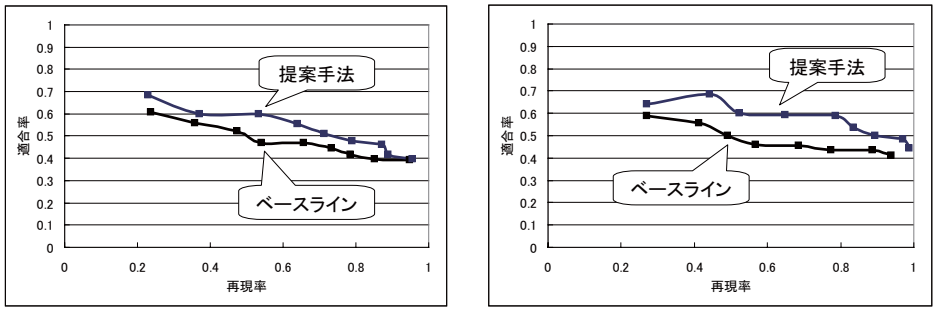

図 3: 特定トピックのブログサイト検索の評価結果 (4 キーワード分)

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[9]．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，一キーワードあたり，約 200 のブログサイトを取得することができた．また，提案手法ではこれらのブログサイトをキーワードの出現数順に並び替えるが，並び替える前の，API の出力順にブログサイトをランキングしたものをベースラインとした．ここで，日英の Yahoo!検索 API でブログサイトをドメイン指定してキーワードを検索した際に求められる検索結果の数を検索キーワードの出現回数とした．

また，本稿の実験に使用した検索キーワードは Wikipedia のエントリのタイトルから幅広い分野の，日本に関したトピックを選定した (表 1)．用意したキーワードの内，ドラゴンボール，Wii，新世紀エヴァンゲリオン，靖国神社，の 4 キーワードを選び，それぞれ，上位 30 位と以下等間隔に 30 ブログサイトをサンプリングし，手動で評価した．また，手動評価の際，特定トピックについてある一定数以上のブログ記事があれば正解とし，一定期間特定トピックについて書かれているということは考慮していない．

4.4 評価結果

ベースラインおよび提案手法について，再現率，適合率の推移を図 3 に示す．図 3 では 4 キーワード分の評価結果を 1 本のプロットによってまとめて示しているが，この結果では提案手法がベースラインを上回っている．

次に，以下では，提案手法を改善する余地について考察する．最も多くみられた問題点としては，システムの出力するキーワードの出現数が実際のブログサイト内でのキーワードの出現数よりも多くなってしまうということである．キーワードの出現数には Yahoo!API の検索結果の数を用いているが，Yahoo!API でブログサイトを検索した際の結果に同一記事がいくつも現れてしまい，検索誤りになるということが多く起こった．また，ブログサイトの検索は Yahoo!Web 検索 API で行っているが，Yahoo!Web 検索 API はブログサイトの本文と，プロフィールなどのサイドカラムに表示される情報，コメント，アフィリエイトなどの区別をせずに検索を行う．そのため，アフィリエイトなどのノイズが多く混入するということがあげられる．今後，これら問題の解決の為に，ブログサイトの本文，コメント，アフィリエイト等のサイドカラムに記載されている情報を区別して検索を行う必要があると考えられる．

表 1: 検索に使用したキーワード

分野	検索キーワード
アニメ	ドラえもん，ポケモン，鉄腕アトム，機動戦士ガンダム，ドラゴンボール，新世紀エヴァンゲリオン，セーラームーン
音楽	パフィー，X Japan，Dir en grey，ジャズ，交響曲
動物	犬，猫，ハムスター，ジャイアントパンダ，チワワ
企業	ソニー，カシオ，任天堂，ホンダ，トヨタ，三洋電機，キャノン
商品	PS3，PSP，iPod，Wii，ニンテンドー DS
歴史・文化	自衛隊，原爆，寿司，自民党，天皇，富士山
社会問題	靖国神社，年金，NOVA，捕鯨，テロ
施設	博物館，水族館，ミュージカル，遊園地，ディズニーランド
スポーツ	ボクシング，亀田興毅，亀田大毅，プロ野球，中村紀洋，イチロー，福留孝介，松坂大輔，井川慶，相撲，プロレス，K-1

また，より多くのブログサイトを取得するために，一つの検索キーワードだけでなく同義語や関連語を含めた検索が必要と考えている．これについては，4.5 節で詳しく述べる．

4.5 Wikipedia を用いた検索質問拡張

TREC-2007 のブログ検索タスクの一つ Blog Distillation タスク [Macdonald07] では，Wikipedia のハイパーリンクを用いた手法 [Elsas07] が最高の性能を達成している．このことをふまえ，我々はタイトルのみでのブログサイトの検索では不十分と考え，Wikipedia の 1 つのエントリを用いて検索質問を拡張する実験を行っている．この実験は Wikipedia のエントリの本文中にある強調文字とハイパーリンク，さらに，エントリタイトルと同名

[9]本稿の評価実験ではベースラインとの比較を行う都合上，用意したキーワードの中から 4 つ選び，複数ドメインを一度に指定して検索を行っているが，現在，用意した 60 キーワードを使用して，各ドメインごとに 1000 記事を取得する実験を行っている．

表 2: 検索質問拡張語候補

日本語トピック名 (英語トピック名)	検索質問拡張語候補 (日本語ブログ)	検索質問拡張語候補 (英語ブログ)
ドラゴンボール (Dragon Ball)	ドラゴンボール Z，ピッコロ，ベジータ，孫悟空，フリーザ，鳥山明，超ドラゴンボール Z，サイヤ人，ドラゴンボール GT	Dragon Ball Z, Dragon Ball GT, anime, Bulma, Dragon Ball AF, Codename, Super Saiyan, Captain Ginyu, Buu Saga, China
Wii (Wii)	Wii Fit，Wii Sports，おどるメイドインワリオ，カドゥケウス Z2 つの超執刀，SD ガンダムスガッドハンマーズ，ドンキーコングたるジェットレース，ドラゴンクエストソード，マリオストライカーズ，ゼルダの伝説トワイライトプリンセス，スイングゴルフ	Asia, Audio, Video, Wii Remote, Mii, Virtual Console, Americas, Animal Crossing, Atari2006, Wii Points
新世紀エヴァンゲリオン (Neon Genesis Evangelion)	エヴァンゲリオン，綾波レイ，スレイヤーズ，使徒，機動戦士ガンダム，セカンドインパクト，パチスロ，惣流・アスカ・ラングレー，残酷な天使のテーゼ，碇シンジ	mecha, manga, Evangelion, Rebuild or Evangelion, Angel, Yoshiyuki Sadamoto, live-action movie, Death and Rebirth, 1.0 You Are, Fly Me to the Moon
靖国神社 (Yasukuni Shrine)	A 級戦犯，合祀，神社，英霊，終戦記念日，遊就館，8 月 15 日，戦死，昭和天皇	Government of Japan, Democratic Party of Japan, Emperor Akihito, East Asia, Crime against peace, Class A War Criminal, Emperor Hirohito, First Sino-Japanese War, Communist Party of China, Boshin War

のカテゴリがある場合はその子となるエントリのタイトルを検索質問の拡張語候補として抜き出した．さらに，Wikipedia でリダイレクト設定されている同義語も拡張語の候補とする．最後に，Wikipedia のエントリから抜き出した，検索トピック X の拡張語の候補となる Y の関連度を求め，順位付けを行った．関連度としては以下の尺度 [佐々木 06] を用いた．

$$\text{関連度}\,(X,Y) = \frac{X \text{ AND } Y \text{ の検索ヒット数}}{X \text{ OR } Y \text{ の検索ヒット数}}$$

この手法により，多いもので 300 件，また，少ないものでも 40 前後の拡張語を取得することができた．本手法で求められた検索質問拡張語上位 10 件の例を順に表 2 に示す．

5 関連研究

ブログの分類に関する研究として，ブログ記事のカテゴリの中から，一般性があり，説明力のあるカテゴリを選びタグとして，マルチタグを付与する研究 [大倉 06] がある．また，ブログ記事をドメインで分類する研究 [橋本 07] などがある．これらの研究におけるドメインやタグは，本研究で用いている Wikipedia エントリよりも粗いものである．本研究では，Wikipedia エントリ程度のより詳細な粒度のトピックを用いて，ブログサイトを検索するという新たなタスクを導入している．

その他には，ニュースサイトとブログを関連づけることで，ブロガーの嗜好にあったニュースサイトを推薦し，ブロガーの嗜好を利用してブログをフィルタリングする研究 [小原 05] がある．

6 まとめと今後の課題

本稿では Wikipedia とブログ集合の対応付けのために，トピックの出現回数の多いブログを検索することでブログ集合を検索する検索実験行った結果を報告した．実験の結果から現在の検索手法の改善すべき点を述べた．

しかし，本稿で行った検索ではまだ，ノイズも多く混入してしまい，Wikipedia のトピックに対応するブログサイトを十分に収集できているとは言いがたい．また，Wikipedia エントリのタイトルのみを使用して検索することだけでは不十分であると考えられる．より精度良く検索を行うために，今後 TREC の Blog Distillation タスク [Macdonald07] の成果なども取り入れて，ノイズ除去，検索質問拡張などを行っていく必要がある．

参考文献

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

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

[中崎 08] 中崎寛之, 川場真理子, 宇津呂武仁, 福原知宏：同一トピックの日英ブログサイト検索による二言語対照ブログ分析, 言語処理学会第 14 回年次大会論文集 (2008).

[佐々木 06] 佐々木靖弘, 佐藤理史, 宇津呂武仁：関連用語収集問題とその解法, 自然言語処理, Vol. 13, No. 3, pp. 151–175 (2006).

[橋本 07] 橋本力, 黒橋禎夫：基本ドメイン情報の構築, 言語処理学会大 13 回年次大会発表論文集, pp. 1105–1108 (2007).

[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).

[大倉 06] 大倉務, 清田陽司, 中川裕志：Folksonomy の機械化：Blog 記事へのマルチタグ付与, 言語処理学会大 12 回年次大会発表論文集, pp. 360–363 (2006).