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

川場真理子† 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 〒305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学第三学群工学システム学類 〒305-8573 茨城県つくば市天王台 1-1-1
††† 東京大学 人工物工学研究センター 〒277-8568 千葉県柏市柏の葉 5-1-5

あらまし 本研究ではある特定のトピックについて検索をしたときに，そのトピックについて有用な情報が書かれているブログサイトを探すことを目的とする．手法として，特定トピック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，特定トピックを表すキーワードの出現数順にリランキングする方法を用いる．この方法によるランク付けと，商用ブログ検索エンジンによって得られるブログサイトランク付けとの比較結果について報告する．また，この手法を用いて，Wikipediaエントリに対応したトピックのブログサイトを検索するタスクについての検討の現状を報告する．また，一つの応用例として，Wikipediaの日本語エントリに対応する英語エントリでブログサイトを検索し，二言語間でブログサイトの記述内容の対照分析を行う．

キーワード ブログ，トピック分析，Wikipedia，blog distillation，言語横断情報検索

Blog Distillation for Linking Wikipedia Entries to Blog Feeds

Mariko KAWABA†, Hiroyuki NAKASAKI††, Takehito UTSURO†, and Tomohiro FUKUHARA†††

† Grad. Sch. Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
†† College of Eng. Sys., Third Cluster of Colleges, University of Tsukuba, Tsukuba, 305-8573, Japan
††† Research into Artifacts, Center for Engineering, University of Tokyo Kashiwa, Chiba 277-8568, Japan

Abstract This paper proposes an approach to blog distillation, i.e., searching for blog feeds that are principally devoted to a given topic. We study this task for the purpose of regarding each of Wikipedia entries as a topic and linking it blog feeds. First, in order to collect candidates of blog feeds for a given query, in this paper, we use existing Web search engine APIs, which return a ranked list of blog posts, given a topic keyword. Next, we re-rank the list of blog feeds according to the number of hits of the topic keyword in each blog feed. We also apply the proposed blog distillation framework to the task of cross-lingually analyze multilingual blogs collected with a topic keyword. Here, we cross-lingually and cross-culturally compare less well known facts and opinions that are closely related to a given topic. Preliminary evaluation results support the effectiveness of the proposed framework.

Key words blog, topic analysis, Wikipedia, blog distillation, cross-lingual IR

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．

以下に本稿の構成を示す．2. はブログの概要と既存ブログ検索のサービスの現状ついて述べる．また，3. では Wikipedia の構造ついて述べる．さらに，4. は Wikipedia のエントリ名でブログサイトを検索する実験について述べ，5. は本実験の応用として，Wikipedia の日本語エントリに対応する英語エントリでブログサイトを検索し，二言語間でブログサイトの記述内容の対照分析を行う．6. で関連研究について述べ，最後にまとめを行う．

2. 商用ブログ検索サービスの現状

2.1 ブログの概要

ブログとはウェブ上に個人が公開する日記の一種である．1 つのブログはいくつもの記事の集合から成り，1 記事ごとにトラックバックというリンクやユーザー同士のコメントをつけることができる．また，多くの人がその時の関心事を書き記すために，その期間の社会的な関心の動向が得られるという利点がある．

2.2 商用ブログ検索サービス

ブログを探すための検索サービスは既に多く存在する．ブログの検索サービスには大きく分けて，カテゴリ式ブログ検索サービスと，キーワード式ブログ検索サービスは 2 つがある．以下の節ではこれらの検索サービスとその問題点について述べる．

2.2.1 キーワード入力式ブログ検索サービス

キーワード入力式ブログ検索サービスとは，通常の検索サービスのようにユーザが自由にキーワードを入力してブログを探す事の出来る検索サービスの事である．これらの検索サービスの代表的なものとしては Google ブログ検索 [3] や Yahoo!ブログ検索 [2] 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．Wikipedia のエントリに対応させたいブログサイトはたまたまそのトピックについての記事がある人気の高いブログサイトではなく，トピックについて詳細に書かれた情報が多く記載されているブログである．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．

2.2.2 カテゴリ式ブログ検索サービス

カテゴリ式ブログ検索サービスとは，ブログが検索サービスを提供している会社によって用意されたカテゴリに分類されていてユーザが好きなカテゴリを選んでブログを探す形式のものを指す．これらの検索エンジンの代表的なものとしてテクノラティ [4] 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．

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 のブログ検索タスクの一つ Blog Distillation タスク [5] は，

図 1　Wikipedia の構造

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，[5] 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，各々の参加者の最高の性能を達成している．このことをふまえて，本稿では，Wikipedia のエントリのタイトルのみを用いて検索質問を作成し，検索実験を行った．

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．しかし，現在のブログ検索サービスでは，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ために，被リンク数は多くないが，特定トピックについて濃い情報を載せているブログサイトが検索されにくい．本研究の目的を達成するためには，トピックについて濃い情報を載せているブログサイトの集合を得る必要がある．よって，被リンク数の多い，人気度の高いブログサイトを優先的に検索するのではなく，検索トピックについて多く述べられているブログサイトを優先的に検索する必要がある．そこで，本稿では，検索トピックがブログサイトにどれだけ出現し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定するという手法を用いる．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た後，検索されたブログサイト集合を検索トピックの出現数が多い順にソートする．

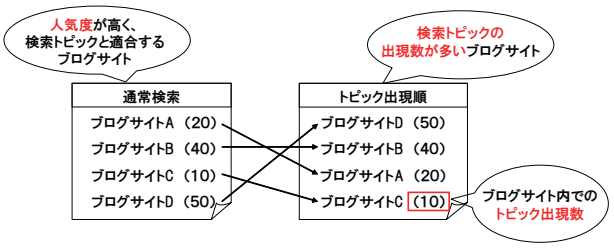

図 2　特定トピックに一致するブログの検索手法

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の検索には，Yahoo!Japan 検索 API を，英語ブログの検索には米 Yahoo!検索 API を利用し，日本語ブログでは大手 11 社，英語ブログでは大手 12 社のブログ会社のドメインに限って検索を行った．

日本語ブログ会社

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

英語ブログ会社

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

検索の際には，複数のドメインを一度に指定して検索し，1,000 件の記事を取得する (注1)．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，1 キーワードあたり約 200 前後のブログサイトを取得することができた．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べ替えるが，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，日米の 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．

検索キーワード

本研究では，あるトピックに対する日英のブログサイトの記述内容を，二言語間で対照分析するというタスクに対して，本研究で提案するブログサイト検索手法を適用する．そこで，評価実験に使用した検索キーワードとしては，Wikipedia のエントリのタイトルを対象として，日本に関する幅広い分野のトピックで，かつ，日本語・英語共にある程度の数のブログサイト集合が得られるようなトピックを約 60 選定した．表 1 に選定したキーワードを示す．これらの約 60 キーワードの内，「ド

(注1)：本稿の評価実験では，ベースラインとの比較において，ベースラインにおける順位付けを利用する必要があるという制約から，各ドメインごとに検索を行うのではなく，複数ドメインを一度に指定して検索を行っている．現在，この検索とは別に，各ドメインごとに 1,000 記事ずつ検索して，これらの和集合に対して，提案手法の順序並べ替えを行うという方法で評価実験を行っている．

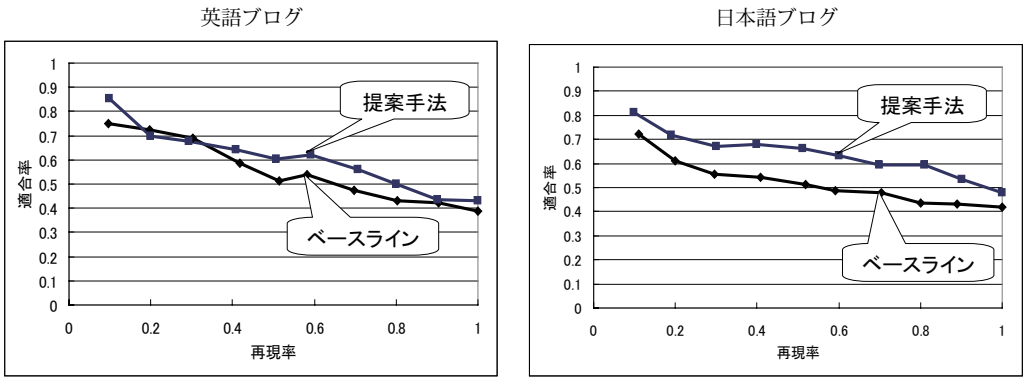

図 3　特定トピックのブログサイト検索の評価結果 (4 キーワード分)

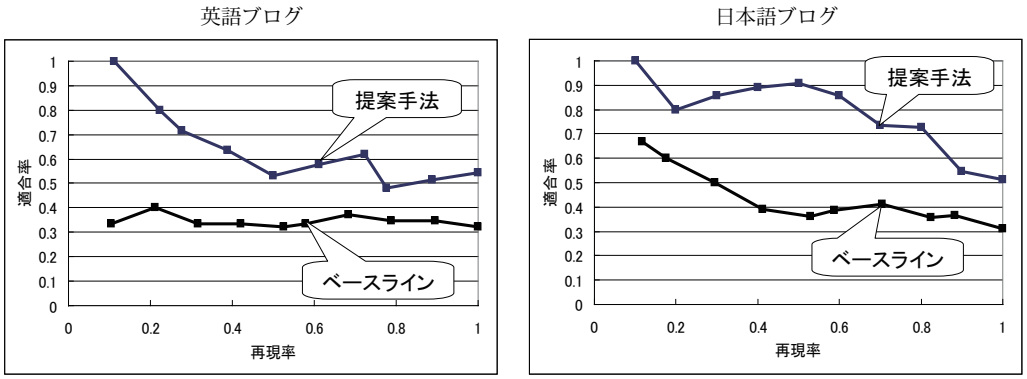

図 4　特定トピックのブログサイト検索の評価結果 (トピック：靖国神社 (Yasukuni Shrine))

ラゴンボール」,「Wii」,「新世紀エヴァンゲリオン」,「靖国神社」,「捕鯨」(注2)の 5 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 3，4 に示す

図 3 では，4 キーワード分の評価結果を一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．英語ブログにおいては，ベースラインが提案手法を上回っている部分もあるが，この問題は，次節で説明する検索質問拡張を行うことで解消できると考えている．また，Yahoo!検索 API では，被リンク数やキーワードの出現数など，高性能なランキング尺度によってブログ記事の順位付けがされているのに対して，提案手法は，ブログサイト内でのキーワード出現数のみを用いてブログサイトを順位付けしている．したがって，このような簡便な方式でも，ブログサイト検索というタスクにおいて，商用検索エンジン API の性能を上回る場合があるという結果は，特筆すべきであると考えている．

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．提案手法がベースラインに負けている例として Wii やドラゴンボールなどの商品的なトピックがあげられる．

これらの原因としてはいくつかあげられるが，中でももっと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 ではブログ記事を検索した際の結果に同一記事がいくつか現れ

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，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	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	原爆，寿司，自民党，富士山
社会問題	靖国神社，年金，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，イチロー，松坂大輔，相撲，プロレス，K-1

(注2)：捕鯨に関しては，提案手法によって収集できたブログ数が 30 ブログ前後であり，また捕鯨について述べられていたブログ数が 30 ブログ中 3 ブログしか現れなかったために，評価結果のグラフは省略した．しかし，各ドメインあたり 100 ブログ記事を検索した場合，約 400 のブログサイトを取得することができ，また，上位に多くの捕鯨に関するブログを確認することができた．

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)	(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

てしまう．そのため，特定トピックをキーワードとして検索した際に，同一記事が何度も検索されてしまい，結果，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 はブログ記事の本文と，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．特に，ドラゴンボール，新世紀エヴァンゲリオン，Wii などの関連製品が発売されているものについては，本文中に検索キーワードが書かれていないにも関わらず，アフィリエイトにキーワードが含まれているために，検索キーワードに一致しないブログサイトを検索してしまうという誤りが見られた．

また，そのトピックについて詳しく書いているブログサイトであるにも関わらず，ブログサイトの記事数自体が少ないために，キーワードの出現数が低くなってしまい，他のブログサイトよりも下にランキングされてしまうという例も見られた．

今後，これら問題の解決の為に，ブログ記事の本文，コメント，プロフィール，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．

さらに，表記揺れの問題により，検索上位にランキングされなかったものや，そのトピックの関連商品について多く述べられているが，そのトピック名については述べられていないものなどもいくつか見られた．この問題の解決のためには，一つの検索キーワードだけではく，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 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 [5] では，Wikipedia のハイパーリンクを用いた手法 [6] 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．この実験はブログ検索タスクで行われた索引語拡張 [6] に加えて，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．これらの情報を Wikipedia から抜き出したあと，検索する上でノイズになりそうな，記号，半角 1 文字を取り除き，さらに括弧で囲まれている文字の場合は括弧を取り除くといった前処理を行った．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 [7] 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．

ドラゴンボール，新世紀エヴァンゲリオンなどは，作品名やキャラクター名などが上位に現れた．また，Wii については，日本語の場合 Wii のゲームソフトが大半を占めるが，英語の場合，Wii のサービスなどについてのキーワードが上位に現れた．さらに，靖国神社については，靖国神社の説明や靖国神社の参拝に関するようなキーワードが多く出現した．

まだ，これらの拡張語を使用した検索は行っていないが，今後これらの拡張語を利用した検索実験を行う予定である．

5.　日英ブログの言語対照分析

本稿では，収集した日英のブログサイトをそれぞれ評価実験と同じ方法で 60 個サンプリングし，記事中の内容を分析し比較した [8], [9]．以下にその手順と結果を述べる．

表 3　日英ブログ記事から抽出した共起語比較

日本語トピック名 (英語トピック名)	日本語名詞句 (日ブログ中頻度)/英訳 (英ブログ中頻度) [日英出現確率比]	英語単語・二単語連語 (英ブログ中頻度)/日本語訳 (日ブログ中頻度) [英日出現確率比]
ドラゴンボール (Dragon Ball)	データカードダス (143)/carddass(0)[∞], 食玩 (58)/英訳無し [∞], 最新グッズ (360)/latest goods(0)[∞], クリリン (106)/Kuririn(4)[5.35]	Japanese anime(67)/日本のアニメ (0)[∞], Japanese manga(99)/日本の漫画 (0)[∞]
Wii (Wii)	ショッピングチャンネル (97)/shopping channel(4)[36.3], ゲームソフト (136)/game software(6)[33.9]	hack(66)/改造 (0)[∞], cheats(45)/チート (0)[∞]
新世紀エヴァンゲリオン (Neon Genesis Evangelion)	パチンコ画像 (87)/pachinko image(0)[∞], 確変 (7)/英訳無し [∞], スロット (11)/slot(3)[6.3], 声優 (37)/seiyuu(18)[3.5]	giant mecha(7)/巨大メカ (0)[∞], anime series(25)/アニメシリーズ (0)[∞], Japanese manga(36)/日本の漫画 (0)[∞], Yoshiyuki Sadamoto(5)/貞本義行 (0)[∞], Hideaki Anno(16)/庵野秀明 (1)[9.3]
靖国神社 (Yasukuni Shrine)	靖国問題 (81)/英訳無し [∞], 合祀 (124)/英訳無し [∞], 歴史認識 (50)/historical recognition(1)[16.4], 売国奴 (8)/treason(1)[2.6], 日本国憲法 (20)/Japanese Constitution(3)[2.2], 反日 (91)/anti-Japanese(47)[0.6]	war shrine(64)/戦争神社 (1)[195.3], Japanese militarism(24)/日本軍国主義 (2)[36.6], Japan-China relations(12)/日中関係 (1)[36.6], patriotic education(6)/愛国心教育 (3)[6.1]
捕鯨 (Whaling)	責任転嫁 (90)/buck-passing(0)[∞], 極右 (5)/extreme right-wing(0)[∞], 文化帝国主義 (213)/cultural imperialism(8)[25.5], ナショナリズム (22)/Nationalism(8)[4.6]	Animal rights(73)/動物の権利 (0)[∞], Animal welfare(26)/動物保護 (0)[∞], Japanese whaling(130)/日本の捕鯨 (0)[∞], endangered species(78)/絶滅危惧種 (0)[∞], humpback whale(82)/ザトウクジラ (2)[24.4], Greenpeace(263)/グリーンピース (48)[3.3]

5.1　日英ブログ言語対照分析のための統計的尺度

本研究では，対照分析の方法として，各言語ブログに出現する共起語を用いる．これを用いる理由は，片方のブログでは高い確率で出現し，その対訳が相手言語ブログではあまり出現しない共起語をみることで，内容分析を容易にすることができると考えているからである．そこで，抽出した共起語に対して片方の言語ブログにおける出現確率とその相手言語ブログにおける対訳の出現確率の比を尺度として適応する．以下でその尺度について述べる．

まず，日本語ブログ記事から名詞句を抽出し，英語ブログ記事からは一単語または二単語連語を抽出し，それぞれの頻度統計と出現確率を求める．また，日本語名詞句 X_J の日本語ブログにおける出現確率 $P_J(X_J)$ と，英語単語・二単語連語 Y_E の英語ブログにおける出現確率 $P_E(Y_E)$ を以下のようにそれぞれ定義する．

$$P_J(X_J) = \frac{X_Jの出現頻度}{対象日本語ブログサイト集合内の総形態素数}$$

$$P_E(Y_E) = \frac{Y_Eの出現頻度}{対象英語ブログサイト集合内の総単語数}$$

また，抽出した語句の訳語が相手言語ブログに出現するか調べるために，Wikipedia の言語間リンクを使用して語句の訳語を求める．Wikipedia で語句の対訳を取得できない場合は，英辞郎 (注3) で語句の対訳を取得する．最後に，抽出した語句の出現率と対訳語句の出現率から，相手言語ブログと比較した出現確率比を求める．本研究では，抽出した日本語名詞句 X_J と X_J の英訳 X_E の出現確率比 $R_J(X_J, X_E)$ と，英語単語・二単語連語 Y_E と Y_E の和訳 Y_J の出現確率比 $R_E(Y_E, Y_J)$ を以下のように定義した．

$$R_J(X_J, X_E) = \frac{P_J(X_J)}{P_E(X_E)},\quad R_E(Y_E, Y_J) = \frac{P_E(Y_E)}{P_J(Y_J)}$$

5.2　同一トピックの日英ブログから抽出した共起語比較

本節では，日本語ブログと英語ブログから抽出した共起語に対して，5.1 で述べた出現確率比を適用して，日本語ブログから抽出した共起語を日本語ブログからみた出現確率比の多い順，英語ブログから抽出した共起語を英語ブログからみた出現確率比の多い順にそれぞれ並び替えた．その結果の一部を表 3 に示す(注4)．商品となるトピックについて書かれた日本語ブログサイトでは，トピックに関連した商品を連想させる名詞句がいくつかみられた．それに対して，英語ブログサイトでは，トピックのジャンルや製作者の名前など，トピックの漫画やアニメの紹介記事を連想させるような語が観測された．Wii についての日英ブログサイトに関しては，日英共にゲーム紹介の記事を連想させる語が観測されたが，英語ブログサイトに特有の語として，Wii の改造を意味する”hack” が観測され，このような意味の語は日本語ブログサイトでは観測されなかったために”hack” の出現確率比が大きくなった．また，社会問題に関するトピックについて書かれた日英ブログサイトでは，互いに意見が対立していることを連想させる語が多くみられた．捕鯨を例として挙げると，英語ブログサイトでは，日本の捕鯨に強く反対していることから，“Japanese whaling” という二単語連語が大きな出現確率比で観測された．一方，日本語ブログサイトでは，極端な反捕鯨行為を指す「文化帝国主義」「極右」という名詞句が観測された．また，英語ブログサイトでは，鯨の具体的な種類の名称への言及が多数観測されたが，日本語ブログサイトでの観測はわずかであったため、それらの英語ブログサイトからみた出現確率比は大きくなった．このように，片方の言語ブ

(注3)：http://www.eijiro.jp/

(注4)：出現確率比∞は，語の対訳が相手言語ブログで観測されなかったことや，その語の対訳が見つからなかったことを示す．

表 4　日英ブログの記述内容の抜粋と言語対照分析

日本語トピック名 (英語トピック名)	簡単な説明 現象や意見の違い (日本語ブログ)	(英語ブログ)
ドラゴンボール (Dragon Ball)	日本の漫画作品．40ヶ国以上で翻訳されている．	
	多くのブログがドラゴンボールカードダスについて書かれていた．また，ゲームのレビューなどがいくつかあったが，著作権違反になるために，動画はほとんど見 られなかった．また，ドラゴンボールのキャラクターの形を模した弁当の画像を多く載せているブログもいくつかあり，同人誌などを作成しているブログも数件みられた．	ドラゴンボールカードダスについてはほとんど述べられていなかった．また，ゲームのレビューはあまり多くないが，ゲームやアニメの動画へのリンクなどが張ってあるブログが多く見られた．また，ドラゴンボールのファンであり，ドラゴンボールに関すること全般に渡って書かれているブログや，訪問者へドラゴンボールに関するアンケートを行っているブログもいくつか みられた．
Wii (Wii)	任天堂から発売された据え置き型ゲーム機．世界中で販売されている．	
	Wii の改造などは日本の法律違反になるために，Wii の改造について述べられているブログは全く見られなかった．ゲームタイトルの一覧があり，それぞれに詳細な感想，評価などが述べられているブログが多くみられた．また，2007 年 10 月に発売された Wii Fit を使用したダイエットの記録をつけたブログなども多く見られた．関連商品へのアフィリエイトも多く見られた．	いくつかのブログに Wii で自作のソフトウェアを動かす方法や，Wii の改造について述べていた．また，多くのブログにゲームの動画や Wii の写真などが載せられていて，ゲームについての感想などが書かれているブログも数個見られた．
新世紀エヴァンゲリオン (Neon Genesis Evangelion)	日本のアニメ．英語版がヨーロッパやラテンアメリカ，アジアなどで放送されている．	
	多くのブログでアニメや映画の感想が書かれていた．また，パチンコを趣味とする人のブログでエヴァンゲリオンのパチンコについての感想や攻略法を書いたブログが見られた．また，エヴァンゲリオンの関連商品にアフィリエイトを張った，スプログもいくつか見られた．著作権の問題からか，動画が載せられているブログはあまり見られなかった．	多くのブログにアニメの動画やアニメの画像を編集した動画が載せられていた．また，自作のデスクトップ画像の配布を行っているブログややフィギュアの画像を多く載せているブログなどが見られた．また，いくつかのブログにアニメの感想などが書かれていた．エヴァンゲリオンのパチンコをしたブログは全く見られず，また，スプログもほとんど現れなかった．
捕鯨 (Whaling)	捕鯨に，反対か，賛成かの意見が書かれたブログが見られた．	
	多くのブログが日本の行為に対し肯定的であり，捕鯨に賛成していて，反捕鯨団体を激しく非難しているブログなどもいくつか見られた．また，日本の新聞，雑誌，テレビなどのメディアが捕鯨に関するニュースをどのように伝えているかという分析を客観的に行っているブログも見られた．捕鯨に関する記事を書いているブロガーは国粋主義，右寄りのブロガーが多くみられた．	多くのブログが日本の捕鯨に反対していて，反捕鯨運動を呼び掛けているブログなども多く見られた．また，ホエールウォッチングをしているブログがいくつか見られた．
靖国神社 (Yasukuni Shrine)	東京にある神道の神社．国会議員や内閣総理大臣が参拝することが問題になっている．	
	全体的に靖国神社参拝に肯定的なブログが多くみられた．また，会合など開き，靖国参拝を呼び掛けているブログも多く見られた．また，靖国神社についてのブログでは国粋主義，右寄りの考えを持つブロガーが多く見られ，意見性の強い記事も多く見られた．	大半のブログが日本の国会議員の靖国神社参拝に否定的な意見を述べていた．また，日本人の過去の歴史に対する意識を批判しているブログが見られた．靖国神社参拝だけではなく，靖国神社そのものに関する批判も多く見られた．中には靖国神社の事を「Yasukuni War Shrine」と表記しているブログも見られた．

ログからみた出現確率比の高い共起語をブログ記事から抽出することで，各言語のブログサイトで何が書かれているのか，ある程度把握できることがわかった．

5.3　日英ブログサイトの記述内容比較

各言語の特徴的な語を踏まえた上で，日英ブログサイトの内容の対照分析を行い，その記述内容を抜粋した結果を表 4 に示す．この結果から，商品・作品等に関するキーワードと，社会問題のような意見性が出やすいキーワードでは，対照分析の結果に差が出るということがわかった．具体的にはドラゴンボール，Wii，新世紀エヴァンゲリオンのような商品となるキーワードで検索されたブログサイトでは，日本とアメリカの社会的な興味，需要の違いを反映した違いが多く観測された．ドラゴンボールを例に挙げると，日本ではアニメの放送も終了していることやウェブ上の動画に対する規制が厳しいこともあり，日本語ブログサイトにおいてはドラゴンボールに関連したゲームやカードダスの商品について記載している記事が多くみられた．しかし，英語ブログサイトでは，ウェブ上の映像に対する規制が日本ほど厳しくないこともあるためか，ドラゴンボールのアニメ動画を記事に載せているブログサイトがいくつかみられた．一方，捕鯨や靖国神社のような社会問題に関するキーワードで検索されたブログサイトの場合は，日本語ブログサイトと英語ブログサイトで全く逆の意見が見られた．日本語のブログサイトでは，国粋主義，右寄りのブロガーによって書かれた，日本の行為に対する肯定的意見の多いブログサイトが多く，英語のブログサイトでは，反日のブロガーによって書かれた，日本の行為に対する否定的な意見の多いブログサイトが多くみられた．この傾向は，捕鯨と靖国神社の両キーワードでそれぞれ強く見られた．

6.　関 連 研 究

本研究の関連研究は大きく分けてブログに関する研究と Wikipedia に関する研究がある．以下でそれぞれの関連研究について述べる．

6.1　ブログに関する研究

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 [10] がある．また，他にも，ブログをドメインで分類する研究 [11] 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．

また，ブログサイトの検索としては TREC2007 年 BlogDistillation タスクの他にも，ブログ著者が詳しい知識を持っている分野を推定し，その知識の深さに基づいた Web コンテンツのトラスト評価を行う研究 [12] などがある．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研究 [13] などがある．この研究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，その辞書のトピックからブログサイトを検索しているという点で本研究とは異なる．

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 [14] がある．また，多言語でのブログの研究には，日韓中英のブログ内で，キーワードのバーストの時系列の変化を各言語間で調べるという研究がされている [15]．本稿で行った日英ブログの比較対照実験では，ブログの内容を見ており，キーワードのバーストの時系列の変化を調べるというようなことは行っていない．

6.2 その他の関連研究

本節ではブログ以外の関連研究について述べる．Wikipedia に関する研究に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 [16] があり，この研究は，Wikipedia にある程度分類分けされた情報を対応付けている．その他に，Wikipedia から固有表現を取出す研究 [17], [18]，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などがなされている [19]．本稿の日英ブログ対照比較に用いた対訳は，この手法 [19] と同様，Wikipedia の言語間リンクを使用している．また，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研究 [20] もある．この研究では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かの理解を図ろうとしている．

7. まとめと今後の課題

本稿では Wikipedia とブログサイト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サイトを検索することでブログサイト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．また，検索によって集めたブログサイト集合を用いて日英ブログサイトの対照分析を行い，日英ブログサイトを同一キーワードで検索した際に，検索キーワードの持つ属性によって日本語と英語のブログサイトの内容に違いがみられることがわかった．

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 [5] 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．

また，日英ブログサイトの対照比較実験においても，本稿の実験で行った商品と社会問題の 2 つの種類のキーワードでの分析だけでは不十分であり，今後様々な種類のキーワードを検索した結果の分析を行っていく予定である．

文　献

[1] http://kizasi.jp/. Kizasi.jp.

[2] http://blog search.yahoo.co.jp/. Yahoo!ブログ検索.

[3] http://blogsearch.google.co.jp/. Google ブログ検索.

[4] http://www.technorati.jp/. Technorati.

[5] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2007 blog track. In *Proc. TREC-2007 (Notebook)*, pp. 31–43, 2007.

[6] J. Elsas, J. Arguello, J. Callan, and J. Carbonell. Retrieval and feedback models for blog distillation. In *Proc. TREC-2007 (Notebook)*, pp. 170–175, 2007.

[7] S. Sato and Y. Sasaki. Automatic collection of related terms from the Web. In *Proc. 41st ACL, Companion Volume*, pp. 121–124, 2003.

[8] 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.

[9] M. Kawaba, H. Nakasaki, T. Utsuro, and T. Fukuhara. Cross-lingual blog analysis based on multilingual blog distillation from multilingual wikipedia entries. In *Proc.ICWSM*, 2008.

[10] 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.

[11] 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.

[12] 竹原幹人, 中島伸介, 角谷和俊, 田中克己. Web 情報検索のための blog 情報に基づくトラスト値の算出方式. 日本データベース学会 Letters (DBSJ Letters), Vol. 3, No. 1, pp. 101–104, 2004.

[13] 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.

[14] 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.

[15] 福原知宏, 宇津呂武仁, 中川裕志. 複数言語間の語彙出現傾向比較による言語横断型ウェブログ関心解析システムの開発. 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. 40–43, 2007.

[16] 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.

[17] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proc. EMNLP-CoNLL*, pp. 708–716, 2007.

[18] J. Kazama and K. Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proc. EMNLP-CoNLL*, pp. 698–707, 2007.

[19] 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. 情報処理学会, 2008.

[20] 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.