オブジェクト間の意外な共通点の発見

佃　洸摂[†,††]　大島　裕明[†]　加藤　誠[†]　田中　克己[†]

† 京都大学大学院情報学研究科社会情報学専攻　〒 606–8501 京都府京都市左京区吉田本町
†† 日本学術振興会 特別研究員
E-mail: †{tsukuda,ohshima,kato,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本稿では，オブジェクトと属性値を入力として与えたときに，オブジェクトと属性値の関係の認知度を推定するための手法を提案する．提案手法を用いることで，たとえばオブジェクトを “中国”，属性値を “ワイン” としたとき，“中国” と “ワイン” は実際の関連度は高いがその関係の認知度は低い，といったことが求められる．提案手法では，(1) オブジェクトの認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる．(2) オブジェクトの多くの類似オブジェクトと属性値の関係の認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる．という仮説のもと関係の認知度を推定する．さらに，提案手法を用いた，オブジェクト間の意外な共通点の発見について考察を行う．

キーワード　意外性, オブジェクト集合, 知識抽出

1. はじめに

Google(注1) や Yahoo(注2)，Bing(注3) などの既存の Web 検索エンジンでは，ユーザが入力したクエリとの適合度や人気度によって Web ページをランキングし，検索結果を返す．クエリに適した文書の検索を可能にするために，これまで様々な検索手法が提案されてきた．たとえば，BM25 [1] はテキスト情報に基づいた手法であり，HITS [2] や PageRank [3] は Web ページ間のリンク構造に基づいた手法である．

しかし，これらの手法を用いて意外な情報を発見するのは困難である．そのため，我々はこれまで入力として与えられた 1 語のクエリに関する意外な情報を発見するための手法を提案してきた [4]．提案手法では，たとえば “落合博満” というクエリ（オブジェクト）に対して “ガンダム” という語（属性値）を意外な語として発見し，それに基づいて “落合博満はガンダムマニアである” という意外な情報を発見していた．その際，“長嶋茂雄” や “野村克也”，“王貞治” などの “落合博満” の多くの同位語らしい語は “首位打者” という語と関連があるため，“落合博満” と “首位打者” の関係はありふれたものであり意外ではないが，“落合博満” の多くの同位語らしい語は “ガンダム” と関連がないため，“落合博満” と “ガンダム” の関係は意外である，というように，オブジェクトの多くの同位語らしい語が対象の属性値と関係があるか（本稿ではこれを “類似関係” と呼ぶ）を基に，語間の関係の意外度を測っていた．

しかし，オブジェクトの多くの同位語らしい語が同じ属性値と関連があるからといって，必ずしもその関係が良く知られているとは限らない．たとえば，“三十三間堂” というオブジェクトと “入母屋造” という属性値の関係を考える．“三十三間堂” の同位語らしい語の中には “醍醐寺” や “仁和寺”，“東寺” などがあり，これらはいずれも “入母屋造” と関連があるが，その関係は一般にはあまり知られていないと考えられる．つまり，実際の関連度と関係の認知度は必ずしも一致しないため，これまでのアプローチとは異なった方法でオブジェクトと属性値の関係の認知度を測る必要がある．

類似関係に依らず語間の関係の認知度を正しく推定できれば，多くのオブジェクトが同じ属性値と関係があるが実は知られていないこと，のようなこれまでの提案手法では発見が難しかったオブジェクト間の意外な共通点の発見も可能になると考えられる．意外な共通点には様々なものが存在する．情報を提示するユーザの興味を引くか否かという観点と，情報が有用であるか否かという観点を考えたときに，ユーザの興味を引く有用な情報を発見し提示することは以下のような状況で有益になり得る．たとえば，京都にある 4 つの寺に観光に行こうとしているユーザに対して，その 4 つの寺の意外な共通点を提示することで，寺に対するユーザの興味をより喚起することができる．他にも，ある 5 人の戦国武将について勉強をしているユーザに対して，その 5 人の戦国武将の意外な共通点を提示することも勉強内容に対する興味の喚起につながりうる．

本研究の最終的な目的は，オブジェクト集合 $O = \{t_{o_1}, t_{o_2}, \cdots, t_{o_n}\}$ と O の全てのオブジェクトに共通する属性値 t_a を入力として与えたときに，O にとっての t_a の意外度を測ることである．この目的を実現するためには，上述したように各 $t_{o_i} \in O$ と t_a の関係の認知度を正しく求める必要がある．そこで，本稿では特にオブジェクトと属性値の関係の認知度を推定することに焦点を当て，手法の提案と評価を行う．語間の関連度を測ることを目的として，これまでに多数の手法が

(注1)：http://www.google.com
(注2)：http://www.yahoo.com
(注3)：http://www.bing.com

提案されてきたが [5] [6] [7] [8] [9] [10] [11]，提案手法では，これらの手法を拡張し，オブジェクトの認知度および，オブジェクトの類似オブジェクトと属性値の関係の認知度を考慮することでオブジェクトと属性値の関係の認知度を推定する．

我々は “国”，“野菜”，“京都の観光地”，“電機メーカー”，“野球選手” の 5 カテゴリに関する 25 語の属性値を対象としてオブジェクトと属性値の関係の認知度推定に関する評価実験を行った．実験の結果，オブジェクトの認知度および，オブジェクトの類似オブジェクトと属性値の関係の認知度を考慮することで，既存手法に比べて有意に高い精度で関係の認知度を推定できることが明らかになった．また，実験の結果を基に意外な共通点の発見への応用について考察を行った．

本稿の以降の構成は下記の通りである．2. 章では関連研究について述べる．3. 章ではオブジェクトと属性値の関係の認知度推定のアプローチについて述べ，4. 章で関係の認知度推定手法を説明する．5. 章では実験結果を示し，6. 章ではオブジェクト間の意外な共通点の発見に関する考察を行う．7. 章ではまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 意外な情報発見に関する研究

これまでにも意外な情報を発見することを目的とした研究はいくつか行われてきた [12] [13] [14] [15]．Noda ら [15] は Wikipedia(注4) のカテゴリ間の関係を分析することで，(Wikipedia の見出し語，カテゴリ名 1，カテゴリ名 2) で表される 3 つ組の中で意外性のある知識を発見する手法を提案した．彼らの手法を用いることで，たとえば “麻生太郎” という Wikipedia の見出し語を入力語として与えると，“日本の内閣総理大臣” と “オリンピック射撃競技日本代表選手” という 2 語のカテゴリ名が意外性のある語として出力される．すなわち，“麻生太郎は，日本の内閣総理大臣というカテゴリに属している一方で，オリンピック射撃競技日本代表選手というカテゴリにも属している” という情報を発見することができる．彼らの目的は 1 つのオブジェクト（Wikipedia の見出し語）にとって意外な属性値（カテゴリ名）を発見することであり，我々の目的とは異なる．

Nadamoto ら [14] は，ブログや SNS といったコミュニティ型コンテンツを対象として，コミュニティ内の議論においてユーザが気づいていない情報を発見する手法を提案している．彼らの手法では，コミュニティ型コンテンツで，1 語の Wikipedia の見出し語と，それについて語られている文集合を入力として与える．それらの入力に対して，入力として与えた見出し語の Wikipedia の記事内から，コミュニティ型コンテンツで語られていないトピックで，かつコミュニティ型コンテンツで語られているトピックとは最も類似度の低いトピックに関する文集合を発見し出力する．つまり，与えられた文集合には存在しない意外な情報を発見することを目的としている．

Liu ら [12] の研究では，2 つの Web ページを与えたときに，一方の Web ページにしか含まれていない意外な情報を，各 Web ページに含まれるキーワードを比較することで発見する手法を提案している．Web ページをオブジェクトとみなすと，彼らの研究ではオブジェクト間で共通していない意外な情報の発見を目的としていると言える．

Mejova ら [13] は，入力として与えられた 1 語のクエリに対する意外な語を発見することを目的とし，手法を提案している．彼らの手法は，ある文書集合の中で，出現する文書数が少なく，かつクエリとの共起度が高い語はクエリにとって意外な語であるという考えに基づいている．

2.2 語間の関連度推定に関する研究

これまでに，WebJaccard や WebDice，WebOverlap や NGDW，WebPMI など，Web 検索エンジンの検索結果を用いて語間の関連度を計算する手法が多数提案されてきた．Bollegala ら [6] は検索結果数に加えて，検索結果のスニペットから抽出した構文パターンを用いることで高い精度で語間の関連度を求めることを可能とした．検索結果のスニペットを用いた手法としては，関連度を求めたい 2 語の検索結果のスニペットの類似度に基づく手法 [7] や，1 方の語をクエリとした Web 検索結果中のスニペットにもう 1 方の語がどの程度出現するかという情報を用いる手法 [8] などもある．

Wikipedia の情報を用いて語間の関連度を求める研究も行われている [9] [10] [11]．Strube ら [9] は Wikipedia 上での 2 つの記事間のパスの長さや，2 つの記事間での語の重複度合いを元に語間の関連度を求めている．Gabrilovich ら [10] は Wikipedia の見出し語が属するカテゴリ情報に基づいて語間の関連度を求める手法を提案した．Milne ら [11] は 2 語が与えられたときに，各語が見出しとなっている Wikipedia の記事に対してリンクを張っている Wikipedia の記事集合同士の比較と，各語が見出しとなっている Wikipedia の記事がリンクを張っている Wikipedia の記事集合同士の比較に基づいて 2 語の関連度を求めている．Milne らの手法は Strube らの手法よりも高い精度で語間の関連度を求めることができ，Gabrilovich らの手法よりも低いコストで Gabrilovich らの手法と同程度の精度が出ることが示されている．

3. 語間の関係の認知度推定のアプローチ

本章では，人が語間の関係の強さを判断する際に影響を与えると考えられる要因について述べる．語間の関係の認知度推定の最終的な目的は，語の組 (t_1, t_2) が与えられたときに t_1 と t_2 の絶対的な関係の認知度を求めることであるが，本稿ではまず，あるカテゴリ c と c に属する語集合 $T_c = \{t_{o_1}, t_{o_2}, \cdots, t_{o_n}, \}$，語 t_a が与えられたときに，$t_{o_i} \in T_c$ と t_a の関係の認知度を求め，その値が高い順に $t_{o_i} \in T_c$ をランキングすることを目的とする．これにより，たとえば c を “国”，T_c を国名の集合，t_a を “ワイン” とすると，ワインとの関係の認知度が高い順にランキングされた国名のリストが返される．以降では説明のため，$t_{o_i} \in T_c$ をオブジェクト，t_a を属性値と呼ぶ．

1. 章で述べたように，語間の実際の関連度と，その関係がどの程度知られているかという関係の認知度は必ずしも一致しな

(注4)：http://ja.wikipedia.org/

いと考えられる．我々はその原因として，人が語間の関連度について考える際に，以下に述べるように考えることで，実際の関連度と人が感じる関連度に差が生じるのではないかと考えた．

3.1 オブジェクトの認知度

カテゴリを “野球選手”，属性値を “ゴールデングラブ賞” としたときの，野球選手とゴールデングラブ賞の関係の認知度について考える．まず，“落合博満” と “ゴールデングラブ賞” の関係の認知度について考える．実際には落合博満はゴールデングラブ賞を獲得したことがないため，関連度は低い．しかしユーザは “落合博満は有名な野球選手なのでゴールデングラブ賞とも関係があるのでは” と考え，本来の関連度よりも高く見積もられてしまう．次に，“岡田彰布” と “ゴールデングラブ賞” の関係の認知度について考える．実際には岡田彰布はゴールデングラブ賞を獲得したことがあるため，関連度は落合博満の場合より高い．しかしユーザは “岡田彰布はあまり聞いたことのない野球選手なのでゴールデングラブ賞とも関係がないのでは” と考え，本来の関連度よりも低く見積もられてしまう．

これらの考えに基づき，次のような仮説を立てる．

【仮説 1】 オブジェクトの認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる．

3.2 オブジェクトの類似オブジェクトと属性値の関係の認知度

カテゴリを “国”，属性値を “コーヒー” としたときの，国とコーヒーの関係の認知度について考える．まず，“アルゼンチン” と “コーヒー” の関係の認知度について考える．実際はアルゼンチンはコーヒーの生産量や消費量において世界の国の中では上位には入っておらず，関連度はそれほど高くない．しかしユーザは “アルゼンチンと似た国であるブラジルやコロンビア，メキシコはコーヒーと関係が強いのでアルゼンチンもコーヒーと関係が強いのでは” と考え，本来の関連度よりも高く見積もられてしまう．次に，“インド” と “コーヒー” の関係の認知度について考える．実際はインドはコーヒーの生産量において世界の中で上位に入っているため，関連度は高い．しかしユーザは “インドと似た国である中国やパキスタン，スリランカはコーヒーと関係が弱いのでインドもコーヒーと関係が弱いのでは” と考え，本来の関連度よりも低く見積もられてしまう．

これらの考えに基づき，次のような仮説を立てる．

【仮説 2】 オブジェクトの多くの類似オブジェクトと属性値の関係の認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる．

ただし，この仮説に従うならば，上記の例で人が “コロンビア” と “コーヒー” の関係の認知度を考えるときは，“コロンビア” と似た国である “アルゼンチン” と “コーヒー” の関係の認知度も影響を与えている，というようにオブジェクトと属性値の関係の認知度は再帰的に決まるものとなる．

4. 関係の認知度推定手法

本章では 3. 章で述べた 2 つの仮説に基づいて，語間の関係の認知度を推定する手法について述べる．

我々は，これまで提案されてきた語間の関連度を測る手法を拡張することで関係の認知度を推定するというアプローチをとる．そのためにまず，4.1 節では本稿で用いる既存手法を説明する．次に 4.2 節でオブジェクトの認知度を考慮した手法，4.3 節でオブジェクトの類似語を考慮した手法について述べる．最後に 4.4 節でオブジェクトの認知度と類似語を共に考慮した手法について述べる．

4.1 既 存 手 法

本稿では語間の関連度を測る手法として以下の 3 手法を用いる．

4.1.1 WLM 手法

1 つ目は Milne ら [11] により提案された手法であり，語 t_1 と語 t_2 の関連度を測る際に Wikipedia 上で t_1 と t_2 それぞれのリンク先ページの類似度と t_1 と t_2 それぞれのリンク元ページの類似度を基に語間の関連度を測る．

まずリンク先ページの類似度は TF-IDF を基にした手法で求める．リンク元ページを s，リンク先ページを t としたときに，s から t へのリンクの重みは次式により求められる．

$$w(s \to t) = \log\left(\frac{|W|}{|B_t|}\right) \text{ if } s \in B_t,\ \ 0 \text{ otherwise.} \qquad (1)$$

ここで B_t は t のリンク元ページ集合，W は Wikipedia の全ページ集合である．この値を要素とするベクトルを t_1 と t_2 それぞれに対して作成する．つまり式（1）において s は t_1 または t_2 であり，t_1 と t_2 のリンク先ページ集合を F_{t_1}，F_{t_2} とすると t は $F_{t_1} \cup F_{t_2}$ に含まれる全てのページとなる．最後に t_1 と t_2 の各ベクトル間のコサイン類似度 $sim_{forward}(t_1, t_2)$ を求める．

続いて t_1 と t_2 それぞれのリンク元ページ集合の類似度を次式により求める．

$$sim_{backward}(t_1, t_2) = \frac{\log\left(\max\left(|B_{t_1}|, |B_{t_2}|\right)\right) - \log\left(|B_{t_1} \cap B_{t_2}|\right)}{\log\left(|W|\right) - \log\left(\min\left(|B_{t_1}|, |B_{t_2}|\right)\right)}. \qquad (2)$$

最後に，上記の 2 つの式の線形和により t_1 と t_2 の関連度 $wlm(t_1, t_2)$ を求める：

$$wlm(t_1, t_2) = \frac{1}{2} \cdot sim_{forward}(t_1, t_2) + \frac{1}{2} \cdot sim_{backward}(t_1, t_2). \qquad (3)$$

4.1.2 WebPMI 手法

2 つ目は Web 上での 2 語の共起を基に関連度を測る WebPMI [6] [5] を用いる．WebPMI では，語 t_1 と語 t_2 の関連度は次式により求められる．

$$web\_pmi(t_1, t_2) = \begin{cases} 0 & \text{if } hit(t_1, t_2) \leqq c \\ \log_2\left(\frac{hit(t_1 \wedge t_2)/N}{(hit(t_1)/N) \times (hit(t_2)/N)}\right) & \text{otherwise.} \end{cases} \qquad (4)$$

$hit(t)$ は t で Web 検索を行った際の検索結果数を表す．検索結果数の取得には Clue Web 09 の日本語データセットを用いた(注5)．本稿では Bollegala ら [6] に倣い，$c = 5$ とした．N は全 Web ページ数であり，Clue Web 09 の日本語データセット

(注5)：http://lemurproject.org/clueweb09/

では $N = 67,337,717$ であった．

検索エンジンの検索結果数を元に単語間の関連度を測る手法としては WebJaccard や WebDice，WebOverlap や NGD があるが，これらの手法の中では WebPMI が最も精度高く語間の関連度を推定できることが示されている [5], [6]．

4.1.3 Noda 手法

3 つ目も Web 上での 2 語の共起を基に関連度を測る手法であり，野田ら [16] によって提案された手法に基づいたものである．この手法では，語 t_1 と語 t_2 の関連度は次式により求められる．

$$noda(t_1, t_2) = \frac{hit(t_1 \wedge t_2)^2}{hit(t_1) \cdot hit(t_2)}. \tag{5}$$

この手法では，WebPMI よりも 2 語の共起回数が語間の関連度に与える影響が大きくなっている．

4.2 オブジェクトの認知度を考慮した関係の認知度推定手法

3.1 節の仮説 1 に基づき，オブジェクトと属性値の関連度とオブジェクトの認知度を考慮した，オブジェクト t_{o_i} と属性値 t_a の関係の認知度 $rel\_pop(t_{o_i}, t_a)$ を次式により求める．

$$rel\_pop(t_{o_i}, t_a) = rel(t_{o_i}, t_a) \cdot pop(t_{o_i}) \tag{6}$$

ここで $rel(t_{o_i}, t_a)$ はオブジェクトと属性値の関連度であり，4.1 節で述べたいずれかの手法により求める．$pop(t_{o_i})$ はオブジェクトの認知度であり，本稿では Wikipedia 上での t_{o_i} のリンク元のページ数の対数を t_{o_i} の認知度とみなす．我々はこれまでに Wikipedia 上でのリンク元のページ数と認知度の間に高い相関があることを示している [17]．

4.3 類似オブジェクトを考慮した関係の認知度推定手法

4.3.1 類似オブジェクトの取得

本稿では，オブジェクト t_{o_i} の同位語らしい語を t_{o_i} の類似オブジェクトとする．同位語とは共通の上位語を持つ語のうち特に意味の異なる語 [18] である．我々はこれまでに，与えられた語に対する同位語らしい語を求める手法を提案してきた [19]．この手法では，ALAGIN フォーラムから提供されている上位下位関係抽出ツール (注6) を用いる．このツールは，Wikipedia で記事の見出し語やカテゴリ名となっている名詞句をその上位語，下位語関係に基づいて階層化したものであり，約 20 万語の上位語と約 245 万語の下位語を含む．このツールを用いて，まず t_{o_i} の全上位語と，t_{o_i} と 1 つ以上の上位語を共有する同位語集合を取得する．次に，上位語と同位語の間で上位下位関係にある語を枝で結び，二部グラフを作成する．その二部グラフに対して HITS アルゴリズム [2] を拡張した手法を適用し，同位語を t_{o_i} との同位語らしさに応じてランキングする．この手法を用いることで，オブジェクト t_{o_i} に対する同位語 t_{o_j} の同位語らしさ $coordinate(t_{o_i}, t_{o_j})$ を求める．

4.3.2 関係の認知度推定

3.2 節で述べたように，仮説 2 に基づいてオブジェクトと属性値の関係の認知度を推定する場合，オブジェクトとその同位語と属性値の関係の認知度は再帰的に求める必要がある．本稿では，これを実現するために biased PageRank アルゴリズム [20] を用いる．biased PageRank を用いるのは，biased PageRank が以下のような性質を持つためである．

（1） より多くの重要なページからリンクされているページは重要である．

（2） あらかじめ重要であることがわかっているページはリンク構造に依らず重要である．

性質（1）は，仮説 2 において属性値との関係の認知度が高い多くのオブジェクトと同位語らしいオブジェクトは属性値との関係の認知度が高いという部分に対応している．また仮説 2 で “実際の関連度よりも高（低）くなる” と述べているように，関係の認知度は実際の関連度と全く無関係ではなく，性質（2）がオブジェクトと属性値の実際の関連度を考慮していることに対応している．

Biased PageRank を適用するため，まず隣接行列を作成する．$T_c = \{t_{o_1}, t_{o_2}, \cdots, t_{o_n}\}$ としたとき，(i, j) 要素の値が $coordinate(t_{o_j}, t_{o_i})$ であるような n 次正方行列を作成する．次に，各列の和が 1 になるように正規化し，隣接行列とする．つまり，隣接行列の (i, j) 要素の値 $a_{i,j}$ は次式により求められる．

$$a_{i,j} = \frac{coordinate(t_{o_j}, t_{o_i})}{\sum_{t_{o_k} \in T_c} coordinate(t_{o_j}, t_{o_k})}. \tag{7}$$

この隣接行列を用いて，オブジェクト t_{o_i} と属性値 t_a の関係の認知度 $rel\_crd(t_{o_i}, t_a)$ は次式により求められる．

$$\begin{aligned} rel\_crd_{l+1}(t_{o_i}, t_a) =& \alpha \cdot \sum_{1 \leqq j \leqq n} a_{i,j} \cdot rel\_crd_l(t_{o_i}, t_a) \\ &+ (1 - \alpha) \cdot \frac{rel(t_{o_i}, t_a)}{\sum_{t_{o_j} \in T_c} rel(t_{o_j}, t_a)}. \end{aligned} \tag{8}$$

式（8）中の第 1 項，第 2 項がそれぞれ biased PageRank の性質（1），性質（2）に対応している．α はダンピングファクターであり，$0 \leqq \alpha \leqq 1$ の値をとる．α の値が大きいほど同位語と属性値の関係の認知度の影響が大きくなる．実験では α の値の違いによる関係の認知度推定の精度についても評価を行う．

4.4 オブジェクトの認知度と類似オブジェクトを考慮した関係の認知度推定手法

オブジェクトの認知度と類似オブジェクトを共に考慮してオブジェクト t_{o_i} と属性値 t_a の関係の認知度を推定する際は，次式のように biased PageRank においてオブジェクトと属性値の関連度の初期値を 4.2 節で求めた値とする．

$$\begin{aligned} rel\_pop\_crd_{l+1}(t_{o_i}, t_a) =& \alpha \cdot \sum_{1 \leqq j \leqq n} a_{i,j} \cdot rel\_pop\_crd_l(t_{o_i}, t_a) \\ &+ (1 - \alpha) \cdot \frac{rel_{pop}(t_{o_i}, t_a)}{\sum_{t_{o_j} \in T_c} rel_{pop}(t_{o_j}, t_a)}. \end{aligned} \tag{9}$$

5. 実　　験

提案手法の有用性を検証するために実験を行った．実験では以下の 2 点を明らかにすることを目的とする．

（1） オブジェクトの認知度を考慮することはオブジェクトと属性値の関係の認知度を推定するうえで有用であるか．

(注6)：https://alaginrc.nict.go.jp/hyponymy/index.html

表 1　評価に用いたカテゴリと属性値および各カテゴリのオブジェクト数.

カテゴリ	オブジェクト数	属性値
国	169	ワイン，ビール，コーヒー，ピザ，バナナ
野菜	117	ビタミン C，食物繊維，鉄分，タンパク質，カルシウム
京都の観光地	155	織田信長，豊臣秀吉，徳川家康，足利義満，千利休
電機メーカー	481	携帯電話，液晶テレビ，デジタルカメラ，炊飯器，パソコン
野球選手	157	本塁打王，首位打者，盗塁王，MVP，ゴールデングラブ賞

表 2　各カテゴリで関係の認知度の評価に使用したオブジェクトの数.

カテゴリ	オブジェクト数
国	40
野菜	39
京都の観光地	13
電機メーカー	30
野球選手	32

（2） オブジェクトの類似オブジェクトと属性値の関係の認知度を考慮することはオブジェクトと属性値の関係の認知度を推定するうえで有用であるか.

5.1　データセット

関係の認知度の評価には “国”，“野菜”，“京都の観光地”，“電機メーカー”，“野球選手” の 5 カテゴリ，各カテゴリで 5 語ずつの属性値を使用した．本実験では，各カテゴリに属するオブジェクト集合は著者が人手で作成した．4.1.1 項の WLM 手法を用いるため，また 4.3.1 項で述べた方法でオブジェクトの同位語を取得するため，オブジェクトおよび属性値はいずれも Wikipedia に記事が存在する語とした．各カテゴリで用いたオブジェクトの数と属性値を表 1 に示す.

本実験で用いた Wikipedia のデータは 2008 年 6 月 24 日にダンプされた Wikipedia 日本語版のデータベースをダウンロードして利用した．このデータ上での式（1）における $|W|$ の値は $1,342,098$ であった.

5.2　関係の認知度の正解セット

4. 章で述べた手法を評価するためには，オブジェクトと属性値の関係の認知度の正解値が必要となる．本実験ではクラウドソーシング (注7) を用いて関係の認知度の正解セットを作成する．属性値とオブジェクトの関係の有無を評価者に問う際，評価者がオブジェクトのことを全く知らないと関係の有無を推測することも難しい．そこでまず，各カテゴリのオブジェクトについて Wikipedia でのバックリンク数の多い上位 40 語を認知度の高い語として取り出した．さらに，各語のことを知っているかどうかをクラウドソーシングを用いて 1 語あたり 10 名にアンケートをとった．最後に，10 名のうち 5 名以上が知っていると回答した語はある程度認知度のある語として関係の認知度の評価に使用した．各カテゴリで 5 名以上が知っていると回答したオブジェクトの数を表 2 に示す.

関係の認知度の正解セットを作成する際は，(京都の観光地，織田信長）のようなカテゴリと属性値の組に対して 1 つのアンケートを作成する．アンケートでは表 2 に示した対象カテゴリのオブジェクトの一覧を評価者に見せ，属性値と関係があると思うものをいくつでも選択させた．このとき評価者には，関係の有無について本や Web で調べたり，他人に聞いたりしないよう求めた．また，いずれのオブジェクトも属性値と関係が無いと思う場合には “いずれとも関係はない” を選択させた．(京都の観光地，織田信長）の組に対するアンケートのインタフェースを図 1 に示す．カテゴリと属性値の 1 つの組のアンケートに対してクラウドソーシング上で 20 人の回答を集め，各オブジェクトに対して 20 人の中で属性値と関係があると回答した人数をそのオブジェクトと属性値の関係の認知度の正解値とした.

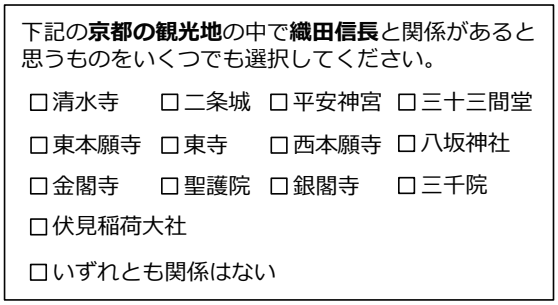

図 1　カテゴリが “京都の観光地”，属性値が “織田信長” の場合のアンケートの例.

5.3　手法・評価指標

関係の認知度の推定手法として，4.1 節で述べた 3 つの既存手法（WLM 手法，WebPMI 手法，Noda 手法），4.2 節で述べたオブジェクトの認知度を考慮した手法（WLM+pop 手法，WebPMI+pop 手法，Noda+pop 手法），4.3 節で述べた類似オブジェクトを考慮した手法（WLM+BPR 手法，WebPMI+BPR 手法，Noda+BPR 手法）4.4 節で述べたオブジェクトの認知度と類似オブジェクトを考慮した手法（WLM+pop+BPR 手法，WebPMI+pop+BPR 手法，Noda+pop+BPR 手法）を用いた．Biased PageRank を用いる手法では，式（8）および式（9）においてダンピングファクターの値を 0.1 から 0.9 まで 0.1 ずつ変化させた手法を用いた.

カテゴリと属性値の組に対して手法の評価を行う際は，各手法で求められる，属性値との関係の認知度が高い順にランキングされたオブジェクトのリストと，5.2 節で作成した正解セット内でのその属性値との関係の認知度が高い順にランキングされたオブジェクトのリストを比較し，スピアマンの順位相関係数を求めた.

5.4　結　　果

5.3 節で述べた各手法に対する全カテゴリの相関係数の平均値および各カテゴリの相関係数の平均値を図 2 に示す．棒グラ

(注7)：本実験ではランサーズ（http://www.lancers.jp/）を使用した.

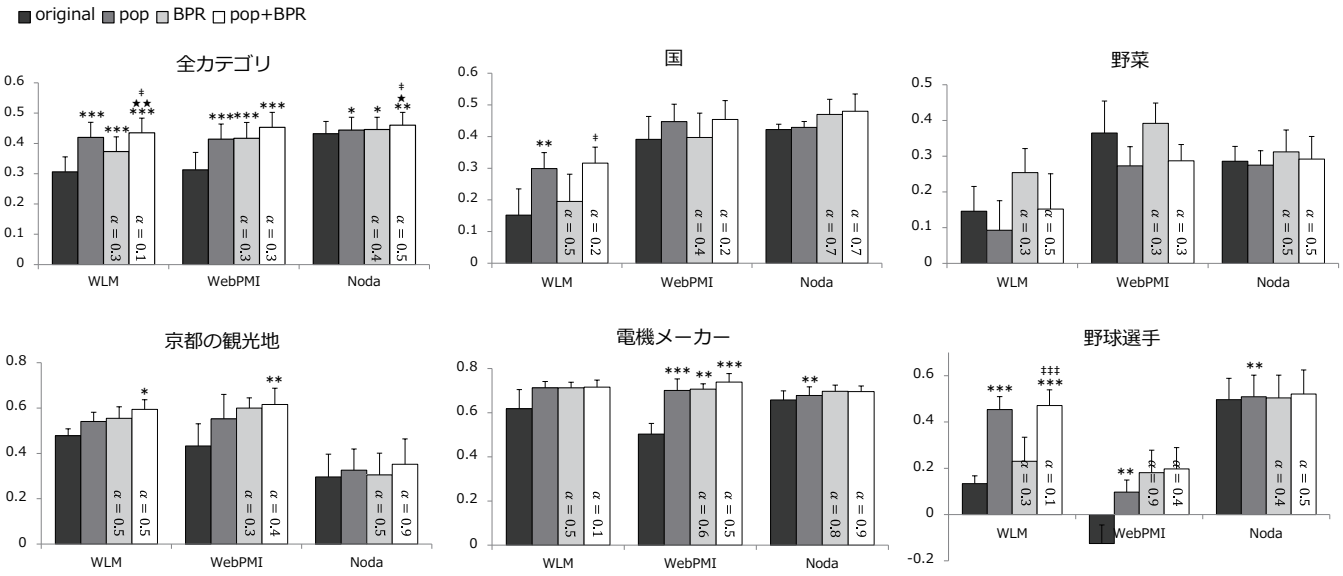

図 2　全手法に対する全カテゴリと各カテゴリの相関係数の平均値．棒グラフ上の α の値はダンピングファクターの値である．*，**，***はそれぞれ original の手法との間に有意水準 10%，5%，1%で有意差があることを表す．同様に★，‡はそれぞれ pop と pop+BPR，BPR と pop+BPR の間の有意差を表す．

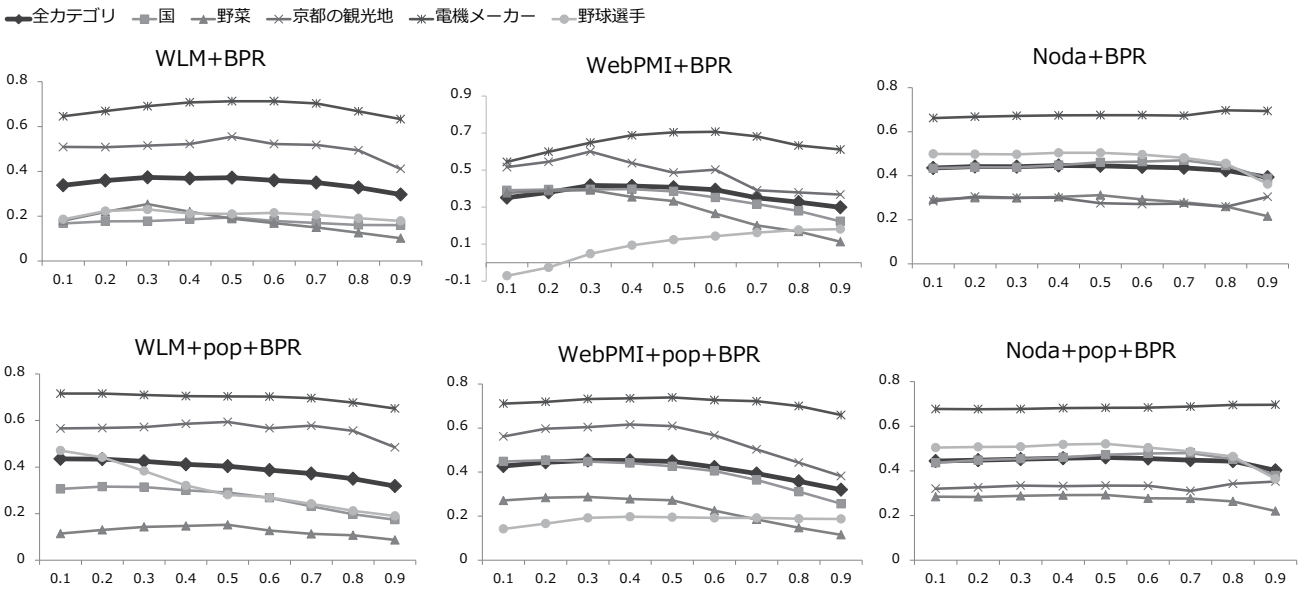

図 3　Biased PageRank を用いた手法で全カテゴリと各カテゴリでダンピングファクターの値を変化させた際の相関係数の平均値．

フの縦軸はスピアマンの順位相関係数を表す．図中で “original” と記述しているものは WLM 手法，WebPMI 手法，Noda 手法のいずれかを指す．Biased PageRank を用いる手法では，ダンピングファクターの値を変化させたときに相関係数の平均値が最も高かった手法のみを記載しており，そのときのダンピングファクターの値は棒グラフ上に記載されている．全カテゴリの平均値を見ると，WLM 手法，WebPMI 手法，Noda 手法のいずれの場合も，オブジェクトの認知度だけを考慮した場合（図 2 中の pop），類似オブジェクトだけを考慮した場合（図 2 中の BPR）ともに既存手法に比べて有意に高い精度でオブジェクトと属性値の関係の認知度を推定できており，3. 章で述べた仮説 1 および仮説 2 は正しいことが示された．WLM 手法と Noda 手法ではオブジェクトの認知度と類似オブジェクトをともに考慮した手法（図 2 中の pop+BPR）がいずれか一方のみを考慮する手法を有意に上回っており，両方の要因を同時に考慮することの有用性を示している．各カテゴリでの結果を見ると，“野菜” カテゴリ以外では，オブジェクトの認知度や類似オブジェクトを考慮することで既存手法より相関係数の値は高くなった．野菜カテゴリでは，評価に用いた多くの野菜が十分に知名度の高いものであったため，オブジェクトの認知度による影響が小さく，仮説 1 が成り立たなかったと考えられる．

次に，biased PageRank を用いた 6 手法について，ダンピン

表 3　カテゴリが “国”，属性値が “ワイン” の際の既存手法と提案手法の結果の比較.

順位	Noda 手法	Noda+BPR 手法（$\alpha = 0.9$）
1	フランス（1）	フランス（1）
2	イタリア（2）	イタリア（2）
3	日本（7）	日本（3）
4	スペイン（3）	スペイン（4）
5	ドイツ（6）	ドイツ（6）
6	中国（-）	イギリス（13）
7	タイ（-）	オーストラリア（8）
8	オーストラリア（8）	ポルトガル（10）
9	インド（-）	インド（-）
10	イギリス（13）	ニュージーランド（17）
相関係数	0.457	0.698

表 4　カテゴリが “電機メーカー”，属性値が “冷蔵庫” の際の既存手法と提案手法の結果の比較.

順位	WebPMI 手法	WebPMI+pop+BPR 手法（$\alpha = 0.9$）
1	ダイキン工業（9）	日立製作所（1）
2	象印マホービン（8）	東芝（3）
3	シャープ（4）	シャープ（4）
4	三菱電機（2）	三菱電機（2）
5	日本ビクター（5）	パナソニック（5）
6	三洋電機（6）	ソニー（10）
7	ケンウッド（-）	富士通（7）
8	パナソニック（5）	三洋電機（6）
9	東芝（3）	日本電気（11）
10	日立製作所（1）	ダイキン工業（9）
相関係数	0.400	0.714

グファクターの値を変化させた際の相関係数について，全カテゴリの平均値と各カテゴリの平均値を図 3 に示す．多くの手法とカテゴリにおいてダンピングファクターの値を大きくしすぎると相関係数が小さくなっていることからも，類似オブジェクトを考慮しすぎうと関係の認知度を推定する上で精度の低下を招くことが分かる．

類似オブジェクトを考慮することが有効に働いた例を表 3 に示す．表 3 にはカテゴリが “国”，属性値が “ワイン” のときの Noda 手法と Noda+BPR 手法で $\alpha = 0.9$ としたときの，ワインとの関係の認知度が高いと推定された上位 10 語ずつを記載している．カッコ内の数値は正解データ内での順位であり，“-” は関係があると回答した評価者がいなかったことを表す．Noda 手法では中国やタイ，インドといった国が 10 位以内に含まれており，これらの国では実際ワインが作られているため，関連度は高いと言える．しかし，いずれの国もワインと関係があると回答した評価者はおらず，関係の認知度は低い．中国は大韓民国や朝鮮民主主義人民共和国との同位語らしさが高く，かつこれらの国は Noda 手法により求められるワインとの関連度が低かったため，Noda+BPR 手法では 17 位に下げることができていた．同様にタイは 26 位に下がっていた．

オブジェクトの認知度と類似オブジェクトを共に考慮することが有効に働いた例を表 4 に示す．表 4 にはカテゴリが “電機メーカー”，属性値が “冷蔵庫” のときの WebPMI 手法と WebPMI+pop+BPR 手法で $\alpha = 0.9$ としたときの，冷蔵庫との関係の認知度が高いと推定された上位 10 語ずつを記載している．提案手法で 5 位以内にある語はいずれも正解データでも 5 位以内に含まれていた語であり，既存手法よりも大幅に精度が改善されていた．ただし，既存手法では上位に “象印マホービン” や “日本ビクター” が含まれていたが，実際にはこれらのメーカーでは冷蔵庫は製造されておらず，関連度は低くなるべき語であった．このように，実際の関連度が正しく求められていない例は他のカテゴリや属性値でも見られたため，Bollegala ら [6] や Gabrilovich ら [10] により提案された手法のような，より精度高く語間の関連度を推定する手法を基にして提案手法の有用性を評価することが今後の課題としてあげられる．

表 5　オブジェクト間の意外な共通点として得られた結果の例.

カテゴリ	属性値	意外度の順位	オブジェクトの組み合わせ
国	ラグビー	1	{ ケニア，南アフリカ共和国 }
		55（最下位）	{ オーストラリア，フランス }
国	石油	3	{ アルジェリア，ミャンマー }
		136（最下位）	{ イラン，日本 }
京都の観光地	足利義満	1	{ 仁和寺，八坂神社 }
		28（最下位）	{ 金閣寺，相国寺 }
芸能人	読売ジャイアンツ	1	{ 堀江由衣，秋元康 }
		190（最下位）	{ 中居正広，石橋貴明 }

6. オブジェクト間の意外な共通点の発見

6.1 ケーススタディ

本章では，実験で得られた知見を基にオブジェクト間の意外な共通点発見への手法の適用を試みる．オブジェクト間の意外な共通点の発見を行う際の入力には様々なものが考えられるが，本稿ではカテゴリ c，オブジェクト数 m，属性値 t_a を入力として与える．この入力に対してシステムは，c に含まれるオブジェクトのあらゆる m 個の組み合わせに対して t_a を共通点として持つことの意外度を計算し，意外度の高い順にランキングされたオブジェクト集合のリストを返す．カテゴリ c で m 個のオブジェクトを含む集合 $O_m = \{o_1, o_2, \cdots, o_m\}$ が t_a を共通点として持つことの意外度は次式により求める．

$$unexp(c, O_m, t_a) = \prod_{o_i \in O_m} (1 - rel\_pop\_cred(o_i, t_a)) \quad (10)$$

本稿では $m = 2$ とし，$rel\_pop\_cred(o_i, t_a)$ には 5. 章の実験において全カテゴリの平均値が最も高かった Noda+pop+BPR 手法でダンピングファクターを 0.5 としたものを用いた．

上記の方法により得られた意外な共通点の例を表 5 に示す．たとえば表 5 では “ラグビー” と関係のある国の組み合わせの中で最も意外度の高いものは “ケニア” と “南アフリカ共和国” であり，最も意外度の低いものは “オーストラリア” と “フランス” であった．“ケニア” と “南アフリカ共和国” はいずれもラグビーが盛んであるが，一般にはそのようなイメージがないため，筆者にとっては意外な共通点であった．

6.2 意外な共通点発見のための課題

最後に，本稿では考慮しなかったが，今後オブジェクト間の意外な共通点の発見に取り組むうえで考慮すべき要素の中でも特に重要であると考えられるものを 2 つあげる．

1 つ目はオブジェクト間の類似度である．たとえば 2 つの国から成る集合 { フランス，イタリア } がある共通の意外な属性値を持ち，また別の 2 つの国から成る集合 { フランス，インドネシア } がある共通の意外な属性値を持ち，式（10）によって計算されるそれらの意外度は等しいとする．しかしこの場合，後者の方が人の感じる意外度は前者より高くなると考えられる．なぜなら，フランスとイタリアは似た国であるイメージがあるため，何らかの共通点を持っていることは意外でないのに対して，フランスとインドネシアは似た国であるイメージがないため，共通点を持っていること自体が意外であるためである．このように，オブジェクト間の類似度が低くなるほど共通点の意外度が高くなるように考慮する必要がある．オブジェクト間の類似度を測る方法としては，オブジェクトが見出しの Wikipedia の記事間の類似度を計算したり，オブジェクト間の同位語らしさを計算するといった方法が考えられる．

2 つ目は属性値が記述される文脈の類似度である．本稿ではオブジェクトと属性値がどのような文脈で関係があるのかを考慮していなかったため，たとえば “フランス” と “コーヒー” に関係があるというのが “フランスではコーヒーの生産量が多い” のか，“フランスではコーヒーの消費量が多い” のか，あるいはそれ以外の関係なのかがわからない．オブジェクト間の意外な共通点を発見する場合には，各オブジェクトと属性値の関係の文脈も類似している方が望ましいと考えられるため，文脈が高くなるほど共通点の意外度が高くなるように考慮する必要がある．文脈の類似度を測る方法としては，オブジェクトが見出しの Wikipedia の記事中で属性値が含まれる文を抽出し，文間の類似度を計算するといった方法が考えられる．

7. ま と め

本稿では，入力として与えられたオブジェクトと属性値の関係の認知度を推定するための手法を提案した．手法を提案するにあたり，我々は人が語間の関係を考える際に，(1) オブジェクトの認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる，(2) オブジェクトの多くの類似オブジェクトと属性値の関係の認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる，という仮説を立てた．5 カテゴリで合計 25 語の属性値に対して関係の認知度推定の評価実験を行った結果，提案手法が既存手法より有意に高い精度で関係の認知度推定ができており，2 つの仮説が正しいことを示した．さらに，提案手法を応用することで，オブジェクト間の意外な共通点の発見の可能性について考察した．今後は，オブジェクト間の意外な共通点の発見について定量的な評価を行う予定である．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金（課題番号 24240013，24680008，12J03993）によるものです．ここに記して謝意を表します．

文　　献

[1] S. E. Robertson and S. Walker: “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval”, Proc. of ACM SIGIR 1994, pp. 232–241 (1994).
[2] J. M. Kleinberg: “Authoritative sources in a hyperlinked environment”, J. ACM, **46**, pp. 604–632 (1999).
[3] S. Brin and L. Page: “The anatomy of a large-scale hypertextual web search engine”, Proc. of WWW 2007, pp. 107–117 (1998).
[4] K. Tsukuda, H. Ohshima, M. Yamamoto, H. Iwasaki and K. Tanaka: “Discovering unexpected information on the basis of popularity/unpopularity analysis of coordinate objects and their relationships”, Proc. of SAC 2013, pp. 878–885 (2013).
[5] G. Lu, P. Huang, L. He, C. Cu and X. Li: “A new semantic similarity measuring method based on web search engines”, W. Trans. on Comp., **9**, 1, pp. 1–10 (2010).
[6] D. Bollegala, Y. Matsuo and M. Ishizuka: “A web search engine-based approach to measure semantic similarity between words”, IEEE Trans. on Knowl. and Data Eng., **23**, 7, pp. 977–990 (2011).
[7] M. Sahami and T. D. Heilman: “A web-based kernel function for measuring the similarity of short text snippets”, Proc. of WWW 2006, pp. 377–386 (2006).
[8] H.-H. Chen, M.-S. Lin and Y.-C. Wei: “Novel association measures using web search with double checking”, Proc. of ACL-44, pp. 1009–1016 (2006).
[9] M. Strube and S. P. Ponzetto: “Wikirelate! computing semantic relatedness using wikipedia”, Proc. of AAAI 2006, pp. 1419–1424 (2006).
[10] E. Gabrilovich and S. Markovitch: “Computing semantic relatedness using wikipedia-based explicit semantic analysis”, Proc. of IJCAI 2007, pp. 1606–1611 (2007).
[11] D. Milne and I. Witten: “An effective, low-cost measure of semantic relatedness obtained from wikipedia links”, Proc. of WIKIAI 2008, pp. 25–30 (2008).
[12] B. Liu, Y. Ma and P. S. Yu: “Discovering unexpected information from your competitors’ web sites”, Proc. of ACM SIGKDD 2001, pp. 144–153 (2001).
[13] Y. Mejova, I. Bordino, M. Lalmas and A. Gionis: “Searching for interestingness in wikipedia and yahoo!: answers”, Proc. of WWW 2013, pp. 145–146 (2013).
[14] A. Nadamoto, E. Aramaki, T. Abekawa and Y. Murakami: “Content hole search in community-type content”, Proc. of WWW 2009, pp. 1223–1224 (2009).
[15] Y. Noda, Y. Kiyota and H. Nakagawa: “Discovering serendipitous information from wikipedia by using its network structure”, Proc. of ICWSM 2010, pp. 299–302 (2010).
[16] 野田武史, 大島裕明, 小山聡, 田島敬史, 田中克己：“主題語からの話題語自動抽出とこれに基づく web 情報検索”, 日本データベース学会論文誌 DBSJ Letters, **5**, 2, pp. 69–72 (2006).
[17] 佃洸摂, 大島裕明, 山本光穂, 岩崎弘利, 田中克己：“語の認知度と語間の関係の非典型度に基づく wikipedia からの意外な情報の発見”, p. to appear (2013).
[18] 大島裕明, 小山聡, 田中克己：“Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見”, 情報処理学会論文誌. データベース, **47**, 19, pp. 98–112 (2006).
[19] 佃洸摂, 大島裕明, 田中克己：“上位下位概念辞書を用いた同位語・上位語のランキング手法の提案”, Web とデータベースに関するフォーラム (WebDB Forum 2013) (2013).
[20] T. H. Haveliwala: “Topic-sensitive pagerank”, Proc. of WWW 2002, pp. 517–526 (2002).