社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (12)
2008－ＦＩ－ 91 (12)
2008／6／20

Wikipedia のカテゴリネットワークを用いた概念のベクトル化手法

白川 真澄[†1] 中山 浩太郎[†2]
原 隆浩[†1] 西尾 章治郎[†1]

分類辞書（タクソノミー）は，文書分類や情報検索などのアプリケーションにおいて幅広くその有用性が実証されてきた．しかし分類辞書の自動構築に関する従来研究では，自然言語処理の技術的限界やノイズデータに起因する精度低下の問題がある．そこで本稿では，大規模 Web 百科事典の Wikipedia に構築されたカテゴリ構造（ネットワーク）を用いて，概念をベクトル化する手法を提案する．

Concept Vectorization Methods using Wikipedia Category Network

Masumi Shirakawa,[†1] Kotaro Nakayama,[†2] Takahiro Hara[†1]
and Shojiro Nishio[†1]

The availability of the taxonomy, which is a kind of category-sorted dictionary, has been demonstrated by various applications such as document classification and information retrieval. However, existing works on automatic taxonomy construction have the problem of decreasing the accuracy due to the technical limitation of statistical NLP (Natural Language Processing) and noise data. In this work, we propose concept vectorization methods using the category network structured in Wikipedia, a large scale Web encyclopedia.

1. はじめに

国語学や言語学，分類学等の研究領域において，概念を一つの大系として整理することは重要であり，国語辞典や対訳辞書，百科事典など，様々な辞書が人の手によって製作され，学習などに役立てられてきた．特に，今後の Web が意味を考慮した次世代 Web へと変革するにあたって，分類辞書（タクソノミー）は重要な基盤技術である．分類辞書とは，概念がどのようなカテゴリに所属しているかを木構造や DAG(Directed Acyclic Graph) 構造などで表した辞書であり，計算機が意味解析をするために必要であると考えられる．しかし，既存の分類辞書の自動構築手法では，現在の自然言語処理技術では統計的な解析手法が主流であり意味を考慮していないことや，Web マイニングにおけるノイズデータに起因する精度低下の問題があり，精度を上げるためには網羅性を犠牲にしなければならない．

そこで本研究では Wikipedia に注目する．Wikipedia は，Wiki をベースにした大規模 Web 百科事典である．Wiki をベースにしているため，誰でも Web ブラウザを通じて記事内容を変更できることが大きな特徴であり，幅広い分野の記事（概念）を網羅している．現在では，一般的な概念だけでなく，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野をカバーし，普遍的な概念から新しい概念に至るまで，非常に膨大なコンテンツが網羅されている．その記事数は既に 180 万（2007 年 6 月英語のみカウント）を超えており，世界最大の百科事典である Britannica の記事数が，全 60 巻で約 65,000 記事であることと比較した場合，実に 30 倍近い数の記事が網羅されていることになる．

また Wikipedia では，ユーザによって構築された大規模なカテゴリ構造が存在する．ほとんどの概念は一つまたは複数のカテゴリに所属しており，カテゴリ同士も所属関係を持った形で密接にリンクしてカテゴリのネットワークを形成している．

Wikipedia の記事やカテゴリ構造の有用性は過去の研究で実証されてきた．そこで本研究では，Wikipedia のカテゴリ構造を利用して，精度の高い分類辞書を構築することを目的とする．しかし，前述のとおり Wikipedia のカテゴリ構造はネットワーク構造であるため，木構造のように，あるカテゴリに所属する概念群を取得するといった処理ができない．これは，場合によってはネットワーク構造を伝って多量の概念を取得してしまう可能性があるためである．そこで，本研

†1 大阪大学，Osaka University
†2 東京大学，Tokyo University

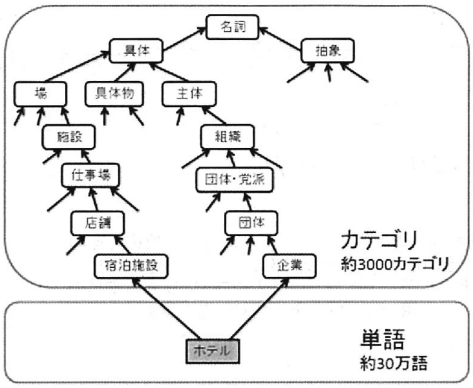


図 1　日本語語彙体系の名詞のカテゴリ体系の例

究では概念のベクトル化手法（BVG 法）を提案する．BVG 法では，概念から再帰的に親カテゴリ（所属リンク）を探索し，概念のカテゴリへの所属の強さを，カテゴリまでの経路（パス）の多さと長さに基づいて数値化する．そして，カテゴリを基底とするベクトルによって概念の特徴を表現する．さらに，ベクトル化手法の精度向上を目的とし，2 つの拡張手法（SPI 法，SCE 法）を提案する．SPI 法では，細分化された専門分野におけるカテゴリ構造は部分的にほぼ完全な木構造となっていることに注目し，経路を短縮することで専門分野における特徴抽出の精度向上を目指す．SCE 法では，経路長の増加に伴って特徴が分散する問題を緩和するため，サブカテゴリの設定によって経路長を削減する．

2. 関 連 研 究

2.1 分類辞書（タクソノミー）

分類辞書（タクソノミー）とは，概念または単語がどのようなカテゴリに所属しているかを，木構造や，複数の親を許す DAG 構造などによって表した辞書である．図 1 は，日本語語彙体系[6] の名詞のカテゴリ体系を示した図である．単語「ホテル」は，親カテゴリ「宿泊施設」「企業」を持っており，「店舗」や「仕事場」，「施設」などのカテゴリに所属している．現在，日本語語彙体系をはじめ，WordNet[8] や MeSH[*1] など，分類辞書として実際に利用されているものは手動で構築されたものが多い．しかし，手動構築は，初期コストの問題や，新たな概念の追加・修正作業のコストの問題，網羅性の問題がある．そのため，分類辞書の自動構築に関する研究が盛んに行われてきた．

分類辞書の自動構築に関係する研究は，Brown らの研究[3]，McMahon らの研究[7]，Cutting らの研究[4] などがある．これらの研究では，分類辞書の自動構築は自然言語処理を用いた Web マイニングによって行われているが，現在の自然言語処理技術では統計的な解析手法が主流であることや，Web マイニングにおけるノイズデータの存在のため，手動分類と比べて精度は低い．

Brewster の研究[2] では，分類辞書の自動構築の基準として一貫性（普遍性），多義語の分類，計算量，カテゴリのラベル付け，汎用性を挙げている．また，分類辞書の自動構築に関係する既存の研究[3),4),7)] について特徴を解析しているが，これらの基準を単独で満たす分類辞書の自動構築手法はなく，結論として一つの手法の弱点を補うために複数の手法を組み合わせる必要性を主張している．このことから，分類辞書の自動構築に絶対的な手法が存在しないことがうかがえる．

2.2 Wikipedia マイニング

本研究では，大規模な Web 辞典である Wikipedia に注目し，Wikipedia マイニング，すなわち Wikipedia を Web マイニングの対象として解析することで，精度の高い分類辞書を自動で構築することを目的としている．

Wikipedia マイニングは，2006 年から活発になった新しい研究領域である．Strube らの研究[12]，Milne らの研究[9]，Gabrilovich らの研究[5]，そして中山らの研究[10] などでは，概念間の関連度を抽出している．Völkel らの研究[13] では，Wikipedia の拡張アーキテクチャとして，リンクに意味情報を付与する仕組みを提案し，Wikipedia 上でのオントロジ構築を目指している．Ruiz-Casado らの研究[11] では，Wikipedia のエントリと WordNet[8] のエントリ間の類似度を計算して両者の概念をマッピングし，一般的な辞書の WordNet を Wikipedia のエントリで拡張することで，両者の長所を活かした辞書の構築を提案している．

これらの研究が示すとおり，通常の Web ページと比べて，Wikipedia が Web マイニングの対象として有効であることは明らかである．これは，Wikipedia が知識抽出のコーパスとして重要な特徴を持っているためである．Wikipedia の特徴については次章で詳述する．

3. Wikipedia の特徴

Wikipedia は，知識抽出のコーパスとしてみたときに，その網羅性だけでなく，密なリンク構造，質の高いリンクテキスト，URL による語彙の一意性，多様なリンク構造といった特徴を持っている[10]．本章では，この中でも特に本研究に関係の深い特徴について詳述する．

*1 http://www.nlm.nih.gov/mesh/

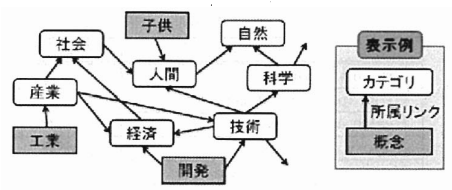


図 2　Wikipedia のカテゴリ構造の例

3.1　密なリンク構造

2006 年 9 月の段階における英語版 Wikipedia には，約 4,998 万のリンク（リダイレクトリンク，言語リンクを除く）が約 168 万のページ内に存在する．これは，一ページあたり平均 29.62 のリンクを持つ計算となる．しかも，これらのリンクはサイト内に対するリンクのみをカウントしたものであり，サイト外へのリンクは含まれていない．これは，Wikipedia は閉じられた語彙空間の中で密なリンク構造を持っており，リンク構造を解析することで有用な情報を抽出できる可能性が高いことを示している．

3.2　コンテンツの網羅性

従来の辞書では，一般的な語からトップダウン的に追加されていくのが通常であり，一般的でない語や専門的な語は辞書に追加されるのが遅れる，もしくはいつまでも登録されないのが一般的である．しかし Wikipedia では，インターネットを通じてリアルタイムに記事が公開・アップロードされ，リンクが構築されていくため，極めて即時性および網羅性が高い．

3.3　URL による語彙の一意性

URL により語彙の一意性が確立されている点は，Wikipedia の大きな特徴の一つである．電子辞書では，通常一つの見出し語が一つのページに割り当てられており，その中で複数の意味について詳述される．一方，Wikipedia では一つの URL（ページ）に一つの概念が割り当てられており，多義性が URL によって解決されている点が大きな特徴である．

3.4　多様なリンク構造

Wikipedia には，ページからページへ移動するための通常のリンク以外に，カテゴリリンクやリダイレクトリンク，言語リンクなど，いくつかの特殊なリンクが存在する．カテゴリリンクについては，次節のカテゴリ構造で説明する．リダイレクトリンクは，ある記事が参照されたときに別の記事へとリダイレクトする機能を提供するリンクである．言語リンクは，多言語をカバーしている Wikipedia において，別言語の同じ概念やカテゴリのページを参照するためのリンクである．

3.5　カテゴリ構造

Wikipedia では，記事（概念）の分類関係をリンクによって表現している．このリンクは，カテゴリリンクと呼ばれ，概念とカテゴリ，およびカテゴリ間の所属関係が表現される．また，カテゴリには専用のカテゴリのページ（URL）が用意され，カテゴリリンクによってカテゴリ間の所属関係が表現されている．カテゴリリンクは，リンクの方向によって所属リンクと被所属リンクに分けられる．

Wikipedia のカテゴリ構造は，全体としてみると一種の木構造をしている．しかし，一つのページが複数の親カテゴリを持つことも可能であり，一部にはループも存在する．そのため，実際には完全な木構造ではなく，図 2 のようなネットワーク構造となっている．Wikipedia の英語版（2006 年 9 月）を調査したところ，約 80 万のカテゴリリンクが存在していた．これは，WordNet[8] の親子関係が 10 万弱であることと比べて 8 倍以上のカテゴリリンクを有していることになる．これらのカテゴリ構造も，記事同様ユーザによって編集され，管理されている．

このような Wikipedia のカテゴリ構造は，一種の分類辞書（タクソノミー）としての役割を有しており，カテゴリを絞り込みながら記事を検索するような機能を実現するために利用されている．Wikipedia が提供しているカテゴリ検索システム「CategoryTree」[*2] では，カテゴリを検索することや，カテゴリの階層構造をブラウジングすることが可能である．しかし，あるカテゴリに所属する概念を全て取得することは不可能である．これは，Wikipedia のカテゴリ構造がネットワーク構造であり，木構造のように，単純にあるカテゴリに所属する概念を抽出できないためである．

4.　概念のベクトル化

前章で述べたとおり，Wikipedia のカテゴリは複雑なネットワーク構造を持つため，通常のカテゴリ化手法によって概念を分類することはできない．そこで，Wikipedia のカテゴリ構造に特化した概念のベクトル化手法を提案する．また，ベクトル化の精度を向上させるために，2 つの拡張手法を提案する．

4.1　概念ベクトル

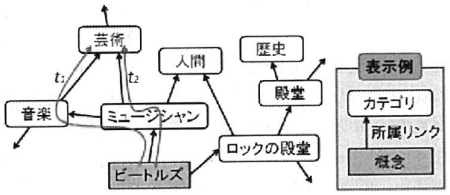
前述のとおり，Wikipedia では一つのページ（URL）が一つの概念またはカテゴリに対応している．また，一つのページは一つもしくは複数のカテゴリに所属することが可能である．したがって，概念がどのようなカテゴリに所属しているかという情報は，Wikipedia のカテゴリを検索することや，カテゴリの階層構造をブラウジングすることで取得できる．しかし，Wikipedia のカテゴリは複雑なネットワーク構造であるため，概

*2 http://en.wikipedia.org/wiki/Special:CategoryTree

	芸術	歴史	地理	思想	人間	自然	·····
ジャズ	0.375	0.125	0	0	0	0	·····

	芸術	歴史	地理	思想	人間	自然	·····
ビートルズ	0.375	0.125	0	0	0.5	0	·····

図 3　「ジャズ」と「ビートルズ」の概念ベクトル



$$I(\text{ビートルズ},\text{芸術}) = \frac{1}{d(t_1)} + \frac{1}{d(t_2)} = \frac{1}{2^3} + \frac{1}{2^2} = 0.375$$

図 4　Basic Vector Generation（BVG）法の例

念と全く関係のないカテゴリからでも，カテゴリの階層構造をブラウジングすることでその概念のページに到達できる．例えば，「動物」というカテゴリを起点としてカテゴリを探索していくと，「哺乳類」「人間」「社会」「法律」とカテゴリの階層構造をブラウジングすることで，「法律」という「動物」にあまり関係のないカテゴリに到達する．これはつまり，ページの遷移回数の増加に伴い，概念がどのようなカテゴリに所属しているかという情報が曖昧化していることを示している．

文書分類に関する研究[1)] では文書の特徴を表すために，意味・カテゴリを基底とする文書ベクトルとして表現している．これらの研究は，文書の特徴の抽出手法としてベクトルを用いる手法が有効であることを示している．本研究では，この考え方を概念に対して適用する．つまり，概念がどのようなカテゴリに所属しているかという情報を，所属しているか否かではなく，どの程度所属しているかという度合（所属の強さ）で表し，カテゴリを基底とする概念ベクトルとして表現する．例えば図 3 に示すように，概念「ジャズ」は，カテゴリ「芸術」に高い値を持つ概念ベクトル，概念「ビートルズ」は，カテゴリ「芸術」とカテゴリ「人間」に高い値を持つ概念ベクトルとして表現する．このように概念をベクトル化することで，概念とカテゴリの多対多の関係を，所属の強さを持った形で表現できる．

4.2 Basic Vector Generation（BVG）法

Wikipedia のカテゴリ構造を用いた基本的な概念のベクトル化手法として，Basic Vector Generation（BVG）法を提案する．

Wikipedia は，概念が一つまたは複数のカテゴリに所属し，また，カテゴリ同士が所属関係を持った形でリンクしているネットワークであるため，概念をノード集合 W，カテゴリをノード集合 V，所属リンクをエッジ集合 E とする有向グラフ $G = \{W, V, E\}$ で表現できる（図 4）．このとき，概念（w_i）のカテゴリ（v_j）への所属の強さを計測する問題を考えた場合，所属の強さは以下の二つの要素に依存すると考えられる．

（1）　概念 w_i からカテゴリ v_j へのパスの多さ

（2）　概念 w_i からカテゴリ v_j への各パスの短さ

つまり，概念 w_i からカテゴリ v_j へのパスが多ければ多いほど，またそのパスの長さが短ければ短いほど強く所属すると考えられる．ここで，パスとは所属リンクを伝って概念 (w_i) からカテゴリ (v_j) へと移動可能な経路を示す．

そのため，w_i から v_j への全パス $T = \{t_1, t_2, ..., t_n\}$ が与えられたとき，概念 w_i のカテゴリ v_j への所属の強さ $I(w_i, v_j)$ を以下の式により表現する．

$$I(w_i, v_j) = \sum_{l=1}^{n} \frac{1}{d(t_l)} \tag{1}$$

d はパス t_l のホップ数に応じて増加する関数であり，単調増加関数や指数関数を利用することができる．

図 4 では，概念「ビートルズ」のカテゴリ「芸術」への所属の強さを計測している（d は指数関数 2^{t_l}）．パス t_1 の長さが 3，パス t_2 の長さが 2 であるため，所属の強さは 0.375 となる．

4.3 BVG 法の主観評価実験

ベクトルの基底を表 1 の Wikipedia で主要なカテゴリとして用いられている 11 種類のカテゴリとし，パス t_l のホップ数に応じて増加する関数 d を指数関数 3^{t_l}，探索の最大ホップ数 H を 4 として，BVG 法に対する主観評価の実験を行った．最大ホップ数 H や関数 d の式は，予備実験により適当なパラメータに決定した．

表 2 に，本手法で抽出した概念ベクトルの例を示す．なお表 2 では，ベクトルの基底に用いる主要なカテゴリを，表 1 に示す日本語の略称として表記する．主観

表 1　Wikipedia の主要な 11 種類のカテゴリ

主要なカテゴリ	日本語の略称
Art and culture	文化
Geography and places	地理
Health and fitness	健康
History and events	歴史
Mathematics and logic	論理
Natural sciences and nature	自然
People and self	人間
Philosophy and thinking	哲学
Religion and belief systems	思想
Social sciences and society	社会
Technology and applied sciences	技術

表 2　BVG 法の主観評価の結果

概念	文化	地理	健康	歴史	論理	自然	人間	哲学	思想	社会	技術
Adam Smith	0.01	0	0	0	0	0.04	0.14	0.03	0.09	0.31	0
AIDS	0	0	0.20	0	0	0.06	0.06	0	0	0.07	0
Albert Einstein	0.01	0	0	0.07	0	0.07	0.53	0.06	0.01	0.38	0.06
Anarchism	0.01	0	0.01	0	0.05	0.09	0	0.28	0.05	0.58	0.01
Arctic	0	0.04	0	0	0	0	0	0	0	0.02	0
Buddhism	0.09	0	0	0	0	0.01	0	0.01	0.22	0.31	0.01
Computer	0	0	0	0	0	0.04	0	0	0	0.11	0.12
Fish	0.01	0	0.01	0	0	0.62	0.01	0	0	0	0
French Revolution	0	0	0	0.12	0	0	0	0	0	0.11	0
Hospital	0	0	0.40	0	0	0.12	0.10	0	0	0.20	0.06
Island	0	0.01	0	0	0	0.17	0	0	0	0	0
Japan	0	0	0	0	0	0.01	0	0	0	0.03	0
Jazz	0.12	0	0	0	0	0	0.01	0	0	0	0
Linear algebra	0	0	0	0	0.01	0	0	0	0	0	0
Lion	0	0	0	0	0	0	0	0	0	0	0
Mountain	0	0.03	0	0	0	0.27	0	0	0	0.03	0
Neural network	0	0	0	0	0.06	0.22	0.04	0.03	0	0.33	0.11
Television	0.10	0	0	0	0.01	0	0.05	0	0	0.10	0.02
World War II	0	0	0	0.01	0	0	0	0	0	0	0

評価の結果，多くの概念について精度よくベクトル化できており，概念のベクトル化が機能していることが確認できたが，いくつかの問題点も明らかになった．まず，直感的に考えれば明らかに所属しているカテゴリへの所属の強さが抽出できていない（0 となっている）場合が存在する点である．例えば，概念「Lion」は動物であるため，直感的にはカテゴリ「自然」に所属するべきだが，ベクトルではカテゴリ「自然」への所属の強さは 0 となっている．これは，専門分野ではカテゴリが細分化され，カテゴリ「自然」までのパスのホップ数が大きくなっていることが原因であった．この問題の解決法として，探索の最大ホップ数 H を大きく設定する方法が考えられるが，H の増加に伴い，特徴の分散が大きくなることを予備実験で確認している．これは所属リンクの通過数に伴い特徴が分散することを意味し，精度低下の要因となっており，もう一つの問題点である．表 2 では，カテゴリ「社会」への所属の強さが比較的表れやすくなっているが，これは，Wikipedia のカテゴリネットワークにおいて，カテゴリ「社会」が多くの一般的なカテゴリからリンクされており，概念から「社会」までのパスが全体量として多いために特徴の分散が数値として表れたものである．

BVG 法の主観評価の実験結果から問題点を以下にまとめる．

(1)　専門分野でカテゴリが細分化されている場合，その特徴を抽出できない．

(2)　所属リンクの通過数に伴い，特徴の分散が大きくなる．

これらの問題の解決を図るため，以降で BVG 法の拡張手法について述べる．

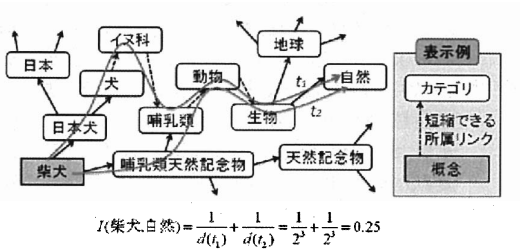


図 5　Single Parent Integration（SPI）法の例

4.4 Single Parent Integration（SPI）法

前述のとおり，BVG 法では専門分野でカテゴリが細分化されている場合，パスのホップ数が大きくなるため，その特徴を抽出できないという問題がある．そこで，Single Parent Integration（SPI）法を提案する．これは，専門分野においてカテゴリが細分化されている場合，その専門分野では部分的にはほぼ完全な木構造となっていることを考慮し，所属リンクがただ一つの場合にパスの長さを短縮する手法である．この手法は，親カテゴリを一つしか持たない場合は意味の分散が生じないという考えに基づいている．

BVG 法と同様に，概念をノード集合 W，カテゴリをノード集合 V，所属リンクをエッジ集合 E とする有向グラフ $G = \{W, V, E\}$ を考える（図 5）．SPI 法では，あるノード v_i（または w_i）からノード集合 V への所属リンク e_k がただ一つの場合，e_k のパスの長さを仮想的に 0 とみなしてエッジ集合を E' と改め，$G' = \{W, V, E'\}$ とする．そして，G' に対して BVG 法を適用する．

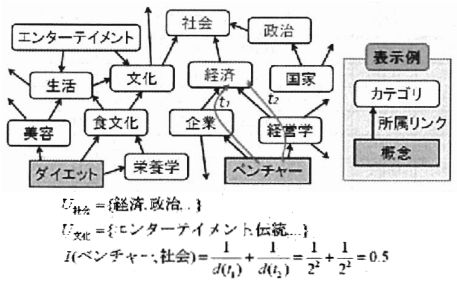


図 6 Sub-Category Expansion（SCE）法の例

図 5 の例では，概念「柴犬」のカテゴリ「自然」への所属の強さを計測している．カテゴリ「犬」は親カテゴリとして「イヌ科」のみを持つため，所属リンクを短縮可能で，パスの長さを 0 とみなす．同様に「イヌ科」は「哺乳類」,「哺乳類」は「動物」,「動物」は「生物」のみを親カテゴリとして持つため,「犬」から「生物」までのパスの長さは仮想的に 0 とみなされる．これにより，概念「柴犬」からカテゴリ「自然」までのパス t_1，t_2 ともに長さが 3 となり，BVG 法と比べて「自然」への所属の強さが大きく表れる．

4.5 Sub-Category Expansion（SCE）法

BVG 法では所属リンクの通過数に伴い特徴が分散するという問題がある．そこで，Sub-Category Expansion（SCE）法を提案する．この手法では，ベクトルの基底となるカテゴリごとにサブカテゴリを設定し，BVG 法や SPI 法を適用した際に，サブカテゴリを基底のカテゴリとみなす．その結果，所属リンクの通過数を減少させ，特徴の分散を緩和できると考えられる．なお，サブカテゴリは，Wikipedia のカテゴリから自由に選択できるものとする．

概念をノード集合 W，カテゴリをノード集合 V，所属リンクをエッジ集合 E とする有向グラフ $G = \{W, V, E\}$ を考える（図 6）．SCE 法では，ベクトルの基底となるカテゴリ v_j についてサブカテゴリのノード集合 $U_j = \{u_1, u_2, ..., u_n\}$ を設定する．そして概念 w_i のカテゴリ v_j への所属の強さ $I(w_i, v_j)$ を，w_i から $v_j \cup U_j$ への全パス $T = \{t_1, t_2, ..., t_n\}$ が与えられたとみなし，BVG 法や SPI 法を用いて算出する．

SCE 法において重要なのは，サブカテゴリの選択である．図 6 の例では，ユーザがカテゴリ「社会」に対して「政治」や「経済」をサブカテゴリに設定し，カテゴリ「文化」を別の基底とすることで，ユーザが想定した「社会」への所属の強さがベクトルに反映される．また，複数の基底のカテゴリに重複してサブカテゴリとして選択することも考えられる．例えば，カテゴリ「科学」は自然科学や社会科学など幅広い意味を持つため,「自然」と「社会」のサブカテゴリとして二重に選択することも可能である．

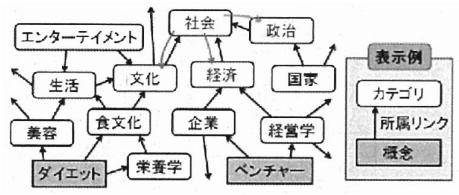


図 7 BVG 法を用いたサブカテゴリの自動選択（$S = 0.5$, $d = 2^{t_l}$）

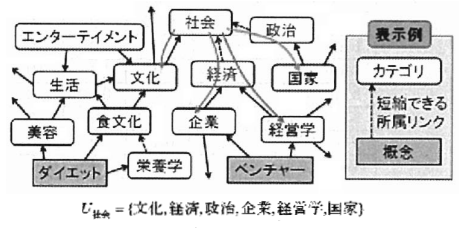


図 8 SPI 法を用いたサブカテゴリの自動選択（$S = 0.5$, $d = 2^{t_l}$）

サブカテゴリを自動選択する方法として，BVG 法や SPI 法をトップダウン式に適用する方法が考えられる．これらの方法では，概念ではなく，カテゴリを BVG 法，SPI 法でベクトル化し，ベクトルの基底のカテゴリへの所属の強さが閾値以上のカテゴリを，その基底のカテゴリのサブカテゴリとして選択する．サブカテゴリの自動選択の例として，BVG 法を用いた自動選択方法について説明する．カテゴリをノード集合 V，所属リンクをエッジ集合 E とする有向グラフ $G = \{V, E\}$ を考える（図 7）．以下の手順で v_i のサブカテゴリを選択する．全てのノード v_j について，v_j から v_i への全パス $T = \{t_1, t_2, ..., t_n\}$ が与えられたとき，カテゴリ v_j のカテゴリ v_i への所属の強さ $I(v_j, v_i)$ を BVG 法を用いて抽出する．そして所属の強さ I に閾値 S を設定し，$I(v_j, v_i)$ が S 以上の v_j をサブカテゴリ u_k とし，サブカテゴリのノード集合 $U_i = \{u_1, u_2, ..., u_n\}$ を設定する．

実際には図 7，図 8 のように，ベクトル化手法をトップダウン式に適用し，基底のカテゴリから子カテゴリを探索，つまり被所属リンクを伝って所属の強さを算出してサブカテゴリを選択できる．なお，図 7，図 8 では閾値 S を 0.5，パス t_l のホップ数に応じて増加する関数 d を指数関数 2^{t_l} としている．

5. 評価実験

本章では，提案手法の有効性を評価するために行っ

抽出	芸術	歴史	地理	思想	人間	自然	·····
	0.375	0.125	0	0	0.5	0	·····

解答	芸術	歴史	地理	思想	人間	自然	·····
	2	1	0	1	2	0	·····

$$\cos(\text{抽出}, \text{解答}) = \frac{0.375\times 2 + 0.125\times 1 + 0.5\times 2}{\sqrt{0.375^2 + 0.125^2 + 0.5^2}\sqrt{2^2 + 1^2 + 1^2 + 2^2}} = \frac{1.875}{0.637377439\times 3.16227766} = 0.93026051$$

図 9　抽出したベクトルと解答ベクトルのコサイン尺度

た実験の結果について述べる．

5.1 実験環境

評価実験は，評価に用いる解答ベクトルの取得と，各手法で抽出したベクトルと解答ベクトルの類似度を算出する二つのフェーズから構成される．本実験ではベクトルの基底として，表 1 の Wikipedia の主要な 11 種類のカテゴリを利用した．

まず，提案手法で抽出したベクトルの評価に用いる解答ベクトルを取得するために，20 代男女 20 人の被験者に質疑応答を行った．質疑応答の手順を以下に示す．

(1)　ベクトルの基底となるカテゴリ A を提示する．
(2)　被験者が，カテゴリ A に所属すると思う概念 B を連想する．
(3)　被験者が，ベクトルの基底となる 11 種類のカテゴリ各々に概念 B が所属しているか否かを 3 段階（0: 所属しない，1: どちらともいえない，2: 所属する）で判定する．
(4)　判定結果から解答ベクトルを生成する．

一人の被験者に対して，11 種類のカテゴリを各一回ずつ提示して 11 の概念を連想してもらい，20 人の被験者から延べ 220 の概念に対する解答ベクトルを取得した．その後，220 の概念に対して 4 章で説明したベクトル化手法を適用し，抽出したベクトルと解答ベクトルの類似度を，コサイン尺度を用いて算出した．具体的には，解答ベクトルを取得した 220 の概念について，BVG 法，SPI 法，SCE 法を用いてベクトル $\boldsymbol{r}$ を抽出し，解答ベクトル $\boldsymbol{s}$ とのコサイン尺度 $\cos(\boldsymbol{r}, \boldsymbol{s})$ を以下の式により求めた．図 9 はコサイン尺度算出の具体例である．

$$\cos(\boldsymbol{r}, \boldsymbol{s}) = \frac{\boldsymbol{r}\cdot\boldsymbol{s}}{\|\boldsymbol{r}\|\|\boldsymbol{s}\|} = \frac{\sum_{i=1}^{m} r_i s_i}{\sqrt{\sum_{i=1}^{m} r_i^2}\sqrt{\sum_{i=1}^{m} s_i^2}} \quad (2)$$

BVG 法，SPI 法では探索の最大ホップ数 H を 4，パス t_l のホップ数に応じて増加する関数 d を指数関数 3^{t_l} として評価した．また，SCE 法ではサブカテゴリの選択方法を，人による選択（Wikipedia のユーザが選択），BVG 法を用いた選択（閾値 S は 0.1），SPI 法を用いた選択（閾値 S は 0.1）の 3 パターンとし，それぞれについて，BVG 法，SPI 法を適用した手法を評価した．

表 3　コサイン尺度の平均値と中央値

ベクトル化手法	平均値	中央値
BVG 法	0.554	0.621
SPI 法	0.595	0.620
SCE 法（Wikipedia users）+BVG 法	0.664	0.704
SCE 法（Wikipedia users）+SPI 法	0.661	0.689
SCE 法（BVG 法）+BVG 法	0.624	0.662
SCE 法（BVG 法）+SPI 法	0.604	0.625
SCE 法（SPI 法）+BVG 法	0.623	0.650
SCE 法（SPI 法）+SPI 法	0.602	0.631

SCE 法の括弧内はサブカテゴリの選択方法を表す．

5.2 実験結果

実験結果を表 3 に示す．

まず SPI 法では，BVG 法と比べてコサイン尺度の平均値が高くなっている．BVG 法は，専門分野でカテゴリが細分化されているために，本来抽出されるべき所属の強さが抽出できないケースが存在する．SPI 法は，専門分野の特徴抽出を目的としているため，BVG 法では抽出できなかった所属の強さを抽出でき，平均値に影響を与えていると考えられる．一方で，コサイン尺度の中央値にあまり変化がみられないのは，SPI 法が逆効果となる場合が存在するためである．例えば，「経済」は一般的なカテゴリであるが，実質的にはカテゴリ「社会」のみに所属すると考えられる．この例では，SPI 法が目的としていない一般的なカテゴリに適用され，所属の強さが過剰に抽出される．

SCE 法の結果をみると，総じてコサイン尺度が高いことが判明した．特に，コサイン尺度の平均値，中央値ともに最も高かったのは，Wikipedia のユーザが選択したサブカテゴリを用い，BVG 法によってカテゴリへの所属の強さを抽出する SCE 法であった．この SCE 法のサブカテゴリは，Wikipedia のユーザ同士が議論し，精査したものが選択されており，一般的な人々の意図が反映されたサブカテゴリである．このサブカテゴリを用いることで被験者との類似度であるコサイン尺度が高くなるということは，提案したベクトル化手法が所属の強さを抽出することに関して有効に機能していることを示している．また，サブカテゴリを自動選択する SCE 法についても，サブカテゴリを使用しない場合と比べて，コサイン尺度の平均値および中央値が高くなっており，SCE 法が有効であるといえる．

6. まとめと今後の課題

本研究では，Wikipediaのカテゴリ構造を用いて，概念のカテゴリへの所属の強さを抽出し，ベクトルとして表現する手法を提案した．BVG法では，Wikipediaのネットワークのカテゴリ構造において，概念から所属リンクを伝って親カテゴリを探索し，カテゴリまでのパスの数とそのホップ数に基づいてカテゴリへの所属の強さを計測する．SPI法は，BVG法を拡張し，所属リンクを一つしか持たない場合にその所属リンクを短縮する．SCE法は，ベクトルの基底のカテゴリごとにそれぞれサブカテゴリを設定し，サブカテゴリも含めたカテゴリまでの全てのパスから，BVG法やSPI法を用いてカテゴリへの所属の強さを計測する．

主観評価の実験結果から，BVG法による概念のベクトル化が機能していることを確認した．さらに，評価実験の結果から，SCE法の有効性を確認した．また，SPI法は，BVG法の問題である専門分野における特徴抽出が可能であるが，過剰に所属の強さを抽出してしまう場合があることが分かった．

本研究で構築した概念ベクトルは，文書分類や情報検索などのアプリケーションの基盤技術として利用可能である．そこで今後は，構築した概念ベクトルをこれらのアプリケーションに適用することで，有効性や実用性を示していく予定である．文書分類では，ユーザが分類したいカテゴリをベクトルの基底に設定し，文書に出現する全ての語の概念ベクトルを抽出する．そしてそれらの概念ベクトルを利用して，文書全体のベクトルを算出する．その結果，文書ベクトルの各カテゴリへの所属の強さから，文書を分類できる．また，情報検索では，文書分類と同様にWebページなどの文章から文書ベクトルを算出し，検索クエリの語の概念ベクトルとの類似度が高いページからユーザに提示する．ベクトルの類似度だけでなく，語の関連語を用いた拡張クエリによる検索などと組み合わせると，より精度の高い検索が期待できる．

謝辞　本研究の一部は，文部科学省特定領域研究（18049050）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の助成によるものである．ここに記して謝意を表す．

参考文献

1) Becker, J. and Kuropka, D.: Topic-based vector space model, *Proc. of International Conference on Business Information Systems (BIS)*, pp.7–12 (2003).
2) Brewster, C.: Techniques for automated taxonomy building: Towards ontologies for knowledge management, *Proc. of Computational Linguistics UK Research Colloquium (CLUK)* (2002).
3) Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D. and Lai, J. C.: Class-based n-gram models of natural language, *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479 (1992).
4) Cutting, D. R., Karger, D. R., Pedersen, J. O. and Tukey, J. W.: Scatter/Gather: A cluster-based approach to browsing large document collections, *Proc. of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp.318–329 (1992).
5) Gabrilovich, E. and Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis, *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pp.1606–1611 (2007).
6) 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林　良彦：日本語語彙大系，岩波書店 (1997).
7) McMahon, J. G. and Smith, F. J.: Improving statistical language model performance with automatically generated word hierarchies, *Computational Linguistics*, Vol. 22, No. 2, pp. 217–247 (1996).
8) Miller, G. A.: WordNet: A lexical database for English, *Communications of the ACM (CACM)*, Vol.38, No.11, pp.39–41 (1995).
9) Milne, D., Medelyan, O. and Witten, I. H.: Mining domain-specific thesauri from Wikipedia: A case study, *Proc. of ACM International Conference on Web Intelligence (WI)*, pp.442–448 (2006).
10) 中山浩太郎，原　隆浩，西尾章治郎：Wikipediaマイニングによるシソーラス辞書の構築手法，情報処理学会論文誌，Vol.47, No.10, pp.2917–2928 (2006).
11) Ruiz-Casado, M., Alfonseca, E. and Castells, P.: Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets, *Proc. of International Atlantic Web Intelligence Conference (AWIC)*, pp.380–386 (2005).
12) Strube, M. and Ponzetto, S.: WikiRelate! computing semantic relatedness using Wikipedia, *Proc. of National Conference on Artificial Intelligence (AAAI)*, pp. 1419–1424 (2006).
13) Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H. and Studer, R.: Semantic Wikipedia, *Proc. of International World Wide Web Conference (WWW)*, pp.585–594 (2006).