Wikipedia エントリとブログサイトの対応付けによる 日本語ブログ空間のトピック分布推定

川場　真理子[†]　中崎　寛之[†]
宇津呂　武仁[†]　福原　知宏[††]

本研究は Wikipedia のエントリをブログサイトと対応付け，Wikipedia カテゴリ空間におけるブログサイトの分布の推定を行うことを目的とする．本稿では，各 Wikipedia エントリについて，詳細な記述をしているブログサイトが存在するかどうかの推定を行った．検索ヒット数が一定数以上となるトピックに対しては，そのトピックについて詳細な記述をしているブログサイトが存在すると仮定し，Wikipedia の約 30 万エントリに対してブログ検索を行い，検索ヒット数を得た．その結果，検索ヒット数が 1 万 ~50 万の範囲であれば，そのエントリと関連性の深いブログサイトが一定数存在する事が分かった．また，Wikipedia カテゴリ空間におけるブログサイトの分布を調べるためには，Wikipedia カテゴリに対して適切な粒度を設定し，その粒度の単位でブログサイトの有無を観測する必要がある．そこで，ブログサイトが存在する Wikipedia エントリの割合に基づいて Wikipedia カテゴリの併合を行う手法を適用することにより，関連するブログサイトが存在し，かつ適切な粒度の Wikipedia カテゴリを発見することができた．

Estimating Topic Distribution of Japanese Blogosphere by Linking Wikipedia Entries to Blog Feeds

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

This paper studies how to estimate distribution of topics in Japanese Blogosphere, where about 300,000 Wikipedia entries are used for representing a hierarchy of topics. First, in order to estimate whether there exists at least one blog feed closely related to a given topic, we use the number of hits of the topic keyword in the blogosphere. We empirically examine the range of the number of hits and conclude that the range should be 10,000 ∼ 500,000. According to our manual evaluation of this range, about 70% of Wikipedia entries can be linked to at least one blog feed, which partially justifies our claim. Next, we study how to discover Wikipedia categories with Wikipedia entries, where more than 30 ∼ 40% of them can be linked to blog feeds closely related to the corresponding topic. Through our manual evaluation of the discovered Wikipedia categories, we can conclude that the proposed technique is effective in discovering categories linking to many blog feeds that are closely related to certain topics in those categories.

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

† 筑波大学大学院 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
†† 東京大学 人工物工学研究センター
Research into Artifacets, 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

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．

文献[1] 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，文献[1] 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．この問題に対応するために，文献[2] の手法では 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．その結果，文献[2] の手法は文献[1] での問題に対し，有効であった．

文献[2] により，そのトピックに密接に関連したブログサイトを対応付ける要素技術が確立した．本稿では，この手法を用いて，Wikipedia カテゴリ空間におけるブログサイトの分布を求めた．本稿の目的を達成するためには，Wikipedia エントリに密接に関連したブログサイトの有無を調べる必要があるが，文献[2] の手法を Wikipedia の全エントリに用いると，膨大な計算時間が必要となる．そこで，検索ヒット数が一定数あるトピックに対しては，それに関連するブログサイトが存在すると仮定した．この仮定をもとに，Wikipedia エントリをブログ検索し，得られたヒット数を利用して，Wikipedia エントリに対応するブログサイトの有無の推定を行った．その結果，ヒット数が 1 万から 50 万の範囲のエントリには，そのエントリについて詳細な記述をしたブログサイトが多く分布している事が分かった．また，Wikipedia エントリを意味のある Wikipedia カテゴリにまとめる作業を行った．具体的には，検索ヒット数が 1 万から 50 万の範囲のエントリが一定数ある Wikipedia 下位カテゴリを，上位カテゴリに併合するという手法を用いた．その結果，各エントリを適切な粒度の Wikipedia カテゴリに対応付けることができた．

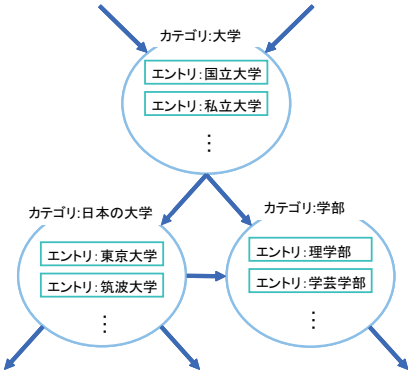

図 1 Wikipedia の構造

以下に本稿の構成を述べる．2 節では Wikipedia について述べ，3 節では Wikipedia のトピックにブログサイトを対応付けるための，特定トピックのブログサイトと検索について述べる．さらに，4 節では Wikipedia エントリに対応するブログサイトを検索し，各エントリに対するブログサイトの有無を調べた．その結果について述べる．また，5 節では Wikipeida カテゴリの妥当性と各カテゴリに対応するブログサイトの分布の推定を行った結果について述べる．6 節では関連研究に述べ，最後にまとめを行う．

2. Wikipedia

2.1 カテゴリ・エントリの階層的構造

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 40 万，英語で約 200 万のエントリ (2007 年 11 月現在) がある．また，本稿の実験では,「過去ログ」「日付」のようなノイズになりそうなエントリを除外した 305,986 エントリを対象としている．

Wikipedia は図 1 に示すように，カテゴリがグラフ構造になっており，任意の位置にあるカテゴリの節点が任意の個数のエントリを持つ．日本語 Wikipedia では，エントリを一つ以上持つカテゴリが，29,970 カテゴリ存在する．また，カテゴリ節点間の最長リンク数は 10 である．

本稿では，Wikipedia の階層構造の，根に相当するカテゴリの子にあたる 8 つのカテゴリ「学問・技術・自然・社会・地理・人間・文化・歴史」を第一層のカテゴリと定義する☆また，第一層のカテゴリから 1 ステップで辿る事

☆ 階層構造の根の子に相当するカテゴリとしては，本稿に記した 8 個以外に「総記」カテゴリが存在するが,「総記」カテゴリにリンクするエントリ・カテゴリは「過去ログ」「履歴」のような Wikipedia に独特のものである．よって，本稿の実験においては「総記」カテゴリ，および,「総記」カテゴリのみにリンクするカテゴリを除外している．

の出来るカテゴリ約 300 個を，第二層のカテゴリと定義する．さらに，第二層のカテゴリから 1 ステップでたどることのできるカテゴリを第三層のカテゴリ，第三層のカテゴリから 1 ステップで辿ることのできるカテゴリを第四層のカテゴリと定義する．また，第二層のカテゴリ以降は同じ階層のカテゴリにも親子関係がある場合がある．本稿では，Wikipedia の第一層カテゴリからの最短距離を用いて，各カテゴリの階層を決定した．

2.2 Wikipedia エントリと上層カテゴリの対応付け

本稿では，任意の日本語 Wikipedia のエントリを，そのエントリから最短の第一層もしくは第二層カテゴリに対応付けた．Wikipedia の各エントリから，第一層もしくは第二層カテゴリを幅優先で再帰的に探索する．エントリから，第一層もしくは第二層カテゴリのいずれかに到達すると探索を終え，辿りついたカテゴリとエントリが対応付けられる．また，同じ距離に対象カテゴリが複数ある場合は重複を認め，同距離に複数のカテゴリが無い場合は，三位までの最短カテゴリを対応付けた．

3. 特定トピックのブログサイトの検索

本稿では Wikipedia エントリ e に対応するブログサイトを検索する．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，Wikipedia エントリ名 $t(e)$ を想定して説明を進める．

本研究では，Wikipedia の中に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．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詳し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，トピックについて詳し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がどれくらい述べられているかを，検索トピックの出現回数で判断する．具体的には図 2 に示すように

Wikipedia 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，エントリ名の出現数順にブログサイトを並び替える．

また，ブログサイトを検索するために，ブログの検索に，Yahoo!Japan 検索 API を利用し，大手 11 社☆ のドメインを対象とし，英語のブログサイト検索には，米 Yahoo! の検索 API を利用し，大手 12 社☆☆のブログ会社のドメインを対象に検索を行った．検索の際には，複数のドメインを一度に指定して検索し，500 件の記事を取得する．

☆ 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

☆☆ 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

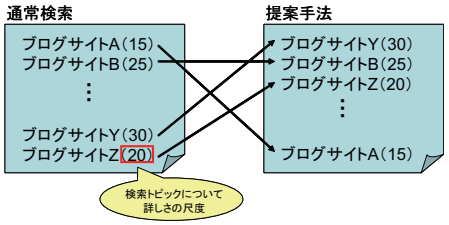

図 2 特定トピックに一致するブログの検索手法

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 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．

4. Wikipedia エントリに対応するブログサイトの有無の推定

4.1 検索エンジン API を用いたブログサイト検索および所要時間

本研究の目的を達成するためには，Wikipedia エントリに対応するトピックについて詳細な記述のあるブログサイトの有無を判定し，Wikipedia カテゴリ体系におけるブログサイトの分布を調べる必要がある．しかし，2 節で述べたように Wikipedia は日本語で約 30 万のエントリを持つ．3 節の手法を用いると 1 トピックにつき約 300 回，API へアクセスを行う．Yahoo!Japan の検索 API は 1IP アドレスにつき一日 50,000 回までのアクセス制限がある．そのため，日本語 30 万エントリすべてを検索するためには 120,000,000 回の API アクセスが必要であり，1 つの IP アドレスを使用すると 2,400 日必要となる．

このような理由から，30 万エントリすべてに対してブログサイトの有無を判定することは困難である．そこで，ブログ検索の検索ヒット数が多いトピックは，そのトピックについて詳細な記述のあるブログサイトが存在すると仮定し，検索ヒット数を用いて，Wikipedia エントリに対応するトピックのブログサイトの分布を近似した．検索ヒット数を求めるには，検索時間が 1 トピックあたり約 1.5 秒，API アクセスは 30 万回であるから，8 日で終えることが可能である．

4.2 トピック名の検索ヒット数を用いたブログサイトの有無の推定

Wikipedia のエントリを無作為に選んで，ヒット数と Wikipedia エントリに対応するトピックのブログサイトの有無の相関性を調べたところ，検索ヒット数が多いものは「人」「ブログ」などの一般語が多く含まれ，逆に検索ヒット数が少ないものはあまり人に知られていない地名や人名などが多く見られた．また，検索ヒット数が 1 万から 50 万のエントリのトピックには,「養子縁組」「デパ地下」「盲導犬」などのブログサイトが存在するトピックが多いことがわかった．

ヒット数と，トピックに対応するブログサイトの有無の間で相関性があることがわかったため，Wikipedia エ

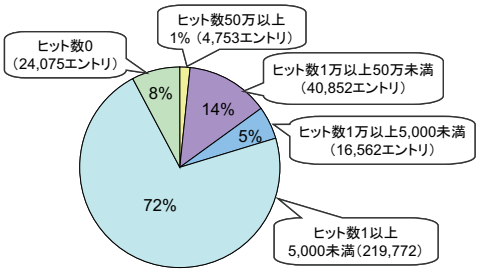

図 3 Wikipedia エントリにおけるブログヒット数の分布 (総数 305,986)

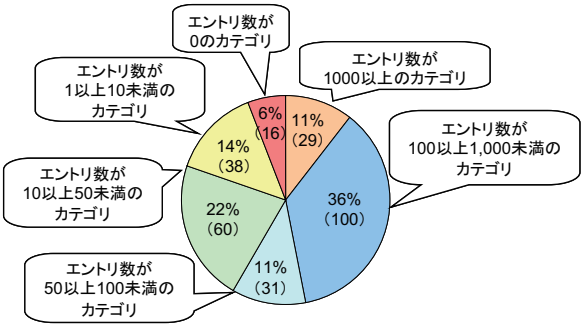

図 4 第二層カテゴリにおいてブログヒット数が 1 万 ~50 万の Wikipedia エントリ数の分布

エントリすべてに対して，ブログ全体におけるキーワードの検索ヒット数を求めた．すると，ヒット数が 1 万から 50 万のエントリは 40,852 個あり，全体の 14 パーセントであった．検索ヒット数による Wikipedia エントリの分布を図 3 に示す．

また，2.2 節の手順に基づいて，1 万 ~50 万のブログヒット数のエントリを第一層のカテゴリに分類すると，図 4 のように分布した．

4.3 ブログサイト有無推定結果の人手評価

4.3.1 評 価 手 順

検索ヒット数が 1 万から 50 万の範囲のエントリに対応するトピックにブログサイトがどの程度存在するかを調べるために，ヒット数が 1 万から 50 万のエントリをサンプリングして検索実験を行った．以下にその手順を述べる．

2.2 節の手順に基づいて，エントリを第二層のカテゴリに分類した．さらに，意味のあるまとまりになっており，各カテゴリの持つエントリ数が 50 以上あるカテゴリの中から 18 個☆を人手で選んだ☆☆．さらに，18 個の各カテゴリが持つエントリを，それぞれヒット数で降順にソートし，等間隔に 40 個ずつサンプリングした．エントリのサンプリング手順を図 5 に示す．18 カテゴリ× 40 エントリ= 720 エントリをサンプリングする事が出来た．これら 720 個のエントリのタイトルを検索トピックとし，3 章で述べた手法を用いてブログサイトを検索した．

☆ 娯楽，生物，レジャー，コンピュータ，政治，趣味，医療，健康，法，社会問題，戦争，社会制度，情報社会，自然保護，安全技術，自然災害，福祉，公衆衛生学

☆☆ 本稿では 18 個のカテゴリを人手で選んだが，カテゴリを無作為に選んでサンプリングする実験も，今後行う予定である

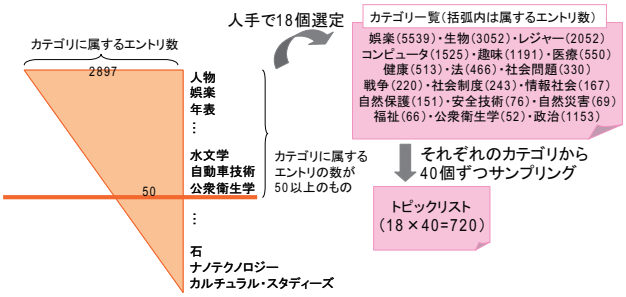

図 5 Wikipedia エントリのサンプリング手順

表 1 ブログサイトの有無推定結果の評価基準

評価	基準
A	トピックについて詳しいブログサイトが 10 件以上ある
B	トピックについて詳しいブログサイトが 5 件以上ある
C	トピックについて詳しいブログサイトが 1 件以上ある
D	トピックの上位概念についてのブログサイトがある
E	トピックについて詳しいブログサイトがない

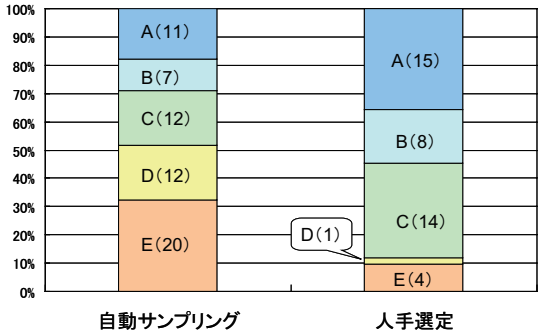

図 6 ブログサイトの有無推定結果の人手評価結果 (() 内の数字はトピック数)

4.3.2 日本語 Wikipedia エントリでの評価

720 トピックからの自動サンプリングしたエントリと人手選定したエントリの全体での 5 段階評価の割合を図 6 に示す．人手評価をした結果，社会問題や教育，環境などの分野にはまとまった量のブログサイトがあることが分かった．また，自動サンプリングしたエントリから得られたブログサイトには「生ゴミ」について書かれたブログサイトなどがあり，生ゴミの処理方法からゴミ問題まで広く語られていた (図 7)．人手選定したエントリから得られたブログサイトとしては,「団塊の世代」につい

図 7　生ゴミについて書かれたブログサイト

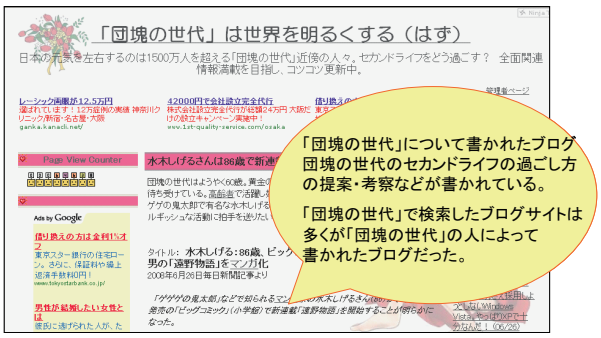

図 8　団塊の世代について書かれたブログサイト

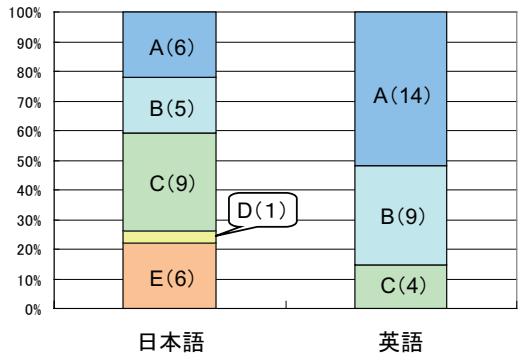

図 9　日英で対訳のある 27 トピックに対するブログサイト有無推定結果の人手評価 (() 内の数字はトピック数)

て書かれたブログサイトなどが見られた．主に，団塊の世代のブロガーによって書かれており，退職後の過ごし方などについて語られていた (図 8)．

また，複数のトピックをまたいで表れるブロガーなども存在した．例えば，「地球温暖化」と「砂漠化」のような地球環境に関するトピックや，「北方領土問題」「反日教育」「BSE 問題」などの外交問題に関するトピックには同一のブロガーがトピックを跨いで現れた．また「人工授精」や「養子縁組」などのトピックでは，両トピックともに「不妊」について述べられていた．このように，共通して現れるブロガーは多くないが，ブログサイトでの話題が共通しているトピックも存在した．

また，評価が D のトピックには「割 (相撲)」や「ライン (競輪)」などスポーツの技の名前などが見られた．これらは上位概念のトピックで検索する事でブログサイトを得られる可能性がある．また，評価が E のトピックには「クラブ・アメリカ」といった店の名前などがみられた．

4.3.3　英語ブログでの評価—英語への言語間リンクを持つエントリの場合

本研究の応用の一つとして，同一トピックにおける日英ブログの言語対照分析があげられる[4]．そこで，日英間で同一のトピックについて検索を行い，トピックに対応するブログサイトの有無の比較を行う．

日本語でヒット数が 1 万から 50 万のエントリの中で，Wikipedia の言語間リンクで繋がっている英語エントリは約 18,000 エントリ存在した．これらの 18,000 エントリに対して，ブログ空間での分布を知るために，ブログ検索しヒット数を求めた．その結果，英語ブログサイト検索のヒット数が 1 万から 80 万の範囲の約 6,000 エントリに，ブログサイトのトピックとなりそうなエントリが多く存在する事が分かった．この 6,000 エントリの内，人手評価を行った日本語 Wikipedia 約 100 エントリに対応するものは 27 エントリ存在した．この 27 エントリを人手で 5 段階評価した．評価結果を図 9 に示す．日本語でヒット数が 1 万から 50 万あり，英語でヒット数が 1 万から 80 万ある 27 エントリは全て ABC のいずれかの評価がつき，DE の評価が付くものは見られなかった．

日本語ではブログサイトが検索できなかったが，英語でブログサイトが検索できたトピックには，「盗作 (plagiarism)」，「パンデミック (pandemic)☆」などがある．

日本語では「盗作」についてのブログサイトは検索できなかったが，英語では「plagiarism」についてのブログサイトがいくつか検索できた．「plagiarism」についてのブログサイトでは，論文の盗作や，ネット上の記事の盗作について述べられていた．これは，日本と海外での「盗作」に対する問題意識に差があるためだと考えられる．

また，英語では「pandemic」について述べられたブログサイトが多く見られたが，日本語ではごく少数のブログサイトが「パンデミック」について述べていた．これは，海外では既にパンデミックの対策がされているところがあり，多くの人に知られている言葉であるが，日本では，まだ一般的な言葉ではないためであると考えられる．今後，このトピックは日本でも多くの人の話題に上る可能性があると推測されるため，数ヶ月後にブログサイトを収集すると，「パンデミック」について述べているブログサイトが増えている可能性がある．

4.4　ブログサイトごとのヒット数に関する分析

4.3 節では，Wikipedia のエントリに対応するトピックのブログサイトの有無の数によって，そのエントリの 5 段階評価を行った．本研究の目的を達成するためには，各エントリに対応するトピックのブログサイトの有無の自動判定が必要である．

そこで，ここでは，ブログサイトの有無の自動判定の手がかりの一つとして，ブログサイトごとの検索ヒット

☆ ある感染症や伝染病が世界的に流行することを表す用語．(Wikipedia より抜粋)

数の分布について分析する．4.3 節では，図 6 において，ブログサイトの有無推定結果を人手で 5 段階に評価した．これに対して，ここでは，5 段階の各評価ごとに，各ブログサイトにおける検索ヒット数の分布を求めた (図 10)．その結果，評価 A 及び B のトピックでは，検索ヒット数が 50 以上あるブログサイトが全体の 2 割ほどあり，逆に評価が D や E となるトピックでは検索ヒット数が 1 以上 10 未満のブログサイトが 8 割ほどあった．検索ヒット数が 50 以上のブログサイトの割合が多いトピックは，そのトピックについて書かれたブログサイトも多い．一方，検索ヒット数が 50 以上のブログサイトの割合が少なく，かつ検索ヒット数 10 未満のブログサイトの割合が高いトピックは，対応するブログサイトも少ないと考えられる．

また，本稿ではブログサイトの検索に，トピック名の出現数のみを利用している．しかし，ブログサイトの検索において，Wikipedia から得られる同義語・関連語などを利用する事で，より性能よくブログサイトの検索が可能である[2)]．今後，ブログサイトにおけるトピックのヒット数の分布や，Wikipedia から得られる同義語・関連語の検索ヒット数の分布の情報などを素性として利用し，トピックに対応するブログサイトの有無の判定の機械学習を行う予定である．

5. Wikipedia カテゴリ空間におけるブログサイトの分布の推定

5.1 予 備 調 査

ブログサイトと Wikipedia エントリを対応づけ，ブログ空間でのエントリの分布を知るためには，Wikipedia カテゴリに対して適切な粒度を設定し，その粒度の単位でブログサイトの有無を観測する必要がある．しかし，カテゴリの粒度が細かすぎると，全体像を見渡すのが困難になり，粗すぎるとまとまりの意味が薄れてしまう恐れがある．そのため，適切な粒度のカテゴリをエントリと対応付けるための予備調査として，Wikipedia の第四層までの上層カテゴリを中心に，エントリを対応付けた．さらに，各カテゴリについて，カテゴリが持つエントリの絶対数，およびカテゴリが持つエントリのうち，検索ヒット数 1 万から 50 万までのエントリの割合を求めた．

これらのカテゴリをサンプリングして予備調査を行ったところ，カテゴリが持つエントリの絶対値が 10 以下のカテゴリは粒度が細かすぎる傾向があった．また，ヒット数 1 万から 50 万までのエントリの割合が高いカテゴリは，意味のある適切な粒度となっており，かつそれらのカテゴリが持つエントリに対応付けられるトピックに対してブログサイトが多く存在することが予想された．

5.2 ブログサイト分布に基づく Wikipedia カテゴリの適切な粒度の決定

5.1 節より，カテゴリ c の持つエントリの絶対値が 11 以上☆でヒット数 1 万から 50 万のエントリが存在する割

表 2 Wikipedia カテゴリに対応するブログサイトの有無の人手評価

評価	基準
A	カテゴリに対応付けられたエントリのうち，ブログサイトがあると推定されるものが半数以上ある
B	カテゴリに対応付けられたエントリのうち，ブログサイトがあると推定されるものが数個ある
C	カテゴリに対応付けられたエントリのうち，上位概念のブログサイトがあると推定されるものがある
D	カテゴリに対応付けられたエントリのうち，ほとんどのエントリにブログサイトがないと推定される
E	カテゴリとエントリの対応付けが間違っている

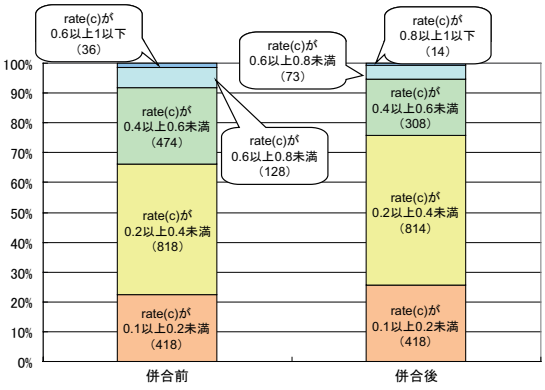

図 11 併合前後の Wikipedia カテゴリの $rate(c)$ の分布 ($LBD_{rate} = 0.4$，() 内の数字はカテゴリ数)

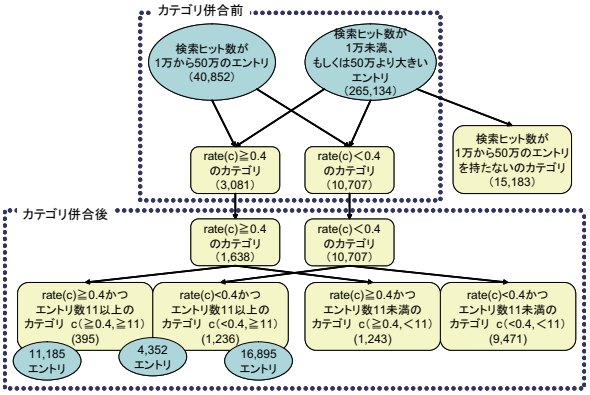

図 12 カテゴリ併合前後のカテゴリ数・エントリ数の推移 ($LBD_{rate} = 0.4$)

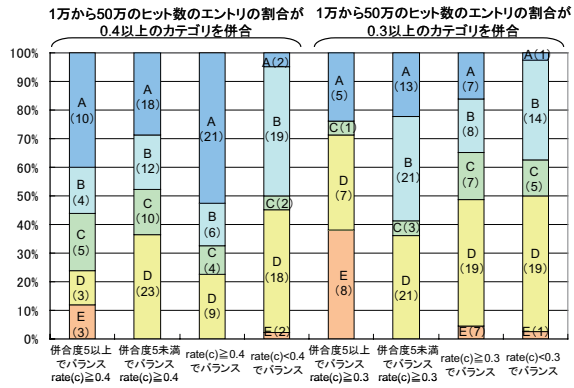

図 13 Wikipedia カテゴリに対応するブログサイトの有無の人手評価 (カテゴリ併合後，$LBD_{rate} = 0.3, 0.4$，() 内の数字はカテゴリ数)

☆ 絶対値が 10 以下のカテゴリも対象とした場合，$rate(c)$ が高いカテゴリの大半が，エントリ数 1 や 2 のブログサイトとなってしまい，

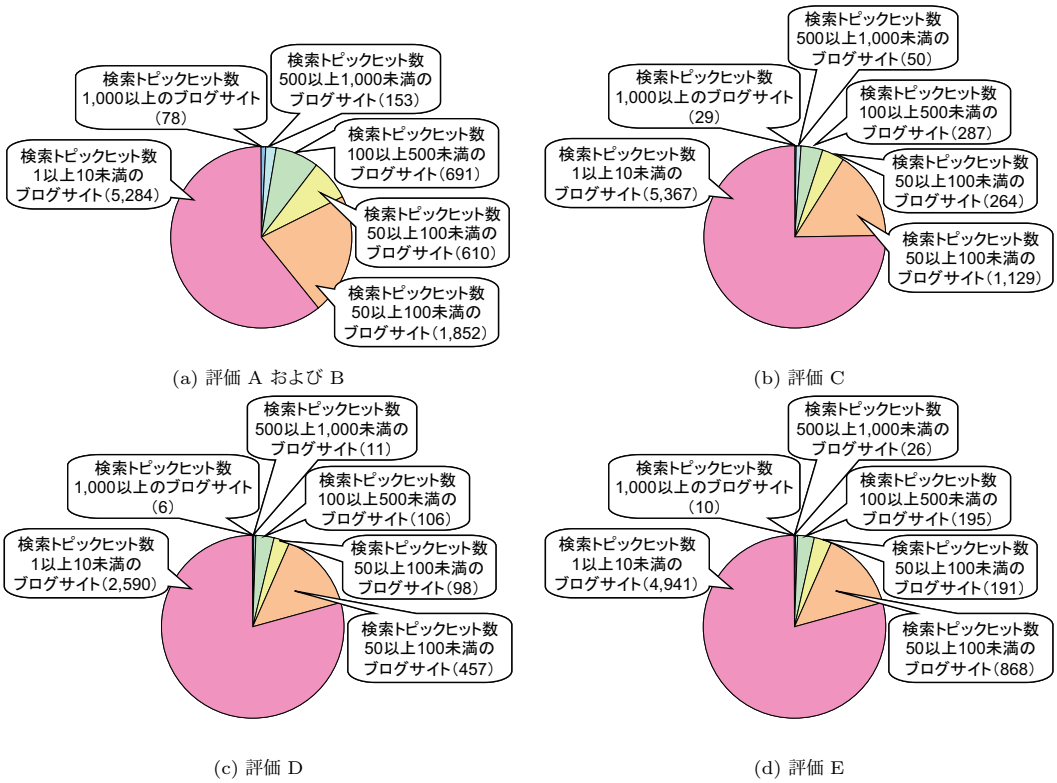

図 10 各ブログサイトにおける検索ヒット数の分布 (評価 A,B,C,D,E ごと)

合 (以下，$rate(c)$ とする) の高いカテゴリに，ブログサイトが多く存在するエントリが対応付けられている事が分かった．また，Wikipedia には約 30 万のカテゴリが存在するが，これらのカテゴリのいくつかは，粒度が細かすぎるために，より上位の概念を持つカテゴリと併合する必要がある．そこで，$rate(c)$ が高いカテゴリを併合する事で，Wikipedia カテゴリの適切な粒度を決定する．以下に手順を述べる．

(1) Wikipedia のカテゴリ c が持つエントリの集合を $ents(c)$ とする．
また，カテゴリ c の持つエントリのうち，検索ヒット数が 1 万から 50 万のエントリの割合 $rate(c)$ を以下の式で表す．ただし，$bhits(e)$ は，エントリ e に対応するトピックのブログ検索ヒット数である．

$$rate(c) = \frac{|\{e|e \in ents(c), 10000 \leq bhits(e) \leq 500000\}|}{ents(c)}$$

併合したカテゴリを集合として記録するために $desc(c)$ を用いる．$desc(c)$ の初期値は，カテゴリ c のみから構成される集合 $\{c\}$ とする．
また，併合したカテゴリの数を併合度とし $|desc(c)|$ で表す．

(2) $rate(c)$ に対する下限値を LBD_{rate} として，カテゴリ c，および，カテゴリ c の子カテゴリ c' について，c，c' とも，$rate(c) \geq LBD_{rate}$，$rate(c') \geq LBD_{rate}$ を満たすあらゆる親子カテゴリの組に対して，カテゴリの併合を行う．
$ents(c) \leftarrow ents(c) \cup ents(c')$
また，併合したカテゴリを $desc(c)$ に追加する．
$desc(c) \leftarrow desc(c) \cup desc(c')$

5.3 Wikipedia カテゴリに対応するブログサイトの有無の人手評価

5.2 節の手順で，最終的に残されたカテゴリのうち，$ents(c)$ が 11 以上のカテゴリを，ヒット数 1 万から 50 万のエントリの割合 $rate(c)$ で整列し，等間隔に 80 カテゴリをサンプリングした．サンプリングしたカテゴリを表 2 の評価基準に基づいて人手で 5 段階評価した．ブログサイトの有無の推定基準としては，エントリタイトルが一般語・固有名であれば，ブログサイトが無いと推定した．また，人名に関しては，オリンピック選手のような誰もが知っているような有名人以外は Wikipedia を参照し，Wikipedia エントリ本文のテキスト長などを考慮して推定を行った．本稿の評価では LBD_{rate} は 0.4 と 0.3 の 2 種類の場合を評価した．LBD_{rate} を 0.4 に設定した場合の，併合前後の $rate(c)$ の分布を図 11 に示し，カテゴリ・エントリ数の推移を図 12 に示す．

Wikipedia カテゴリに対応するブログサイト有無の人手評価の結果を図 13 に示す．また，各 $LBD_{rate} = 4$ の場合の併合前後のカテゴリ・エントリの例を表 3 に示す．

評価 A のカテゴリは併合前のカテゴリの多くが，ブログサイトがあると推定されるトピックと関連性の強いエ

意味のあるまとまりが得られなくなる．そこで，本稿ではカテゴリの持つエントリの絶対値が 11 以上のカテゴリを対象とした．

表 3 各評価 (A,B,C,D,E) におけるカテゴリの併合前後の Wikipedia カテゴリ・エントリ

評価	カテゴリ	併合度	併合前カテゴリ/エントリ
A	コレクション	5	骨董品/骨董市・有田焼，トレーディングカード/カードダス・デルトラクエスト
B	インターネットサービス	5	ウェブホスティング/インフォシーク・GeoCites，動画/ニコニコ動画・ストリーミング配信
C	電子機器	10	懐中時計/オメガ・ウォルサム，プリンター/インクジェット・トナー
D	物理化学の現象	1	物理化学の現象/落下・爆破
E	太陽系の惑星	68	ミネラルウォーター/六甲のおいしい水・コントレックス，月探査/月面着陸・月面基地

ントリを持っている．また評価 B のカテゴリでは，エントリに対応するトピックのブログサイトがあると推定されるエントリを持つカテゴリと，持たないカテゴリが併合されることで，ブログサイトと対応付ける事の出来るエントリの割合が減ってしまったものが見られた．また，評価 C のカテゴリに属するエントリは，上位概念ならブログサイトと対応付ける事の出来るものが多く見られた．評価 D のカテゴリについては，一般語をエントリに持つカテゴリが多く見られた．さらに，評価 E となるカテゴリについては，併合前の個々のカテゴリは意味のあるまとまりになっているが，併合しすぎた結果，カテゴリに対して適切でないエントリが多くなってしまっている．

$rate(c)$ の高いカテゴリは A，B の割合が高く，適切な粒度でカテゴリが対応付けられている事が分かる．しかし，D と判定されたカテゴリも 22.5 パーセント存在した．これは，現在のアルゴリズムだと，カテゴリを併合させる過程で止める方法が無いために，粒度が粗くなってしまう場合があるためだと考えられる．

また，LBD_{rate} を 0.3 にした場合，LBD_{rate} が 0.4 の時と比較して，D の割合が増えた．これは，LBD_{rate} が 0.4 の場合と比較して，$rate(c)$ が低いカテゴリも親カテゴリに併合されるために，最終的に出来上がったカテゴリにノイズが多く混入してしまい，カテゴリの粒度が粗くなってしまうためであると考えられる．

また，LBD_{rate} が 0.3，0.4 それぞれの場合で併合度を求め，降順に整列した．これらから，等間隔に 80 カテゴリをサンプリングし，同様に人手評価を行った．人手評価を行った結果，併合度と評価の相関は見られなかったが，LBD_{rate} が 0.4，0.3 の両方で，併合度が大きいと評価 E の割合が大きくなるという現象が見られた．

これらの結果より，$rate(c)$ がある程度高いカテゴリのみを併合させたほうが，より適切な粒度のカテゴリが得られるということが分かった．また，LBD_{rate} の値に関わらず，途中でカテゴリの併合を止める手法が必要である．

6. 関 連 研 究

ブログサイトの検索に関する関連研究として，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[5] 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[3] 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．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[6] 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．

7. お わ り に

本稿では，ブログ空間における Wikipedia のエントリの分布を，各エントリのブログ検索ヒット数で近似した．その結果，ヒット数が 1 万から 50 万の範囲のエントリには 7 割前後のブログサイトが対応づけられることがわかった．また，ブログ空間における，Wikipedia エントリの分布推定を行うためには，エントリを適切な粒度で意味のあるまとまりに分類することが必要不可欠である．そのために，各エントリを Wikipedia のカテゴリに適切な粒度で対応づける手法を提案した．現在，日本語 Wikipedia30 万エントリについてブログサイト検索を行っている．本稿ではヒット数を 1 万から 50 万に限定したが，1 万から 50 万以外のエントリについても，ブログサイトを対応づけ，評価を行う．

参 考 文 献

1) 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．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.
2) 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.
3) 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.
4) 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, pp. 115–118, 2008.
5) 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.
6) 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.