# Wikipediaをトピック体系とす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‡
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**アプローチをとる．

本論文では，[川場 08] 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．

## 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

*Estimating Topic Distribution of Japanese Blogsphere with Wikipedia as a Topic Hierarchy

†Mariko Kawaba, Hiroyuki Nakasaki, 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

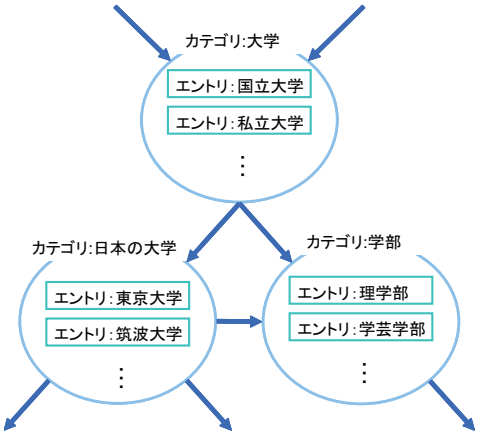


図 1: Wikipedia の構造

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．

### 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

[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．

[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．

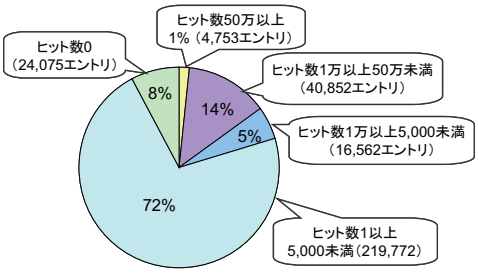


図 2: Wikipedia エントリにおけるブログヒット数の分布 (総数 305,986)

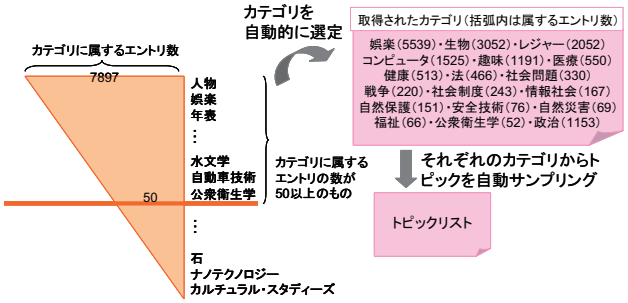


図 3: Wikipedia エントリのサンプリング手順

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．

### 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．

## 4 機械学習によるブログサイトのトピックの自動判定

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ることが可能であることがわかった．しかし，

[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

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 |

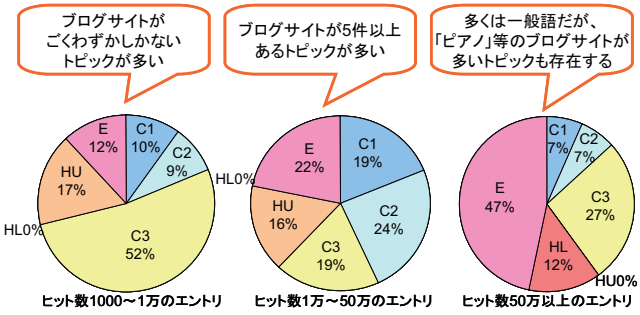


図 4: 各レンジの Wikipedia エントリの分布

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．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．

### 4.1 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[8]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

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 |

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

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 | 3(+6) | 55.8 | 71.6 | **63.7** |
| 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 | 1+3 | 66.3 | 72.0 | **69.2** |
| 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 | 3+6+8 | **87.5** | 65.0 | **76.3** |
| 適合率 1 位 (信頼度閾値 0.2) | 3+6 | 86.3 | 54.9 | 70.6 |

ヒット数に加えて，Wikipedia から得られる関連語を利用した．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[9]．

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．

[9]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．

### 4.2 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．

## 5 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．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．

## 6 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．また，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．

## 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).

[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).

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