情報処理学会論文誌　数理モデル化と応用　Vol.9　No.3　1–7 (Dec. 2016)

確率的トピックモデルに基づく仲間はずれさがし問題の解法

佐藤 進也[1,a)]

受付日 2016年2月19日，再受付日 2016年5月6日，
採録日 2016年5月18日

概要：理解できることを「わかる」というが，これは「分かる」などと綴る．事物の集まりを分けること，分けるための基準を持つことは理解の基本であり，分類により同類の集まりを作り出すことは抽象化の第一歩である．人間の類別する能力を評価するための問題に，小学校入試において出題される「仲間はずれさがし」がある．この問題は，提示された複数の事物の中から，他の物とは区別されるべきものを 1 つ選び出すというものである．本論文では，文書集合から抽出した潜在トピックを分類の基準として用いる解法を提案する．本手法を実際に仲間はずれさがし問題に適用したところ，76.9%という正解率を得た．

キーワード：クラスタリング，潜在トピック，シソーラス，コーパス

A Solution to the “Find the Odd One” Problem Based on a Probabilistic Topic Model

Shin-ya Sato[1,a)]

Received: February 19, 2016, Revised: May 6, 2016,
Accepted: May 18, 2016

Abstract: Classification and classification criteria discovery are a fundamental operations in understanding things and concepts. It is also a basic operation for abstraction. “Find the odd one” problem is a question in elemenary school entrance examinations for measuring ability of classification. In the problem, examinees are requested to choose one item out of a group of given items that can be distinguished from others. In this paper, a solution to the find the odd one problem is proposed, where the odd one is identified on the basis of latent topics extracted from a relevant document set. It is shown that the proposed method could achieve 76.9% of correct answers rate in an evaluation experiment.

Keywords: clustering, latent topic, thesaurus, corpus

1. はじめに

理解できることを「わかる」というが，これは「分かる」などと綴る．事物の集まりを分けること，分けるための基準を持つことは理解の基本であり，類別により同類の集まりを作り出すことは抽象化の第一歩である．

人の類別する能力を評価する問題は入学試験でも出題されている．その小学校入試での代表例が仲間はずれさがし問題である．この問題では，**図 1** にあるような複数の事物が並べられた絵が提示され，他の物とは類別されるべきものを 1 つ選び出すことが求められる．図 1 の場合，ただ 1 つ鳥類でないコウモリが正解である．

本論文では，この仲間はずれさがし問題を，語（概念）の分類問題としてとらえ，語に関連する文書集合の解析により解を見つけ出す手法を提案する．より具体的には，本手法は文書集合から潜在トピックを抽出し，それを分類の基準とすることにより本問題を解く．

本論文での議論は以下のようにすすめる．まず 2 章で問題を定式化する．そのうえで，3 章で関連研究と仲間はずれさがし問題への適用可能性・限界について述べる．次に，4 章で提案手法について説明する．さらに，本手法の評価方法と結果を 5 章で示す．最後に，本研究の位置づけと意義について改めて 6 章で議論する．

[1] 日本工業大学
Nippon Institute of Technology, Saitama 345–8501, Japan
[a)] shin-ya.sato@acm.org

© 2016 Information Processing Society of Japan

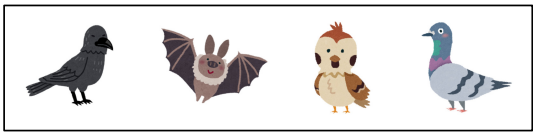

図 1　仲間はずれさがし問題の例
Fig. 1　Example of the “find the odd one” problem.

2.　問題の定式化

ここでは画像認識処理，すなわち，もともとの入試問題で提示される絵に描かれている事物を認識する処理については議論の対象外とし，事物の関係を把握し，類別することに焦点を当てる．すなわち本問題を以下のように語（概念）の分類問題として定式化する．

仲間はずれさがし問題

与えられた事物（の名称）の集合 $I = \{x_1, ..., x_n\}$ を $\{x\}$ と $I - \{x\}$ に分けることが妥当な基準に基づいた分類となるような I の要素 x を選び出せ．

図 1 の例では，$I = \{$カラス, コウモリ, スズメ, ハト$\}$ であり，このうち 1 つだけ鳥類に属さないコウモリが仲間はずれさがし問題の解である．この定式化における「妥当な」分類基準としては，例にあげたような分類体系以外にも，（日用品などの）用途の違いや，特定の状況（年間行事や物語など）の構成要素となっているか否か，など様々なものが考えられる．この基準を見つけ出すことが，本問題を解く重要なポイントである．分類基準の妥当性についての基準を設定することは，それ自体に多くの議論を要する難しい問題である．ここでは，その判断を人間に委ねることにする．具体的には，人間が作成した問題は妥当な基準に基づくものと見なし，その答えを見つけ出すことができるか否かで問題解決手法の効果を判断する．なお，以降，例に用いたこの I を I_{bird} と記す．

3.　関連研究

3.1　クラスタリングによる語の分類

仲間はずれさがし問題を解く方針として，シソーラスのような分類体系を利用したり，教師データから分類基準を学ぶアプローチも考えられる．しかし，先に述べたように，分類基準は多様であり，あらかじめ適切なものを決めておくことはできない．このような状況に柔軟に対応するためには，データから分類基準を発見するクラスタリングのアプローチが適していると考えられる．

クラスタリングについては，今までに非常に多くの研究がなされており，様々な手法が提案されてきた [3]．広く利用されている手法として，階層的クラスタリング [3] や k-means [6] などをあげることができる．これらはいずれも，分類対象の 2 つの要素間の距離（あるいは，その逆の度合いを測る指標である関連度）を定義し，その距離に基づいて分類を行う．語の分類にこれらの手法を適用するためには，語間の距離を定義する必要がある．距離の定義には共起解析 [4] やベクトル空間モデル [9] などが利用されてきた．

共起解析は，語の文書あるいは文書の一部分における共起の度合い（共起度）を調べるものである．共起度を計算する手法も複数提案されている．その一例が 2 つの語 x, y の共起度を

$$J(x,y) = \frac{|D(x) \cap D(y)|}{|D(x) \cup D(y)|}$$

で与える Jaccard 係数である．ここで，$D(x)$ は，x が出現する文書の集合である．

ベクトル空間モデルに基づく方法では，まず，語 x について書かれた文書 $d(x)$ を選ぶ．そして，$d(x)$ に出現する語の出現頻度分布（あるいは tf・idf [9] のような，出現頻度に基づいて算出される語の重要度の分布）$v_{d(x)}$ を計算する．語 x, y の類似度は，それぞれの分布をベクトルと見なしたときの向きの一致度

$$cos(x,y) = \frac{v_{d(x)} \cdot v_{d(y)}}{|v_{d(x)}||v_{d(y)}|}$$

で関連度を測る．ここで，$v_1 \cdot v_2$ は，2 つのベクトル v_1, v_2 の内積である．

語の関係を判定するためには，関係を定量的に把握できるかたちで語が表している意味・内容を表現しなければならない．共起解析に基づく手法，ベクトル空間モデルに基づく手法の双方とも，その表現のために，語に関わりのある文書を利用している．語の関連性をその使われ方の類似性から判定している，と考えることができる．

3.2　既存クラスタリング手法の仲間はずれさがしへの適用

本節では，既存クラスタリング手法の仲間はずれさがし問題への適用可能性について具体例を示しながら議論する．

まず，x について書かれた Wikipedia のページを $d(x)$ とし，Ward 法による階層的クラスタリング [3] を適用して I_{bird} をクラスタリングした結果（デンドログラム）を**図 2** に示す．コウモリが最後にクラスタに組み込まれており，期待どおりの答えが得られている．

もう 1 つの例として，同じ手法を用いて $I = \{$イヌ, キジ, サル, 鬼, お化け$\}$ をクラスタリングした結果を**図 3** に示す．これは，お伽話「桃太郎」に登場しない「お化け」を解とする仲間はずれさがし問題である．しかしながら，図 3 では $\{$イヌ, キジ, サル$\}$ と $\{$鬼, お化け$\}$ の 2 つのグループに分類されている．この結果は仲間はずれさがし問題の正解は与えていないが，「1 つだけ仲間はずれ」という条件がなければ，また別の観点（実在する生物か否か）に基づいた妥当な分類と考えられる．そもそも，このような結果が

© 2016 Information Processing Society of Japan

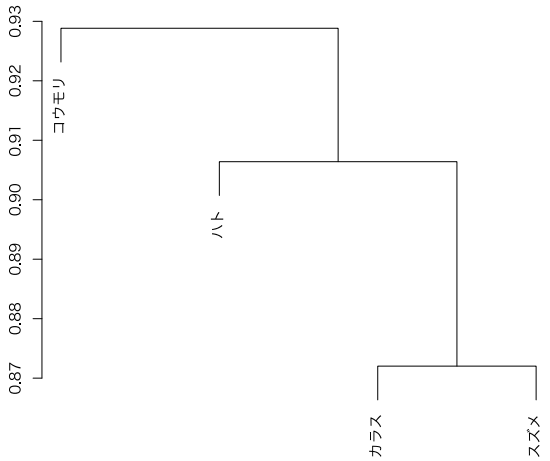

図 2　{カラス, コウモリ, スズメ, ハト} のクラスタリングの結果を示すデンドログラム

Fig. 2　Dendrogram showing the result of clustering {crow, bat, sparrow, pigeon}.

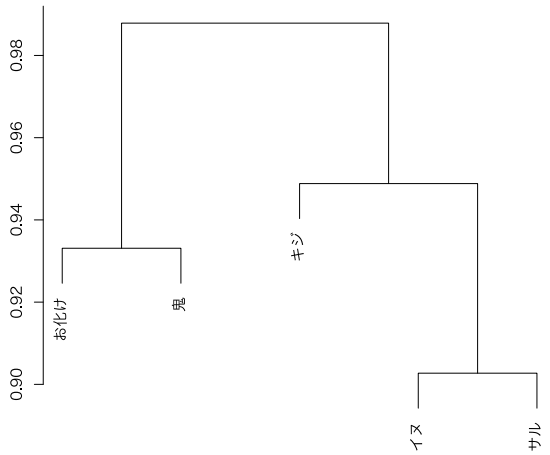

図 3　{イヌ, キジ, サル, 鬼, お化け} のクラスタリングの結果を示すデンドログラム

Fig. 3　Dendrogram showing the result of clustering {dog, pheasant, monkey, ogre, ghost}.

得られるのは，$d(x)$ として，言葉の意味を概念体系の一要素として x を説明している Wikipedia のページを選択したからであると考えられる．すなわち，$d(x)$ の比較に基づく分類では，分類観点は $d(x)$ において x が言及されている文脈によって与えられるのである．このことから，分類観点を変えるには $d(x)$ の選び方を変えればよいことが分かる．しかし，目的に応じて $d(x)$ を選ぶ具体的な方法は，著者の知る限り，確立されていない．

なお, 語のクラスタリングにより多義性を解消したり [10], シソーラスを構築したりする研究 [2] がある．この場合，仲間はずれさがし問題を解くのとは異なり，分類観点は固定されている．

3.3　トピックモデル

語間の関係を直接調べるのではなく，特定の話題（トピック）との関わり方を調べ，その結果に基づきクラスタリングを行う方法も知られている．その 1 つがトピックモデルに基づく方法である．トピックモデルの代表的な手法が LDA（Latent Dirichlet Allocation）[1] である．LDA では，文書には複数のトピックが混在しており，各トピックと語の間に確率的な関係があるものとして，文書・トピック・語の 3 者の関係をモデル化する．このモデルに基づき，文書集合からトピックを抽出するとともにトピックと語の関係を推定する．本研究では，LDA を用いた仲間はずれさがし問題の解法を提案する．

4.　提案手法

本章では提案手法について述べる．まず，改めて LDA のより詳しい説明を行う．そのうえで，LDA を利用している本手法の手順を説明する．

4.1　LDA

文書集合 C を選び，その要素（すなわち，文書）を d_i（$i = 1, ..., n$）とする．一般的に，1 つの文書の内容は複数のトピックを含んでいると考えられる．また，同一のトピックを複数の文書で共有していることもよくある．これを LDA では，文書 d_i には有限個（K 個）の重み付けられたトピック $\{z_j\}$（$j = 1, .., K$）が混在していると考える．z_j の重みを θ_{ij} とすると，$\{\theta_{ij}\}_{j=1,..,K}$ は d_i のトピック分布を表している．

さらに，各トピック z_j ごとに語 w_k の現れやすさ（出現確率）ϕ_{jk} が定まっていると考える．現れやすさはトピックと語の関連度の高さと考えてもよいだろう．説明の便宜上，トピック z_j における語 w の出現確率を $\Phi_{z_j}(w)$ と書く．定義から，$\Phi_{z_j}(w_k) = \phi_{jk}$ である．

これらを組み合わせ，LDA では，文書ごとのトピックの分布と，トピックごとに定まっている確率に従い語が生成されると考え，実際に観測できる語の出現頻度分布から，θ_{ij} と ϕ_{jk} を推定する．推定方法には，変分ベイズ推定 [12] や崩壊型ギブスサンプリング [5] などがある．

4.2　提案手法の処理手順

本節では，本手法が語の集合 I から仲間はずれを見つけ出す手順について述べる．まず，全体の流れを示した後，各処理の内容について詳しく説明する．

4.2.1　概要

提案手法は以下のステップからなる．4.2.2 項以降，各ステップの詳細について述べる．

(1) 文書集合 C の構成
　I に基づいて LDA を適用する文書集合 C を構成する．
(2) 潜在トピックの抽出
　C に LDA を適用し，潜在トピックを抽出する．

© 2016 Information Processing Society of Japan

(3) 適合するトピックと解の選出
　得られた潜在トピック z_i と語の分布 ϕ_{ij} をもとに解となる語の候補を選び出す．
(4) 複数の推定結果に基づく解の評価
　条件を変えながら (3) の解の選出を複数回行い，得られた結果から最も妥当な解を選出する．

4.2.2　文書集合 C の構成

後に評価結果を示すように，文書集合 C を構成する文書の取捨選択が正解率に大きく影響する．これは，文書の取捨選択が，LDA によって得られる潜在トピックに影響するからである．

仲間はずれさがし問題は，正解が示されれば誰もが納得するような，いわば常識を問う問題である．よって，C としては，常識が話題にあがっているような文書集合を構成すべきである．では，常識とは何であり，どのようにとらえるべきものなのだろうか．本研究でもいまだ明確な答えを示すことはできていないが，

- 日常生活の中でよく語られること
- 日常生活の様々な場面
- 人々の体験

に関する記述がこの範疇に属すると考えている．上記の条件を満たす文書としては，日常生活の中で遭遇した疑問や解決したい問題が扱われる Q&A サイトのページや，日々の生活の中で見聞きしたこと，体験したことを記録したブログ記事をあげることができる．それらのページ（URL）を取得するため，本研究では Web 検索エンジンを利用した．多くの検索エンジンでは，検索対象を特定のサイトに限定することができる．このオプションを指定し，$x \in I$ をクエリとして検索した結果得られる URL から C を構成する．以上をまとめると，文書集合を構成する擬似コードは以下のようになる．ここで，webSearch(x, $site$, h) は，検索対象を $site$ に限定し x で検索して得られる上位 h 件の Web ページの集合を返す手続きである．

```
procedure biuldCorpus(I, site, h) {
  C = φ;
  foreach x in I {
    r = webSearch(x, site, h);
    add(C, r);
  }
  return C;
}
```

4.2.3　潜在トピックの抽出

次に，得られた文書集合 C から LDA を用いて潜在トピックを抽出し，各トピックと語の関係を推定する．そのために，まず各 $d_i \in C$ の文に形態素解析を行い，語の品詞種別や出現頻度を調べる．その結果に基づき，語の取捨選択を行う．具体的には，本手法では，名詞，動詞，形容詞以外を除去する．さらに，文書ごとに tf・idf 値に基づいて語を降順にソートし，$N+1$ 位以降を除く．このようにして得られた，語の集合としての文書に LDA を適用する．具体的には，トピック数 K を選び，崩壊型ギブスサンプリングによりトピック分布，語の分布を推定する．

4.2.4　問題に適合するトピックと解の選出

推定の結果，C に潜在するトピック z_i とトピックにおける語の分布 ϕ_{ij} が得られる．このとき，$u(I, z_i)$ を以下のように定義する．

$$u(I, z_i) = |U(I, z_i)|,$$
$$U(I, z_i) = \{x \in I | \Phi_{z_i}(x) = 0\}$$

ここで $\Phi_{z_i}(x)$ は語 $x \in I$ のトピック z_i における出現確率であった．よって，$U(I, z_i)$ は I の中で z_i に関係ないと推定される語の集合と考えられる．さて，仲間はずれさがし問題の「1 つを除き同類である」は「ただ 1 つを除いて，あるトピック z_i と関連性を有する」と解釈できるが，この条件は先ほど定義した u を使って，

$$u(I, z_i) = 1$$

と書ける．そして，LDA の結果に基づき仲間はずれを選び出す手続きは以下のように定義できる．

```
procedure guessTheOddOne(I, {z_i}, {Φ_{z_i}}) {
  X = φ;
  foreach z in {z_i} {
    u = 0;
    odd = NULL;
    foreach x in I {
      if (Φ_z(x) == 0)
        u++;
        odd = x;
      }
    }
    if (u == 1) add(X, odd);
  }
  return X;
}
```

4.2.5　複数の推定結果に基づく解の評価

LDA の結果は，文書ごとに抽出する語の数 N やトピック数 K などのパラメータの値に依存する．また，トピック抽出のアルゴリズムは決定論的ではないので，同一条件下でも試行ごとに値が異なる可能性がある．本研究では，パラメータの組合せを複数生成し，各組合せごとに推定を複数回（M 回）行い，それらの結果から総合的に仲間はずれを判断する．このアプローチには，適切な解の探索範囲を広げる一方で，確率的に発生する不適切な解を除去するとともに，複数の妥当な解のうちより尤もらしいものを選び出す効果が期待できる．

この手順を書き下すと下記の擬似コードのようになる．ここで，$params$ はパラメータの組合せの集合，LDA(C,

© 2016 Information Processing Society of Japan

p) は，あるパラメータの組合せ p に基づいた C からのトピック抽出の手続きである．また，countOccurrences(x, L) は，リスト L 中に x が出現する頻度をカウントする手続きである．要約すると，この findTheOddOne という手続きは，パラメータを取り替えながら複数の推定を行い，仲間はずれとして選ばれた回数が最も多い語を最終的に解として選び出している．

```
procedure findTheOddOne(I, site, H, params, M) {
  C = buildCorpus(I, site, H);
  L = φ;
  foreach p in params {
    for (m = 0; m < M; m++) {
      ({z_i}, {Φ_i}) = LDA(C, p);
      X = guessTheOddOne(I, {z_i}, {Φ_i});
      add(L, X);
    }
  }
  return arg max_x countOccurrences(x, L);
}
```

5. 評価

5.1 評価実験の内容

提案手法で実際に複数の仲間はずれさがし問題を解き，有効性を評価した．評価に使用した 13 題の問題を**表 1** に示す．それぞれの問題は，4〜5 個の選択肢で構成され，その中の 1 語があらかじめ正解として選ばれている．表で下線が引かれている語が正解である．この 13 問題は，小学校入試の過去の問題と，この評価のために作成した問題からなる．作成にあたっては，小学校の入試を意識し，常識の枠を超えた特別な知識を必要とする問題にならないよう注意した．

これらの問題に手続き findTheOddOne を適用して解を推定した．文書集合を構成するための検索の対象を指定する *site* には，Q&A サイト，ブログ，指定なし，という条件をそれぞれ設定した．より具体的には，検索時にそれぞれ**表 2** にあるオプションを指定した．また文書集合を構成する検索結果の数 H は 100 とした．*params* に指定するパラメータのバリエーションとしては，K は $\{10, 20\}$ のいずれか，N は $\{100, 200\}$ のいずれかから選んだ．また，$M = 5$ とした．

表 1　評価に用いた仲間はずれさがし問題
Table 1　Problems used for the evaluation experiment.

(a) カラス, コウモリ, スズメ, ハト
(b) タヌキ, トカゲ, ヘビ, ワニ
(c) カブトムシ, クモ, テントウムシ, バッタ
(d) うちわ, 蚊取線香, 手袋, 風鈴
(e) 鉛筆, 財布, ノート, はさみ
(f) イヌ, 鬼, お化け, キジ, サル
(g) 絵馬, お守り, 短冊, 破魔矢
(h) カッター, ノコギリ, はさみ, 包丁
(i) カバ, キリン, ゾウ, ネコ
(j) カラス, スズメ, ハト, ペンギン
(k) ゴキブリ, ネコ, ネズミ, ハエ
(l) タンポポ, チューリップ, ツツジ, ヒガンバナ
(m) 小松菜, キャベツ, トマト, 白菜

5.2 評価結果

表 3 は，推定の結果を一覧にしたものである．比較のために，3.2 節で述べた階層的クラスタリングの適用結果を表の 1 行目に示した．2 行目移行の各行は提案手法を適用した結果であり，それぞれ *site* の条件に対応している．すなわち，文書集合として Q&A ページやブログ記事をそれぞれ用いた場合（表中の Q&A，BLOG），あるいは，条件を設定せず任意のページを使用した場合（ANY）の結果を示している．表の o，x はそれぞれ推定が正解，不正解であったことを示しており，– は仲間はずれの語を見つけられなかったことを示している．また，最後の Σ の列に示したのは正解できた問題の数である．

この一覧から正解率を求め，**表 4** にまとめた．問題 (f)

表 2　検索対象を指定するオプション
Table 2　Options for limiting the search results.

検索対象	検索オプション
Q&A サイト	site:oshiete.goo.ne.jp/qa
ブログ	site:blogs.yahoo.co.jp
指定なし	—

表 3　実験結果
Table 3　Result of the experiment.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
既存手法	o	x	-	-	x	-	o
Q&A	o	x	o	o	o	o	o
BLOG	o	o	o	x	o	o	o
ANY	o	o	x	-	o	x	o

	(h)	(i)	(j)	(k)	(l)	(m)	Σ
既存手法	-	x	o	x	x	x	3
Q&A	o	o	o	x	x	o	10
BLOG	o	o	o	o	x	x	10
ANY	-	-	o	-	x	x	5

表 4　実験結果のまとめ
Table 4　Summary of the the experiment result.

手法	正解率 (%)
無作為選択	24.6
既存手法	23.1
Q&A	76.9
BLOG	76.9
ANY	38.5

© 2016 Information Processing Society of Japan

だけが 5 択で他は 4 択問題であるので，選択肢から無作為に語を選んだ場合の正解率の期待値は 24.6%である．これをベースラインとし実験結果を評価する．

まず，単純に階層的クラスタリングを適用した場合，正解率は 23.1%であり，ベースラインを下回った．3.2 節で，階層的クラスタリングの結果が分類としては適切でありながら，仲間はずれさがし問題の解とはなりえない場合があることを述べた．評価実験においても，このことが低い正解率の一因となっていた．

一方，提案手法を適用し，Q&A ページあるいはブログ記事を用いた場合の正解率は 76.9%であった．ベースラインを大きく上回っているこの結果は本手法の有効性を示している．ただし，文書集合として任意の Web ページを用いた場合の正解率は 38.5%であった．Q&A ページあるいはブログ記事を用いた場合との明確な差は，本手法における文書の取捨選択の重要性を示している．なお，この場合でも無作為な選択よりは高い正解率が得られている．

5.3　文書集合と潜在トピック

本節では，評価実験の結果をより詳しく調べて得られた，成否の要因に関する知見について述べる．

本手法で正解が得られない主要な原因に，文書集合 C におけるトピックの欠損がある．正解を見つけ出すためには，まず，その分類観点と合致するトピックが C から抽出されなければならないが，そもそもそのようなトピックを扱っている文書が C に含まれていない場合もありうるのである．具体例として，問題 (f) の場合を考える．もし「桃太郎」というお伽話に関するトピックが抽出できれば，そのトピックとは関係のない「お化け」を解として選ぶことができるだろう．逆に，「桃太郎」を話題にしている文書が C の中に存在しなければ，この問題に正解できない可能性がある．このような観点から，問題 (f) については，特に Q&A を用いた場合，正解するのは難しいと予測していた．しかし，表 3 にあるように，本手法で正解が得られており，実際，文書集合の中に「桃太郎」について述べている文書を確認した．この実験結果は，分類対象 I の各要素をクエリとして検索した結果から文書集合を構成するという方法で，分類に必要なトピックをおおかたカバーできていることを示している．ただし，さらに様々な問題に対応できるようにするには，文書集合の構成方法をさらに工夫する必要があると考えられる．

さて，問題 (b) の分類は「爬虫類に属する」という条件を基準としていると考えられる．この問題に対し，本手法でブログ記事を用いた場合正解が得られている．これは，この分類観点と合致するトピックが C の中に存在していたことを意味する．しかし，実際に話題になっていたのは，動物の分類論ではなく，装身具や装飾雑貨の材料としての動物の皮革（つまり，ワニ革など）であった．これはつまり，種類に属する動物の特徴（この場合，爬虫類の特徴的な皮革）を日常生活の中で利用しており，その利用に関する話題の中で言及されるか否かが動物の分類と整合しているということである．

6.　考察

提案手法の詳細と効果が明らかになったところで，改めて本研究の位置づけについて議論したい．

本論文では仲間はずれさがしという一問題の解法について議論しているものの，本研究は，文書集合からの日常生活における文脈（コンテキスト）の把握という一般的な問題に関する取り組みの 1 つとして位置づけられる．本論文で議論してきた潜在トピックがコンテキストに対応する．

日常生活の様々なコンテキストを把握することは，実世界に関する多面的な理解を得ることである．これは，画像認識 [8]，ユビキタスコンピューティングでの状況把握 [7] や，検索結果を複数の観点から整理するファセット検索 [11] などの発展に寄与し，実世界での行動支援や，情報への効率的アクセスを可能にする．

日常生活で我々を取り巻くコンテキストは多種多様である．たとえば「かぼちゃ」に関するものとしては，調理，栽培，ハロウィンのようにただちに思いつく “メジャー” なコンテキストもあれば，「離乳食」のようないわれてそういえばと思うものや，「小動物の飼育」[*1] のように説明されてはじめて分かる “マイナーコンテキスト” とでも呼ぶべきものもある．これらはいずれも，頻度の差はあれ，日常生活の中で誰もが出会う可能性のあるコンテキストである．マイナーコンテキストに関する情報は，マイナーであるがゆえに，メジャーなコンテキストに関する情報に埋もれがちであるが，そういった情報にも利用価値はある．これらはむしろ，入手し難い希少価値のある情報ということもできるだろう．そして，埋もれがちな情報を抽出するためには，単純な統計処理を超えた手法の開発が必要となる．本研究はその要求に対して手がかりを与えるものである．

7.　むすび

本論文では，Web 上の文書集合から抽出した潜在トピックを類別の基準として仲間はずれさがし問題を解く手法を提案し，その有効性を示した．Web から知識を抽出する研究は Web マイニングという 1 つの分野を形成し，そこで多くの取り組みが行われている．本研究もその範疇に属するものであり，特に，常識に関する知識を Web 上の文書から抽出する一手法として位置づけられる．

今回は単純に正解を当てられるか否かで評価し，なぜ仲間はずれに選んだかという推定理由については問わなかった．今後は，推定理由を何らかの形で示すことを検討し，

[*1] 種子が小動物の餌になる．

© 2016 Information Processing Society of Japan

典型的な解だけでなく，「意外な」解も推定理由とともに示す方法の開発に取り組んでいきたい．

参考文献

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.31, pp.993–1022 (2003).

[2] Hodge, V.J. and Austin, J.: Hierarchical word clustering — automatic thesaurus generation, *Neurocomputing*, Vol.48, No.1-4, pp.819–846 (2002).

[3] Jain, A.K., Murty, M.N. and Flynn, P.J.: Data clustering: A review, *ACM Computing Surveys*, Vol.31, No.3, pp.264–323 (1999).

[4] Li, H. and Abe, N.: Word Clustering and Disambiguation Based on Co-occurrence Data, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics — Volume 2, ACL '98*, Stroudsburg, PA, USA, pp.749–755, Association for Computational Linguistics (1998).

[5] Liu, J.S.: The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem, *Journal of the American Statistical Association*, Vol.89, No.427, pp.958–996 (1994).

[6] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations, *Proc. 5th Symposium on Math, Statistics and Probability*, pp.281–297 (1967).

[7] Perkowitz, M., Philipose, M., Fishkin, K. and Patterson, D.J.: Mining Models of Human Activities from the Web, *Proc. 13th International World Wide Web Conference*, pp.573–582, ACM Press (2004).

[8] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E. and Belongie, S.: Objects in Context, *ICCV*, IEEE, pp.1–8 (2007).

[9] Salton, G.: *Automatic Information Organization and Retrieval*, McGraw-Hill (1968).

[10] Schütze, H.: Automatic word sense discrimination, *Computational Linguistics*, Vol.24, No.1, pp.97–123 (1998).

[11] Tunkelang, D.: *Faceted Search*, Morgan and Claypool Publishers (2009).

[12] Waterhouse, S., MacKay, D. and Robinson, A.: Bayesian methods for mixture of experts, *Advances in Neural Information Processing Systems*, pp.351–357, MIT Press (1995).

佐藤 進也　（正会員）

1986 年東北大学理学部数学科卒業．1988 年同大学大学院修士課程修了．同年日本電信電話株式会社入社．NTT 未来ねっと研究所等を経て，2014 年より日本工業大学情報工学科教授．博士（情報理工学）．協調作業支援，Web 情報検索・マイニング，複雑ネットワーク等の研究に従事．訳書『スモールワールド』（ダンカン・ワッツ著，東京電機大学出版局，共訳）．ACM，ISOC，電子情報通信学会，人工知能学会各会員．

© 2016 Information Processing Society of Japan