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

恣意的に名前付けされたオブジェクトの識別手法

Identification of Loosely-Named Objects

高橋 良平† 小山 聡‡ 田中 克己‡
Ryohei TAKAHASHI Satoshi OYAMA
Katsumi TANAKA

料理や事件といったオブジェクトは，書き手によって恣意的に名前が付けられるため，同一オブジェクトに対して無数の名前が存在する．そのため，あるWebページがどのオブジェクトについての記述であるかを識別することは容易ではない．本論文では，まず多くの人が使用している名前のパターンをもとにオブジェクト名を抽出し，次にそのオブジェクトに分類されるための規則を作成し，最後にその規則をもとに各 Web ページが言及しているオブジェクトを同定する，という三段階でこの問題を解決する手法を提案する．

Since objects like dishes and events are loosely-named by authors, some objects have a lot of names. So, it is not easy to identify the object that is described in a Web page. In this paper, we propose a method to solve this problem in the following three steps. First, object names are extracted by using the naming patterns that many people use. Second, rules to identify each object are generated. Third, objects that are written in each Web page are identified by the rules.

1. はじめに

近年，従来からの検索エンジンのようにWebページ単位で検索するのではなく，オブジェクト単位で検索する“オブジェクト検索”に関する研究が多数行われている[1][2]．オブジェクト検索では，オブジェクトについての記述を含む複数ページからスキーマに従って情報を抽出し，オブジェクト単位で情報を集約してユーザに提示する．例えば人物を対象としたオブジェクト検索の場合，同一人物に関する情報をWebから取得し，生年月日や連絡先などの属性を集約してユーザに提示する．人物だけでなく，論文や商品についても，このようなオブジェクト単位の検索は行われている．

このようなことを実現するためには，あるWebページの記述と，他のWebページの記述が，同一オブジェクトについて書かれているかどうかを判断する必要がある．これは，“オブジェクト識別”と呼ばれている問題である．オブジェクト識別に関する研究は従来から行われているが，その多くは名前の曖昧性の解消に関するものであった．例えば，人物を対象としたオブジェクト識別の場合，同じ名前の人物について書かれているページであっても同姓同名の別人である可能性があり，それらを実際の人物ごとに分離するという問題である．ある人物に付けられる名前は，別名を持つ場合などの例外を除くと一つのみであるので，“名前が異なれば異なるオブジェクトである”という仮定が成り立つ．そのため，まずデータを名前ごとに分け，その後，同じ名前を持つデータについてクラスタリングを行うことで同姓同名の問題を解消するという手法が一般的である[3]．

人物などのオブジェクトを識別する際の問題点は，主に名前の曖昧性の解消であった．しかし，それ以外のオブジェクトを識別する際には，別の問題が起こることもある．本論文で提起する問題は，名前が恣意的に付けられるオブジェクトの識別問題である．

例えば料理レシピの場合，同じような料理レシピに対して，“本格タイ風グリーンカレー”や“我が家のグリーンカレー”といったように，様々な名前が付けられている．また，ニュースサイトでは，同一の事件に対して“○○市の事件”や“△△さんの事件”といったようにニュースサイトごとに異なる名前が付けられる．これらの名前は，書き手によって恣意的に付けられるものであるので，同一のオブジェクトに対して名前が多数存在する．このような恣意的に名前付けされたオブジェクトを識別する際には，次の二つの問題点がある．

一つ目は，名前が恣意的であるために，付けられた名前の中に，集約されるオブジェクトとしてふさわしくないものも含まれるという問題である．例えば，“グリーンカレー”と名付けられた料理レシピは多数あるが，グリーンカレーはタイ料理の一つであり，オブジェクトとして集約するのが妥当であると考えられる．一方，“我が家のカレー”と名付けられた料理レシピもいくつかあるが，“我が家”というのは単にその人の家庭で作ったものという意味に対応する場合や，いわゆる“家庭料理としてのカレー”とも解釈できる場合もあり，オブジェクトとして認識すべきかどうかは議論の余地がある．本論文ではこのような場合はオブジェクトとして認識・集約されるべきではないという立場をとる．

二つ目は，同一オブジェクトに多数の名前が存在するという問題である．例えば“我が家のカレー”と名前が付けられていても，実際にはグリーンカレーである場合もある．このような場合でも正しく判定できないと，実際にはグリーンカレーのものがグリーンカレーと判断されず，集約の際に情報が少なくなってしまったり，“グリーンカレー”で検索した際に再現率が下がってしまうなどの問題が起こってしまう．

本論文の目的は，恣意的に名前が付けられたオブジェクトを対象として扱い，それぞれのページがどのオブジェクトについて書かれたものであるかを同定することである．そのためには，世の中にどのようなオブジェクトが存在するのかということと，それぞれのオブジェクトの持つべき性質は何であるのかの２つを自動で取得する必要がある．

料理レシピなどの場合，どの程度似ていれば同一オブジェクトとみなすかが，オブジェクトにより異なる．例えば，肉の種類はグリーンカレーであるかどうかには重要ではないが，チキンカレーであるかどうかには重要である．そのため，あらかじめ材料などの属性の類似度の閾値を決めて，同一かどうかを判断するということは難しい．

本論文では，多くの人が同一と考えているものを同一オブジェクトとみなす．具体的には，多くの人の名前の付け方のパターンをもとに分類階層を抽出し，これ以上細かく分類されていないものをオブジェクトとして集約の単位とする．

† 学生会員 京都大学大学院情報学研究科社会情報学専攻 博士前期課程 takahasi@dl.kuis.kyoto-u.ac.jp
‡ 正会員 京都大学大学院情報学研究科社会情報学専攻 {oyama, tanaka}@dl.kuis.kyoto-u.ac.jp

2. 関連研究

オブジェクト識別に関する研究のうち，名前の曖昧性の解消についての研究は従来から多数行われている．例えば，[4][5][6]の研究では，属性情報などをもとに同姓同名を分離し，実際の人物ごとに集約を行っている．

また，同一オブジェクトの別名について扱っている研究もある．[7]の研究では，ある人物の正式な名称以外の呼び名（ニックネームなど）をWebから抽出している．

本論文で扱う，恣意的に名前付けされるオブジェクトは，同一オブジェクトに名前が多数存在するという点で，別名や表記ゆれの研究などに似ている．しかし，別名や表記ゆれは，同一の名前が多くの人によって使われるのに対し，恣意的に付けられた名前の場合は，一人の人しか使っていない場合も多いという点で異なる．例えば，“京都大学”の別名である“京大”は多くの人が使用する名前であるが，“グリーンカレー”のレシピに付けられた“本格タイ風グリーンカレー”という名前は一人しか使用していないということもある．また，別名の場合はせいぜい数個しか存在しないのに対し，恣意的に付けられた名前の場合は，名前が無数に存在しうるという点でも異なる．

名前が恣意的に付けられるオブジェクトのうち，事件についてのオブジェクト識別についての研究は既に行われている．[8][9]の研究では，全ニュース記事に対して，名前の情報を用いず，記事の内容だけで階層的にクラスタリングを行うことで同一の事件に関する記事を集め，さらに要約の自動生成も行っている．具体的には，様々なクラスタリング手法について，閾値を動かして実験することで，ニュース記事に最適な手法を見つけている．また，名前については，クラスタの中で最も典型的な記事についている名前を与えている．

本論文では，このようにページの内容だけを用いるのではなく，付けられた名前のパターンと，ページの内容の両方を用いて，オブジェクトの識別を行う．以降では料理レシピを具体例として説明する．

3. 本論文の目的

本論文では，オブジェクトについて書かれたWebページが多数あり，各Webページでオブジェクトに名前が恣意的に付けられている場合に，各Webページに書かれたオブジェクトが実際には何であるのかを同定するための手法を提案する．料理の場合，Web上の多数の料理レシピについて，それぞれのレシピがどの料理について書かれたものかを同定するということである．

3.1 提案手法の流れ

前述したように，どの程度似ていれば同一オブジェクトとみなされるかがオブジェクトによって異なるため，事前に属性の類似度の閾値を決めてどのオブジェクトであるかを同定することは難しい．そこで本論文では，まず多くの人の名前の付け方のパターンをもとに分類階層を抽出し，これ以上細かく分類されていないものをオブジェクトとして扱う．

提案する手法の流れは，大きく分けて以下の三段階から構成される．(図 1)

(1)名前の頻出パターンをもとに，分類階層を抽出する

(2)一番下の階層をオブジェクトとし，オブジェクトとなるために属性が満たすべき規則を作成する

(3)規則をもとに，各ページがどのオブジェクトであるかを同定する

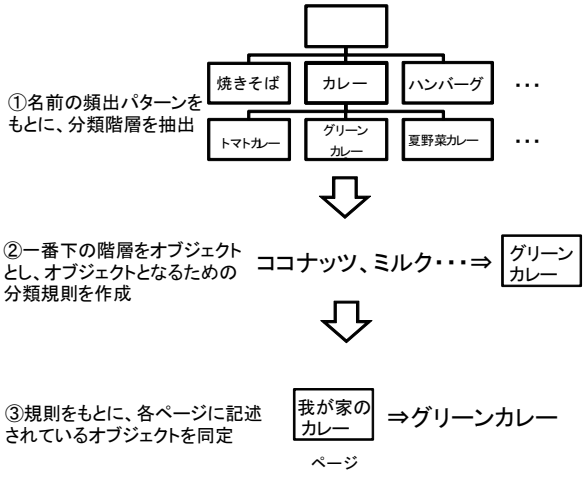

図 1　提案手法の流れ

Fig.1　Overview of the proposed method

3.2 抽出対象とする分類

本論文では，以下の３つの性質を持つものを“分類”として抽出する．

(1)その分類に含まれるデータ数が十分にある

(2)分類に含まれるデータにある程度共通の特徴がある

(3)その特徴が既存の分類の特徴とは異なっている

(1)の条件は，その分類がオブジェクトとしてどの程度認識されているかの指標である．その分類があまり認識されていないということは，その分け方が意味のある分け方だと考えている人が少ないということを表していると考えられる．

(2)の条件は，共通性の低いものを取り除くためである．分散の大きいものも分類と考えることもできるが，オブジェクトの識別という観点からこれはふさわしくないと考える．

(3)の条件は，冗長な分類を取り除くためである．すでに分類として認識されているものがある場合，それと同じ特徴をもつ分類を新たに作る必要はないためである．

たとえば料理の場合，“グリーンカレー”に分類されるレシピにはいくつかの共通の性質があり，通常のカレーとも異なっており，数も十分にあるため，分類名としてふさわしいとして扱う．一方，“我が家のカレー”の場合，それぞれのレシピは通常のカレーとは異なり，数も十分にあるが，“我が家のカレー”の間には共通の性質はほとんどないため，“我が家のカレー”は分類名としてはふさわしくないとして扱う．

4. 提案手法

4.1 提案手法の流れ

分類の抽出は以下の２つのステップによって行う．

＜抽出手順＞

Step1：名前の一部が共通のものを集め分類の候補とする

Step2：分類の候補に含まれるデータについて，分類内の相関と，他の分類との差異を計算することで，分類であるかどうかを決定する

まず Step1 で，名前をもとにデータ数が十分にあること（分類の条件１）を満たすものを集めてきて分類の候補とし，Step2 で共通の性質を持つか（条件２），既存の分類と異なる性質であるか（条件３）を満たしているかを調べ，全てを満たしているものを分類として抽出する．これらについて，順番に説明していく．

なお，以下では“データ”という語を用いるが，各ページに書かれたオブジェクトの名前と属性情報の組という意味で用いる．つまり，一つのページが一つのデータに対応する．

4.1.1 分類候補の抽出

まず，データの名前に出現する部分文字列の出現頻度を用いて分類名候補を抽出する．これは，ある程度のデータには正しい分類名が含まれているという仮定に基づいている．また，“複数語からなる分類名では，後ろに出現する語ほど上位階層を表す”という仮定を用いて，階層的に分類する．たとえば，“グリーンカレー”は“カレー”の一つ下の階層になるように分類する．具体的な手法は以下のとおりである．

＜抽出手順＞

(1)木の根にすべてのデータを入れる

(2)各データの名前 N_i を単語ごとに分割し，$n_{ik}n_{i(k-1)}\cdots n_{i1}$ とする（k は名前に含まれる単語の総数)

(3)n_{i1},$n_{i2}n_{i1}$,…,$n_{ik}n_{i(k-1)}\cdots n_{i1}$ の節点のそれぞれにデータを入れる（節点がなければ新しく作る）

(4)n_{i1} の子に $n_{i2}n_{i1}$，$n_{i2}n_{i1}$ の子に $n_{i3}n_{i2}n_{i1}$…となるように枝を張る

(5)以上のことを全データの名前について行った後，節点に含まれるデータの数が閾値 θ 以上の節点だけを残す

このようにすることで，図 2 のような階層構造が得られる．そのそれぞれの節点が分類候補となる．

4.1.2 分類の判定

Step1 で抽出される分類候補に含まれるデータは，名前に共通性があるだけである．次に，それぞれの分類候補が条件２と条件３を満たしているかを調べる．階層の上位のものから順番に分類候補を選択し，分類であるかを順次決定していく．以下では，選択された分類候補が，分類かどうかを判定するアルゴリズムを示す．

まず，選択した分類候補（集合 A と呼ぶ）に含まれる全てのデータの属性から特徴ベクトル w を求める．特徴ベクトル w の属性 t_j の重み w_j の値は以下の式で求める．

$$w_j = \begin{cases} \mathrm{DF}_A(t_j)\ (\mathrm{DF}_A(t_j) > \alpha N) \\ 0\ (otherwise) \end{cases} \qquad (1)$$

ここで，$\mathrm{DF}_A(t_j)$ は集合 A の中で属性 t_j を含むデータの総数，N は集合 A に含まれる全データの数，α は 1 未満の係数である．つまり，集合内の属性に一定の割合 α より多く出現するものだけを取り出し，共通の特徴としている．

もう一つ，比較対象として，親ノードのデータから集合 A に含まれるデータを除いたもの（集合 B と呼ぶ）に含まれるすべてのデータの属性に関する特徴ベクトル wを求める．

特徴ベクトル wの属性 t_jの重み w'_jは以下の式で求める．

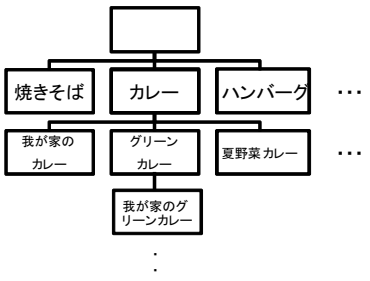

図2　階層的な分類候補

Fig.2　Hierarchical classification of candidates

表1　集合間の属性の出現頻度に有意差があるかを調べる際のカイ２ 乗検定（自由度3の例）

Table 1　Chi-square test for the differences of attribute frequencies between two groups

	属性１	属性２	属性３	属性４	計
集合A	w_1	w_2	w_3	w_4	b_1
集合B	w'_1	w'_2	w'_3	w'_4	b_2
計	a_1	a_2	a_3	a_4	S

$$w'_j = \begin{cases} \mathrm{DF}_B(t_j)\ (w_j > 0) \\ 0\ (otherwise) \end{cases} \qquad (2)$$

ここで，$\mathrm{DF}_B(t_j)$ は集合 B の中で属性 t_j を含むデータの総数である．つまりここでは，集合 A の中に一定数以上出現する属性だけについて，集合 B 内での DF 値を求めている．これは，どちらの集合内でも少数しか出現しないような属性によって有意差が出ることを防ぐためである．

次に，集合 A と集合 B における属性の出現頻度に有意差があるかどうかを，有意水準 p，自由度 d-1 のカイ２乗検定を用いて調べる．なお，d は特徴ベクトル w の中の 0 でない要素の数のことである．そして，ここで有意差があると判断されれば，集合 A は適切な分類と判断できる（条件３）．具体的には次式によりカイ２乗値を求める（表 1)．

$$\chi^2 = \sum_{i=1}^{d}\sum_{j=1}^{2} \frac{(w_{ij} - a_i b_j / S)^2}{a_i b_j / S} \qquad (3)$$

ここで，

$w_{i1} = w_i,\ \ w_{i2} = w'_i,\ \ a_i = w_i + w'_i$

$$b_1 = \sum_{k=1}^{d} w_k\,, b_2 = \sum_{k=1}^{d} w'_k, S = b_1 + b_2$$

である．

この式で求めたカイ２乗値が有意水準 p，自由度 d-1 のカイ２乗値よりも大きければ有意差があると判断できる．有意差がない場合は，対象の分類候補が一つ上の階層とはあまり変わらないという意味であるので，新たにこの分類を作る必要はないと判断され，子ノードとともに分類木から削除する．これを繰り返し，最終的に残ったものが分類階層となる．

4.2 分類規則の作成

この階層木の一番下の階層をオブジェクトとして扱う．次に，各オブジェクトに分類されるための規則を求める．前節で集合間の有意差を調べたが，今度はどの属性で出現頻度に有意差があったのかを調べるために，各属性ごとに集合 A と集合 B とでの出現頻度の違いを表 2 のような有意水準 p，自由度１のカイ２ 乗検定で調べ，有意差があった属性だけを分類規則として記憶する．具体的には次式によりカイ２ 乗値を求め，各属性が分類規則に含まれるかを判定する．比較対象となる集合Ｂは先ほどと同様，一つ上の階層のものである．

表2　属性1が分類規則に含まれるか判定するカイ２乗検定

Table 2　Chi-square test to decide whether the identification rule includes attribute 1

	属性１を含む	属性１を含まない	計
集合A	w_1	w_2	b_1
集合B	w'_1	w'_2	b_2
計	a_1	a_2	S

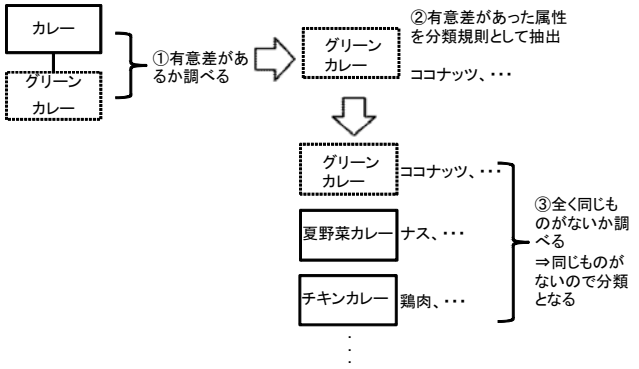

図3 グリーンカレー（点線）が分類として判定される過程
Fig.3 The process to decide whether "green curry" is an appropriate classification

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(w_{ij} - a_i b_j / S)^2}{a_i b_j / S} \qquad (4)$$

ここで，
$w_{i1} = w_i,\ w_{i2} = w'_i,\ a_i = w_i + w'_i$
$b_1 = w_1 + w_2, b_2 = w'_1 + w'_2, S = b_1 + b_2$
である．

有意差のある属性を，全て分類Aの分類規則として記憶する．この分類規則を既に分類と判定されたものの分類規則と比較し，同じであれば同一の分類であるとみなして統合する．

一番上位の階層の分類について調べた後は，一つ下の階層についても先ほどと同様に調べる．これを一番下の階層まで再帰的に繰り返すと，適切な分類だけが残る．

前節と本節の内容を，図 2 の階層木から，“グリーンカレー”が分類として判定される過程で説明する（図 3）．

まず，“グリーンカレー”に分類されている全レシピを集め（集合 A），その中で一定の割合以上のレシピに含まれる語をもとに特徴ベクトルを作る．なお，Step1 での木の作り方から，“我が家のグリーンカレー”のように“グリーンカレー”の下の階層に含まれるレシピも“グリーンカレー”の分類に含まれている．

“グリーンカレー”の比較対象となるのは，“グリーンカレー”以外の全てのカレーであり（集合 B），集合 A の中で一定数以上出現する語だけを用いて，集合 A と B の間に有意差があるかを調べる（図 3①）．有意差がなければ集合 A は分類ではないとみなして階層木から削除する．実際には有意差があるので，どの属性について有意差があるのかを調べ，有意差がある属性を分類規則として記憶する．例えば，ココナッツやミルクなどが分類規則となる．（図 3②）．次に，“グリーンカレー”の分類規則を，兄弟ノードである“夏野菜カレー”や“チキンカレー”の分類規則と比較する（図 3③）．もしどれかと完全に一致すれば二つを同一の分類とみなし統合されるが，実際は同一の分類規則をもつものは存在しないので，最終的にグリーンカレーは分類と判定される．

4.3 分類規則を用いたデータの同定

以上により，オブジェクト名とそれに分類されるための規則が求まった．最後に，各ページがどのオブジェクトについての記述なのかを属性の情報のみで同定する．この際，データに付けられた名前の情報は利用しない．それは，名前が恣意的に付けられているため，あるオブジェクト名が付けられていても，それが間違いである可能性があるためである．

前節で求めた分類規則は，その分類を特徴づける属性が抽出されている．そこで，各ページが，分類規則に含まれる属性を全て含んでいるかによって，その分類に分類されるべきかどうかを判定する．例えば“グリーンカレー”の分類規則には，ココナッツやミルクなどがある．そこで，それぞれのレシピがこれらの属性を含んでいれば“グリーンカレー”と判断する．以上のようにして同定を行う．

なお，その際，あるページが複数の分類の分類規則を満たすことも考えられるが，その場合も複数の分類に同定してよい．例えば，ある料理レシピが“トマトカレー”でもあり“チキンカレー”でもある場合もあるためである．

5. 実験と考察

実験では，恣意的に名前付けされたオブジェクトの具体例として，料理について取り上げ，得られる料理名の数とその精度の関係に関する実験と，各レシピがどの料理について書かれたものなのかを同定する実験の 2 つを行った．

実験で使用したデータは，投稿型レシピサイトの一つであるクックパッド[10] から，カレー，ハンバーグ，ヤキソバのレシピ計 5,894 件を取得したものである．その中には全部で 4,826 種類の名前が存在した．各レシピは名前，材料や作り方などで構成されている．食材，調理器具，調理動作などが各レシピの属性となるように，各レシピの本文を形態素解析器 MeCab [11] を用いて名詞と動詞を切り出し，それらをレシピの特徴ベクトルとした．つまり，“ナス”などの材料や“煮る”などの動作などが，特徴ベクトルの属性となる．なお“焼きそば”と“ヤキソバ”のような表記ゆれを統一するために必要に応じて辞書を作成した．

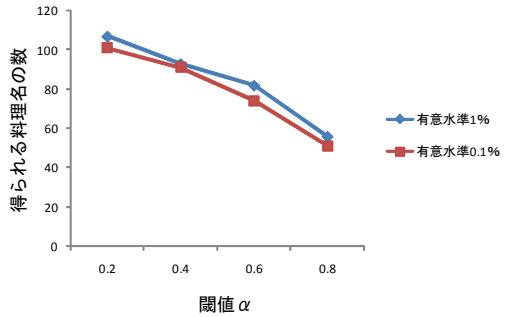

図4 閾値，有意水準と得られる料理名の数の関係
Fig.4 Relationships among threshold values, significance levels and numbers of dish names

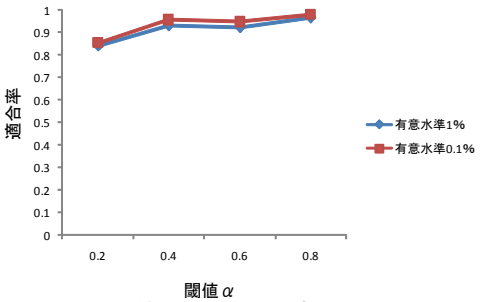

図5 閾値，有意水準と適合率の関係
Fig.5 Relationships among threshold values, significance levels and precision

表3　α=0.4, 有意水準 0.1%の時に得られたカレー料理名
Table 3　Extracted names for curry dishes when α=0.4, p =0.1%

ドライカレー	エビカレー
チキンカレー	ゴーヤカレー
キーマカレー	大根カレー
グリーンカレー	イエローカレー
スープカレー	ミンチカレー
トマトカレー	ヨーグルトカレー
タイカレー	ツナカレー
豆カレー	キャベツカレー
夏野菜カレー	キノコカレー
ヒキ肉カレー	ミートボールカレー
牛スジカレー	トウフカレー
ナスカレー	チーズカレー
和風カレー	オムカレー
シーフードカレー	ミソカレー
焼カレー	オカラカレー
ホウレンソウカレー	ゴボウカレー
ココナッツカレー	サバカレー
ポークカレー	甘口カレー
レッドカレー	豆ドライカレー
インドカレー	オカラドライカレー
ビーフカレー	和風ドライカレー
ヒキニクカレー	豆キーマカレー
カボチャカレー	

5.1 得られる料理名の数と精度の関係

一つ目の実験では，4 節で挙げた手法に使われる各種パラメータ α，θ，p を変えることによって，得られる料理名の数がどのように変わるかについて調べた．なお，α は各分類の何割以上に含まれていれば分類共通の属性とみなすかの値，θ は分類とみなす最低のインスタンス数，p はカイ 2 乗検定を行う際の有意水準である．θ を動かすことによって得られる分類数の違いは，他のパラメータとは無関係であるので，今回は θ を 5 に固定して α と p を動かし，得られる料理名の数の違いを調べ，図 4 と図 5 に示した．また α= 0.4，有意水準 0.1 %の時に得られたもののうち，上位階層がカレーであるものを表 3 に示す．

図 4 の縦軸は，全レシピについて提案手法を適用し，最終的にオブジェクトと判断されたもののうち，人手で正解と判断されたものの数である．また，図 5 の縦軸の適合率は，提案手法が出力した結果のうち，人手で正しいと判断された割合である．有意水準は，一般的によく使われる 1%と 0.1%の 2 つで行い，α は 0.2,0.4,0.6,0.8 の 4 つについて行った．

まず，有意水準 p に関してであるが，有意水準を低くすると適合率は上がるが得られる料理名の数が少なくなるという傾向はあるものの，大きな差は出ないことがわかる．

次に，α について考察する．α の値は分類の何割以上に含まれていれば分類共通の属性とみなすかについての値である．つまり，α 未満の割合でしか現れない属性語は，検定の際には使用されない．α が大きい場合には，その分類にかなり頻繁に現れる属性しか使用しないため，適合率は上がるものの得られる料理名の数は少ない．一方，α が小さい場合には，検定に使用する語が増えるために有意差が現れやすくなるが，その分類の特徴とは直接関係のない属性に影響されてしまい，適合率が少し下がってしまうことがある．

また，本手法と比較するために，単純に 5 回以上出現する名前のパターン全てを料理名として出力するというベースライン手法を用意した．ベースライン手法の適合率は 73.8 %であったため，提案手法により適合率が約 20 %ほど上がると言える．一方，ベースライン手法で得られた料理名の数は 124 であったため，提案手法では 20 個程度得られなかったことがわかる．その理由は，5.3 節で説明する．

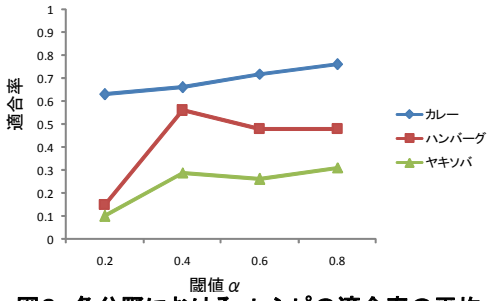

図6　各分野における，レシピの適合率の平均
Fig.6　Average precision for recipes in each category

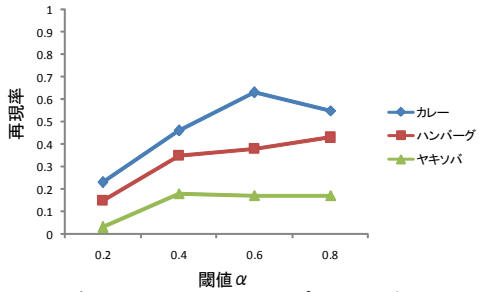

図7　各分野における，レシピの再現率の平均
Fig.7　Average recall for recipes in each category

5.2 各レシピが言及しているオブジェクトの同定

各レシピが言及しているオブジェクトを同定する際には，分類規則を使用する．今回は，比較の際に出現頻度に有意差があった属性を取り出し，あるレシピの中にその属性がすべて現れれば，そのオブジェクトとして同定する．

前節の実験により，有意水準はそれほど結果に影響を与えないということがわかったので，有意水準は 0.1 %に固定し，α の値を先ほどと同様に動かし，同定の精度がどのように変化するのかを調べた．精度の評価方法は以下のように行った．

まず，カレー，ハンバーグ，ヤキソバの 3 分野について，レシピを 30 件ずつ用意する．次に，各レシピが言及しているオブジェクトを人手で同定しそれを正解とする．この際，あるカレーのレシピが，“ポークカレー”でも“ゴーヤカレー”でもあるという場合もあるが，その場合はその全てを正解とする．次に，システムに各レシピを入力し，提案手法により同定させる．そして，以下の 2 つの尺度により評価する．

・レシピの適合率
　システムが同定したオブジェクト名のうち，正解の割合
・レシピの再現率
　正解のうち，システムが同定したオブジェクト名の割合

分野ごとに，30 個のレシピの適合率・再現率の平均を取り，その結果を図 6,7 に示した．図から分かる通り，カレーの分野ではある程度の精度であるが，他の二つの分野では，かなり低くなってしまっている．

5.3 実験結果の考察

提案手法では，それぞれの分類候補が分類であるかを判定する際に，一つ上の階層の分類と，対象の分類候補の 2 つの間の属性の出現頻度に有意差があるかをカイ 2 乗検定を用いて判定を行った．その際，分類候補に比較的頻繁に現れる属性のみを用いて検定を行っている．そのため，“肉が入っていないこと”が条件である“野菜カレー”のように，特定の属性が含まれていないということが条件となるような分

類の場合，分類規則はうまく抽出することができない．それは，“野菜カレー”には“肉”という属性がほとんど存在せず，現在の手法のままでは比較の際の属性として用いられないためである．これを解決するために，集合間の比較の際に使用する語を増やすことが考えられる．つまり，分類候補内の頻出属性だけを用いるのではなく，比較対象の分類の頻出属性も用いるものである．こうすれば，“肉”も比較の際の属性に含まれるため，“野菜カレー”が抽出される可能性がある．

また，各レシピがどの分類についての記述であるかを同定するために，分類規則を求めている．現在の手法では，比較の際に有意差があった属性を分類規則とし，その全てを含んでいるレシピのみをその分類に同定している．しかし 5.2 節のとおり，その同定の精度は分野によっては低くなってしまう．分類規則に含まれる属性が一つのときには，その属性が含まれるかどうかで分類の同定ができる．例えば“トマトカレー”はカレーにトマトが入っていればよく，本手法で正しく同定できる．カレーでは，このような例が多かったために，他の分野と比べて適合率・再現率が高くなったと考えられる．しかし，現実の分類規則は，このような単純なものばかりではない．例えば，“和風ハンバーグ”では，醤油を使うものと大根おろしを使うものの大きく２つが存在する．この場合，どちらかの材料が含まれていればよいが，現在の手法では両方含んでいなければ和風ハンバーグと同定されないため，レシピの再現率が下がってしまう．また，塩が入っているヤキソバが塩ヤキソバではなく，ソースが入っていてはいけない．現在では，塩が入っているヤキソバを全て塩ヤキソバとしているため，適合率が下がってしまった．

6. まとめと今後の課題

本論文では，恣意的に名前付けされたオブジェクトの識別問題という新しい問題を提起した．このようなオブジェクトを識別する場合，オブジェクトに付けられた名前が恣意的であり，多種多様な名前が存在するため，名前の情報だけで２つのページが同一オブジェクトについての記述であるかを判断することができない．また，オブジェクトの粒度も容易に決定することができない．このために本論文では，多くのユーザの名づけパターンから、分類の粒度を決定する手法を提案した．

実験は，オブジェクトを抽出する実験と，各レシピに書かれたオブジェクトを同定する実験を行った．前者は，抽出漏れは多少あるものの，抽出されたものでは 90%程度の精度が得られた．一方後者の実験では，一番精度の低いヤキソバの分野では 20%程度となってしまった．その原因は，同定の規則の作り方が単純すぎるためであり改善の必要がある．

また本手法の問題点として，分類の抽出方法が名前の付けられ方に依存していることがある．そのため，名前には出現しないような分類は抽出できず，今後の課題である．

本論文で提案した手法の評価実験では料理レシピデータを対象にしたが，提案手法自体は名前と属性の組が与えられれば識別できる一般的な手法であり，今後は，事件名データなど多様なデータに対して適用していきたいと考えている．

[謝辞]

本研究の一部は，科学研究費補助金（課題番号 18049041，18049073，21700106），京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトによるものです．ここに記して謝意を表します．

[文献]

[1] Z. Nie, J.-R. Wen, and W.-Y. Ma, “Object-level Vertical Search,” CIDR 2007, pp.235-246.
[2] M. Wang, Z. Li, L. Lu, W.-Y. Ma, and N. Zhang, “Web Object Indexing Using Domain Knowledge,” KDD 2005, pp.294-303.
[3] S. Oyama and K. Tanaka, “Distance Metric Learning from Cannot-be-linked Example Pairs, with Application to Name Disambiguation,” S. Basu, I. Davidson and K. Wagstaff Eds., In Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapter 15, pp. 357-374, Chapman & Hall/CRC Press, 2008.
[4] 森純一郎，松尾豊，石塚満，“Web からの人物に関するキーワード抽出，”人工知能学会論文誌，Vol.20，No.5，pp.337-345, 2005.
[5] R. Kimura, S. Oyama, H. Toda, and K. Tanaka, “Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction,” ICWE 2007, pp.400-414.
[6] 上田洋，村上晴美，“Web 上の同姓同名人物を分離して人物属性情報を表示するシステム，” 2007 年度人工知能学会全国大会(第 21 回) 論文集, 2007.
[7] 外間智子，北川博之，“Web データを用いた人物の呼称抽出，” DBSJ Letters, Vol.5, No.2, pp. 49-52, 2006.
[8] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, C. Sable, B. Schiffman, and S. Sigelman, “Tracking and Summarizing News on a Daily Basis with Columbia’s Newsblaster,” HLT 2002, pp.280-285.
[9] V. Hatzivassiloglou, L. Gravano, and A. Magnati, “An investigation of linguistic features and clustering algorithms for topical document clustering,” SIGIR 2000, pp.224-231.
[10] クックパッド, http://cookpad.com/
[11] MeCab, http://mecab.sourceforge.net/

高橋 良平 Ryohei TAKAHASHI
京都大学大学院情報学研究科博士前期課程在学中．2009 年京都大学工学部情報学科卒業．主に Web マイニングの研究に従事．日本データベース学会学生会員．

小山 聡 Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助教．2002 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI 各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京大工博．主にデータベース，マルチメディアコンテンツの処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．