人工知能学会研究会資料
SIG-SWO-A803-12

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

Estimating Topic Distribution of Japanese Blogsphere based on Wikipedia Topic Hierachy

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

[1] 筑波大学大学院 システム情報工学研究科
[1] Graduate School of Systems and Information Engineering, University of Tsukuba
[2] 東京大学 人工物工学研究センター
[2] Research into Artifacts, Center for Engineering, University of Tokyo

Abstract: 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. Then, we apply SVMs to the task of judging whether, given a topic, each of blog feeds is closely related to the given topic. Based on the learned SVMs model, we further automatically judge whether there exists at least one blog feed closely related to a given topic. Finally, 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.

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

*連絡先： 筑波大学大学院 システム情報工学研究科
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

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．

本論文では，[川場 08a] の手法を用いて，Wikipedia をトピック体系として日本語ブログ空間のおけるブログサイトの分布を求めた．また，検索ヒット数が一定数あるトピックは，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，ブログサイトが多く分布するトピックの有無をより正確に推定するためには，個々のブログサイトを判定する必要がある．そこで，Wikipedia エントリから得られる知識を素性として機械学習 (Support Vector Machines(SVM) [Vapnik98]) によってブログサイトのトピック判定を行う方式を提案する．また，各トピックに対して収集された全ブログサイトに対して，トピックとの対応についての判定を行った結果に基づいて，トピックごとにブログサイトの有無の判定を行い，その結果を評価した．さらに，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．

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百科事典であり，日本語で約 55 万エントリ存在する (2009 年 1 月現在)．本論文の実験では 2007 年 11 月の段階での日本語約 40 万エントリから,「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．

本論文では，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[5]また，第一層のカテゴリから 1 ステップで辿る事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．本論文では，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．

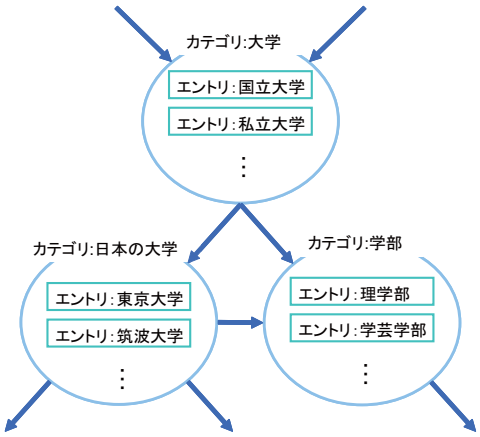


図 1: Wikipedia の構造

2.2 Wikipedia エントリと上層カテゴリの対応付け

本論文では，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 Wikipedia エントリのタイトルのヒット数を用いた日本語ブログサイトの有無の推定

3.1 概要

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

[5]階層構造の根の子に相当するカテゴリとしては，本論文に記した 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，本論文の実験においては「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．

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．この結果，ヒット数 1 万から 50 万の範囲のエントリにブログサイトが多く分布することがわかった．そこで，本節では，この傾向を定量的に検証するために，エントリ名のヒット数とブログサイトの有無に相関があるか否かを分析した結果を示す．

3.2 評価対象の Wikipedia エントリおよびブログサイト

以下の節では，まず，評価対象となる Wikipedia エントリおよびブログサイトを選定する手順について述べる．

3.2.1 Wikiepdia エントリの選定手順

まず，本論文では，前節の観察に基づいて，Wikipedia エントリに対して，タイトルのヒット数が 1 万以下，1 万から 50 万，50 万以上の 3 つの範囲を設けて，各範囲ごとに Wikipedia エントリを選定することとする[6]．

次に，Wikipedia のエントリ内から無作為にカテゴリを選び，それらのカテゴリに属するエントリを数個 (無作為に) サンプリングした．サンプリング手順を図 3 に示す．サンプリングの結果，ヒット数 50 万以上を 13 エントリ，ヒット数 1 万から 50 万以上を 82 エントリ，ヒット数 1 万以下を 87 エントリ，それぞれサンプリングすることができた．

3.2.2 ブログサイトの収集

次に，前節で選定した各 Wikipedia エントリ e について，人手評価の対象とするブログサイトを収集する．以下ではエントリ e に対応して用いる検索クエリとして，Wikipedia エントリ名 $t(e)$ を用いる．ここで，検索されるべきブログサイトは，Wikipedia エントリ e に対応するトピックについて詳細な記述が多いブログサイトである．このことを実現するために，本論文では，検索クエリとして用いる Wikipedia エントリ名 $t(e)$ の，ブログサイト内での出現数を用いて，Wikipedia エントリ e のトピックとの対応度合いを測定する．具体的には，Wikipedia エントリ名 $t(e)$ 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て，その上位 20 ブログサイトを対象として，Wikipedia エントリ e とのトピックの対応を人手で評価した．ここで，ブログサイトを検索するために，Yahoo!Japan 検索 API を利用し，大手 11 社[7] のドメインを対象とした．

[6]参考情報として，Wikipedia エントリすべてに対して，ブログ空間全体におけるエントリ名の検索ヒット数を求め，検索ヒット数による Wikipedia エントリの分布を求めた結果を図 2 に示す．この結果においては，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%であった．

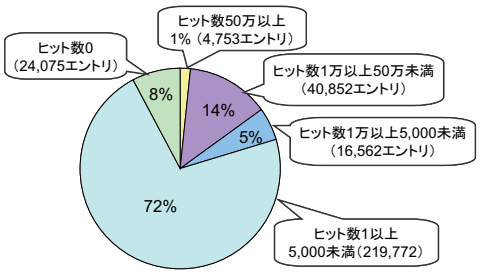


図 2: Wikipedia エントリにおけるブログヒット数の分布 (総数 305,986)

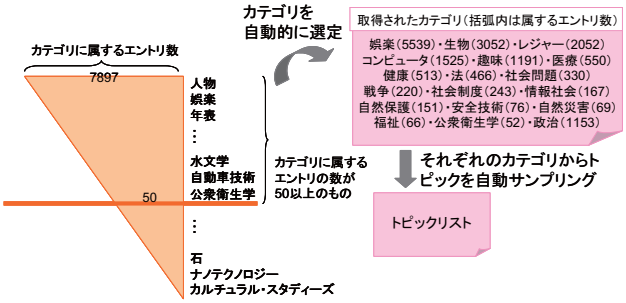


図 3: Wikipedia エントリのサンプリング手順

3.3 評価結果

ブログサイト単位でのトピックの判定結果に基づいて，表 1 の評価基準を用いて，トピック単位での評価を行った．その結果，ヒット数 1 万から 50 万の範囲に，ブログサイトが存在するトピックが多く分布していた．よって，トピックのヒット数と Wikipedia エントリの対応するブログサイトの有無には相関性があることがわかった．トピックの評価の分布をヒット数のレンジごとに示したものを図 4 に示す．ここで，検索クエリとなったトピックに対応するブログサイトの数は，ヒット数 50 万以上のトピックでは，209 ブログサイト中 51 ブログサイト，ヒット数 1 万から 50 万の範囲のトピックでは，1150 ブログサイト中 326 ブログサイト，ヒット数 1 万以下のトピックでは，1125 ブログサイト中 204 ブログサイトであった．

表 1: ブログサイトの有無推定結果の評価基準

評価	基準
C1	トピックについて詳しいブログサイトが 10 件以上
C2	トピックについて詳しいブログサイトが 5 件以上
C3	トピックについて詳しいブログサイトが 1 件以上
HU	トピックの上位概念についてのブログサイトがある
HL	トピックの下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

[7]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

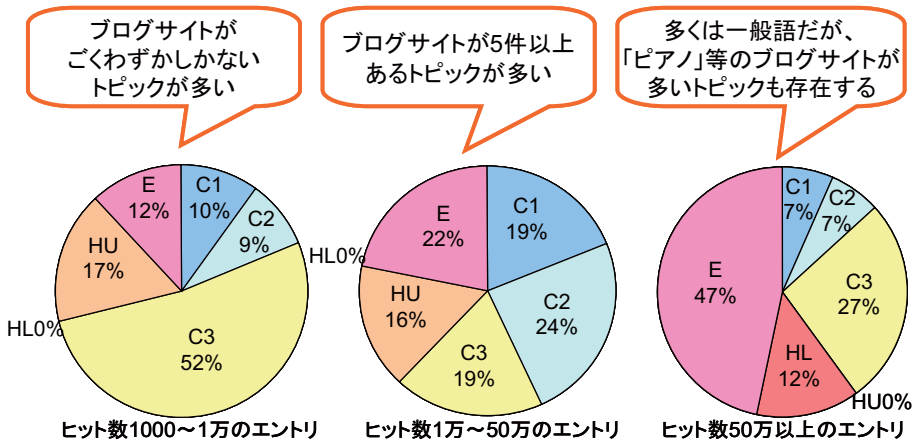


図 4: エントリ名のヒット数のレンジごとでの「ブログサイトの有無」の分布

4 Wikipediaエントリのタイトルのヒット数を用いた日英ブログサイトの有無の比較

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 [中崎 09]．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．また，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[8]のブログホスト会社を対象とした．

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 5 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，HL，HU，E 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)[9]」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．

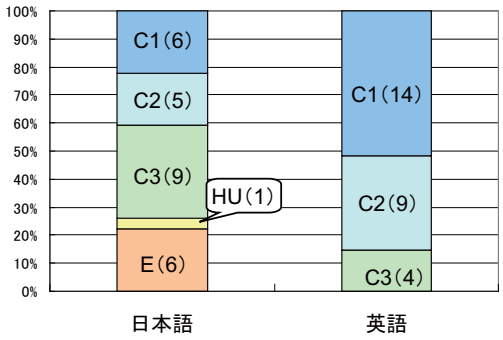


図 5: 日英で対訳のある 27 トピックに対するブログサイト有無推定結果の人手評価 (() 内の数字はトピック数)

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．

5 ブログサイトごとのヒット数に関する分析

3 節の分析結果から，トピックのヒット数を用いて Wikipedia エントリに対応するブログサイトの有無を粗く推定す

[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.co

[9] 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)

表 2: ブログサイト内での関連語のヒット数・種類数を用いた素性一覧

ID	素性
1	ピックのヒット数
2	H 関連語のヒット数の総和
3	M 関連語のヒット数の総和
4	L 関連語のヒット数の総和
5	H 関連語の種類数
6	M 関連語の種類数
7	L 関連語の種類数
8	全関連語の種類数

ることが可能であることがわかった．しかし，トピックの判定をより正確に行うためには，各々のブログサイトについてトピック判定を行う必要がある．

そこで，本節では，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数の分布について分析する．3 節では，図 4 において，ブログサイトの有無推定結果を人手で 6 段階に評価した．これに対して，本節では，3 節において評価対象としたトピックおよびブログサイトに対して，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 6)．ただし，図 4 に示す三段階のヒット数レンジによって，ブログサイトを区別はせず，全ブログサイトを一括して評価した．

その結果，評価 C1 及び C2 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 HL，HU 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．

6 機械学習によるブログサイトのトピックの自動判定

本節では，Wikipedia から得られるトピックの関連語を利用して，ブログサイトのトピック判定を自動で行った．具体的には各々のブログサイトに対して，トピックの関連語のヒット数や関連語の出現種類数を素性とする機械学習 (Support Vector Machines (SVM)) を適用した．

6.1 学習および判定手順

本節では SVM を用いて，ブログサイトがトピックについて書かれたものかどうかを判定する．SVM のツールとして TinySVM[10]を用いた．また，訓練および評価事例を $\langle b_e, c\rangle$ と記述する．ここで，b_e は Wikipedia エントリ名 $t(e)$ をトピックとして検索されたブログサイト，c は b_e がそのトピック $t(e)$ について書かれたものかどうかを示す．b_e が正解の場合 $c = +$ となり，そうでない場合 $c = -$ となる．

また，素性としては，エントリ名のブログサイト内ヒット数に加えて，Wikipedia から得られる関連語を利用した[11]．Wikipedia のエントリから得られる関連語としては Wikipedia エントリ中のリンクテキスト，太字，リダイレクト語がある．また，加えて，エントリと同名の Wikipedia カテゴリがあった場合，その Wikipedia カテゴリの持つ子エントリのエントリ名も関連語として利用した．このようにして関連語を取得した結果，一トピックあたり平均 15 個の関連語が得られた．これらの関連語を API で検索し，関連語のヒット数および，各ブログサイト内での関連語のヒット数を取得した．ここで，関連語の持つヒット数を 50 万以上，1 万から 50 万，1 万以下の 3 つの範囲に分け，それぞれ H 関連語，M 関連語，L 関連語とした．これらの情報を用いて設計した素性を表 2 に示す[12]．

また，分離平面からの距離を信頼度とし，信頼度が一定の範囲以下であるものを除外した．信頼度を用い

表 3: SVM を用いたブログサイトのトピックの自動判定の評価結果 (%)

(a) ヒット数 1 万以下のトピック

条件	素性	適合率	再現率	F 値
ベースライン	1	62.5	36.1	49.3
F 値 1 位 (信頼度閾値なし)	3(+6)	55.8	71.6	**63.7**
(信頼度閾値なしの場合の適合率 1 位)	3+7	69.4	42.6	56.0
適合率 1 位 (信頼度閾値 0.9)	1+8	**80.0**	16.4	48.2

(b) ヒット数 1 万 ~50 万のトピック

条件	素性	適合率	再現率	F 値
ベースライン	1	59.8	76.9	68.4
F 値 1 位 (信頼度閾値なし)	1+3	66.3	72.0	**69.2**
(信頼度閾値なしの場合の適合率 1 位)	3+8	73.3	46.3	59.8
適合率 1 位 (信頼度閾値 0.9)	1+8	**83.9**	20.3	52.1

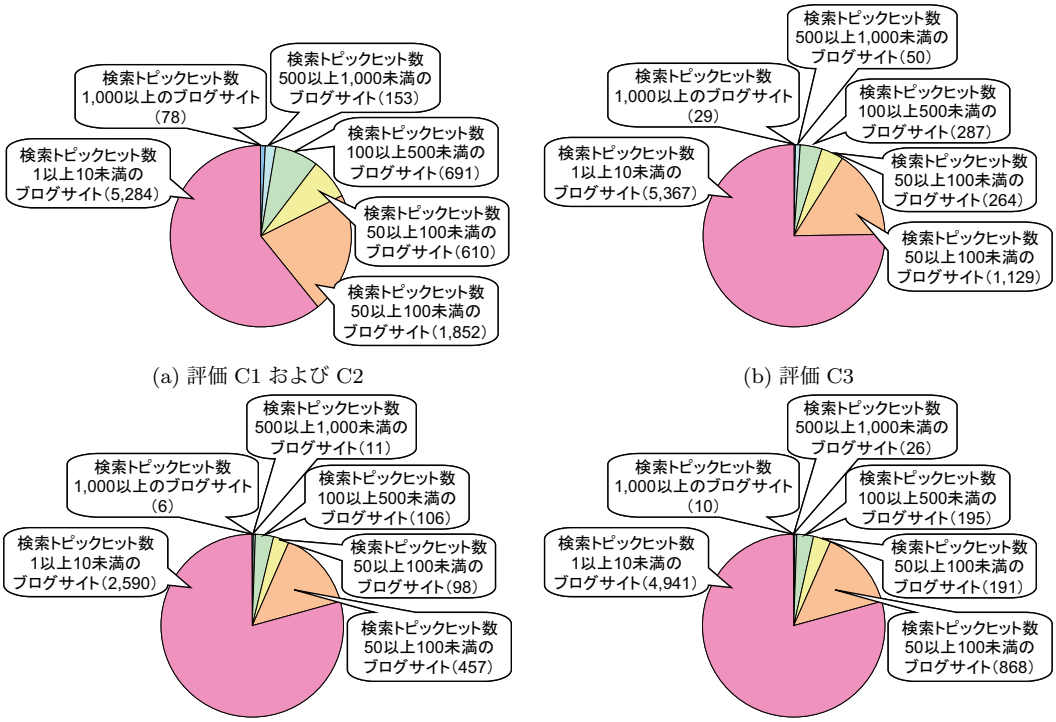
(c) ヒット数 50 万以上のトピック

条件	素性	適合率	再現率	F 値
ベースライン	1	65.3	45.7	55.5
F 値/適合率 1 位 (信頼度閾値なし)	3+6+8	**87.5**	65.0	**76.3**

[10] http://chasen.org/~taku/software/TinySVM/

[11] ブログサイトの検索において，Wikipedia から得られる関連語のヒット数などを利用することにより，より性能よくブログサイトの検索が可能である [川場 08b].

[12] 素性 ID=1~4 の素性は，5 段階のレンジに分けて，各レンジに該当するか否かを個別の二値素性とした．



(a) 評価 C1 および C2

(b) 評価 C3

(c) 評価 HL および HU

(d) 評価 E

図 6: 各ブログサイトにおける検索ヒット数の分布 (評価 C1,C2,C3,HU,HL,E ごと)

て候補を絞りこむと，再現率が下がってしまうが，本研究ではトピックごとについて詳しく書かれたブログがあるかないかということを正確に判定する必要があるため，再現率よりも適合率を重視する．信頼度は F 値が最低でも 50 前後になる範囲で，適合率が最大ににになるところを閾値とした．

訓練および評価事例には，3 節で評価したブログサイトを利用した．また，ヒット数 50 万以上，1 万から 50 万，1 万以下の各範囲で，ブログサイトの正例・負例が同数になるように調整した．そのため，ヒット数 50 万以上での訓練および評価事例は 102 個，ヒット数 1 万から 50 万では 652 個，ヒット数 1 万以下では 408 個となった．これらに対して，それぞれ 10 分割交差検定を行った．カーネル関数として，二次多項式カーネルを採用した．

6.2 評価結果

実験を行った結果を表 3 に示す．ヒット数 1 万以上のトピックに関してはいずれもベースラインよりより高い性能を達成している．これは，もともとのトピックにある程度のヒット数があり，関連語の情報も多く得られたためであると考えれられる．一方，ヒット数 1 万以下のトピックに対しては，相対的に性能が低くなったが，これは，もともとのヒット数が少ないために，関連語のヒット数などの情報を十分に得ることができなかったためであると考えられる．今後は，ヒット数が少ない範囲のトピックについても適合率を上げるために，Wikipedia の本文テキストの情報や，ブログサイトの記事単位の情報を素性として利用する．

6.3 トピックごとのブログサイトの有無の推定

本節では，機械学習によって得られたブログサイトの判定結果を用いてトピックごとのブログサイト有無推定を行った．6.1 節では，ヒット数 50 万以上，1 万〜50 万，1 万以下のどの範囲についても，正例，負例の数が 1 対 1 になるようにデータセットの調整を行った．その結果，ヒット数 50 万以上のデータセットでは，全 209 ブログサイト中 105 ブログサイトが用いられ，ヒット数 1 万〜50 万以上のデータセットでは，全 1150 ブログサイト中 652 ブログサイトが用いられ，ヒット数 1 万以下のデータセットでは，全 1125 ブログサイト中 408 ブログサイトが用いられた．一方，本節では，6.1 節で利用したデータセットで訓練したモデルを用いて，6.1 節では対象としなかったブログサイトも含めた全ブログサイトを評価対象とした．

さらに，全ブログサイトでの識別精度として，以下の評価値を求めるとともに，トピックごとに測定した識別精度の平均値を求めた．

$$\frac{|\{\langle b_e, c\rangle \in \text{評価事例集合} \,|\, \mathrm{sign}(f(b_e)) = c\}|}{|\,\text{評価事例}\ \langle b_e, c\rangle\ \text{の集合}\,|}$$

表 4: トピックに対するブログサイトの有無の自動判定: 評価結果 (ブログサイトあり=C1∼C3) (%)

(a) ヒット数 1 万以下のトピック (素性 1+8)

条件	ブログサイトのトピックの自動判定：識別精度			トピックに対するブログサイト有無の自動判定	
	正例:負例=1:1	全ブログサイト	全ブログサイト (トピックごとの平均)	識別精度	ブログサイトあり判定 (適合率/再現率/F 値)
ベースライン (ブログサイトあり=エントリ名のヒット数が 50 万以下)	–	–	–	71.0	71.0/100.0/**85.5**
信頼度の閾値なし	56.1	43.0	44.6	70.7	75.8/89.3/82.6
適合率 1 位 (閾値 1.0)	57.8	72.8	44.6	49.3	**88.0**/39.3/63.7
F 値 1 位 (閾値 0.3)	56.6	43.9	44.9	**72.0**	76.1/91.1/83.6

(b) ヒット数 1 万 ∼50 万のトピック (素性 3+8)

条件	ブログサイトのトピックの自動判定：識別精度			トピックに対するブログサイト有無の自動判定	
	正例:負例=1:1	全ブログサイト	全ブログサイト (トピックごとの平均)	識別精度	ブログサイトあり判定 (適合率/再現率/F 値)
ベースライン (同上)	–	–	–	62.0	62.0/100.0/**81.0**
F 値 1 位 (閾値なし)	64.6	72.3	65.6	**72.4**	73.1/88.4/80.8
適合率 1 位 (閾値 1.4)	61.9	79.5	67.0	53.6	**92.9**/30.2/61.6

(c) ヒット数 50 万以上のトピック (素性 3+6+8)

条件	ブログサイトのトピックの自動判定：識別精度			トピックに対するブログサイト有無の自動判定	
	正例:負例=1:1	全ブログサイト	全ブログサイト (トピックごとの平均)	識別精度	ブログサイトあり判定 (適合率/再現率/F 値)
ベースライン (同上)	–	–	–	59.0	0/0/0
F 値 1 位 (閾値なし)	76.5	80.5	78.2	75.0	70.0/100.0/**85.0**
適合率 1 位 (閾値 0.7)	76.7	84.2	74.0	**76.0**	**83.3**/71.4/77.4

表 5: トピックに対するブログサイトの有無の自動判定: 評価結果 (ブログサイトあり=C1,C2) (%)

(a) ヒット数 1 万以下のトピック (素性 1+8)

条件	ブログサイトのトピックの自動判定：識別精度			トピックに対するブログサイト有無の自動判定	
	正例:負例=1:1	全ブログサイト	全ブログサイト (トピックごとの平均)	識別精度	ブログサイトあり判定 (適合率/再現率/F 値)
ベースライン (表 4 と同じ)	–	–	–	19.0	19.0/100.0/59.5
F 値 1 位 (閾値なし)	56.1	43.0	44.6	82.7	52.4/78.6/**65.5**
適合率 1 位 (閾値 1.1)	57.1	77.7	65.1	**90.9**	**100.0**/7.1/53.6

(b) ヒット数 1 万 ∼50 万のトピック (素性 1+3)

条件	ブログサイトのトピックの自動判定：識別精度			トピックに対するブログサイト有無の自動判定	
	正例:負例=1:1	全ブログサイト	全ブログサイト (トピックごとの平均)	識別精度	ブログサイトあり判定 (適合率/再現率/F 値)
ベースライン (同上)	–	–	–	43.0	43.0/100.0/71.5
F 値 1 位 (閾値なし)	67.8	64.4	63.7	**76.8**	68.4/86.7/**77.6**
適合率 1 位 (閾値 1.0)	69.6	81.2	69.6	75.0	**100.0**/30.0/65.0

(c) ヒット数 50 万以上のトピック (素性 3+6+8)

条件	ブログサイトのトピックの自動判定：識別精度			トピックに対するブログサイト有無の自動判定	
	正例:負例=1:1	全ブログサイト	全ブログサイト (トピックごとの平均)	識別精度	ブログサイトあり判定 (適合率/再現率/F 値)
ベースライン (同上)	–	–	–	14.0	0/0/0
F 値 1 位 (閾値なし)	76.5	80.5	78.2	**91.7**	75.0/100.0/**87.5**
適合率 1 位 (閾値 1.8)	75.3	87.6	76.4	90.9	**100.0**/33.3/66.7

また，SVM によるブログサイトの判定結果を利用して，トピックに対するブログサイトの有無の自動判定を行った．また，トピックの判定には，表 1 と同じ条件を用い，トピックについて書かれたと判定されたブログサイトが 10 以上あれば C1，5 以上 10 未満あれば C2，1 以上 5 未満あれば C3，トピックについて書かれたと判定されたブログサイトがなければ E とした．本節では，C1，C2，C3 のいずれかの場合に，ブログサイトが存在すると判定する場合と C1 または C2 のいずれかの場合のみ，ブログサイトが存在すると判定する場合の二通りについて評価を行い，識別精度および，再現率・適合率・F 値を求めた．トピックに対するブログサイトの有無の自動判定における識別精度の式を以下に示す．

$$\frac{\text{ブログサイト有無の判定が正解したトピック数}}{\text{信頼度閾値以上のブログサイトが存在したトピック数}}$$

評価結果を表 4 および表 5 に示す．ただし，図 4 に示す，検索ヒット数のレンジごとの，ブログサイトの有無の分布を考慮して，ヒット数 50 万未満のトピックを正解，ヒット数 50 万以上のトピックを不正解としたものをベースラインとした．

C1∼ から C3 までをブログサイトありとする場合では，F 値などはベースラインと比較して十分な性能を達成できてはいないが，本研究のタスクにおいて重要である適合率は，ベースラインと比較して高い性能を達成している．また，Wikipedia エントリのトピックの分布をはかる場合には，C3 よりも C1,C2 の判定を正確に行う必要がある．C1，C2 をブログサイトありとする場合では，信頼度の閾値にかかわらず，80∼90%の識別精度を達成している．また，信頼度の閾値を設けることで，閾値を設けない場合と比較して高い適合率を達成している．今後，Wikipedia の本文テキストの情報やブログサイトの記事単位の情報などの素性を増やすことによって，更なる改善が期待される．

7 Wikipedia カテゴリ空間におけるブログサイトの分布の推定

7.1 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．

表 6: 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

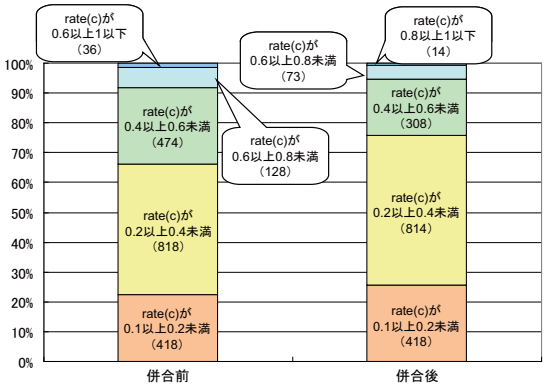


図 7: 併合前後の Wikipedia カテゴリの $rate(c)$ の分布 ($LBD_{rate} = 0.4$，() 内の数字はカテゴリ数)

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．

7.2 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

7.1 節より，カテゴリ c の持つエントリの絶対値が 11 以上[13]でヒット数 1 万から 50 万のエントリが存在する割合 (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

[13]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，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．

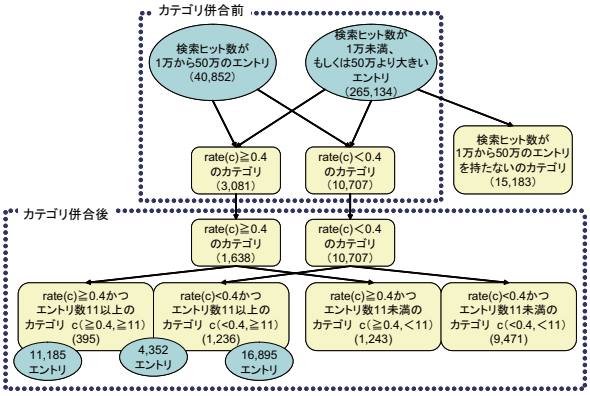


図 8: カテゴリ併合前後のカテゴリ数・エントリ数の推移 ($LBD_{rate} = 0.4$)

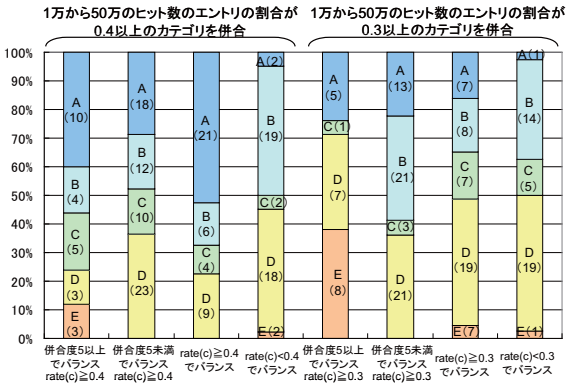


図 9: Wikipedia カテゴリに対応するブログサイトの有無の人手評価 (カテゴリ併合後，$LBD_{rate} = 0.3, 0.4$, () 内の数字はカテゴリ数)

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．
   $desc(c) \leftarrow desc(c) \cup desc(c')$

7.3 Wikipedia カテゴリに対応するブログサイトの有無の人手評価

7.2 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 6 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 7 に示し，カテゴリ・エントリ数の推移を図 8 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 9 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 7 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

表 7: 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

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．

8 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 [竹原 04] 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，TREC の 2007 年度の Blog Distillation タスク [Macdonald07] では，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．本研究のタスクにおいてもこれらのタスクで用いられた手法の適用を検討する予定である．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 [田村 07] 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．

9 おわりに

本論文では，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．さらにブログサイトのトピック判定の自動化を行うために，SVM を用いて各トピックの持つブログサイトの評価を行った．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．

参考文献

[川場 08a] 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による日本語ブログ空間のトピック分布推定, 情報処理学会研究報告, Vol. 2008, No. (2008–NL–187), pp. 83–90 (2008).

[川場 08b] 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).

[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).

[中崎 09] 中崎寛之, 川場真理子, 山崎小有里, 宇津呂武仁, 福原知宏：同一トピックの日英ブログにおける文化間差異の発見支援, DEIM フォーラム論文集 (2009).

[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).

[田村 07] 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).

[Vapnik98] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).