Web からの語集合間の特定関係抽出とその可視化

Identifying Relationships of Term Sets Extracted from the Web and their Visualization

稲川 雅之♥ 大島 裕明♦
小山 聡♦ 田中 克己♦

Masayuki INAGAWA Hiroaki OHSHIMA
Satoshi OYAMA Katsumi TANAKA

本稿では，人物や作品などのオブジェクトに対して，ある語集合がオブジェクトを特定するという関係を Web から抽出する手法を提案する．本手法では，オブジェクト名を入力し，それを特定する語集合を出力することができる．逆に，語集合を入力とし，それが特定するオブジェクトを出力することも可能である．これにより，語集合からオブジェクトへのマッピングができるため，例えば名前の分からないオブジェクトを探す場合などに用いることができる．また，オブジェクトを一般化して語集合に拡張することにより，語集合が語集合を特定するという関係を得ることができる．本稿ではさらに，このような特定関係をグラフとして表現することで可視化を行う．可視化により，オブジェクト間あるいは語集合間の関係を把握することや，入力語集合の周辺の話題を探索するといったことが直感的に可能となる．

In this paper, we propose a way to extract a relationship that a term set identifies an object such as a person or a book from the Web. Our system makes it possible to obtain term sets which identify a given object. Conversely, we can also obtain the object which is identified by a given term set. Those term sets can be useful, for example, when looking for the object whose name is forgotten since we can make the mapping from term sets to an object. Moreover, we can extract a relationship that term sets identify a term set by generalizing an object to a term set. We also propose a method to visualize such identifying relationships by using graph representation. This method enables users to comprehend relationships between term sets and search surrounding topics of an input term set intuitively.

1. はじめに

近年，インターネットとWeb検索エンジンの普及により，多くの調べ物を検索エンジンを用いて手軽かつ迅速に行うことが可能になった．しかし，未だに既存の検索エンジンが苦手とする分野も多い．現在の検索エンジンは1つあるいは複数のキーワードを入力し，そのキーワードを含むWebページの集合を結果として返すものが一般的である．このようなページベースの検索システムは，うろ覚えの人物名を周辺情報から探したり，作品や地名などのオブジェクトの情報を収集したりといった目的に適しているとは言えない．何故なら，うろ覚えのオブジェクトを探し出すのに最適なキーワードを思いつけるとは限らないからである．また，たとえユーザが入力したクエリで検索されたページの中に求めるオブジェクトについての情報が含まれていたとしても，検索結果中のページに含まれるオブジェクトについての記述の中から重要な部分を見つけ，それらの情報を集約するという作業が必要になる．

本研究では，これらの問題に対する1つのアプローチとして，語集合がオブジェクトを特定するという関係をWebから抽出することを考える．ここで「特定する」とは，ある語集合が与えられた時に，それに対応するオブジェクトを人が一意に判断できるという意味である．

オブジェクトは語として表現され得るため，多義性や表記揺れなどの問題を考えなければ，語集合からオブジェクトへの特定関係は，語集合から語への特定関係と捉えることができる．ここで，特定されるものを語から語集合に拡張すれば，語集合から語集合への特定関係，すなわち語集合間の特定関係を考えることになる．本稿では，この語集合間の特定関係を抽出する手法を提案する．

語集合間の特定関係の応用としては次のようなものが考えられる．1つは名前の分からないオブジェクトを探すためのインデックスとして用いることである．うろ覚えで名前が思い出せないようなオブジェクトを探そうとした時，あらかじめオブジェクトに対応する語集合が分かっていれば，その語集合からオブジェクトを見つけることが可能となる．実際にオブジェクト名の検索システムを作る場合は，女優や観光地名などの特定のドメインに属するオブジェクトを集めて，それらをそれぞれ特定する語集合を求める，すなわちインデキシングを行っておけば，そのドメイン内のオブジェクト名発見が実現できる．

また，あらかじめインデックスを作成することはせず，ユーザが与えたクエリからその場で特定関係を求め，それを提示することでユーザの情報探索を支援することも考えられる．このアプリケーションの1つとして，語集合の特定関係のグラフ表現による可視化システムを構築した．このシステムを用いれば，ユーザが入力したクエリ（語あるいは語集合）からの特定関係を直感的に把握することが可能となる．さらに，クエリと特定関係にある語集合と特定関係にある別の語集合を簡単な操作で求めることが出来，特定関係グラフを次々と広げていくことができる．これは，ユーザが与えたクエリを出発点とする関連概念の探索と捉えることができる．

特定関係の抽出手法についてであるが，特定関係には当然ながら向きが存在するため，クエリを特定するものを抽出するのか，クエリから特定されるものを抽出するのかによって手法が異なってくる．例えば「人間失格 斜陽 作家」が「太宰治」を特定するという関係を例にとると，前者は「太宰治」をクエリとするものであり，後者は「人間失格 斜陽 作家」をクエリとするものである．本研究で提案する手法は，特定関係の抽出手法に対称性を持たせることによって，どちらの向きにでも適用可能な手法となっている．

具体的な特定関係の抽出は次の2段階に分けて行う．1段階目ではクエリからそれを特定する／それから特定される語集合の候補を求める．2段階目では求められた候補のそれぞ

♥ 学生会員 京都大学大学院情報学研究科修士課程
inagawa@dl.kuis.kyoto-u.ac.jp
♦ 正会員 京都大学大学院情報学研究科
{ohshima, oyama, tanaka}@dl.kuis.kyoto-u.ac.jp

れについて，クエリを特定する／クエリから特定される語集合として適している度合を評価し，この値に従って語集合をランキングしたものを最終的な出力とする．

本稿の構成は以下の通りである．1章で関連研究について述べ，2章で語集合間の特定関係とは何かについて述べ，3章で候補語集合の抽出手法について述べ，4章で候補語集合との特定関係の評価指標について述べ，5章で実験とその結果について述べ，6章で特定関係の可視化について述べ，7章では本研究のまとめと今後の課題について述べる．

2. 関連研究

語間の関係を発見する研究は広く行われている[1]～[3]．これらの研究は，文書集合から下位語や類語，関連語などを抽出するものである．また，Webを利用した語間の関係性抽出手法も数多くの先行研究がある．Shinzatoらの研究[4]はHTMLで記述されたWebページの文書構造を利用し，上位・下位関係を獲得するものである．Ohshimaら[5]は言語パターンを用い，Web検索エンジンのインデックスを利用して同位関係にある語を抽出している．また，Nakayamaら[6]はWeb上の百科事典であるWikipediaから関連シソーラスを構築する手法を提案している．

これらの研究は，複合語などのフレーズを対象とするものもあるが，基本的に1語と1語の関係を扱うものである．一方本研究は，2つ以上の語の組み合わせ，すなわち語集合が別の語集合と結びつくといった関係を扱うことができる．また，我々が提案するような「特定」関係を抽出する研究も他にはないと言える．

3. 語集合間の特定関係

本章では，「語集合間の特定関係」の定義および抽出手法の概要について述べる．

3.1 語集合の特定関係の定義

本節では「語集合」とは何か，また，語集合間の特定関係とは何かについて述べる．

「語集合」とは文字通り語の集合であり，要素となる語に順序はないものとする．また，1つの語も大きさ1の語集合と考えることが可能である．

「語集合Aが語集合Bを特定する」という関係をA→Bと表す．このときAを特定語集合，Bを被特定語集合と呼ぶものとする．

語集合Aが語集合Bを「特定する」とは，語集合Aを提示したとき，語集合Bが唯一つ導き出されると人が判断できることを言う．これは「連想する」あるいは「想起する」と言い換えることも出来るが，「特定する」と言う場合はこれらの意味よりも導き出されるものが限定されることになる．1つの特定語集合に対してそれが特定する被特定語集合は1つだけであるが，1つの被特定語集合に対してそれを特定する特定語集合は複数存在し得る．すなわち，「語集合が語集合を特定する」という関係は多対一対応であると考えられる．ただし，後に述べるシステムにおいては，クエリとして与えられた特定語集合に対してそれから特定される被特定語集合を出力する際，定義上は1つに決まらなければならないが，応用上の有用性を確保するため，被特定語集合の候補を評価値の高いものから複数出力することを前提としている．

特定語集合として適切でないものとして，一般的な語の集まりが挙げられる．例えば，｛夏目漱石｝という（1語の）語集合に対して｛小説，作家｝という語集合が得られたとする．この語集合の要素である「小説」も「作家」も，どちらも「夏目漱石」についてよく述べられる重要なキーワードではあるが，これらの語だけで「夏目漱石」を特定することは不可能である．つまり特定語集合としては，その被特定語集合について主要なキーワードであるかどうかだけでなく，その語集合でどの程度の強さで被特定語集合を絞り込めるかということも重要である．

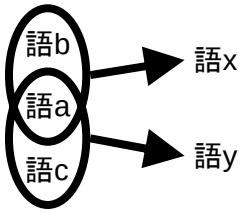

図1 特定関係のイメージ
Fig.1 Image of identifying relationships

図1は語集合の特定関係のイメージをシンプルに表現したものである．簡単のため被特定語集合が1語の例を挙げている．図中の矢印は矢印元の語集合が矢印先の語を特定するということを表す．この図において，語aは語xおよびyのどちらとも関連する語であると思われる．よって，この語aただ1語だけではどちらの語も特定することは出来ない．しかし，語aに別の語bを加えた語集合を考えると，この語集合は語xを特定することが出来る．また，語aに語cを加えた語集合を考えると，この語集合は語yを特定することが出来る．

このように，1語ではどの語集合も特定できないが，別の語を加えることによりある語集合を特定できる場合があることが考えられる．オブジェクト検索の視点から見れば，ユーザが語aというクエリを与えた時，オブジェクトxおよびyという2つの候補が存在するが，語bと語cを提示し，どちらかを選ばせることにより，ユーザが探しているものがどちらであるのかを判断することが出来る．

3.2 特定関係の抽出手法

本研究において提案する手法は，ある語集合を入力とし，その入力語集合との特定関係を抽出するものである．以降ではこの入力語集合をクエリと呼ぶ．ここで，特定関係には向きが存在するため，クエリと特定関係を持つ語集合には2つの種類がある．

・順方向：クエリは特定語集合であり，クエリから特定される被特定語集合を求める

・逆方向：クエリは被特定語集合であり，クエリを特定する特定語集合を求める

本手法は，語集合間の特定関係の評価指標に対称性を持たせることで，ほぼ同じ手順で両方向の特定関係を求めることが可能となっている．

クエリと特定関係を持つ語集合の抽出手法は次の2段階に分けて行う．

（1）候補語集合の抽出
（2）候補語集合の評価

これらの段階を4章と5章でそれぞれ述べる．

4. 候補語集合の抽出

候補語集合の抽出は大まかに以下のような手順で行う．この手順は順方向・逆方向共に同じである．

（1）クエリが含まれるWebページを取得する．
（2）クエリの周辺テキストに出現する語にクエリとの距

離に応じた重みを付け，ページの特徴ベクトルを生成する．

（3）各ページの特徴ベクトルから語集合がクエリの周りに出現する頻度を計算し，この値が大きいものを候補語集合として抽出する．

（3）はクエリの近くに頻出する語集合を抽出するということだが，このような語集合を求める理由は，クエリの周辺によく出現する語集合はクエリとの関連が強いと考えられるからである．以下では各手順について詳しく述べる．

4.1 クエリを含むWebページの取得

クエリの周辺テキストを取得するため，Web検索エンジンを用いてWebページ検索を行う．Web検索エンジンへの入力はクエリ（本手法への入力語集合）である．検索エンジンとしてはYahoo! JAPANがAPIを提供しているウェブ検索Webサービス[1]を用いた．

4.2 Webページの特徴ベクトル生成

クエリが出現するWebページを取得したら，クエリの周辺に出現する語にクエリからの距離を考慮した重みを与えることによって各ページの特徴ベクトルを生成する．距離の概念を導入してはいるが，この特徴ベクトルは基本的には頻度ベクトルである．クエリとの距離を考慮する理由は，クエリの近くにある語の方がよりクエリとの関連が強いと考えられるためである．

まず，特徴ベクトルの1つの次元となる語を抽出する．4.1節によって得られたWebページのテキストを形態素解析器にかけ，日本語文の意味的な最小単位である形態素に分割する．本研究では形態素解析器としてMeCab[7]を用いた．

分割された形態素のうち，品詞が名詞であるものを残す．ただし代名詞は除外する．また，連続する複数語の名詞は1語の名詞として扱う[2]．これは人名を含めた複合名詞を正しく扱うことが出来るようにするためである．さらに「人」や「子」のように，経験的にノイズになりやすいと分かっている一般的過ぎる語をストップワードとして除外する．

以上のような操作で抽出された語に対する特徴ベクトルの重み（頻度）を計算する．

まず，クエリと語との距離を定義する．ある語の出現aに対して，クエリQとの距離は，クエリに含まれる各語との距離の平均であり，

$$\mathrm{d}(a,Q)=\frac{\sum_{q\in Q}\mathrm{dist.}(a,q)}{|Q|} \quad (1)$$

と定義される．ここで，$\mathrm{dist.}(a,q)$は出現aとクエリ中の語qとの距離であり，具体的には，Webページから抽出した語をテキスト中での出現順に並べ，qに近い語から順に1，2，3...と距離を与えていく．qがページ中に複数出現する場合は最も近い値をとる．

クエリQについてのページP_kの特徴ベクトル$\boldsymbol{v}_{PkQ}$の語w_iに対する重み$v_{PkQ}(w_i)$は以下のように定義される．

$$v_{P_kQ}(w_i)=\sum_{a\in A_{ki}}\frac{1}{\log(\mathrm{d}(a,Q)+1)} \quad (2)$$

ここで，A_{ki}はページP_kにおける語w_iの出現の集合である．

式(2)によって各ページに出現する各語に対する重みを計算し，特徴ベクトルを生成したら，最後に各ページの特徴ベクトルを長さ1に正規化する．これはページの文章量やクエリの出現回数による影響をなくすためである．ただし，クエリとの距離が離れるほど重みに加算される値は小さくなっていくので，この正規化による結果の変化はそれ程大きくない．

4.3 語集合の周辺頻度

ページの特徴ベクトルを生成したら，次は語集合を生成し，その語集合の各ページに対する重みを求め，その総和を計算する．

4.3.1 周辺頻度の定義

語集合の周辺頻度(Surrounding Term Set Frequency)とは，対象とするクエリの近くにその語集合が出現する度合（頻度）を表す値である．この周辺頻度を以下の手順で計算する．

クエリQに対して，ある語集合WのページP_kに対する重みweight(W, P_k, Q)を以下のように定義する．

$$\mathrm{weight}(W,P_k,Q)=\min_{w_i\in W}v_{P_kQ}(w_i) \quad (3)$$

すなわち，語集合Wの要素である語のうち，ページP_kに対する重みが最も小さいものの重みをP_kに対するWの重みとするということである．

語集合Wの各ページに対する重みの和をクエリQに対するWの周辺頻度とする．クエリQに対するWの周辺頻度stsf(W,Q)は以下のように求められる．

$$\mathrm{stsf}(W,Q)=\sum_k\mathrm{weight}(W,P_k,Q) \quad (4)$$

ただし，P_kはクエリQが出現するページである．

前述のように，この周辺頻度は語集合WがクエリQの周辺にどの程度の頻度で出現するかを表す値である．

4.3.2 語集合の生成

式(4)を用いて語集合の周辺頻度を求めるのであるが，単純にクエリの周辺テキストに出現する語の全ての組み合わせを作り，その周辺頻度を求めると，容易に組み合わせ爆発を起こす．そこで，相関ルール抽出アルゴリズムApriori[8]の手法を参考にし，枝刈りを行いながら語集合の生成と周辺頻度の計算を行うことにする．

まず閾値θを定める．これはAprioriアルゴリズムにおける最小サポートに対応する．枝刈りの考え方は，この閾値を下回る周辺頻度を持つ語集合を生成しないというものである．

大きさ1の語集合から始め，以下を繰り返す．

（1）大きさnの語集合のうち，周辺頻度が閾値θを下回るものを捨てる．

（2）大きさnの語集合に1語追加することにより大きさn+1の語集合を生成する．この際，生成された語集合の全ての真部分集合は捨てられていない，すなわち周辺頻度が閾値を上回っているものとする．

（3）生成された大きさn+1の語集合を対象として（1）に戻る

以上の繰り返しが終了するのは，生成された語集合が1つもない場合や，あらかじめ設定された語集合の大きさの上限に達した場合である．

以上の操作によって生成された，閾値以上の周辺頻度を持つ語集合を候補語集合とする．また，候補の数が多い場合は周辺頻度の大きいものから上位K件を候補とする，といったことも可能である．

5. 候補語集合の評価

前章で抽出された各候補語集合について，クエリに対する特定するもの／されるものとしての評価指標を定義し，これ

[1] Yahoo!デベロッパーネットワーク http://developer.yahoo.co.jp/

[2] ただし「さん」などの人名に付く接尾語は除く

を計算することにより，特定関係の強さを評価する．

5.1 特定関係の要件

「語集合Aが語集合Bを特定する」という関係が成り立つための要件は以下の2点である．

（1）語集合Aで語集合Bを絞り込める

（2）語集合Aが語集合Bについて主要な話題を表している

（1）は，語集合Aを提示されたとき，高い割合で語集合Bと結びつくことを意味する．例えば，「斜陽　人間失格」という語集合を提示されれば，多くの人は「太宰治」という（1語の）語集合を思い浮かべるだろう．この絞り込める度合を評価する指標を「確定度」と呼ぶ．（2）は，語集合Bにとっての語集合Aの話題の大きさを評価するものである．「斜陽　人間失格」という語集合は「太宰治」にとってそれなりに大きな話題を表すものである．この，特定語集合の話題の大きさを評価する指標を「メジャー度」と呼ぶ．

5.2 確定度・メジャー度の定義

確定度およびメジャー度は，語集合を含むWebページの重なりを用いて計算する．

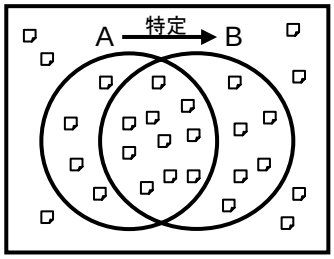

図2　語集合AおよびBを含むWebページの重なり
Fig.2　Overlap of pages including term set A and B

語集合Aが語集合Bを特定するという関係A→Bを考える．図2は，語集合Aおよび語集合Bを含むWebページの重なりをベン図を用いて表したものである．図中のA，Bはそれぞれ語集合A，Bを含むWebページ集合である．これらの集合の大きさ，すなわちWebページ数を用いて確定度とメジャー度を以下のように定義する．

確定度　$\dfrac{|A\cap B|}{|A|}$

メジャー度　$\dfrac{|A\cap B|}{|B|}$

確定度は，特定語集合Aを含むページのうち，さらに被特定語集合Bも含むページの割合である．一方，メジャー度は被特定語集合Bを含むページのうち，さらに特定語集合Aも含むページの割合である．これらはどちらも語集合AとBの片側共起度であると言える．

実際に語集合を含むWebページ数を求めることは困難であるため，ここでは語集合をクエリとしてWeb検索エンジンで検索を実行した際の検索総件数（ヒット数）をその近似値として用いる．

語集合$A=\{a_1,\ldots, a_m\}$が語集合$B=\{b_1,\ldots, b_n\}$を特定するという関係A→Bを考えたときの確定度determ(A, B)とメジャー度major(A, B)は以下のように求められる．

$$\mathrm{determ}(A,B)=\frac{DF(a_1\wedge\cdots\wedge a_m\wedge b_1\wedge\cdots\wedge b_n)}{DF(a_1\wedge\cdots\wedge a_m)} \quad (5)$$

$$\mathrm{major}(A,B)=\frac{DF(a_1\wedge\cdots\wedge a_m\wedge b_1\wedge\cdots\wedge b_n)}{DF(b_1\wedge\cdots\wedge b_n)} \quad (6)$$

ここで$DF(a_1\wedge\cdots\wedge a_m)$は，語$a_1\cdots a_m$全てでAND検索した際の検索総件数である．Web検索エンジンとしては4.1節と同様Yahoo! JAPANが提供するAPIを利用した．

5.3 特定度の定義

前節で定義した確定度とメジャー度を用いて，特定関係A→Bの強さを表すスコアである特定度identify(A, B)を以下のように定義する．

$$\mathrm{identify}(A,B)=\frac{(1+\beta^2)\times\mathrm{determ}(A,B)\times\mathrm{major}(A,B)}{\mathrm{determ}(A,B)+\beta^2\times\mathrm{major}(A,B)} \quad (7)$$

ここで，βは0から正の無限大までの値をとるパラメータであり，メジャー度に対して確定度を何倍重視するかを表す値である．5.1節で述べたように，メジャー度よりも確信度を重視するため，βは1よりも大きく設定することになる．実験的にはβ=5程度が良いと思われる．

式(7)は，情報検索においてランキングの評価尺度などに良く用いられるF値[9]，[10]の一般式を参考に考案したものである．F値の一般式とは，適合率（Precision）と再現率（Recall）を用いて，以下の式で定義される[3]．

$$\mathrm{F}=\frac{(1+\beta^2)\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\beta^2\times\mathrm{Recall}} \quad (8)$$

すなわち，F値の適合率を確定度で，再現率をメジャー度でそれぞれ置き換えたものが特定度である．

5.4 特定度の対称性

4章で述べたように，候補語集合の抽出手法は順方向・逆方向共に同じである．クエリ（入力語集合）から順方向と逆方向の特定関係を共に抽出するには，クエリQと候補語集合W_iの全てに対して，以下の2つを行う．

・順方向：

特定関係$Q\rightarrow W_i$に対する特定度identify(Q,W_i)の計算

・逆方向：

特定関係$W_i\rightarrow Q$に対する特定度identify(W_i,Q)の計算

ここで，式(5)および(6)より，確定度とメジャー度には対称性が成立する．すなわち，

$$\mathrm{determ}(A,B)=\mathrm{major}(B,A) \quad (9)$$

が成り立つ．よって，順方向の特定度を計算する際に求めた確定度とメジャー度を入れ替えることで，逆方向の特定度を求めることが出来る．また，

$$\mathrm{identify}(B,A)=\frac{(1+\beta^2)\times\mathrm{major}(A,B)\times\mathrm{determ}(A,B)}{\mathrm{major}(A,B)+\beta^2\times\mathrm{determ}(A,B)} \quad (10)$$

$$=\frac{(1+(1/\beta)^2)\times\mathrm{determ}(A,B)\times\mathrm{major}(A,B)}{\mathrm{determ}(A,B)+(1/\beta)^2\times\mathrm{major}(A,B)} \quad (11)$$

が成り立つ．式(11)は，式(7)のβを$1/\beta$で置き換えたものである．よって，逆方向の特定度は順方向の特定度の計算式のβの値を逆数にすることによっても求められる．これらの2つの性質は本質的には同じものである．

6. 実験

｛夏目漱石｝をクエリとして逆方向と順方向の特定関係を求めた結果をそれぞれ表1，表2に示す．また，｛Mac，最新OS｝の順方向を表3に示す．これらは特定度順にランキングしたものである．参考のため，表1と表2にはそれぞれの語集合に対する確定度・メジャー度と，候補語集合の抽出の際に

[3]一般には分母においてβ^2がかかるのはPrecisionの方であるが，ここでは式(7)との対応付けのために変形している．よって式(8)におけるβの意味も一般に言われるものとは反転している．

用いた周辺頻度も示す.

実験における条件は以下の通りである.

・周辺頻度を計算するために取得するWebページ数は100
・周辺頻度による枝刈りの閾値θ=0.6
・生成する語集合の大きさの上限は5
・特定度のパラメータβ=5

表1 「夏目漱石」を特定する語集合（逆方向）
Table 1 Term sets which identify "Soseki Natsume"

順位	特定語集合	特定度	確定度	メジャー度	周辺頻度
1	こころ，文豪	0.3163	0.5243	0.0290	2.3865
2	作品，草枕	0.3160	0.5219	0.0291	0.8262
3	小説，草枕	0.3139	0.5341	0.0278	1.2744
4	小説，坊っちゃん	0.3134	0.4496	0.0366	0.7744
5	世界，草枕	0.3046	0.4680	0.0313	0.7385
6	草枕，猫	0.2945	0.5801	0.0221	1.6063
7	夢十夜	0.2901	0.3693	0.0456	1.6030
8	草枕，日本	0.2892	0.4059	0.0353	0.6247
9	草枕，本	0.2866	0.4236	0.0315	0.6487
10	作品，坊っちゃん	0.2768	0.3838	0.0347	0.9537

表2 「夏目漱石」から特定される語集合（順方向）
Table 2 Term sets identified by "Soseki Natsume"

順位	被特定語集合	特定度	確定度	メジャー度	周辺頻度
1	文学	0.2175	0.3655	0.0195	1.1796
2	小説，日本	0.1901	0.3044	0.0183	0.6962
3	作品，文学	0.1875	0.2286	0.0341	0.7603
4	小説，本	0.1672	0.3007	0.0138	0.6078
5	明治	0.1602	0.2836	0.0135	1.9415
6	作家	0.1587	0.3093	0.0120	1.5345
7	小説，心	0.1580	0.2200	0.0196	0.6374
8	作品，小説	0.1568	0.2482	0.0154	1.1633
9	小説，世界	0.1516	0.2641	0.0130	0.6756
10	小説	0.1488	0.3985	0.0089	2.9661

表3 「Mac，最新OS」から特定される語集合（順方向）
Table 3 Term sets identified by "Mac" and "latest OS"

順位	特定語集合	特定度
1	leopard，インストール	0.1987
2	leopard，アップデート	0.1731
3	leopard，機能，発売，発表	0.1725
4	leopard，過去，機能	0.1655
5	leopard，搭載	0.1650

6.1 考察

表1で示した逆方向の特定関係は，比較的良い精度で抽出できていると言える．表1では「こころ」や「草枕」など，夏目漱石の作品名の組み合わせや，「小説」や「作品」などのドメインを限定する語などを加えた語集合が抽出されており，これらの語集合を提示した時，多くの人は第一に夏目漱石を思い浮かべるだろう．

一方，表2の「夏目漱石」から特定される語集合は，夏目漱石に対して頻繁に言及されるような語集合が得られているが，「夏目漱石」がこれらの語集合を特定するとは言い難い．そもそも，「夏目漱石」という語が何らかの語集合を一意に導き出すとは言えないと思われる．一方，もう1つの順方向の結果である表3ではMac OS Xの最新版である「leopard」と他の語との組み合わせが得られており，この結果を集約する等の加工を行えば，「leopard」という語を得ることは可能である．つまり，クエリによって順方向の特定関係が得られるべきものと得られるべきではないものが存在すると考えられる．しかし現在の手法ではその切り分けができていない．この問題はユーザにどちらの方向の特定関係を求めたいかを指定させることによってある程度解決できると思われるが，何らかの方法で自動的な切り分けができれば，より柔軟な応用が可能になる．この自動的な切り分け手法の考案も含め，特定関係を考えた時にクエリがどのような性質を持ち得るかを今後検討していきたい．

7. 可視化

本章では，前章までの手法によって得られた特定関係をグラフ表現を用いて可視化する手法と，その手法により構築したシステムについて述べる．

7.1 特定関係の可視化

特定関係をグラフ表現を用いて可視化する．同種の可視化手法はプログラミングライブラリであるSlothLib[11]の機能として提案されている．ただし，本研究では語の集合間の関係を扱うため，一般的なグラフ表現ではない．

特定関係グラフの構成要素は以下の3つである．

・語ノード
・語集合ノード
・有向エッジ

語集合の要素である各語はそれぞれ語ノードで表される．これはラベルのついた四角形で表現する．語集合は語集合ノードで表され，これは要素である各語ノードを囲うような線で表現する[4]．特定関係は語集合ノードから語集合ノードへの有向エッジで表される．ただし，エッジの端点は語集合ノードが含む語ノードが作る多角形の中心点である．また，語集合ノードの中心からその語集合が含む各語ノードに向かって短い線分を描く．これはエッジの端点がどの語を含む語集合ノードであるかを分かり易くするためである．

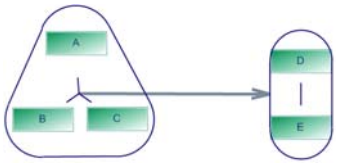

図3 特定関係のグラフ表現
Fig.3 Graph expression of a identifying relationship

語集合間の特定関係のグラフ表現の例を図3に示す．このグラフは語集合｛A，B，C｝が語集合｛D，E｝を特定するという関係，すなわち特定関係｛A，B，C｝→｛D，E｝を表現するものである．

7.2 可視化システムによる関連概念探索

ユーザが入力したクエリに対して，前章までの手法で抽出された特定関係を7.1節のように可視化するシステムを作成した．このシステムを用いれば，入力したクエリと他の語集合との特定関係を直感的に把握することができる．また，グラフ中の語ノードや語集合ノードをダブルクリックすることで，その語や語集合と特定関係にある語集合を求めてグラフに追加表示させることが出来る．また，ユーザにとって不要な語や語集合のノードを削除することも可能である．

このシステムにより，ユーザの興味のある語を選んで次々と特定関係を広げていくことで，関連概念の探索をグラフ上のマウス操作のみで直感的に行うことができる．

関連概念探索の例を図4に示す．この例では，ユーザは「太

[4] ただし，見易さのために1語の語集合については線を描かずに，語ノードをそのまま語集合ノードとする

宰治」というクエリを入力し，「太宰治」を特定する語集合を得た．この特定関係グラフから「人間失格」という語ノードをダブルクリックすることにより，さらに「人間失格」を特定する語集合を得，次に「斜陽」をダブルクリックすることで「斜陽」を特定する語集合を得た．この例では全て逆方向の特定関係を求めているが，もちろん順方向の特定関係を得ることも可能である．

図 4　可視化システムを用いた関連概念探索の例
Fig.4　Example of searching associated concepts by the visualization system

8. まとめと今後の課題

本稿では，ある語集合と特定関係にある語集合を抽出する手法を提案した．入力語集合から求められる特定関係には順方向と逆方向の2種類あり，特定関係の評価指標に対称性を持たせることで，ほぼ同じ操作により両方の特定関係を求めることが出来る．

今後の課題としては，まず確定度およびメジャー度の定義の再検討が挙げられる．今回提案した手法は，特定度の計算に対称性を持たせるため，確定度・メジャー度の定義が片側共起度という非常にシンプルなものになっている．これは特定度計算の際に両方向を同様に扱える一方，評価値としての精度には疑問が残る．今後は対称性を維持しつつ有効性・新規性のある評価指標の考案を目指す．

また，本手法により抽出された特定関係を用いたアプリケーションの構築も今後の課題である．7章において，特定関係をグラフ表現を用いてユーザに提示することで関連概念の探索が可能であると述べたが，これ以外の有用性のあるアプリケーションを構築することも必要であると考えている．

[謝辞]

本研究の一部は，京都大学グローバル COE 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，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」，および，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号 18049041），ならびに，計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073），および，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己）によるものです．ここに記して謝意を表すものとします．

[文献]

[1] M. A. Hearst: “Automatic Acquisition of Hyponyms from Large Text Corpora”, Proceedings of the Fourteenth International Conference on Computational Linguistics, pp. 539-545 (1992).
[2] K. W. Church and P. Hanks: “Word Association Norms, Mutual Information, and Lexicography”, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 76-83 (1998).
[3] D. Lin: “Automatic Retrieval and Clustering of Similar Words”, Proceedings of the 36th annual meeting on Association for Computational Linguistics, pp. 768-774 (1998).
[4] K. Shinzato and K. Torisawa: “Acquiring Hyponymy Relations from Web Documents”, Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL04), pp. 73-80 (2004).
[5] H. Ohshima, S. Oyama and K. Tanaka: “Searching Coordinate Terms with Their Context from the Web”, Proceedings of WISE 2006, pp. 40-47 (2006).
[6] K. Nakayama, T. Hara and S. Nishio: “Wikipedia Mining for an Association Web Thesaurus Construction”, Proceedings of WISE 2007, pp. 322-334 (2007).
[7] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.jp/.
[8] R. Agrawal and R. Srikant: “Fast Algorithms for Mining Association Rules”, Proceedings of VLDB'94, pp. 487-499(1994).
[9] C. J. V. Rijsbergen: “Information Retrieval”, Butterworth-Heinemann, Newton, MA, USA (1979).
[10] G. Hripcsak and A. S. Rothschild: “Agreement, the F-Measure, and Reliability in Information Retrieval”, Journal of the American Medical Informatics Association, **12**, pp. 296-298 (2005).
[11] 大島裕明，中村聡史，田中克己：“SlothLib: Web サーチ研究のためのプログラミングライブラリ”，日本データベース学会 Letters, **6**, 1, pp. 113-116 (2007).

稲川 雅之　Masayuki INAGAWA
京都大学大学院情報学研究科社会情報学専攻修士課程在学中．2007 年京都大学工学部情報学科卒業．主に情報検索の研究に従事．日本データベース学会学生会員．

大島 裕明　Hiroaki OHSHIMA
京都大学大学院情報学研究科社会情報学専攻特定助教．2007 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）．主にウェブ，情報検索，データベースの研究に従事．情報処理学会，電子情報通信学会，日本データベース学会，ACM 各会員．

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．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．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．