1H2-5

# 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

Blog Distillation from Multilingual Wikipedia Entries and Japanese/English Cross-Lingual Blog Analysis

川場 真理子[*1] 中崎 寛之[*1] 宇津呂 武仁[*1] 福原 知宏[*2]
Mariko Kawaba Hiroyuki Nakasaki Takehito Utsuro Tomohiro Fukuhara

[*1]筑波大学大学院システム情報工学研究科
Grad. Sch. Systems and Information Engineering, University of Tsukuba

[*2]東京大学 人工物工学研究センター
Research into Artifacts, Center for Engineering, University of Tokyo

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.

## 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**アプローチをとる．

[川場 08] では，Wikipedia エントリ名のブログサイト内での出現回数をブログサイトの順位付けに使用した．その結果，被リンク数などで順位付けされる検索 API の出力順位より，良い性能を達成することが出来た．しかし，[川場 08] の手法では，ブログサイト内にエントリ名の同義語があった場合などに，同義語の出現数を順位付けに反映できないという問題や，ノイズが混入してしまうという問題も見られた．そこで，本稿では Wikipedia から得られる情報を利用してエントリの同義語や関連語を取得し，同義語や関連語のブログサイト内での検索ヒット数をブログサイトの順位付けに利用した．同義語や関連語の情報を利用したブログサイトの順位付けに利用する実験を行った結果を報告し，本稿の手法が [川場 08] での問題に対し，有効であることを示す．

## 2. Wikipediaを用いた同義語・関連語の収集

### 2.1 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．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

連絡先: 川場 真理子，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 1-1-1，029-853-5427

∗1 http://kizasi.jp
∗2 http://blog-search.yahoo.co.jp
∗3 http://blogsearch.google.co.jp
∗4 http://www.technorati.jp

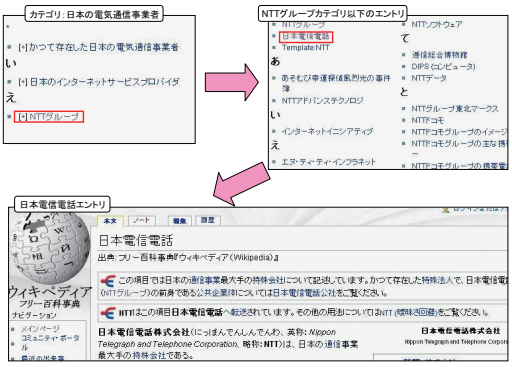


図 1: Wikipedia の構造

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．

### 2.2 同義語・関連語の収集

[川場 08] では，Wikipedia のエントリ $e$ に対して，エントリ名 $t(e)$ のみを検索トピックとして使用して検索を行った．以下では，エントリ $e$ に対して用いる検索トピックとしては，Wikipedia エントリ名 $t(e)$ を想定して説明を進める．ここで，本稿ではエントリ名以外にも，Wikipedia から収集した同義語や関連語を使用する．以下にブログサイトの順位付けに使用する同義語・関連語の収集法について述べる．

まず，同義語の候補を Wikipedia のエントリ $e$ のリダイレクトから取得し，ノイズになりそうなものを人手で取り除いた．エントリのタイトル $t(e)$ と上記の手順により作成された同義語の集合を合わせた集合を $T(e)$ とする．関連語の候補は，Wikipedia エントリの本文中の太字，強調文字とした．さらに，エントリの子が存在する場合は子エントリのタイトルも関連語の候補に加えた．最後に，Wikipedia エントリのタイトル $t(e)$ と関連語の候補 $r$ の関連度を求め $r$ の順位付けを行った．関連度には以下の尺度 [佐々木 06] を用いた．

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

関連語は一つのトピックから約 0~200 語取得できるが，その中で関連度の高いものを 5~10 個使用した．さらに，明らかに誤っているものを人手で取り除き，最終的に得られた関連語の集合を $R(e)$ とする．

## 3. Wikipedia エントリに対応するブログサイトの検索

### 3.1 ブログサイトの収集

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．そこで本稿ではそのブログサイトに検索トピックに関する事柄がどれくらい述べられているかという尺度を設定し，その尺度の大きさで検索トピックについて書かれたブログサイトかどうかを判定するという手法を用いる．具体的には，図 2 に示すように，

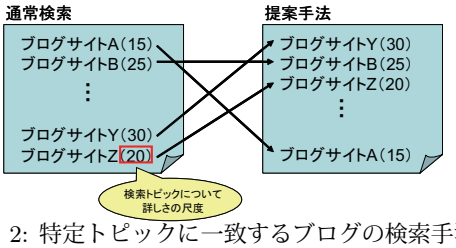


図 2: 特定トピックに一致するブログの検索手法

**エントリ名を検索クエリとした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あらかじめ用意した尺度で並び替える．**

どれくらい検索トピックについて述べられているかという尺度については，3.2 節で詳しく述べる．

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[*5]，英語ブログでは大手 12 社[*6]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 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．

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[*7]を選定した．本稿ではこれらのキーワードの中からドラゴンボール，新世紀エヴァンゲリオン，Wii，靖国神社の 4 キーワードを選び，評価実験を行った．

### 3.2 ブログサイトの順位付け

[川場 08] ではトピック名のブログサイト内での出現回数の多い順にブログサイトの順位付けを行った．しかし，表記揺れなどがあった場合，そのトピックについて書かれているにも関わらず，ランキング上位に上がってこないという問題があった．そこでこの問題を解決するために，本稿では 2.2 節で述べた同義語・関連語の情報を使用して，収集したブログサイトの順位付けを行った．本実験で使用した同義語及び関連語を表 1 及び表 2 に示す．

[川場 08] では Wikipedia のエントリタイトル $t(e)$ のブログサイト内での出現数 (以下 $Score_t$ と表し「エントリ名のみ」と呼ぶ) でのブログサイトの順位付けを行った．本稿ではこれに

*5 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

*6 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.co

*7 パフィー，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，ハローキティ，ジャズ，交響曲，犬，猫，ハムスター，ジャイアントパンダ，チワワ，ソニー，カシオ，任天堂，ホンダ，トヨタ，三洋電機，キャノン，PS3，PSP，Wii，iPod，ニンテンドー DS，靖国神社，原爆，寿司，自民党，民主党，富士山，年金，捕鯨，テロ，博物館，水族館，ミュージカル，遊園地，ディズニーランド，ボクシング，イチロー，松坂大輔，相撲，プロレス，K-1

表 1: 同義語

| 日本語トピック名 (英語トピック名) | 同義語 (日本語ブログ) | (英語ブログ) |
|---|---|---|
| ドラゴンボール (Dragon Ball) | DRAGON BALL | なし |
| Wii (Wii) | Wii，Nintendo Wii，ニンテンドー Wii | Nintendo revolution，Nintendo wii |
| 新世紀エヴァンゲリオン (Neon Genesis Evangelion) | 新世紀エバンゲリオン，Evangelion | Evangelion |
| 靖国神社 (Yasukuni Shrine) | 靖国，靖國神社，東京招魂社 | Yasukuni |

表 2: 関連語

| 日本語トピック名 (英語トピック名) | 関連語 (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，ドラゴンボール GT，サイヤ人，ドラゴンボール AF，かめはめ波，天下一武道会，仙豆，ナメック星，筋斗雲，界王拳 | Dragon Ball Z，Dragon Ball GT，anime，Akira Toriyamaf，Super Saiyan，film，franchise |
| Wii (Wii) | 任天堂，発売日，Wii リモコン，セガ，バーチャルコンソール，Revolution，ハドソン，Wii ウェア | Wii Remote，Strage，Mii，Virtual Console，Memory，Wii games，Wii Points，Internet Channel，Wii Balance Board |
| 新世紀エヴァンゲリオン (Neon Genesis Evangelion) | エヴァンゲリオン，エヴァ，綾波レイ，使徒，新世紀エヴァンゲリオン まごころを，君に，セカンドインパクト，パチスロ，惣流・アスカ・ラングレー，エヴァンゲリオン新劇場版 | mecha，manga，Angel，1.0 You Are |
| 靖国神社 (Yasukuni Shrine) | 英霊，良識の府，政教分離，九段下，九段，東京招魂社，忠魂，忠霊 | なし |

加え，同義語・関連語の情報を利用したブログサイトの順位付けを行った．以下に詳細を述べる．

本稿では，同義語・関連語を OR 検索，もしくは AND 検索しブログサイトの順位付けに使用した．OR 検索を用いた順位付けとしては一つのブログ記事内での同義語・関連語の異り数を重複せずにカウントするものと，重複してカウントするものの 2 種類がある．重複せずにカウントするものは，同義語・関連語のあらゆる要素を OR 検索したヒット数 (以下 $Score_{rOR}$ と表し,「エントリ名 OR 関連語 (重複なし)」と呼ぶ) を順位付けに使用する．重複してカウントするものは，同義語・関連語のそれぞれの要素をブログ内検索したヒット数の総和 (以下 $Score_{rORd}$ と表し「エントリ名 OR 関連語 (重複あり) と呼ぶ) を順位付けに使用する．また，AND 検索はエントリタイトルもしくはその同義語 $t(t \in T(e))$ と関連語 $r(r \in R(e))$ のあらゆる組合わせの AND 検索のヒット数の総和 (以下 $Score_{AND}$ と表し「エントリ名 AND 関連語」と呼ぶ) を順位付けに使用する．以下にこれらの式を示す．

$$Score_t = Hits(t(e))$$
$$Score_{rOR} = Hits(OR(R(e) \cup T(e) \text{ の各要素}))$$
$$Score_{rORd} = \sum_{t \in T(e)} Hits(t) + \sum_{r \in R(e)} Hits(r)$$
$$Score_{AND} = \sum_{t \in T(e), r \in R(e)} Hits(t \text{ AND } r)$$

また，API の出力順にブログサイトを並べたものをベースラインとする．

## 4. 評価

### 4.1 手順

本稿の実験では 3.1 節で述べた手法を用いて検索したブログサイトを，3.2 節で述べた方法で順位付けした．順位付けされたブログサイトの上位 30 ブログサイトと以下等間隔にサンプリングした 30 ブログサイトを手動で評価し，その性能を比較する実験を行った．実験に使用した 4 キーワードの再現率・適合率の推移を平均したプロットを図 3 に示す．

### 4.2 考察

#### 4.2.1 エントリ名のみを用いた順位付け

エントリ名の頻度のみを順位付けに使用する場合，被リンク数はあまり多くないが，そのトピックについて詳しく書いているブログサイトを上位にあげることができる．しかし，日英のブログ両方で見られた問題点として，プロフィールやアーカイブ，などのサイドカラムの情報がノイズになるということが挙げられる．本稿の実験ではブログサイト内でのエントリ名などの出現数に Yahoo!API の検索ヒット数を用いている．Yahoo!API は Web 検索の API であるため，ブログサイトのサイドカラムを本文と区別して扱わない．そのため，サイドカラムの情報はすべての記事についてカウントされてしまう．

特に英語のブログサイトでは，著者の好きなものなどを好きなものリストとして羅列することが多く，ブログサイトの内容と関係なしに，検索に使用したエントリ名が出現する，ということがあり，エントリ名を用いた順位付けがベースラインよりも下回る結果になった．

#### 4.2.2 同義語・関連語の OR 検索のヒット数を用いた順位付け

エントリ名 OR 関連語（重複あり）による順位付けと，エントリ名 OR 関連語（重複なし）による順位付けの結果を比較した．その結果，日英共に，エントリ名 OR 関連語 (重複あり) による順位付けが，エントリ名 OR 関連語 (重複なし) による順位付けを上回る結果になった．これは，検索トピックに関連する語が同じブログ記事に同時に存在すると，重複ありの場合は，その記事にでてきた同義語・関連語の異り数だけカウントされるが，重複なしの場合は 1 回としかカウントされない．そのため，そのトピックについて詳しく書かれたブログサイトが上位に上がることができなかったと考えられる．

英語ブログ　　日本語ブログ

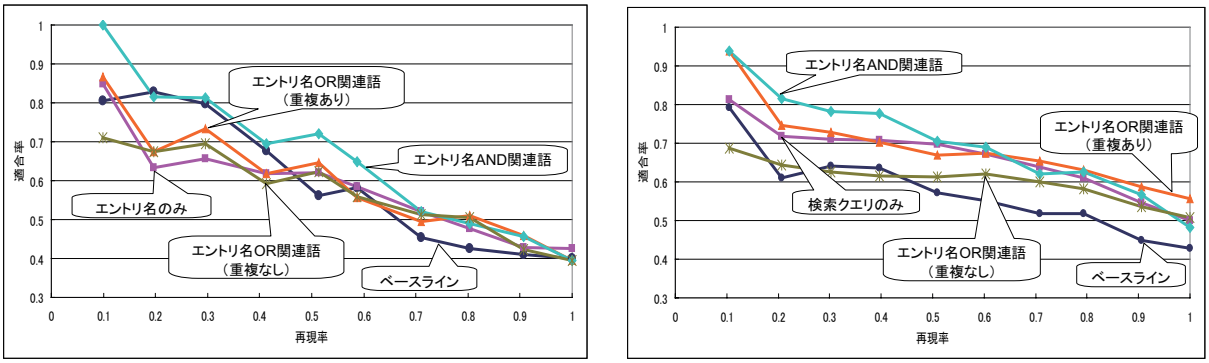


図 3: 特定トピックのブログサイト検索の評価結果 (4 キーワード分)

#### 4.2.3　エントリ名・同義語と関連語の AND 検索のヒット数を用いた順位付け

$Score_{AND}$ を使用した順位付けは，他の手法と比較してより良い性能を達成した．これは，AND 検索を行うことで，4.2.1 節で述べたサイドカラムのノイズがなくなるためであると考えられる．本手法では，[川場 08] で問題となったサイドカラムの影響を少なくすることができた．

#### 4.2.4　まとめ

本稿の実験の結果，エントリ名と関連語のブログサイト内での AND 検索の検索ヒット数を順位付けに使用することで，より良い性能を達成することができることがわかった．

しかし，日英ブログ共に，どの手法でも見られた問題点として，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ものがある．今後，これらの問題の解決の為に，ブログ記事の本文，コメント，プロフィールなどのサイドカラムに記載されている情報を区別して検索する必要があると考えられる．

## 5.　日英ブログの言語対照分析

本研究の，特定のトピックに対するブログサイト検索の一つの重要な応用として，日英ブログの言語対照分析がある．本稿の手法によって，特定のトピックについて詳細な内容が書かれたブログサイトを収集することができる．これを複数の言語について行い，その内容を比較することで，そのトピックに関して，片方の言語に特有の意見や関心を発見することができる．[川場 08] では，日英ブログサイトにおける意見や関心の違いをマイニングする手がかりとして，各言語のブログから共起語を抜き出して比較するアプローチをとった．

取得した共起語を日英で比較し，特徴的な語が現れたブログサイトを調べた結果，商品・作品等に関するキーワードでは社会的な興味，需要の違いを反映した結果が得られ，社会問題に関するキーワードでは相反する意見が見られた．例えば，ゲーム機の Wii だと，日本語ではゲームソフトなどの共起語が現れ，英語では Hack などの共起語が見られた．ブログサイトの中身は，日本語ではゲームのレビューなどが多く，英語では Wii で自作ゲームを動かす，といった Wii の改造についての記事が見られた．また，社会問題の靖国神社だと，日本語では，合祀，反日などの共起語が得られ，英語では War Shrine や Japanese militarism などの共起語が見られた．ブログサイトでは日本語は靖国神社参拝に肯定的な意見が多く，英語では否定的な意見が多くみられた．

## 6.　関連研究

ブログサイトの検索としては TREC2007 年 Blog Distillation タスク [Macdonald07] があげられる．これはある特定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探す，というものである．このタスクはブログの検索を目的としており，知識体系との対応付けは行わないという点で本研究とは異なる．その他にも，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 [竹原 04] 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 [中島 08] 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．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 [吉岡 07] 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，特定のトピックについて書かれたブログサイトの検索を行い，関連語などを用いた尺度で順位付けし，その結果を比較した．今後は，日英でのブログサイトの分布を調べるために，Wikipedia をサンプリングし，大規模な検索実験を行う予定である．

## 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：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, DEWS (2008).

[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, DEWS (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).

[竹原 04]　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).

[吉岡 07]　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).