

属性のマッチングに基づくオブジェクトの相性検索

佃 洗撰^{†,††} 大島 裕明[†] 山本 光穂^{†††} 近藤 賢志^{†††} 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

^{†††} 株式会社デンソーアイティラボラトリ 〒150-0002 東京都渋谷区渋谷 2-15-1 渋谷クロスタワー 28F

E-mail: [†]{tsukuad,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ^{††}{miyamamoto,skondo}@d-itlab.co.jp

あらまし 本稿では、オブジェクトの属性の相性に基づくオブジェクト集合検索を提案する。我々はオブジェクト間の相性を測るためにある観点における属性の“類似性”と“バランス性”という考えを導入する。我々が対象とするクエリは“3つの観光地からなる観光コース”などである。本論文で提案するオブジェクト集合検索により、このようなクエリに対して“三十三間堂、大覚寺、大徳寺はいずれも本殿の様式という観点では入母屋造である”（類似性に基づく検索）や“壬生寺、南禅寺、大仙院はそれぞれ東洋の美術家という観点では長谷川等伯、狩野探幽、狩野元信と関連がある”（バランス性に基づく検索）といった検索が可能となる。

キーワード オブジェクト集合検索, 相性検索, 知識抽出

1. はじめに

Web 検索において、ユーザが検索する対象にはさまざまなものがある。“京都大学 HP”というクエリで検索するユーザが探しているのは単一の Web ページである。“脳梗塞”というクエリで検索するユーザが探しているのは脳梗塞に関する原因や症状、治療法など多岐に渡る場合もある。その場合、単一の Web ページで十分な情報が得られなければ、複数の Web ページを閲覧することになり、結果的にユーザは Web ページ集合を検索していると言える。“iPad Retina”というクエリで検索するユーザが探しているのは、“iPad Retina”という単一のオブジェクトおよびそれに関する情報である。また、“京都 観光コース”というクエリで検索するユーザが探しているのは、“金閣寺”や“清水寺”といった実世界オブジェクトを要素とするオブジェクト集合である。オブジェクト集合を検索する際のユーザの検索意図は様々であるが、例えば以下のような検索意図があげられる。

(1) あるユーザが京都に観光に行こうとしており、3箇所の観光地を巡る観光コースを知りたい。

(2) あるユーザが夕食の献立を決めようとしており、4つの料理からなる献立を知りたい。

(3) あるユーザが京都に観光に行こうとしており、金閣寺と東寺に行くことは決めている。しかし、物足りなさを感じるのあと2箇所観光地を追加して1つの観光コースにするとしたらどの観光地を追加すれば良いか知りたい。

(4) あるユーザが休日に読む小説を購入しようとしており、2冊は購入するものが決まった。もう1冊読む時間がありそうなので、あと1冊追加するとしたらどの小説が良いか知りたい。

(5) あるユーザが京都に観光に行こうとしており、金閣寺、清水寺、三十三間堂、醍醐寺に行きたいと考えている。しかし、

すべての観光地に行く時間はないので、1箇所削除するならどれが良いか知りたい。

(6) あるユーザがクラシック音楽に興味をもち、ある5枚のCDを購入しようとしている。しかし、すべてのCDを購入するのは費用が高すぎるので、5枚の中から2枚削除するならどれが良いか知りたい。

上記以外にも多数の状況が考えられるが、一般的な検索エンジンを利用して上記のような検索を行うのは困難である。また、例えば観光コースを Web 上の情報から検索する場合、Web 上に記載されている観光コースしか検索の対象にならない。

これまでも、単一オブジェクトの検索を目的とした研究は多数行われてきた [1] [2] [3]。それらの研究では、オブジェクトとその属性を精度良く Web から抽出することを目的としているものが多い。オブジェクト集合検索の場合も、オブジェクト集合の構成要素となる各オブジェクトと、各オブジェクトの属性を求める必要がある。これに加えて、オブジェクト集合検索の場合、オブジェクトの組み合わせについて考慮しなければならない。例 (1) の場合、京都の観光地が全部で n 個あったとすると、3つの観光地からなる観光コースの数は、観光地の順序を考慮しないとしても全部で nC_3 個ある。例 (3) の場合も、追加の候補となる2つのオブジェクトの組み合わせの数は膨大である。考えられる全組み合わせを検索結果としてユーザに提示するのはユーザの検索の負担が大きくなり適切でない。

そこで本研究では、オブジェクトとその属性が既知の場合において、集合としての“相性の良さ”を考慮したオブジェクト集合検索を目的とし、そのための手法を提案する。オブジェクト間の相性の良さを決める要因には様々なものがあるが、我々はオブジェクトのある観点における“類似性”と“バランス性”に着目する。まず“類似性”について説明すると、例えば例 (1) において、“金閣寺”、“龍安寺”、“南禅寺”からなる観光コー

スは、“時代”という観点から見るといずれも“室町時代”に建設された建物である。“室町時代に関連のある観光コース”としてこの観光コースをユーザに提示することは、ユーザが観光コースを決める上で有益な情報と言える。“バランス性”については、例（１）において、“東寺”，“仁和寺”，“広隆寺”からなる観光コースは，“仏像”という観点から見るとそれぞれ“明王像”，“如来像”，“菩薩像”の種類の仏像がある。“仏像の様々な種類に関連のある観光コース”としてこの観光コースをユーザに提示することも、ユーザにとって有益な情報である。以上のように、ある観点において共通の属性をもつオブジェクト集合は“類似性”によって相性が良いとし、ある観点において互いに異なる属性をもつオブジェクト集合は“バランス性”によって相性が良いとする。オブジェクト集合の追加および削除についても、追加・削除後のオブジェクト集合の相性の良さに基づいて追加・削除するオブジェクト集合を決定する。“類似性”および“バランス性”については3.章で詳しく述べる。

我々はドメインとして“京都市の重要文化財”を対象に実験を行った。実験では、例（１）や（２）のようにオブジェクト集合の大きさを指定した検索と、例（３）～（６）のようにあるオブジェクト集合に対するオブジェクト集合の追加および削除による検索を行い、検索結果の考察を行った。

本稿の構成は以下の通りである。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集合における“類似性”と“バランス性”という観点から本研究で解くべき問題を定義する。4.章では3.章で定義した問題を解くための手法を説明する。5.章で実験について述べ、6.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 相性に関する研究

心理学の分野では、対人関係における相性の良し悪しに関する研究が行われている[4][5][6][7][8]。これらの研究では、主に2つの観点から相性を規定している。1つ目は“類似性”であり、価値観や態度、性格などの相互の性質が類似しているほど、相互の対人魅力が高まるというものである。Byrne[4]をはじめとする対人魅力研究において、類似性の重要性が明らかにされている。2つ目は“相補性”であり、一方の性質が他方の足りない性質を補う場合に、相互の対人魅力が高まるというものである。相補性に関する研究として、中里ら[7]は内向型の人物は内向型の人物よりも外向型の人物がより魅力的であると感じるということを明らかにしている。また、Winch[8]は、配偶者選択の際には、類似性以上に相補性が重要であるとしている。これらの研究で述べられている“類似性”と“相補性”はそれぞれ、我々の研究における“類似性”と“バランス性”に近いものといえる。

Kakeeら[9][10]は、アルコール飲料と料理との相性に関する研究を行った。彼らは、ビール、日本酒、ワイン、チューハイの4種類のアルコール飲料と様々な料理との相性についてアンケートを実施した。アンケートの結果の主成分分析を行ったところ、料理について“おかず軸”と“あぶらっこさ軸”を考慮

すれば、各飲料との相性の良さを説明できることを明らかにした[9]。また、特にビールと料理の成分値との関係を調べると、脂質、あぶらっこさとは有意に正の相関があり、糖質とは有意に負の相関があることが明らかになった[10]。このように、オブジェクト間の相性を考慮する際には、成分などのような属性に着目することは重要であると考えられる。

2.2 属性の組み合わせに関する研究

Support Vector Machine (SVM) [11]では、Kernel関数を用いることで複数の素性の組み合わせを考慮した学習が可能となる。小山ら[12]はデータ間の関係予測のために、異なる属性の組み合わせに対する重みを学習し予測に用いる手法を提案している。彼らはカーネル関数のテンソル積を用いたカーネルトリックにより、効率的な学習を可能としている。彼らの手法では、SVMを用いて学習しているため、正例と負例を用意する必要がある。オブジェクト間の相性予測のために教師データを用意する場合、正例（相性が良いオブジェクトの組）を用意するのは容易であるが、負例（相性が悪いオブジェクトの組）を十分に用意するのは困難であると考えられる。

2.3 オブジェクト検索に関する研究

Nieらは、ユーザがデジタルカメラの商品名のように、オブジェクト名のクエリを入力したときにWebページを検索結果として返すのでは検索精度が十分ではないため、オブジェクトレベルでの検索を実現する手法を提案した[2][3]。彼らは、Webページの構造に着目し、大量のWebページからまずオブジェクトの属性の候補を抽出する。その後各オブジェクトの属性としての確からしさを求め、確度の高いものをオブジェクトの属性とする。本研究では、我々はWikipedia^(注1)の記事からオブジェクトの属性を抽出するが、彼らの手法を用いてオブジェクトの属性を抽出することも考えられる。Nieらは抽出した属性情報をもとにクエリに対してオブジェクトをランキングする手法も提案している[1]。彼らの手法では、オブジェクトが記述されているWebページの重要度と、オブジェクトの属性間の参照関係に基づいてオブジェクトの重要度を求めている。彼らはオブジェクトとして論文を扱っているため、その属性である論文名や会議名、著者名の間には参照関係が存在する。しかし、一般のオブジェクトには必ずしも明確な参照関係は存在しないため、手法の適用範囲は限られる。

Yumotoらはユーザの求める情報の全容を表すWebページ集合を発見する全容検索を提案した[13][14]。ページをオブジェクトとみなすと、彼らの研究はオブジェクト集合の検索といえる。彼らの目的が、クエリに関する情報を網羅できるだけの少ないページ集合を求めることであるのに対して、我々は特定の観点に対するオブジェクト集合内での類似性やバランス性に基づいてオブジェクト集合を求めることを目的としているという違いがある。

3. 問題定義

本章では、まずオブジェクト集合の“類似性”と“バランス

(注1) : <http://ja.wikipedia.org/>

性”について定義し、次にそれらに基づくオブジェクト集合検索問題を定義する。これらを定義するにあたり、あるドメイン D が与えられたとき、以下の項目については既知であるとする。

- ドメイン D に属するオブジェクト集合 $O_D = \{o_1, o_2, \dots, o_n\}$ 。例えばドメインを“京都市の観光地”とすると“金閣寺”や“清水寺”などが O_D の要素になる。 O_D は検索の対象となるオブジェクト集合であり、あるオブジェクト集合に対して追加または削除するオブジェクト集合も O の部分集合である。

- オブジェクト $o_i \in O_D$ の属性集合 A_{o_i} 。例えば、オブジェクト“金閣寺”の属性集合の要素としては“室町時代”や“足利義満”、“右京区”などがあげられる。

- ドメイン D における観点集合 $V_D = \{v_1, v_2, \dots, v_m\}$ 。ドメイン“京都市の観光地”では“時代”や“戦国大名”、“仏教の宗派”などが V_D の要素になる。

- 観点 $v_i \in V_D$ の属性集合 A_{v_i} 。例えば、“時代”という観点の属性集合の要素としては“平安時代”や“室町時代”、“明治時代”などがあげられる。

3.1 類似性に基づくオブジェクト集合の相性

あるオブジェクト集合 $S \subseteq O_D$ が類似性において相性が良いというのを、以下の2つの条件を満たす観点 $v_i \in V_D$ が1つ以上存在することと定義する。

$$\forall o_k \in S, |A_{v_i} \cap A_{o_k}| = 1. \quad (1)$$

$$\left| \bigcup_{o_k \in S} A_{v_i} \cap A_{o_k} \right| = 1. \quad (2)$$

1つ目の条件は、 S 内のいずれのオブジェクトも観点 v_i の属性性を1つもっていることを表す。2つ目の条件は、その属性性が S 内のすべてのオブジェクトで共通していることを表す。

3.2 バランス性に基づくオブジェクト集合の相性

あるオブジェクト集合 $S \subseteq O_D$ がバランス性において相性が良いというのを、以下の2つの条件を満たす観点 $v_i \in V_D$ が1つ以上存在することと定義する。

$$\forall o_k \in S, |A_{v_i} \cap A_{o_k}| = 1. \quad (3)$$

$$\left| \bigcup_{o_k \in S} A_{v_i} \cap A_{o_k} \right| = |S|. \quad (4)$$

1つ目の条件は、 S 内のいずれのオブジェクトも観点 v_i の属性性を1つもっていることを表す。2つ目の条件は、その属性性が S 内のすべてのオブジェクトで異なることを表す。

3.3 類似性とバランス性に基づくオブジェクト集合の検索

3.3.1 オブジェクト集合検索

本研究で対象とするオブジェクト集合検索では、入力としてオブジェクト集合のサイズ k を受け取る。観光コースの検索の場合、ユーザが観光コースに含めたい観光地の数に相当する。オブジェクト集合のサイズ k が与えられたとき、類似性に基づくオブジェクト集合検索は次の条件を満たすオブジェクト集合 S を発見する問題と定義される：式(1)および(2)を満たす観点 $v_i \in V_D$ が1つ以上存在し、かつ $|S| = k$ 。また、バラ

ンス性に基づくオブジェクト集合検索は次の条件を満たすオブジェクト集合 S を発見する問題と定義される：式(3)および(4)を満たす観点 $v_i \in V_D$ が1つ以上存在し、かつ $|S| = k$ 。

3.3.2 追加オブジェクト集合検索

本研究で対象とする追加オブジェクト集合検索では、入力としてオブジェクト集合 T と、追加するオブジェクトの数 k を受け取る。観光コースの検索の場合、ユーザが既に観光に行くことを決めている観光地集合が T に、追加したい観光地の数が k に相当する。この入力に対して、ある観点 $v_i \in V_D$ の類似性に基づくオブジェクト集合検索は次の3つの条件を満たすオブジェクト集合 S を発見する問題と定義される。

$$\forall o_k \in T \cup S, |A_{v_i} \cap A_{o_k}| = 1. \quad (5)$$

$$\left| \bigcup_{o_k \in T \cup S} A_{v_i} \cap A_{o_k} \right| = 1. \quad (6)$$

$$|T \cup S| = |T| + k. \quad (7)$$

式(7)は T と S の要素間に重複がなく、かつ S の要素数が k であることを表す。

また、ある観点 $v_i \in V_D$ のバランス性に基づく追加オブジェクト集合検索は次の3つの条件を満たすオブジェクト集合 S を発見する問題と定義される。

$$\forall o_k \in T \cup S, |A_{v_i} \cap A_{o_k}| = 1. \quad (8)$$

$$\left| \bigcup_{o_k \in T \cup S} A_{v_i} \cap A_{o_k} \right| = |T \cup S|. \quad (9)$$

$$|T \cup S| = |T| + k. \quad (10)$$

3.3.3 削除オブジェクト集合検索

本研究で対象とする削除オブジェクト集合検索では、入力としてオブジェクト集合 T と、削除するオブジェクトの数 k を受け取る。観光コースの検索の場合、ユーザが既に観光に行くことを決めている観光地集合が T に、削除したい観光地の数が k に相当する。この入力に対して、ある観点 $v_i \in V_D$ の類似性に基づく削除オブジェクト集合検索は次の3つの条件を満たすオブジェクト集合 S を発見する問題と定義される。

$$\forall o_k \in T \setminus S, |A_{v_i} \cap A_{o_k}| = 1. \quad (11)$$

$$\left| \bigcup_{o_k \in T \setminus S} A_{v_i} \cap A_{o_k} \right| = |T \setminus S|. \quad (12)$$

$$|T \cap S| = k. \quad (13)$$

式(13)は、 S が T の部分集合であり、 S の要素数が k であることを表す。

また、ある観点 $v_i \in V_D$ のバランス性に基づく削除オブジェクト集合検索は次の3つの条件を満たすオブジェクト集合 S を発見する問題と定義される。

$$\forall o_k \in T \setminus S, |A_{v_i} \cap A_{o_k}| = 1. \quad (14)$$

$$\left| \bigcup_{o_k \in T \setminus S} A_{v_i} \cap A_{o_k} \right| = 1. \quad (15)$$

$$|T \cap S| = k. \quad (16)$$

4. 提案手法

3. 章で定義した各問題を解決するためには, 3. 章で既知としていた O_D , A_{o_i} , V_D , A_{v_i} を求める必要がある. これらを求めることができれば, 各問題に対する解となるオブジェクト集合の集合は一意に決まる. 次節以降では, O_D , A_{o_i} , V_D , A_{v_i} を求める方法について述べる.

4.1 ドメインに属するオブジェクト集合の取得

本研究では, あるドメインに属するオブジェクト集合を取得するために, ALAGIN フォーラムから提供されている上位語階層データ^(注2)を用いる. このデータは, 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, 223,772 個の上位語と 2,751,046 個の下位語から成る. これを用いることで, ある語の下位語集合を求めることが可能となる. 例えば, “京都市の重要文化財” という語 (ドメイン) に対して, “清水寺” や “金閣寺” など合計 168 個の下位語が得られる.

4.2 オブジェクトの属性集合の取得

本研究で我々がオブジェクトの属性を収集する情報源として選択したのはオブジェクト o_i を見出し語とする Wikipedia の記事である. 情報源としては, o_i をクエリとしたときの検索結果なども考えられるが, 一般に o_i をクエリとした Web 検索の結果中には, クエリとは無関係な文章も多数含まれており, そのような文書集合の中から属性を収集するとノイズとなる語も多く含まれてしまう. 一方, Wikipedia の記事には見出し語に関連のある情報のみが記述されているため, ノイズとなる語が比較的含まれにくいという利点がある. これに加えて, Wikipedia の記事中では, Wikipedia に記事が存在する語にはリンクが作成されるが, Wikipedia のポリシーとして, 見出し語と関連のない語についてはリンクは作成しないことになっている. そのため, ある見出し語の記事中でリンクが作成されている語があった場合, その語は見出し語の属性であるとみなせる. 以上のような理由から, 本研究では o_i が見出し語となっている Wikipedia の記事中でリンクが貼られている全ての語をそのオブジェクトの属性語集合として収集する.

4.3 ドメイン内の観点集合の取得

ドメイン内で考慮する観点としての適切さを考えると, 例えば O_D 内の複数のオブジェクトが “鎌倉時代” や “江戸時代”, “明治時代” のような “時代” という観点に関する属性をもっているのであれば, その観点はオブジェクト集合の類似性やバランス性を考慮するうえで有用といえる. 一方で, O_D のいずれのオブジェクトも “冷蔵庫” や “電子レンジ” のような “電化製品” という観点に関する属性をもっていない場合, そのような観点は有用ではない. つまり, ドメイン D における観点集合は, O_D 内の全オブジェクトの属性に応じて決めるべきである. 本研究では, 大きく 2 つの段階に分けてドメイン内の観点集合を求める. まず, O_D 内の全てのオブジェクトの全属性集合, すなわち $\bigcup_{o_i \in O_D} A_{o_i}$ に対してクラスタリングを行う. 次に, 各

クラスタの名称を求める. 次項以降で, それぞれの手法について述べる.

4.3.1 属性集合のクラスタリング

O_D 内の全てのオブジェクトの全属性集合 $\bigcup_{o_i \in O_D} A_{o_i}$ を A_D とすると, A_D 内の属性のクラスタリングを行うためには, まず任意の 2 つの属性 $t_i \in A_D$ と $t_j \in A_D$ の距離を定義する必要がある. 本研究では, 属性間の距離を次式により定義する.

$$d(t_i, t_j) = \frac{1}{|H_{t_i} \cap H_{t_j}|}. \quad (17)$$

ここで, H_{t_i} は 4.1 節で述べた上位語階層データにおける t_i の上位語集合である. つまり, 2 つの属性がより多くの上位語を共有しているほど, その 2 つの属性は類似しており, 距離は近いと考える. 2 つの属性 t_i と t_j が上位語を共有していない場合, つまり $|H_{t_i} \cap H_{t_j}| = 0$ の場合, 十分大きな値 h を用いて $d(t_i, t_j) = h$ とする.

クラスタリングには最長距離法を用いる. 最長距離法は階層的なクラスタリング手法であり, 2 つのクラスタ C_1 と C_2 間の距離は次式で定義され, この距離が最も近い 2 つのクラスタを逐次的に併合する.

$$D(C_1, C_2) = \max_{t_1 \in C_1, t_2 \in C_2} d(t_1, t_2). \quad (18)$$

階層的クラスタリングでは, クラスタ数が 1 になるまでクラスタリングが行われるため, 本研究では $D(C_1, C_2)$ の最小値が閾値 r より大きくなった時点でクラスタリングをやめるようにする. クラスタリング後, クラスタのサイズが 2 以上のものを集め, これをドメイン D におけるクラスタ集合 C_D とする.

4.3.2 属性集合のラベリング

クラスタ $C_i \in C_D$ のラベルは, C_i に含まれる全ての属性に共通の上位語である必要がある. そのような上位語が複数あるときに, 最も適切なクラスタのラベルの性質について考える. 例えば, C_i に含まれる語が “俵屋宗達”, “狩野探幽”, “長谷川等伯” の 3 語であり, これら全てに共通する上位語が “男性” と “江戸時代の画家” であったとする. 本研究の目的が, オブジェクト集合としての相性の良さを観点とともにユーザに提示することであることを考えると, 例えばあるコースが “男性という観点においてバランス性が満たされている” と提示するよりもなぜなら, “江戸時代の画家という観点においてバランス性が満たされている” と提示する方がユーザにとって有益であると考えられる. なぜなら, “男性” と “江戸時代の画家” に関して, それぞれの下位語の数を比較すると後者の方が少なく, “江戸時代の画家” という観点をもつ語は生起しづらく, “江戸時代の画家” を提示する方がユーザの得られる情報量が多いためである.

以上より, 本研究では次式によりクラスタ C_i のラベル l_i を求める.

$$l_i = \operatorname{argmin}_{t_j \in H_{C_i}} |L_{t_j}|. \quad (19)$$

ここで, $H_{C_i} = \bigcup_{t_j \in C_i} H_{t_j}$, つまり C_i 内の全ての語に共通する上位語集合であり, L_{t_j} は 4.1 節の上位語階層データにより得られる t_j の下位語集合である.

(注2) : <http://nlpwww.nict.go.jp/corpus/>

表 1 “京都市の重要文化財” ドメインのオブジェクトの属性数に関するデータと属性の例.

属性数の 最小値	属性数の 最大値	属性数の 平均値	オブジェクト	属性数	属性の例
1	931	73	東寺	176	千手観音, 土一揆, 京都市, 新潮社, 真言宗, 寛永
			八坂神社	156	東山区, 神社建築, 豊臣秀吉, 古事記, 天照大御神
			京都国立博物館	79	国宝, 雪舟, 文化庁, 古今和歌集, 七条駅, 俵屋宗達

表 2 クラスタリングを停止する閾値に対するクラスタ数とサイズ.

閾値	総クラスタ数	最小クラスタ サイズ	最大クラスタ サイズ	平均クラスタ サイズ
$\frac{1}{2}$	576	2	128	5.7
$\frac{1}{3}$	539	2	58	4.5
$\frac{1}{4}$	436	2	55	4.1
$\frac{1}{5}$	349	2	43	3.8
$\frac{1}{6}$	280	2	43	3.5

4.4 観点の属性集合の取得

観点 $v_i \in V_D$ に対する属性集合は, 4.3 節で得られるクラスタリングの結果をもとに, v_i のラベルが付与されたクラスタ内の属性集合とする.

5. 実 験

提案手法の有効性を確かめるため, “京都の観光コースを決める” というユーザの意図を想定して実験を行った. 既存の上位語階層データを使用するという手法の制約上, ドメイン名は上位語階層データに含まれる上位語に限られる. 本実験では京都の観光地を多く含む “京都市の重要文化財” という語をドメイン名として用いた. 5.1 節では, 4. 章で提案した各手法の結果を示す. 5.2 節では, 3. 章で定義した問題に対する結果を示す.

5.1 ドメインにおけるデータ収集

まず, 4.1 の方法により, ドメイン “京都市の重要文化財” に属するオブジェクト集合を取得する. オブジェクトの総数は 168 であった.

次に, 各オブジェクトの属性に関する情報を表 1 に示す. 属性の数はオブジェクトによって大きく異なり, オブジェクトの一般的な知名度が高いほど Wikipedia の記述量が多く, 属性数も多いという傾向が見られた. 本稿では各オブジェクトの属性をその多寡に関わらず全て用いたが, 属性数の少ないオブジェクトについては Wikipedia 以外の情報源から属性を取得するといったことが今後の課題としてあげられる.

属性集合のクラスタリングは, 4.3.1 項で述べた閾値 r の値によって結果が異なる. r の値を $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$ としたときに得られたクラスタ数に関するデータを表 2 に示す. 各閾値によって得られるクラスタリングの結果を著者が確認したところ, $r = \frac{1}{4}$ のときに最も適切なクラスタに分割されていると判断したため, これ以降では $r = \frac{1}{4}$ におけるクラスタリングの結果を用いる. $r = \frac{1}{4}$ のときに得られる各クラスタのラベルとクラスタに属する属性の例を表 3 に示す. 最もサイズの大きいクラスタは, 日本の大学に関するクラスタであった. これは, 本実験で対象としたドメインの中に “龍谷大学” や “同志社大学” というオブジェクトが含まれており, それらの属性集合に多数の

大学が含まれていたためである. しかし, 観光において大学という観点はあまり重要でないと言える. 今後は, オブジェクトとしてより適切なもののみを用いることでノイズとなるような観点を除くことができると考えている. また, “日本の地方公共団体” という観点およびその属性集合も観光においてはあまり重要でないと言える. “久留米市” や “伊賀市” のような地名は, それぞれは 1 つのオブジェクトのみの属性であったとしても, 複数の地名が現れることで大きなクラスタになりうる. これらの語を除くには, DF 値の低い属性は除くなどの方法が考えられる. 他のクラスタについては, ラベル名やクラスタ内の属性は観光コースを構成するうえで有効であると考えられる.

5.2 オブジェクト集合の検索

5.2.1 要素数に基づくオブジェクト集合検索

まず, 要素数を指定したオブジェクト集合検索について述べる. 手法上は, 要素数に制約はないが, 要素数が増えるにつれて考慮すべきオブジェクトの組み合わせの数は指数的に増加する. 本実験では比較的短時間で結果が得られるように要素数を 2 および 3 とした. 要素数が増加した場合に計算量を低減する方法が今後の課題としてあげられる.

要素数を 2 および 3 としたときの検索結果数を表 4 に示す. 例えば要素数が 2 の場合, オブジェクトの全組み合わせ数が 14,028 個であるのに対して, 類似性の観点をもちオブジェクト集合は 10,993 個存在している. 1 つのオブジェクト集合が複数の観点において類似性を満たす場合もあるため, 表中の “総結果数” は例えば類似性を満たす全オブジェクト集合の全観点の数を表している. この結果を見ると, 類似性を満たすからといって全てのオブジェクト集合を検索結果として提示するには検索結果数が多すぎると言える. この問題を解決するには, 類似性を満たすオブジェクト集合を何らかの尺度に基づいてランキングして提示する必要がある. オブジェクト集合のランキングについては 6. 章で述べる.

検索結果として実際に得られたオブジェクト集合の例を表 5 に示す. 提案手法により, アニメや茶人, 本殿の様式など, 多様な観点から類似性やバランス性に基づく観光コースが検索できていることがわかる. また, 各オブジェクトの属性を Wikipedia から抽出しているため, オブジェクトと属性は高い精度で実際に何らかの関連をもっていた.

検索結果として適切でないものには大きく分けて 2 つの傾向が見られた. 1 つ目は, 属性は適切であるが観点が適切でない場合である. 例えば, 表 5 において要素数が 3 で類似性を満たすオブジェクト集合のひとつに “紅葉” を属性としてもつものがある. “紅葉” は観光において観光地の有用な属性である. しかし, この属性に対する観点は “キャラクター” となってい

表 3 “京都市の重要文化財”ドメインにおけるクラスタのラベルとクラスタに属する属性の例.

ラベル	クラスタサイズ	属性の例
日本の大学	55	同志社大学, 上智大学, 関西大学, 早稲田大学, 帝京大学, 近畿大学, 高知工科大学, 拓殖大学
東洋の美術家	19	俵屋宗達, 雪舟, 尾形光琳, 与謝蕪村, 狩野探幽, 長谷川等伯, 堂本印象, 円山応挙, 呉春
日本の地方公共団体	18	久留米市, 伊賀市, 奈良市, 桑名市, 長野市, 沖縄市, 松山市, 津市, 鳥取市, 秋田市, 佐世保市
仏教系の主な伝統宗教	10	臨済宗, 天台宗, 浄土宗, 真言宗, 浄土真宗, 曹洞宗, 日蓮宗, 律宗, 融通念仏宗, 法相宗
日本の神	6	雷神, 風神, スサノオ, ヤマタノオロチ, ツクヨミ, 八咫鳥
本殿の様式	6	入母屋造, 流造, 祇園造, 権現造, 春日造, 寝殿造
戦国甲斐武田氏の城	2	箕輪城, 高遠城

表 4 オブジェクト集合のサイズを指定した検索において得られた結果数.

サイズ	オブジェクトの全組み合わせ数	相性のタイプ	相性の良い組み合わせ数	総結果数
2	14,028	類似性	10,993	20,686
		バランス性	7,173	11,308
3	776,216	類似性	197,649	232,761
		バランス性	24,608	25,810

る. 上位語階層データの中には, “紅葉” の上位語として “季節現象” のようにより適したものが存在し, こちらを提示する方が好ましい. 本稿でのクラスタへのラベル付けに関する提案手法では, 上位語とドメインの関係を考慮していないため, このような問題が生じたと考えられる. 解決策としては, ドメインと各上位語の関連度をあらかじめ求め, 関連度を考慮して適切な上位語を求めるといった方法がある. 2 つ目は, 観点として “地名”, 属性として “京都市” のように, ドメイン内の多くのオブジェクトがもつ属性が類似性の観点として提示される場合である. この問題の解決策としては, 各属性の DF 値を調べ, DF 値が極端に高い属性はオブジェクト間の相性を考慮するうえで有効でないとして除くといったことが考えられる.

5.2.2 オブジェクト集合の追加・削除に基づく検索

オブジェクト集合に対する追加オブジェクト集合検索および削除オブジェクト集合検索の結果の例をそれぞれ表 6 と表 7 に示す. ここでも, 計算量の問題のため, 元のオブジェクト集合のサイズは追加の場合 2 または 3, 削除の場合 4 または 5 とし, 追加・削除するオブジェクト数は 1 または 2 としている.

観光コースを決めるうえで有効な観点や属性を発見できているものが多かったが, 以下のような課題もある. 1 つ目は, 5.2.1 の考察で述べた課題と同様, ドメイン内の多くのオブジェクトがもつ属性が提示されたり, 観点的ラベルが適切でないという問題である. 2 つ目は, 属性として “摂政” や “関白” のように, 抽象的な語が提示されるという問題である. ユーザの検索の支援という観点からは, 抽象的な属性よりも具体的な属性の方が望ましい. したがって, 語の具体度を考慮した属性や観点的提示といった方法が解決策のひとつとして考えられる.

6. ま と め

本稿では, オブジェクト集合におけるオブジェクトの属性間の類似性とバランス性という観点到に着目し, サイズを指定したオブジェクト集合の検索, オブジェクト集合に対するオブジェ

クト集合の追加および削除の問題を定義した. その問題に対して, オブジェクトの属性間に類似性やバランス性による関係が成立しているかを調べるために, オブジェクトの属性集合のクラスタリングを行い, 各クラスタにラベルを付与する手法を提案した.

また, 定義した各問題に対して得られるオブジェクト集合の有用性を調べるために “京都市の重要文化財” を対象ドメインとして実験を行った. 実験の結果, 各問題に対してユーザがオブジェクト集合を選択するうえで有効と考えられる観点および属性を求められることが明らかになった.

今後の課題として, 大きく以下の 2 つがあげられる.

(1) オブジェクトの属性集合

本稿では, あるオブジェクトが見出し語である Wikipedia の記事中の全てのリンク語をそのオブジェクトの属性集合とした. Wikipedia の記事中のリンク語は, オブジェクト間の相性を測るドメインにおいて有効な属性を十分に網羅できているとは言えない. 例えば, “京都の観光地” というドメインにおいて重要な観点的ひとつとして “桜” がある. しかし, Wikipedia から集めた観光地の属性集合には, “桜” を上位語としてもつような属性はほとんど含まれていなかった.

この問題を解決するためには, Wikipedia 以外の情報源からオブジェクトの属性を収集する必要がある. ある語が与えられたときに, その語に関連する語を Web 上から収集することを目的とした研究は多数行われている [15] [16] [17] [18]. 森ら [15] はコンテキストを考慮した関連語抽出の手法を提案している. したがって, 対象ドメインを表すコンテキストとして適切な語を求めて彼らの手法を用いることで, より適切な属性集合を求められる可能性がある.

(2) オブジェクト集合のランキング

実験の考察でも述べたように, オブジェクト集合の類似性とバランス性を考慮したとしても, 検索の解となるオブジェクト集合の数は非常に多い. そのため, 検索結果をユーザに提示する際はオブジェクト集合をランキングする必要がある. オブジェクト集合をランキングするための尺度としては様々なものが考えられるが, 我々はドメインと観点的間, およびオブジェクトと属性間の関連度に基づくランキングを考えている. 例えば, “京都の観光地” というドメインにおいて “仏教” と “大学” という観点的があったとき, “京都の観光地” の関連の強い “仏教” に対する類似性やバランス性を満たすオブジェクト集合を上位に提示する方が適切であると考えられる. また, 同ドメインに

表 5 オブジェクト集合のサイズを指定した検索において得られた結果の例.

サイズ	相性のタイプ	観点	オブジェクト (属性)
2	類似性	日本の神	地主神社 (スサノオ), 大將軍八神社 (スサノオ)
		劇場版アニメ	仁和寺 (名探偵コナン_迷宮の十字路), 東福寺 (名探偵コナン_迷宮の十字路)
		小倉百人一首の家人	退耕庵 (小野小町), 随心院 (小野小町)
	バランス性	新撰組隊士	壬生寺 (近藤勇), 阿弥陀寺 (斎藤一)
		平安時代の人物	安楽寿院 (空海), 妙法院 (最澄)
		平安時代の日記	清水寺 (更級日記), 毘沙門堂 (明月記)
3	類似性	茶人	南禅寺 (小堀政一), 大徳寺 (小堀政一), 養源院 (小堀政一)
		キャラクター	善峯寺 (紅葉), 常寂光寺 (紅葉), 高山寺 (紅葉)
		本殿の様式	三十三間堂 (入母屋造), 大覚寺 (入母屋造), 大徳寺 (入母屋造)
	バランス性	真言系仏教宗派	仁和寺 (真言宗御室派), 東寺 (東寺真言宗), 大覚寺 (真言宗大覚寺派)
		東洋の美術家	壬生寺 (長谷川等伯), 南禅寺 (狩野探幽), 大仙院 (狩野元信)
		日本の儒学者	二尊院 (伊藤仁斎), 仁和寺 (貝原益軒), 長福寺 (広瀬淡窓)

表 6 オブジェクト集合に対する追加オブジェクト集合の検索結果の例. 太字は追加されたオブジェクト.

元のオブジェクト集合	追加数	タイプ	観点	オブジェクト (属性)
{ 伏見稲荷大社, 東福寺 }	1	バランス性	門	伏見稲荷大社 (鳥居), 東福寺 (三門), 八坂神社 (桜門)
			本殿の様式	伏見稲荷大社 (流造), 東福寺 (入母屋造), 平野神社 (春日造)
{ 鹿苑寺, 三十三間堂 }	2	バランス性	東洋の美術家	鹿苑寺 (伊藤若冲), 三十三間堂 (俵屋宗達), 三千院 (竹内栖鳳), 南禅寺 (狩野探幽)
			仏教系の主な伝統宗教	鹿苑寺 (臨済宗), 三十三間堂 (天台宗), 浄土院 (浄土宗), 即成院 (真言宗)
{ 鹿苑寺, 天龍寺 }	1	類似性	室町時代の人物	鹿苑寺 (後醍醐天皇), 天龍寺 (後醍醐天皇), 南禅寺 (後醍醐天皇)
			宗教・文化人	鹿苑寺 (夢窓疎石), 天龍寺 (夢窓疎石), 法観寺 (夢窓疎石)
{ 清水寺, 三十三間堂 }	2	類似性	観音菩薩霊場	清水寺 (洛陽三十三所観音霊場), 三十三間堂 (洛陽三十三所観音霊場), 仲源寺 (洛陽三十三所観音霊場), 大蓮寺 (洛陽三十三所観音霊場)
			織豊政権の大名	清水寺 (豊臣秀吉), 三十三間堂 (豊臣秀吉), 佛光寺 (豊臣秀吉), 春光院 (豊臣秀吉)
{ 八坂神社, 東福寺, 等持院 }	1	バランス性	織豊政権の大名	八坂神社 (豊臣秀吉), 東福寺 (安田国恵瓊), 等持院 (福島正則), 常寂光寺 (小早川秀秋)
			延喜式による上国	八坂神社 (山城国), 東福寺 (駿河国), 等持院 (下野国), 西芳寺 (丹波国)
{ 三千院, 東寺, 清浄華院 }	2	バランス性	江戸時代の征夷大將軍	三千院 (徳川綱吉), 東寺 (徳川家光), 清浄華院 (徳川吉宗), 青蓮院 (徳川家斉), 常照皇寺 (徳川秀忠)
			仏教系の主な伝統宗教	三千院 (天台宗), 東寺 (真言宗), 清浄華院 (浄土宗), 本圀寺 (日蓮宗), 平等寺 (曹洞宗)
{ 東本願寺, 知恩院, 西本願寺 }	1	類似性	路傍の神仏	東本願寺 (観音菩薩), 知恩院 (観音菩薩), 西本願寺 (観音菩薩), 三十三間堂 (弥勒菩薩)
			思想家	東本願寺 (空也), 知恩院 (空也), 西本願寺 (空也), 六波羅蜜寺 (空也)
{ 天龍寺, 東寺, 聖護院 }	2	類似性	室町時代の人物	天龍寺 (後醍醐天皇), 東寺 (後醍醐天皇), 聖護院 (後醍醐天皇), 佛光寺 (後醍醐天皇), 妙顯寺 (後醍醐天皇)
			記念物	天龍寺 (史跡), 東寺 (史跡), 聖護院 (史跡), 大徳寺 (史跡), 二条城 (史跡)

において“三千院”と“豊国神社”を含むオブジェクト集合に対して1つオブジェクトを追加する際, “戦国武将”という観点において“豊臣秀吉”という属性で類似性を満たすオブジェクトが“高台寺”と“方広寺”の2つあったとする. このような場合, 豊臣秀吉と関連度のより高い方を追加候補として上位に提示する方が適切である. ある2語間の関連の強さを測る研究はこれまでも行われているため[19][20], これらの研究で提案され

た手法を用いて語間の関連度を測ることも考えている.

また, 上記の2つ以外にも, 様々なドメインへの提案手法の適用や, ユーザによる評価実験を行う予定である.

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, 243993) によるものです. ここに記し

表 7 オブジェクト集合に対する削除オブジェクト集合の検索結果の例.

元のオブジェクト集合	削除数	タイプ	観点	オブジェクト (属性)
{ 鹿苑寺, 伏見稲荷大社, 仁和寺, 醍醐寺 }	1	バランス性	日本の映画作品	鹿苑寺 (五番町夕霧), 伏見稲荷大社 (梟の城), 仁和寺 (僕の初恋をキミに捧ぐ)
			東洋の美術家	鹿苑寺 (伊藤若冲), 仁和寺 (尾形光琳), 醍醐寺 (俵屋宗達)
{ 伏見稲荷大社, 東福寺, 御香宮神社, 青蓮院 }	2	バランス性	作庭家	東福寺 (重森三玲), 御香宮神社 (中根金作)
			日本の令外官	東福寺 (摂政), 青蓮院 (関白)
{ 大徳寺, 東本願寺, 知恩院, 西本願寺 }	1	類似性	住僧	東本願寺 (善導), 知恩院 (善導), 西本願寺 (善導)
			仏教	東本願寺 (浄土教), 知恩院 (浄土教), 西本願寺 (浄土教)
{ 大徳寺, 妙心寺, 曼殊院, 相国寺 }	2	類似性	室町幕府の征夷大將軍	大徳寺 (足利義満), 妙心寺 (足利義満)
			保健所政令市	妙心寺 (京都市), 相国寺 (京都市)
{ 三千院, 伏見稲荷大社, 南禅寺, 醍醐寺, 鹿苑寺 }	1	バランス性	東洋の美術家	三千院 (竹内栖鳳), 伏見稲荷大社 (伊藤若冲), 南禅寺 (狩野探幽), 醍醐寺 (俵屋宗達)
			仏教系の主な伝統宗教	三千院 (天台宗), 伏見稲荷大社 (真言宗), 南禅寺 (浄土宗), 鹿苑寺 (臨済宗)
{ 妙満寺, 上品蓮台寺, 本禅寺, 天龍寺, 曼殊院 }	2	バランス性	仏教章	妙満寺 (顕本法華宗), 上品蓮台寺 (真言宗智山派), 本禅寺 (法華宗陣門流)
			日本人	上品蓮台寺 (聖徳太子), 天龍寺 (桓武天皇), 曼殊院 (菅原道真)
{ 大徳寺, 東寺, 聖護院, 西本願寺, 鹿苑寺 }	1	類似性	室町時代の人物	大徳寺 (後醍醐天皇), 東寺 (後醍醐天皇), 聖護院 (後醍醐天皇), 鹿苑寺 (後醍醐天皇)
			国指定等文化財	大徳寺 (重要文化財), 聖護院 (重要文化財), 西本願寺 (重要文化財), 鹿苑寺 (重要文化財)
{ 南禅寺, 天授庵, 金地院, 東福寺, 醍醐寺 }	2	類似性	臨済宗の僧	南禅寺 (崇伝), 天授庵 (崇伝), 金地院 (崇伝)
			駿河国の人物	南禅寺 (円爾), 天授庵 (円爾), 東福寺 (円爾)

て謝意を表します.

文 献

- [1] Z. Nie, Y. Zhang, J.-R. Wen and W.-Y. Ma: “Object-level ranking: bringing order to web objects”, Proceedings of the 14th international conference on World Wide Web, WWW ’05, pp. 567–574 (2005).
- [2] Z. Nie, F. Wu2, J.-R. Wen and W.-Y. Ma: “Extracting objects from the web”, Proceedings of the 22nd International Conference on Data Engineering, ICDE ’06, pp. 123–125 (2006).
- [3] Z. Nie, Y. Ma, S. Shi, J.-R. Wen and W.-Y. Ma: “Web object retrieval”, Proceedings of the 16th international conference on World Wide Web, WWW ’07, pp. 81–90 (2007).
- [4] D. E. Byrne: “The attraction paradigm”, Academic Press (1971).
- [5] K. Endou, I. Yamane and H. Hori: “大学生の結婚に対する意識 (1) : 性格特性の相性観について”, Tsukuba psychological research, **12**, pp. 85–91 (1990).
- [6] 岸本康孝: “Jung のタイプ論からみる大学生カップルの相性についての一考察”, Journal of clinical and educational psychology, **31**, 1, p. 109 (2005).
- [7] 田中国夫: “人格類似性と対人魅力-向性と欲求の次元-”, 心理学研究, **46**, pp. 109–117 (1975).
- [8] R. F. Winch: “Mate-selection; a study of complementary needs”, Harper (1958).
- [9] M. Kakee, E. Imai, M. Kasai and K. Hatae: “Questionnaire-based study on the sensory compatibility between four alcoholic beverages and foods”, Journal of cookery science of Japan, **36**, 3, pp. 200–209 (2003).
- [10] M. Kakee, E. Imai, T. Murakami, M. Kasai and K. Hatae: “Questionnaire-based study on the sensory compatibility between beer and various foods”, Journal of cookery science of Japan, **37**, 1, pp. 2–14 (2004).
- [11] V. Vapkin: “The nature of statistical learning theory”, Springer, 2nd edition (1999).
- [12] 小山聡, クリストファー D.: “異なる例からの素性の組合せを用いたベアワイズ分類器の学習”, 人工知能学会論文誌, **20**, 2, pp. 105–116 (2005).
- [13] T. Yumoto and K. Tanaka: “Finding pertinent page-pairs from web search results”, ICADL, pp. 301–310 (2005).
- [14] T. Yumoto and K. Tanaka: “Page sets as web search answers”, ICADL, pp. 244–253 (2006).
- [15] 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 = Transactions of the Japanese Society for Artificial Intelligence : AI, **20**, pp. 337–345 (2005).
- [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”, 情報処理学会研究報告. データベース・システム研究会報告, **2006**, 78, pp. 305–311 (2006).
- [17] J. Huang, H. Wang, Y. Jia and A. Fuxman: “Link-based hidden attribute discovery for objects on web”, Proceedings of the 14th International Conference on Extending Database Technology, EDBT/ICDT ’11, pp. 473–484 (2011).
- [18] Y. Zhu, G. Yin, X. Li, H. Wang, D. Shi and L. Yuan: “Exploiting attribute redundancy for web entity data extraction”, Proceedings of the 13th international conference on Asia-pacific digital libraries: for cultural heritage, knowledge dissemination, and future creation, ICADL’11, pp. 98–107 (2011).
- [19] 榊剛史, 松尾豊, 内山幸樹, 石塚満: “Web 上の情報を用いた関連語のシソーラス構築について”, 自然言語処理 = Journal of natural language processing, **14**, 2, pp. 3–31 (2007).
- [20] D. T. Bollegala, Y. Matsuo and M. Ishizuka: “Measuring the similarity between implicit semantic relations from the web”, Proceedings of the 18th international conference on World wide web, WWW ’09, pp. 651–660 (2009).