DBSJ Journal, Vol.8, No.1
June 2009

Wikipedia 概念体系とブログ空間の間のトピック対応の推定

Estimating Topic Links between Wikipedia Topic Hierarchy and Blogosphere

川場 真理子♥　中崎 寛之♦
横本 大輔♦　宇津呂 武仁♠
福原 知宏#

Mariko KAWABA　Hiroyuki NAKASAKI
Daisuke YOKOMOTO　Takehito UTSURO
Tomohiro FUKUHARA

本論文では，ブログ空間に対してトピックの索引付けを行い，ブログ空間におけるトピック分布を推定することを目的として，Wikipedia 概念体系とブログ空間の間のトピック対応を推定する手法を提案する．まず，予備調査として，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．次に，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))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しその評価を行った．さらに，トピックごとに，そのトピックについて詳細な記述を含むブログサイトが存在するか否かを判定する方式の評価を行った．これらの評価のいずれにおいても，80%以上の適合率を達成することができた．

This paper studies the issue of conceptually indexing blogosphere through the whole hierarchy of Wikipedia entries. With such a conceptual indexing of blogosphere, this paper aims at estimating topic distribution in blogosphere based on the topic hierarchy in Wikipedia. More specifically, this paper proposes how to link Wikipedia entries to blog feeds in Japanese blogosphere, where about 300,000 Wikipedia entries are used for representing a hierarchy of topics. Furthermore, based on the results of judging whether each blog feed is relevant to a given Wikipedia entry, this paper also examines how to judge whether there exist blog feeds to be linked from the given entry. In our experimental evaluation, we achieved over 80% precision in those tasks.

♥ 非会員　日本電信電話株式会社 NTT サイバースペース研究所
♦ 学生会員　筑波大学大学院システム情報工学研究科博士前期課程
♦ 非会員　筑波大学第三学群
♠ 正会員　筑波大学大学院システム情報工学研究科
正会員　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，まず，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．ここで，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)[1])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しその評価を行った．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

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

判定を行い，その結果を評価した．これらの評価のいずれにおいても，80%以上の適合率を達成することができた．

2　評価対象の Wikipedia エントリ

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百科事典であり，日本語で約58万エントリ存在する(2009年4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においては，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．本論文で用いた版の日本語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．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700個を，第二層のカテゴリと定義する．

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．以上の手続きの結果，第一層もしくは第二層のカテゴリのうち，少なくとも一つのWikipediaエントリと対応付けられるカテゴリ約300個を選定した．

2.2　Wikipedia エントリの選定手順

Wikipediaのエントリを無作為に選んで，日本語ブログ空間におけるヒット数(3節で述べる大手11社のブログホストが対象)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．そこで，本論文では，Wikipediaエントリに対して，タイトルのヒット数が1,000から1万，1万から50万，50万以上の三つの範囲を設けて，各範囲ごとにWikipediaエントリを選定することとする[5]．

次に，前節で選定された300カテゴリを用いて，ヒット数の範囲が1万以下，1万から50万，50万以上の三つの範囲のそれぞれから，以下の手順でWikipediaエントリを選定する．まず，ヒット数の範囲が1万以下，および，1万から50万のエントリについては，300カテゴリを経由することにより，できる限り多様なカテゴリからエントリを選定する．具体的には，300カテゴリのうち，ヒット数1,000から1万のエントリを50個以上含むカテゴリを無作為に18個選択し，この18カテゴリから均等にエントリを無作為抽出することにより，合計で75エントリを選定した．同様に，300カテゴリのうち，ヒット数1万から50万のエントリを50個以上含むカテゴリを無作為に18個選択し，この18カテゴリから均等にエントリを無作為抽出することにより，合計69エントリを選定した．一方，ヒット数が50万以上のエントリについては，十分な数のカテゴリを経由したエントリ選定が困難なため，エントリを直接無作為抽出することにより，12エントリを選定した．

[5] 全 305,986 エントリのタイトルのヒット数の分布は，ヒット数 0 が約 8% (24,075 エントリ)，ヒット数 1 から 1,000 が約 56% (172,471 エントリ)，ヒット数 1,000 から 1 万が約 21% (63,835 エントリ)，ヒット数 1 万から 50 万が約 14% (40,852 エントリ)，ヒット数 50 万以上が約 1% (4,753 エントリ)となった．

3　評価対象のブログサイトの収集

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 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 ブログサイトを評価対象として選定した．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 社のブログホスト(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)を対象とした．なお，[2]では，Yahoo!Japan 検索 API の出力順の上位 20 ブログサイトと比較して，エントリ名の出現数順の上位 20 ブログサイトの方が，検索クエリとの関連性が高くなるという評価結果を示している．また，Wikipedia エントリ本文に含まれる関連語を用いることにより，検索性能が改善することが確認されている．

4　Wikipedia エントリのタイトルのヒット数とブログサイトの有無の相関

各トピックに対して収集された上位 20 ブログサイトに対してトピックの判定を行い，その結果に基づいて，各トピックに対して表 1 に示す 6 段階の判定ラベルのいずれかを付与した．ヒット数の範囲ごとに 6 段階の判定ラベルの分布を求めた結果を図 1 に示す．この結果からわかるように，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おり，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があることがわかった．判定が C1 となったエントリの例のうち，ヒット数 1 万から 50 万の範囲のものとしては,「ゴールドカード」,「バイオマス」,「不動産競売」,「サークル K サンクス」が，ヒット数 1,000 から 1 万以下の範囲のものとしては「三原山」,「カイツブリ目」,「WebSphere」,「ボホール島」，ヒット数 50 万以上のものとしては「ピアノ」がある．また，各エントリに対応するブログサイトの数は，ヒット数 1,000 から 1 万の範囲のエントリでは，1225 ブログサイト中 204 ブログサイト，ヒット数 1 万から 50 万の範囲のエントリでは，1150

ブログサイト中 326 ブログサイト，ヒット数 50 万以上のエントリでは，209 ブログサイト中 51 ブログサイトであった．

表 1 上位 20 ブログサイトを用いたブログサイトの有無の推定基準(上位 20 ブログサイト中)
Table 1 Estimating Existence of Blog Feeds: Criterion (with top ranked 20 feeds)

判定ラベル	説明
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 5～9 件
C3	トピックについて詳しいブログサイトが 1～4 件
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

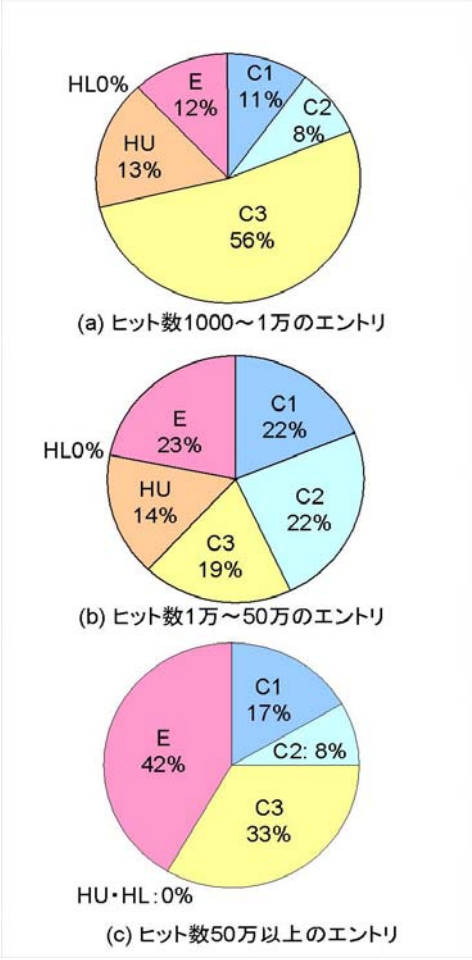

図 1 エントリ名のヒット数の範囲ごとの「ブログサイトの有無」の分布
Fig. 1 Estimating Existence of Blog Feeds: Distribution per Range of Hits of the Wikipedia Entry Titles

5 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．

5.1 学習および評価手順

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[6]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．

本節の評価における評価尺度としては，c=+ となる評価事例に対する適合率・再現率・F 値を用いる．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 35 前後になる範囲で，適合率が最大となるところを閾値とした．また，F 値が最大となる閾値における評価結果も別途示す．

訓練および評価事例としては，3 節で収集したブログサイトを用いる．特に，本節では，ブログサイトのトピック判定タスクにおける性能の上限値を推定することを目的として，ヒット数 1,000 から 1 万，1 万から 50 万，50 万以上の各範囲で，ブログサイトの正例・負例が同数になるように調整した．その結果，ヒット数 1,000 から 1 万での訓練・評価事例は 408 個，ヒット数 1 万から 50 万では 652 個，ヒット数 50 万以上では 102 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，線形カーネルおよび二次多項式カーネルの性能を比較したところ，二次多項式カーネルの方が高い性能が示したため，本論文では二次多項式カーネルを用いた場合の評価結果を示す．

5.2 素性

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．Wikipedia のエントリから得られる関連語としては Wikipedia 中の他のエントリへのリンクのアンカーテキスト，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 31 個の関連語が得られた．これらの関連語をクエリとして検索エンジン API を用いた検索を行い，各関連語の日本語ブログ空間内でのヒット数，および，各ブログサイト内でのヒット数を取得した．ここで，各関連語の日本語ブログ空間内でのヒット数を 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の一覧を表 2 に示す．ここで，t-hits，H-hits，M-hits，L-hits の各素性については，その

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

ヒット数(もしくは総和)を 5 段階の範囲に分けて，各範囲に該当するか否かを表す個別の二値素性を導入した．

表 2 ブログサイトのトピック自動判定のための素性一覧
Table 2 Features of Linking Wikipedia Entries to Blog Feeds

素性ラベル	説明
t-hits	エントリ名の当該ブログサイト内ヒット数
H-hits	H 関連語の当該ブログサイト内ヒット数の総和
M-hits	M 関連語の当該ブログサイト内ヒット数の総和
L-hits	L 関連語の当該ブログサイト内ヒット数の総和
H-num	H 関連語の種類数
M-num	M 関連語の種類数
L-hum	L 関連語の種類数
all-num	全関連語の種類数

表 3 ブログサイトのトピック自動判定の評価結果 (%)
Table 3 Linking Wikipedia Entries to Blog Feeds by SVM: Evaluation Results (%)

(a) ヒット数 1,000～1 万のエントリ

条件	素性	適合率/再現率/F 値
ベースライン	t-hits	62.4/32.6/42.8
F 値 1 位(信頼度閾値なし)	M-hits + H-num	63.0/67.0/**64.9**
適合率 1 位 (信頼度閾値 0.8)	t-hits + L-num	**84.2** /21.4/34.1
信頼度閾値なしの場合の適合率 1 位	M-hits	68.8/38.7/49.5

(b) ヒット数 1 万～50 万のエントリ

条件	素性	適合率/再現率/F 値
ベースライン	t-hits	60.2/77.0/67.5
F 値 1 位 (信頼度閾値なし)	t-hits + H-num	60.5/76.7/**67.6**
適合率 1 位 (信頼度閾値 0.9)	t-hits + M-num	**84.1**/23.0/36.1
信頼度閾値なしの場合の適合率 1 位	t-hits + all-num	70.5/54.2/61.3

(c) ヒット数 50 万以上のエントリ

条件	素性	適合率/再現率/F 値
ベースライン	t-hits	62.4/51.4/56.4
F 値 1 位 (信頼度閾値なし)	t-hits + all-num	82.6/75.5/**78.9**
適合率 1 位 (信頼度閾値 0.5)	t-num + M-hits + M-num + all-num	**90.0**/60.5/72.4
信頼度閾値なしの場合の適合率 1 位		88.5/67.9/76.8

5.3 評価結果

評価結果を表 3 に示す．ここで，ベースラインとしては，素性として t-hits(日本語ブログ空間におけるエントリ名のヒット数)のみを用いた場合の性能を用いた. 適合率一位の場合，および，F 値一位の場合のいずれにおいても，適合率・F 値ともベースラインを上回る性能を達成した．特に，ヒット数の範囲が 1,000～1 万のエントリ，および，50 万以上のエントリの場合には，エントリ名のヒット数自身の情報量が十分ではないため，提案手法によりベースラインの性能を大きく改善することができた．また，信頼度閾値を用いた場合には，80～90%の適合率を達成することができた．今後は，Wikipedia エントリ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ことにより，さらに性能を改善することを試みる．

6 トピックごとのブログサイトの有無の推定

6.1 評価手順

次に，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．5.1 節ではヒット数 1,000 から 1 万，1 万から 50 万，50 万以上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 1,000 から 1 万のデータセットでは，全 1225 ブログサイト中 408 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 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た．一方，本節では，5.1 節で利用したデータセットで訓練したモデルを用いて，5.1 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，SVM によるブログサイトの判定結果を利用して，トピックごとのブログサイトの有無の自動判定を行った．ここで，トピックごとのブログサイトの有無の判定においては，**表 1** の基準を用い，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 件以上あれば C1，5～9 件あれば C2，1～4 件あれば 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った．評価尺度としては，トピック単位の識別精度，および，「ブログサイトが存在する」トピックに対する適合率・再現率・F 値を用いた．これらの評価尺度の定義を以下に示す．ここでも，前節と同様に，適合率・F 値が最大となる素性・閾値における性能を抜粋して示す．

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

$$\text{適合率} = \frac{\text{自動判定＝「ブログサイトあり」となり，かつその判定が正解したトピック数}}{\text{信頼度閾値以上のブログサイトが存在し，自動判定＝「ブログサイトあり」となったトピック数}}$$

$$\text{再現率} = \frac{\text{自動判定＝「ブログサイトあり」となり，かつその判定が正解したトピック数}}{\text{人手による正解＝「ブログサイトあり」であるトピック数}}$$

6.2 評価結果

評価結果を表 4 および表 5 に示す. ただし, 図 1 に示す, 検索ヒット数の範囲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 万以下のトピックの場合は「ブログサイトあり」，ヒット数 50 万以上のトピックの場合は「ブログサイトなし」とした場合の性能をベースラインとした.

表 4 および表 5 のいずれにおいても，信頼度の閾値を設けることで，95～100%の適合率を達成している．F 値については，C1～C3 までをブログサイトありとする場合では，ベースラインと比較して十分に高い性能を達成できてはいないが, C1 および C2 をブログサイトありとする場合の評価結果では，ベースラインを大幅に改善できた．また，識別精度については, 信頼度の閾値を設けることにより, 85～100%の性能を達成している．また，素性としては，M-hits を含む素性の組み合わせにおいて高い性能となる傾向がある．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追加することによって，更に性能が改善することが期待される.

表 4 トピックごとのブログサイトの有無の推定の評価結果 (ブログサイトあり=C1～C3) (%)

Table 4 Binary Judgment on Existence of Blog Feeds to be Linked from an Wikipedia Entry: Evaluation Results ("Blog Feeds Exist = C1～C3) (%)

(a) ヒット数 1,000～1 万のエントリ

条件	素性	適合率/再現率/F 値	識別精度
ベースライン	－	74.7/100.0/85.5	74.7
F 値 1 位/信頼度閾値なしの場合の適合率 1 位	L-hits	79.1/94.6/**86.0**	77.3
適合率 1 位 (信頼度閾値 1.3)	t-hits + M-num	**95.5**/38.2/54.0	**83.0**

(b) ヒット数 1 万～50 万のエントリ

条件	素性	適合率/再現率/F 値	識別精度
ベースライン	－	62.3/100.0/76.8	62.3
F 値 1 位	M-hits + H-num	69.0/93.0/**79.0**	69.6
適合率 1 位 (信頼度閾値 2)	M-hits + M-num + all-num	**100.0**/32.6/49.0	**87.0**
信頼度閾値なしの場合の適合率 1 位	M-hits + M-num (+ all-num)	71.2/86.0/77.0	69.6

(c) ヒット数 50 万以上のエントリ

条件	素性	適合率/再現率/F 値	識別精度
ベースライン	－	0/0/0	41.7
F 値 1 位/信頼度閾値なしの場合の適合率 1 位	(t-hits + M-hits+) M-num (+ all-num)	87.5/100.0/**93.0**	91.7
適合率 1 位 (信頼度閾値 1.4)	M-hits + L-num	**100.0**/71.4/83.0	**100.0**

表 5 トピックごとのブログサイトの有無の推定の評価結果 (ブログサイトあり=C1,C2) (%)

Table 5 Binary Judgment on Existence of Blog Feeds to be Linked from an Wikipedia Entry: Evaluation Results ("Blog Feeds Exist = C1,C2) (%)

(a) ヒット数 1,000～1 万のエントリ

条件	素性	適合率/再現率/F 値	識別精度
ベースライン	－	18.7/100.0/31.5	18.7
F 値 1 位(信頼度閾値 1)	t-hits + all-num	60.0/42.9/**50.0**	87.3
信頼度閾値なしの場合の F 値/適合率 1 位	H-hits のみ or t-hits + M-hits + M-num + all-num	35.5/78.6/48.0	69.3
適合率 1 位 (信頼度閾値 2)	M-hits + M-num + all-num	**100.0**/14.3/25.0	**98.4**

(b) ヒット数 1 万～50 万のエントリ

条件	素性	適合率/再現率/F 値	識別精度
ベースライン	－	43.5/100.0/60.6	43.5
F 値 1 位(信頼度閾値 0.4)	t-hits + H-hits	59.1/86.7/**70.0**	68.1
信頼度閾値なしの場合の F 値 1 位	t-hits + H-hits	56.5/86.7/68.0	65.2
適合率 1 位 (信頼度閾値 1)	M-hits + L-num	**100.0**/33.3/50.0	**84.5**
信頼度閾値なしの場合の適合率 1 位	t-hits + all-num	66.7/53.3/59.0	68.1

(c) ヒット数 50 万以上のエントリ

条件	素性	適合率/再現率/F 値	識別精度
ベースライン	－	0/0/0	75.0
F 値/適合率 1 位(信頼度閾値 0.9～1.3)	t-hits , M-hits, M-num の いずれか 2 つ	**100.0**/100.0/**100.0**	**100.0**
信頼度閾値なしの場合の F 値/適合率 1 位	t-hits/M-hits + M-num	60.0/100.0/75.0	83.3

7 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[3]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[4]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[5]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 に関する研究としては，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]などがある．

また，情報検索分野における関連研究として，階層型ディレクトリに対して文書分類を行う手法の研究[7][8][9]がある．これらの研究では，階層型の各ディレクトリは，単なるラベルで表現されているか，もしくは，キーワードの列として表現されている．また，手法的には，それらの階層的に配置された複数のディレクトリ情報，および，各ディレクトリに分類済みの教師文書等を用いた機械学習手法が適用されている．一方，本研究の範囲では，単一の Wikipedia エントリに対して，ブログサイトのトピックの対応の有無を判定している．本研究では，特に，Wikipedia エントリ中において活用すべき関連語等の知識の有用性に焦点を当てた．また，本研究では，一定の方式で収集したブログサイトに対するトピック対応の有無の判定にとどまらず，各トピックに対応するブログサイトの有無の判定までを研究の対象とした．

8　おわりに

本論文では，ブログ空間に対してトピックの索引付けを行い，ブログ空間におけるトピック分布を推定することを目的として，Wikipedia 概念体系とブログ空間の間のトピック対応を推定する手法を提案した．まず，予備調査として，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を示した．次に，Wikipedia エントリから得られる知識を素性として，SVM 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しその評価を行った．さらに，トピックごとに，そのトピックについて詳細な記述を含むブログサイトが存在するか否かを判定する方式の評価を行った．これらの評価のいずれにおいても，80%以上の適合率を達成することができた．この成果の活用法の一つとして，[10][11]では，Wikiipedia カテゴリ単位でのブログサイトの有無を推定する手法およびその有効性を示している．今後は新たな素性として，Wikipedia の本文テキストや，各ブログサイト記事単位の情報等の有効性を検証する．

[文献]

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

[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] 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)．

[4]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)

[5]Macdonald, C., Ounis, I.and Soboroff, I.: Overview of the TREC-2007 Blog Track. Procedings of TREC-2007 (Notebook), pp. 31–43, (2007).

[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)

[7]Dumais, S.. and Chen, H. Hierarchical Classification of Web Content. Proceedings of the 23rd SIGIR, pp. 256–263, (2000).

[8]Sun, A. and Lim, E.-P.. Hierarchical Text Classification and Evaluation. Proceedings of ICDM, pp. 521–528, (2001).

[9]Adami, G. Avesani, P. and Sona., D.:Clustering Documents in a Web Directory. Proceedings of WIDM, (2003)

[10]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)

[11]川場真理子，中崎寛之，宇津呂武仁，福原知宏．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．人工知能学会研究会資料, SIG-SWO, (2009)

川場　真理子　Mariko KAWABA

日本電信電話株式会社 NTT サイバースペース研究所研究員．2009 筑波大学大学院システム情報工学研究科博士前期課程修了．自然言語処理・情報検索の研究・開発に従事．情報処理学会会員．

中崎　寛之　Hiroyuki NAKASAKI

筑波大学大学院システム情報工学研究科博士前期課程在学中．2008 筑波大学第三学群卒業．自然言語処理・情報検索の研究・開発に従事．日本データベース学会学生会員．

横本　大輔　Daisuke YOKOMOTO

筑波大学第三学群在学中．自然言語処理・情報検索の研究・開発に従事．

宇津呂　武仁　Takehito UTSURO

筑波大学大学院システム情報工学研究科准教授．1994 京都大学大学院工学研究科博士後期課程修了，博士(工学)．自然言語処理・情報検索の研究・開発に従事．日本データベース学会会員．

福原　知宏　Tomohiro FUKUHARA

東京大学人工物工学研究センター特任助教．2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程 単位取得認定退学. 博士(情報工学). 多言語ブログからの関心分析，スパムブログの分析等の研究・開発に従事．日本データベース学会会員．