Vol.2012-DBS-155 No.13
2012/11/19

カテゴリに対する所属度と典型度を考慮した希少なWebページの発見

多田 亮平　湯本 高行　新居 学　佐藤 邦弘

概要：
本研究ではユーザが指定したカテゴリから，有用で非典型的な Web ページを希少な Web ページとし，発見する手法を提案する．本研究では，カテゴリに該当するページを有用であると考え，提案手法では所属度と非典型度の２つを用いて希少な Web ページを発見する．所属度は Web ページがカテゴリに該当する度合いで，予め用意した Web ページとカテゴリの対応データから計算した，語とカテゴリの関係の強さを用いて算出し，適切なしきい値を決定する．非典型度は Web ページがカテゴリ内でどのくらい非典型的かを表す指標である．カテゴリ内で類似する Web ページが多いほど典型的である観点と，カテゴリ内での出現頻度が小さい語を含むほど非典型的である観点の２つを検討した．

1. はじめに

近年，検索システムの精度向上により，ユーザは入力したキーワードに関する典型的な Web ページは得られるようになってきている．その一方で，ユーザが非典型的な情報を探すことは依然として難しい．なぜなら，検索システムにはキーワードを入力する必要があるが，非典型的な情報に対するキーワードは思いつくことが難しく，検索が困難なためである．また，既知のキーワードで検索を行い，その検索結果から多くの Web ページを閲覧することで非典型的な情報を見つけることができるが，ユーザの負担は大きい．

そこで本研究では，有用かつ非典型的な情報を記載した Web ページを**希少な Web ページ**と定義し，ユーザが指定したカテゴリから希少な Web ページを発見する手法を提案する．情報の有用性については様々な観点から考えることができるが，本研究では「ユーザが指定したカテゴリに該当している状態」を有用と考え，所属度として表す．また，非典型性については先行研究の観点を参考に定義し，非典型度として表す．本研究では，希少な Web ページをこれら２つの指標を用いて発見する．

所属度とは，Web ページがユーザの指定したカテゴリに該当する度合いを示す．典型性と所属度の関係を考えたとき，カテゴリ内で典型的な Web ページほど，そのカテゴリに該当する度合いが大きいと考えられる．そのため，非典型的な Web ページにはカテゴリに該当しない Web ページが混在する可能性が大きくなる．ユーザが指定したカテゴリに該当しない Web ページはユーザの検索の意図とは関係がなく，有用でないといえる．そこで，まず指定したカテゴリへ該当することを有用であると考え，Web ページの有用性を所属度を用いて表す．算出方法は，Web ページとカテゴリを対応付けたデータセットを予め用意し，これを用いて Web ページに出現する語とカテゴリの関係の強さを計算する．この語とカテゴリの関係の強さをもとに所属度を算出する．この値によって有用な，指定したカテゴリに該当する Web ページのみを取得する．

次に非典型度とは，Web ページがカテゴリ内でどのくらい非典型的か (典型的でないか) を表す指標である．非典型度については２つの観点から指標を検討する．１つ目は，指定したカテゴリ内で類似する Web ページが多いほど典型的である観点を用いる．２つ目は，指定したカテゴリ内で Web ページの本文中にカテゴリと関係の強い語が多く含まれるほど典型的という観点である．この２つの観点から算出する指標を検討し，非典型度を算出する．

以上の所属度と非典型度を用いることで，希少な Web ページを発見する．提案手法では入力として，ユーザの指定する「カテゴリ」と，提案手法によって指標を算出する対象の「Web ページ集合」を用いる．特に入力に用いる Web ページ集合については希少な Web ページを含んでいる必要があり，この条件を満たすならば RSS の新着記事などを用いてもよい．本研究ではソーシャルブックマークサービスに登録された Web ページ集合を用いる．様々なユーザがブックマーク登録しているデータには希少な Web ページも含まれていると考える．

提案手法では，入力の Web ページ集合に対して所属度を算出する．この所属度の値によって，指定したカテゴリに該当する Web ページ集合のみを取得する．次に，取得した指定カテゴリに該当する Web ページ集合に対して，非典型度を算出する．この順で処理を行うことにより，指定カテゴリに該当する有用，かつカテゴリ内で非典型的な Web ページを「希少な Web ページ」として取得することができる．

2. 関連研究

典型性に基づく情報検索の研究として，佃らの典型度を用いたオブジェクトの検索手法がある [1]．この研究では認知心理学の分野で提案された典型性の観点 [2][3] を参考に，オブジェクトに対する典型度を算出している．佃らはオブジェクトに対する典型度を，「オブジェクトに対する人の認知による観点」と「オブジェクト自体の性質に依存する観点」の 2 つから定義している．本研究の提案手法では，佃らによる「オブジェクトの内容の類似性を用いた典型度の算出方法」と同様の指標を算出し，検討する非典型性を測る指標の 1 つとして用いる．

次に，典型性に対する先行研究 [4],[5] をもとに，様々な観点からオブジェクトの典型性を整理した研究として，藤坂らの研究がある [6]．この研究では，カテゴリと典型性の観点について関係性を明らかにするため，Web を用いた重要な観点の抽出手法を提案している．また，認知心理学や情報推薦において定義された典型性をまとめ，8 つの観点から定義している．本研究ではこの典型性の観点の一部を参考にし，典型性の指標を算出する．また後述する想起率の考えを用いて，評価実験を行った．

これらの研究を参考に，オブジェクト自体の性質に依存した観点と人の認知による観点の 2 つから典型度を考える．まず，オブジェクトの性質を基準とした観点として，central tendency[1] がある．これは，あるカテゴリ内において類似するオブジェクトが多いオブジェクトほど，典型的とする観点である．次に，人の認知を基準とした観点として，frequency of instantiation[1] と想起率 [6] がある．frequency of instantiation は，あるカテゴリ内において人が遭遇しやすいオブジェクトほど典型的とする観点である．また想起性は，あるカテゴリ内において人が思い浮かべやすいオブジェクトほど典型的であるという観点である．

本研究では指定したカテゴリに該当する，希少な Web ページを発見するため非典型度を用いる．本研究では特に「オブジェクト自体の性質に依存する観点」から典型性の指標を算出し，人の認知による観点は非典型度を算出するには不向きと考えたため使用しない．なぜなら，カテゴリ内において人が遭遇しにくいという指標や，思い浮かべにくいという指標に関してはデータが得られないからである．そのため，本研究ではオブジェクト自体の性質に依存する観点のみを用いて，非典型度の指標を算出する．

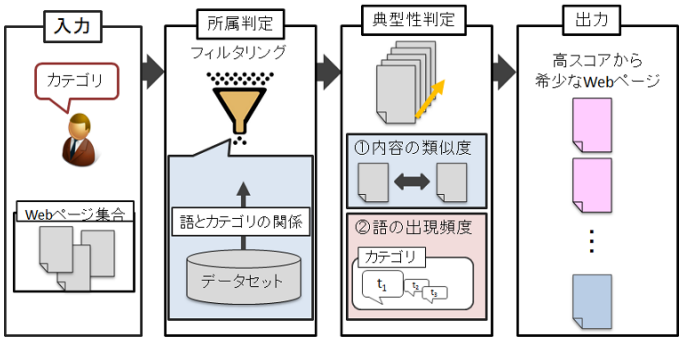

図 1 提案手法の全体像

3. 希少な Web ページ

3.1 希少な Web ページの発見

Web ページの希少性を表すには，以下の 2 つの条件が必要だと考えられる．

(a) 有用である

(b) 非典型的な内容を含む

まず (a) について説明する．1 章で述べたように，指定カテゴリに該当する状態を有用である考え，Web ページの内容が指定カテゴリに該当する度合いを「**所属度**」とする．所属度の算出方法については 4 章に示す．

次に，(b) については「**非典型度**」を用いて表現する．非典型度については 2 つの観点から指標を算出し，検討する．1 つ目の観点は Web ページの本文の類似度に注目し，2 章で述べた佃らの提案した典型性の指標を用いる．Web ページの非典型度は，指定したカテゴリ内の他の Web ページと類似するほど小さいとする．2 つ目は語の出現頻度に基づく指標を用いる．Web ページの非典型度は，出現頻度が全体に比べて指定したカテゴリ内では小さい語を多く含むほど大きいと考える．検討する非典型度の算出方法はそれぞれ，5 章に示す．

これらの「所属度」，「非典型度」を用いて希少な Web ページを発見する処理の概要を図 1 に示す．まず所属度を用いて，ユーザの指定カテゴリに該当する Web ページ集合のみを取得する．Web ページ本文中の語に着目し，指定カテゴリと関係がない Web ページを排除する．次に非典型度を算出し，非典型度の値が大きい順に各 Web ページをランキングする．他の Web ページと違う内容が記載された Web ページの典型度は小さくなるため，上位の Web ページほど非典型的になる．

3.2 希少な Web ページの例

例として “チョコレート”カテゴリの Web ページについて考える．チョコレートに関する多くの Web ページは，販売店舗や有名なチョコレートの商品に対するレビューが記述されていると考えられる．しかしチョコレートに関す

る希少な Web ページの場合，これらのチョコレートのカテゴリの Web ページとは明らかに違う内容を記述している．チョコレートのカテゴリ内での希少な Web ページとしては，「明太子ロッシェ」について記述しているページが挙げられる．明太子ロッシェは明太子とチョコレートを組み合わせた銘菓であり，多くの人は知らないといえる．またチョコレートに関する Web ページの中でも，他とは違う「明太子を使用している」という内容について記述している．

4. Web ページのカテゴリへの所属判定

所属判定では，Web ページが指定カテゴリに該当するかを判別する．まず，Web ページとカテゴリを対応付けたデータセットを用いて，Web ページ本文中の語 t とカテゴリ c の関係の強さを関連度 $Relevance(c,t)$ として算出する．次に，この関連度を用いて入力の Web ページ p が指定カテゴリ c へ該当する度合いを所属度 $Belong(c,p)$ として算出する．これによって，Web ページ p が指定カテゴリ c と関係の強い語を含むほど，所属度 $Belong(c,p)$ は大きくなる．

4.1 語と指定カテゴリの関連度

まず，データセットを用いて指定カテゴリ c と語 t の関係の強さを算出する．本研究では，Web ページとカテゴリを対応付けた学習用のデータとして，ソーシャルブックマークサービス (以下，SBM サービス) から収集したデータセットを用いる．

SBM サービスとは，サービスを利用するユーザが「Web ページに対するブックマーク」をオンライン上で管理できる Web サービスである．本研究では，このサービス上でのブックマーク情報から，特にブックマークされた「Web ページ」と「タグ」をデータセットに用いる．ユーザは複数のタグをブックマークに付与することで，ブックマークした Web ページがどんな情報かを簡単に分類することができる．一般的に，タグには単語や短い文が使用され，多くのユーザは「カテゴリ」としてタグを使用している．例えば Ruby のプログラミングに関する Web ページに対しては，多くのユーザは “プログラミング (programming)” や “Ruby” をタグとして使用している．そこで SBM サービス上のタグをカテゴリとして考え，関連度の算出に用いる．

タグ c が使用された Web ページの本文中に，t が存在する確率が大きいほど，c と t の関連度は大きいと考える．そこで，c が付けられた Web ページ集合中で本文中に t を含む Web ページの数を $DF_c(t)$，データセット全体で本文中に t を含む Web ページの数を $DF_{all}(t)$ とする．この 2 つを用いて，関連度を (1) 式のように定義する．

$$Relevance(c,t) = \frac{DF_c(t)}{DF_{all}(t)} \tag{1}$$

この定義より，カテゴリ c 内だけで多く使用される語 t については，$Relevance(c,t)$ が大きくなり，語とカテゴリの関係を表すことができる．

しかし，(1) 式による関連度の値は，カテゴリ c ごとに大きく変わる可能性があるため正規化する必要がある．そこで，正規化した関連度 $nRelevance(c,t)$ を (2) 式に示す．データセット中の全 Web ページ集合に出現する，全ての語の集合を $T = \{t_1, t_2, ..., t_n\}$ とし，$t \in T$ のそれぞれの関連度 $Relevance(c,t)$ を，関連度が最大値を取る語 t_{max} の関連度 $Relevance(c,t_{max})$ で正規化する．

$$nRelevance(c,t) = \frac{Relevance(c,t)}{Relevance(c,t_{max})} \tag{2}$$

この関連度 $nRelevance(c,t)$ を用いて，Web ページの所属度を算出する．

4.2 Web ページのカテゴリへの所属度

Web ページ p がカテゴリ c に該当するほど，p の本文中に c と関連の強い語を多く含むと考える．“プログラミング” のカテゴリにおける例を述べる．p の本文中にプログラミングに関する話題がない場合，p はプログラミングに該当しないといえる．しかし，p の本文中にプログラミングに関する話題がある場合，プログラミングに関する話題が多いほど p はプログラミングに強く該当するといえる．そこで，「カテゴリ c に関する話題」を「c の話題を構成する名詞」で表現し，c と関連度 $Relevance(c,t)$ が大きい語の数を指標として用いる．

p の本文中に含まれる語の集合を $T(p)$ とし，$T(p)$ 中で c との関連度がしきい値 θ_r 以上の名詞の数を所属度とする．Web ページ p のカテゴリ c への所属度 $Belong(c,p)$ を (3) 式に示す．

$$Belong(c,p) = |\{t_i | nRelavance(c,t_i) \geqq \theta_r \\ , t_i \in T(p)\}| \tag{3}$$

この所属度が θ_b 以上の Web ページ集合を，カテゴリ c に該当する Web ページ集合 P_c として非典型度を算出する対象とする．

5. Web ページの非典型度

ここでは，2 章で述べたようにオブジェクトの性質に基づく観点から Web ページの非典型性を算出する．Web ページ p が指定カテゴリ c に該当する他の Web ページと，どのくらい違う内容を記述しているかを非典型度を用いて算出する．

非典型度については，次の 2 つの観点から算出した指標を検討する．1 つ目は，Web ページの本文の類似度を用いた指標である．c に該当する Web ページ間の類似度を算出し，他と類似するページが多いほど非典型度は小さくな

る．2 つ目は，c と p 本文中に含まれる語の出現頻度を用いた指標である．p の本文が c と関係の弱い語を多く含むほど非典型度が大きくなる．

5.1 本文の類似度に基づく指標

2 章で述べた佃らの手法を用い，c に該当する Web ページ p_c の非典型性の指標を算出する．

まず，各ページ間の類似度を算出する．c に該当する Web ページ集合 P_c の本文をそれぞれ解析し，p_c を「語と TF-IDF 値」を要素としたベクトルとして表現する．このとき，IDF 値は P_c ではなく，データセットに含まれる全 Web ページ集合 P_{all} の範囲で以下のように算出する．

$$IDF(t) = \frac{1}{DF_{all}(t)} \tag{4}$$

また，p_c 本文中の全ての語を用いてベクトルとした場合，語の数が多い Web ページほど類似度が大きくなってしまう．そのため，TF 値が上位 N 件の名詞のみを用いて特徴ベクトルを作成する．

このベクトルを用いて，P_c 中のすべての Web ページの組合せについてコサイン類似度を算出する．算出した類似度を用い，行と列がそれぞれの Web ページに対応した類似度行列を作成し，行方向成分について総和が 1 となるように正規化を行うことで確率遷移行列 $\mathbf{S}$ とする．この $\mathbf{S}$ について PageRank[8] の式を適用することで，c 内における TextRank[9] のスコア $\mathbf{TR_c}$ を (5) 式で算出する．このとき，$\mathbf{u}$ は各ページをランダムに選択する確率ベクトルであり，要素数は $|P_c|$，ベクトルの要素はそれぞれ $\frac{1}{|P_c|}$ とする．

$$\mathbf{TR_c} = d \cdot \mathbf{S} \times \mathbf{TR_c} + (1-d) \cdot \mathbf{u} \tag{5}$$

この $\mathbf{TR_c}$ から，c 内における p_c の非典型性の指標 $Atypicality_{TR}(c, p_c)$ を算出する．(5) 式から，$\mathbf{TR_c}$ は典型的な度合いであり，0 から 1 で表される．そこで p_c に対応する TextRank のスコアを $TR_c(p_c)$ とし，以下の式で非典型度の指標 $Atypicality_{TR}(c, p_c)$ を算出する．

$$Atypicality_{TR}(c, p_c) = 1 - TR_c(p_c) \tag{6}$$

5.2 カテゴリ内での語の出現頻度に基づく指標

この手法では，p が「c 内で記述されることが少ない話題を，どのくらい含んでいるか」という点に着目する．p_c が c と関連が小さい語を含むほど c 内で記述される確率が小さい話題を含んでいると考え，本文に対して関連の小さい語が占める割合が大きいほど非典型的な Web ページであるとする．

まず，p_c の本文から名詞集合 $T(p_c)$ を取り出し，この語の数を $num_t(p_c)$ とする．

$$num_t(p_c) = |T(p_c)| \tag{7}$$

次に，4.1 節で算出した関連度 $nRelevance(c, t)$ を用いて，c と関連の小さい語を判別する．語の集合 $T(p_c)$ 中の DF_{all} が M 以上の語に対して関連度を算出し，語の関連度が θ_{LF} 以下の語の数を $num_{LF}(p_c)$ とする．DF_{all} による制限は，あまり出現しない語は単に誤字である場合や，重要な意味を持たない可能性があるからである．

$$num_{LF}(p_c) = |\{t | t \in T(p_c), DF_{all}(t) \geqq M, \\ nRelevance(c, t) \leqq \theta_{LF}\}| \tag{8}$$

これら 2 つの語の数を用いて，次式で「カテゴリと関連の小さい語」を含む割合を算出する．

$$Atypicality_{LF}(c, p_c) = \frac{num_{LF}(p_c)}{num_t(p_c)} \tag{9}$$

6. 実験

まず，SBM サービスからデータを収集してデータセットを作成し，これを用いて語とカテゴリの関連度を算出した．関連度を用いて各ページの所属度を算出し，指定カテゴリ c に該当する Web ページのみを取得する所属度のしきい値を決定した．これによって c に該当する Web ページのみを取得するフィルタリングを行った．

次に，所属度を用いて取得した，c に該当する Web ページ集合 P_c に対して希少な Web ページを発見する実験を行った．P_c に対して 5.1，5.2 節の手法を適用することで各 Web ページをランキングし，スコアが上位の Web ページに対して「カテゴリに該当するか」，「典型的か」をそれぞれ人手と被験者へのアンケートを元に判定した．これによって，カテゴリから希少な Web ページを得られるかの評価を行った．

6.1 データの収集手法

既存の SBM サービスのデータ収集手法では，多くブックマークされている Web ページや，多くの人が使用するタグを収集していた [7]．しかし，この方法では多くのユーザがブックマークしているような，典型的な Web ページの情報しか得られない．本研究では，カテゴリ内で希少な Web ページを発見することを目的としているため，ブックマーク数が少ない Web ページも取得する手法をとる．

まず，はてなブックマークが提供する，最新・注目記事の HotEntry に注目する．HotEntry は各トピックに対して分野が分かれている．今回の実験では { 社会，政治・経済，スポーツ・芸能・音楽，科学・学問，コンピュータ・IT，ゲーム・アニメ，おもしろ } を用いる．上記の各トピックの RSS 情報を基点として，下の (1)〜(4) の工程を行うことでブックマーク情報を収集する．

(1) RSS 情報[*1]を用い，各トピックの HotEntry 上位 30 件のタイトル，URL を取得する．

[*1] http://b.hatena.ne.jp/hotentry/*Topic*.rss

(2) 30 件の URL 集合から，それぞれの URL を登録しているブックマーク情報 (ユーザ ID，URL，タグ，ブックマーク日時，コメント) をすべて取得し[*2]，データベースへ格納する．

(3) 取得したブックマーク集合からユーザ集合を取得し，各ユーザのブックマークした URL[*3]の最新上位 100 件を取得する．

(4) その URL 集合に対して（2）の処理をもう一度繰り返す．

以上のように検索対象を URL からユーザ，ユーザから URL と変えることで，ブックマーク数が多いページだけでなく，少ないページも取得する．

上記の手法によって，はてなブックマークからブックマーク情報の収集を行った．収集期間は 2011 年 4 月 14 日から，2011 年 10 月 27 日までの期間である．また，手法によって取得した URL の Web ページから，ExtractContent[*4]を用いて本文を抽出した．提案手法では本文の情報が必要であるため，ExtractContent によって本文が取得できた Web ページに関してのブックマーク情報をデータセットとした．その規模を表 1 に示す．

表 1 データセットの規模

データセット情報	件数・種類数
Web ページ数	11874
総単語数	196220
タグ総種類数	28996

6.2 所属度の適切なしきい値の決定

まず，所属度の計算に必要である，語とカテゴリの関連度を算出した．このとき，評価実験を行う上で使用したカテゴリを表 2 に示す．4.1 節の手法を用いて全ての語の集合 T 中の語 t と，表 2 中の各カテゴリ c の関連度 $nRelevance(c,t)$ を算出した．

表 2 入力に使用するカテゴリ

政治	経済
アニメ	ゲーム
音楽	映画
テレビ	エネルギー
twitter	google

次に，4.2 節の手法を用いて所属度を算出し，指定カテゴリの Web ページのみを取得するしきい値を決定する．関連度は 0.5 以上の語は関連があると考え，しきい値 θ_r を 0.5 として所属度を計算した．さらに，c に該当する Web ページ集合と，該当しない Web ページ集合に対して所属度を算出した．このとき，それぞれの Web ページ集合について所属度の分布を比較することで，適切なしきい値を決定する．

具体的には，データセット中でタグ c を付けられた Web ページ集合を取得し，この集合の半分を関連度の算出用 P^{comp}，もう半分を所属度の検証用 P^{test} に分けた．c に該当する Web ページ集合には P^{test}，該当しない Web ページ集合にはタグ c を付けられていない Web ページ集合 P^{other} を用い，P^{test} と同じ数だけ取得した．

次に，P^{comp} から計算した関連度を用いて P^{test}，P^{other} のそれぞれの Web ページに対して所属度を算出した．このとき，P^{test} の Web ページは所属度は大きく，P^{other} の Web ページは所属度は極端に小さくなっているはずである．よって，Web ページの所属度の分布を P^{test} と P^{other} で比較することで適切なしきい値を決定する．

各カテゴリと，所属度の分布を表 3 に示す．また，例として c を “音楽”としたときの所属度と所属度に対応する Web ページ数の関係を図 2 に示す．各要素は%表示とする．

表 3 カテゴリと所属度の分布

関連語数	0	1	2	3 以上
音楽	17	52	26	5
経済	12	63	19	5
映画	26	18	35	22
政治	59	28	12	0
テレビ	16	70	13	1
ゲーム	14	65	17	4
エネルギー	22	57	19	2
アニメ	14	75	10	2
twitter	13	58	24	6
google	12	35	38	14

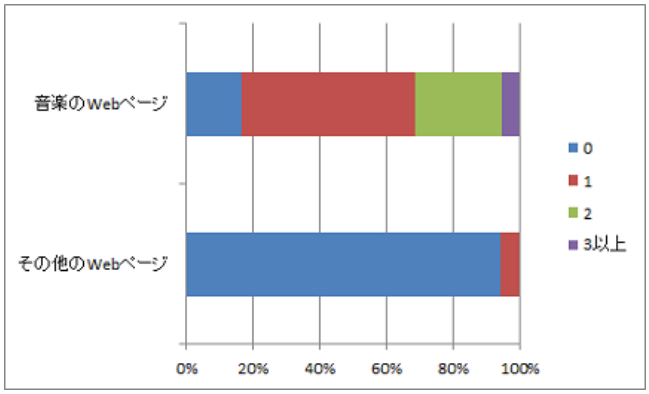

図 2 音楽カテゴリ内外の Web ページ集合に対する所属度の比率

図 2 では色によって所属度を表しており，青色が 0，赤色が 1，緑色が 2，紫が 3 以上とし，それぞれの色の所属度のページが全体の何%を占めるかを示している．また，図 2 の上のグラフは P^{test} での比率，下のグラフは P^{other} での比率である．

表 3，図 2 の上の比率から，c に該当する Web ページ集

[*2] http://b.hatena.ne.jp/entry/json/?url=Url
[*3] http://b.hatena.ne.jp/USER/rss?of$NunberOfItems$
[*4] http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html

合の 83% の所属度が 1 以上となっている．図 2 の下の比率から，c に該当しない Web ページ集合に関しては，94% 所属度が 0 となっている．このことから，所属度のしきい値 θ_b を 1 より大きくすることで，c に該当する Web ページのみを取得できる．しかし，図中の所属度が 2 以上のページ数の比率から，所属度を 1 より大きくしてしまうと 31% しか Web ページを取得できなくなってしまう．

典型的でない Web ページは，カテゴリ内の他の Web ページとは大きく違う内容を記述しており，カテゴリと関連の強い語を多く記載している例は少ないと考えられる．そのため，θ_b を 2 より大きくした場合は希少な Web ページの多くが取得できないため，θ_b を 1 とする．

6.3 希少な Web ページの発見実験

データセットの全ページを入力として用いた場合，計算時間が膨大になるため P^{comp} と P^{test} を入力として用いる．タグ付けされた Web ページを用いることで計算時間を短縮し，さらにタグ付けのミスを排除することで所属判定を正確にできると期待する．

6.2 節で取得した Web ページ集合 P_c に対して，5.1，5.2 節の手法を用いることで各ページ p_c にスコア付けを行った．このとき，各手法によるスコアが上位の Web ページに対して「非典型的な内容か」を評価した．

まず，Web ページに対する典型性の評価方法を説明する．2 章で例に挙げた想起率を参考に，あるカテゴリに対して人が思い浮かびやすいキーワードは典型的であるとする．この観点をもとに，人がカテゴリから思い浮かべたキーワードに関する Web ページは典型的であるとみなす．

具体的には，各被験者に対して表 2 の各カテゴリから思い浮かぶ「キーワード」を 3 つ挙げてもらった．同じ意味のキーワードを集約した結果を表 4，5 に示す．表 4，5 で各カテゴリから挙げられたキーワードは典型的であると考え，Web ページの内容がこれらのキーワードに関する場合は「典型的」，そうでない場合は「非典型的」とみなす．

次に，入力の Web ページ集合に対して 6.2 節の所属判定を行い取得した Web ページ集合 P_c に 5.1，5.2 節の手法を適用した．このとき，それぞれのスコアが上位 10 件の Web ページに対して「c に該当するか」「典型的か」の 2 点を評価した．

その結果を表 6，7 に示す．表 6 は 5.1 節の内容の類似度に基づく手法，表 7 は 5.2 節の語の出現頻度を用いた手法の結果を示している．各表における「該当しない」の列の要素は，上位 10 件の Web ページの中でカテゴリに該当しなかった Web ページ数を示す．また「典型的」，「非典型的」の列の要素は，上位 10 件の Web ページがカテゴリ内で典型的だとみなしたページ，非典型的だとみなしたページの数をそれぞれ示している．

表 6，7 から，“音楽”カテゴリについてはスコアの上位 10 件にカテゴリに該当し，かつ非典型的な「希少な Web ページ」が多く現れており，このカテゴリでは手法は有効に機能していることがわかる．しかし表 6，7 中の他のカテゴリではそもそもカテゴリに該当しない Web ページが多く現れている．そのため，所属判定に問題があることが分かる．

次に，上位 10 件中の Web ページについて，カテゴリに該当した Web ページに対して非典型的な内容の Web ページがあった割合を表 8 に示す．各列の要素は，内容の類似度に基づく手法，語の出現頻度に基づく手法によって非典型的な Web ページが得られた割合である．この値が大きければ，カテゴリに該当する Web ページから非典型的な Web ページが得られる確率が大きい．

表 8 から，カテゴリに該当する Web ページから非典型的な Web ページが得られる確率は，内容の類似度に基づく手法では平均 0.81，語の出現頻度に基づく手法では平均 0.79 となり，高い確率で希少な Web ページが得られると分かる．よって，現段階では非典型度の算出について大きな問題はないが，所属判定に大きな問題があるといえる．

所属判定の問題について，大きく 3 つの要因が考えられる．まず 1 つ目は，SBM サービス上でタグがカテゴリとして使用されていないことが要因となっている．そのため，本来は指定カテゴリと関係がない語の関連度が大きくなり，所属判定に影響したと考えられる．例えば “twitter” と “テレビ”では，タグが「情報源」を表すために使われることが問題である．表 6，7 の出力されたカテゴリに「該当しない」Web ページを見ると，「Twitter で話題の」と記述しているページや，Twitter で多くリツイートされた Web ページに “Twitter” タグが付いている．“テレビ”についても同様に「テレビで話題の」と記述しているページに “テレビ” タグが使用されており，“Twitter”や “テレビ” カテゴリに該当しない情報に対してタグが使用されている．

2 つ目として，SBM サービス上でユーザが「タグの意味を混同している」ことが要因となっている．これによって，指定カテゴリとは別カテゴリの語の関連度が大きくなることで，1 つ目の要因と同様に所属判定に影響したと考えられる．これは “映画”，“ゲーム”，“アニメ”，“政治”，“経済” のタグが当てはまる．“映画” の場合では，小説や映画の内容について記述された Web ページに対して，どちらかのカテゴリの内容を含むと “小説” と “映画” の両方のタグを付けるユーザ層がある．そのため，小説に関する Web ページが “映画” カテゴリの出力に現れている．“ゲーム” と “アニメ”，“政治” と “経済” についても同様で，どちらかの内容を含む Web ページに両方のタグを使用するユーザが複数おり，“アニメ” に “ゲーム” の Web ページや，“政治” に “経済” が現れている例が多くあった．

3 つ目は，「複数の概念を持つタグ」が原因であると考えられる．これによって，2 つ目の要因と同様，別カテゴリ

表 4 被験者がカテゴリから思い浮かべたキーワード 1

政治	アニメ	エネルギー	ゲーム	音楽
金	深夜アニメ	火力	モンスターハンター	ラルク
韓国	年収	原子力	任天堂	アニソン
野田	画質	風力	SONY	MP3
オスプレイ	Sword Art Online	核	Tales Of Xillia2	ONE OK ROCK
消費税増税	るろうに剣心	低燃費	ペルソナ	Fear, and Loathing in Las Vegas
尖閣諸島問題	おおかみこどもの雨と雪	福島原発	Final Fantasy	凛として時雨
鳩山	ゆるゆり	火力	FPS	J-POP
選挙	アイドルマスター	再生可能	beatmania DX	水樹奈々
外交	じょしらく		アイドルマスター	バラード
	エヴァ		RPG	ポップス
	ガンダム		スポーツ	
	サザエさん			

表 5 被験者がカテゴリから思い浮かべたキーワード 2

映画	twitter	経済	google	テレビ
価格	炎上	不景気	検索	3 D
監督	情報	インフレ	ggrks	価格
ポッター	青	簿記	ホームページ	ニュース
るろうに剣心	ハッシュタグ	消費税増税	Gmail	地デジ
おおかみこどもの雨と雪	なう	所得隠し	Android	しゃべくり 007
トータルリコール	リツイート	脱税	Google Earth	視聴率
魔法少女まどか☆マギカ	ツール	日経新聞	初音ミク	フジテレビ
スタジオジブリ	青い鳥	不景気	Google Chrome	韓流
シャフト		円高	マップ	２４時間テレビ
俳優		就職難	中国	薄型
アカデミー賞		ギリシャ		有機 EL
制作費				

の語の関連度が大きくなるため，所属判定に影響したと考えられる．この原因は “エネルギー”カテゴリが当てはまる．電力をはじめとした「エネルギー」の意味と，人間の活力を表す「エネルギー」が同じタグで分類されていた．

また，手法の問題として，「ミスによるタグ付け」と本来の「カテゴリを示すタグ付け」を等価に扱って計算したことが考えられる．

これらが原因でカテゴリと関係のない語の関連度が大きくなってしまい，カテゴリに該当すると判断されることで出力に現れたと考えられる．希少な Web ページを非典型性を用いて発見するため，所属度によるフィルタリングの取りこぼしの影響が非常に大きい．これらのページに関しては，カテゴリの判定手法を再考案する必要がある．

7. まとめ

指定したカテゴリと Web ページ集合から，有用で非典型的な Web ページを希少な Web ページとして発見する手法を提案した．本研究ではカテゴリに該当するページを有用であると考え，所属度と典型度の２つを用いて希少な Web ページを発見する手法を提案した．所属度は Web ページがカテゴリに該当する度合いで，用意したデータセットから計算した語とカテゴリの関連度から算出し，適切なしきい値を決定する．非典型度は Web ページがカテゴリ内でどのくらい非典型的かを表す指標である．カテゴリ内で内容が類似する Web ページが多いほど典型的である観点と，カテゴリ内での語の出現頻度が小さい語を含むほど非典型的である観点の２つを検討した．

表 6 内容の類似度に基づく手法で得られた Web ページの分布

	該当しない	該当する	
カテゴリ		典型的	非典型的
音楽	2	1	7
映画	8	0	2
テレビ	6	0	4
ゲーム	5	1	4
エネルギー	4	3	3
アニメ	6	0	4
twitter	9	1	0
google	8	0	2
政治	1	5	4
経済	6	1	3

関連度をデータセットから算出し評価した結果，関連度の大きい語は被験者に「カテゴリと関係がある」と判断され，関連度の小さい語は被験者によって「カテゴリと関係がない」と判断されたことから，語とカテゴリの関係を関

表 7 語の出現頻度を用いた手法で得られた Web ページの分布

	該当しない	該当する	
カテゴリ		典型的	非典型的
音楽	2	0	8
映画	7	1	2
テレビ	6	1	3
ゲーム	4	1	5
エネルギー	3	7	0
アニメ	6	1	3
twitter	7	2	1
google	4	1	5
政治	9	0	1
経済	9	0	1

表 8 カテゴリに該当するページ中の非典型的なページの割合

カテゴリ	内容の類似度	語の出現頻度
音楽	0.88	1.00
映画	1.00	0.67
テレビ	1.00	0.75
ゲーム	0.80	0.83
エネルギー	0.50	0
アニメ	1.00	0.75
twitter	0	0.33
google	1.00	0.83
政治	0.44	1.00
経済	0.75	1.00
平均	0.74	0.72

連度によって表せたといえる．

次に，この関連度を用いて所属度を算出し，しきい値を決定した．指定カテゴリ外の Web ページの多くは所属度が 0，指定カテゴリ内の Web ページの多くは所属度が 1 以上となることから，所属度のしきい値を 1 と決定した．入力の Web ページ集合から所属度が 1 以上の Web ページを取得することで，指定カテゴリに該当する Web ページ集合のみを取得し，典型度の算出対象とした．

最後に指定カテゴリに該当する Web ページに対して，非典型的さを示すスコアを算出し，希少な Web ページが含まれるかを評価した．人が「カテゴリに所属する」と判断した Web ページから非典型的な Web ページを得られる割合が大きいことから，非典型度を算出する手法は有効であることが分かった．

しかし，スコア上位には指定したカテゴリに該当しない Web ページが多く現れたことから，カテゴリの所属判定の手法に大きな問題があることが分かった．この原因としては，SBM サービス内でタグがカテゴリとして使用されていない場合や，タグの意味を混同して使用していることが原因だと考えられる．これが語とカテゴリの関連度に影響したため，所属判定に問題が見られた．

今後の課題としてカテゴリの所属判定の改良が挙げられる．

謝辞

本研究の一部は，平成 24 年度科研費若手研究 (B)「情報の詳細関係に基づくＷｅｂページの組織化」(課題番号：24700097) によるものです．ここに記して謝意を表すものとします．

参考文献

[1] 佃洸摂, 中村聡史, 山本岳洋, 田中克己: “オブジェクトの典型度分析とその検索への応用”, WebDB Forum 2011, 2G-1-1, 2011.
[2] L.W. Barasalou: “Ad hoc categories”, Memory & cognition, Vol.11, No.3, pp.211-227, 1983.
[3] L.W. Barasalou: “Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories”, Jounal of Experimental Psychology: Lerning, Memory, and Cognition, Vol 11(4), pp.629- 654, 1985.
[4] J. Cohen and K. Basu: “Alternative Models of Categorization: Toward a Contingent Processing Framework”, Journal of Consumer Reserch, Vol. 13, No. 4, pp.455-472, 1987.
[5] B. Loken and J. Word: “Alternative Approaches to Understanding the Determinants of Typicality”, Journal of Consumer Reserch, Vol. 17, No. 2, pp.111-126, 1990.
[6] 藤坂達也, 湯本高行, 角谷和俊: “典型的なオブジェクト選定のためのカテゴリタイプにおける重要な観点の抽出”, DEIM Forum 2012, E7-3, 2012.
[7] 百田信, 伊東栄典: “ソーシャルブックマークに基づく情報発見”, 電子情報通信学会第 19 回データ工学ワークショップ, I1-15, 2008.
[8] S. Brin and L. Page: “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems, vol.30, No.1-7, pp.107-117, 1998.
[9] R. Mihalcea and P.Tarau: “TextRank: Bringing order into texts”, Proceedings of EMNLP 2004, pp.404-411, 2004.
[10] 川中翔：“ソーシャルブックマークにおけるタグの派生関係の抽出”, 東京大学大学院 基盤情報学専攻修士論文, 2009.
[11] 藤村滋, 藤村孝, 片岡良治, 奥雅博：“Blog のタグ間類似度のスコアリング”, DBSJ Letters, Vol.5, No.4, 2007.
[12] S.A. Golder and B.A. Huberman： “The structure of collablorative tagging systems”, Journal of Information Science, 32, 2, pp.198-208, 2006.