情報処理学会論文誌　データベース　Vol.6　No.5　49–61 (Dec. 2013)

属性値の同一性・相補性に着目したオブジェクト集合検索手法の提案とその観光地データへの適用

佃 洸摂[1,†1,a)]　大島 裕明[1,b)]　加藤 誠[1,c)]　田中 克己[1,d)]

受付日 2013年6月21日, 採録日 2013年10月8日

概要：本稿では，オブジェクトの属性値の組合せに基づくオブジェクト集合の検索手法を提案する．オブジェクト集合検索において，考えられるあらゆるオブジェクトの組合せを検索結果として提示するのはユーザの負担が大きくなり適切でない．たとえば，京都市内の 3 つの観光地から構成される観光地集合を検索する場合に，観光地のあらゆる 3 つ組を提示すると膨大な数になる．そこで提案手法では，オブジェクト集合を構成する各オブジェクトの属性値の組合せに着目し，ある観点における属性値の “同一性” と “相補性” という考えを用いる．そして，集合内のすべてのオブジェクトがある観点において同じ属性値を持つ場合（“同一性”），または互いに異なる属性値を持つ場合（“相補性”）に限り，検索結果として提示する．本稿では特に，同一性や相補性を満たす属性値の多重集合の有用性を測ることに焦点を当て，ドメイン内での属性値の生起確率，属性値の認知度，ドメイン名と属性値の関連度の 3 点から属性値多重集合のスコアを求める．我々は京都市内の観光地を対象オブジェクトとして評価実験を行い，ユーザにとって有用な属性値の性質について考察を行った．

キーワード：オブジェクト集合検索，属性値，同一性，相補性

Object Set Retrieval Based on Identity and Exclusivity of Attribute Values and its Application to Tourist Spot Data

Kosets Tsukuda[1,†1,a)]　Hiroaki Ohshima[1,b)]　Makoto P. Kato[1,c)]　Katsumi Tanaka[1,d)]

Received: June 21, 2013, Accepted: October 8, 2013

Abstract: In this paper, we propose methods for object set retrieval on the basis of the combination of object attribute values. In object set retrieval, it is not feasible to show users all the possible combinations of objects. For example, when a user searches for three tourist spots in Kyoto city, an extremely large number of search results are possible if the system shows all combinations of the three tourist spots. To solve the problem, we focus on the combination of object attribute values and propose the concept of “identity” and “exclusivity” of attribute values. We show object sets in a search result only if all objects in an object set have the same attribute value (identity) or different attribute values (exclusivity) for a given viewpoint. We also propose methods for estimating the usefulness of attribute value multisets which fulfill the identity or exclusivity. The usefulness of an attibute value multiset is calculated using three features: occurrence probability of attribute values in a domain, popularity of attribute values and degree of relationship between a domain name and an attribute value. We conduct an experiment targeting tourist spots in Kyoto city and discuss the characteristics of attribute values which are useful for object set retrieval.

Keywords: object set retrieval, attribute value, identity, exclusivity

[1] 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[†1] 現在，日本学術振興会特別研究員（DC1）
Presently with JSPS Research Fellow (DC1)
[a)] tsukuda@dl.kuis.kyoto-u.ac.jp
[b)] ohshima@dl.kuis.kyoto-u.ac.jp
[c)] kato@dl.kuis.kyoto-u.ac.jp
[d)] tanaka@dl.kuis.kyoto-u.ac.jp

1. はじめに

我々が日常的に行う Web 検索はすべて，あるタスクを達成するための検索であるといわれている [10], [19]．たとえば “iPhone5 評判” というクエリは “iPhone5 に関する評判を記述したページを発見する” というタスクを表しており，“東京大学 HP” というクエリは “東京大学のホームページにたどり着く” というタスクを表している．これ以外にもタスクの種類は様々であるが，本研究で着目する検索タスクは，“来週，京都の観光に行くので，その際に訪れる観光地を 3 カ所決める” や “今日の晩御飯で作る料理を 4 品決める”，“太宰治に興味があるので今度の休日に読む太宰治の小説を 3 冊決める” といったものである．これらの検索タスクの共通点として，検索の対象がオブジェクト集合であるという点があげられる．つまり，それぞれ 1 カ所の観光地，1 品の料理，1 冊の小説を 1 つのオブジェクトとするオブジェクト集合となっている．

上記以外にも様々なオブジェクトを対象としたオブジェクト集合検索があげられるが，オブジェクト集合検索を一般の Web 検索エンジンを用いて行う際には，以下のような問題が考えられる．まず，検索の対象が Web ページであるという問題がある．たとえば京都観光の際に訪れる観光地を 3 カ所決めたいユーザが “京都観光 3 カ所” と入力しても，検索結果として観光地集合ではなく Web ページが提示されるため，ユーザは検索の対象であるオブジェクト集合が記述されているかどうか 1 ページずつクリックして閲覧する必要がある．また，検索された Web ページ中では，必ずしもちょうど 3 つの観光地を含む観光地集合が検索されているとは限らない．次に，たとえば京都の 3 つの観光地に訪れたというブログ記事を発見したとしても，その 3 カ所が選ばれた理由が明確に記述されていることは少ないと考えられるため，検索しているユーザの意思決定に有用であるとはいえない．さらに，たとえオブジェクト集合を検索結果として直接提示できる仕組みがあったとしても，たとえば 3 カ所の京都の観光地からなる集合を検索する場合に，3 カ所の観光地の組合せの数は膨大であり，そのすべてを検索結果として提示するのは現実的でない．

そこで本研究では，上記のような問題を解決したオブジェクト集合検索を実現するための手法を提案する．我々が提案するオブジェクト集合検索では，検索結果として Web ページではなくオブジェクト集合を直接提示する．また，ユーザの意思決定の支援と，組合せ数の多さの解決のために，オブジェクト間の関係を考慮し，ある関係を満たすオブジェクト集合のみをその関係性とともに検索結果として提示する．オブジェクト間の関係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属性値の “同一性” と “相補性” に着目する．まず “同一性” について説明すると，たとえば “金閣寺”，“龍安寺”，“南禅寺” の 3 つからなる観光地集合は，いずれも “室町時代” に建設された建物であり，それぞれが属性値として “室町時代” を持っている．“金閣寺，龍安寺，南禅寺はいずれも室町時代と関連がある” という情報をユーザに提示することは，ユーザが意思決定を行ううえで有益な情報になりうる．“相補性” については，たとえば “豊国神社”，“阿弥陀寺”，“知恩院” からなる観光コースは，“歴史上の武将” という観点から見るとそれぞれ，“豊国神社” は “豊臣秀吉” と，“阿弥陀寺” は “織田信長” と，“知恩院” は “徳川家康” と縁があり，それぞれを属性値として持つ．“豊国神社，阿弥陀寺，知恩院はそれぞれ豊臣秀吉，織田信長，徳川家康と関連がある” という情報は，武将に興味のあるユーザにとっては有益な情報となりうる．以上のように，ある観点において共通の属性値を持つオブジェクト集合は “同一性” による関係を持ち，ある観点において互いに異なる属性値を持つオブジェクト集合は “相補性” による関係を持つと考える．このような性質を考慮することで，“豊国神社，阿弥陀寺，知恩院” のような 3 つの観光地から構成される観光地集合に意味を持たせることができ，“阿弥陀寺” のように比較的知名度の低い観光地であっても，武将に興味があれば訪れるきっかけを作ることができる．同一性と相補性の定義については 4 章で詳しく述べる．提案手法におけるユーザの入力は（ドメイン名，オブジェクト集合の要素数，我々の提案する同一性または相補性のいずれか）の 3 つ組である．入力の例としては，（京都府の観光地，3，相補性）や（太宰治の小説，4，同一性）といったものがあげられる．ただし，提案手法の制約上，本研究でドメイン名として入力できる語は ALAGIN フォーラムから提供されている上位語階層データ[*1]に含まれる上位語に限られる．詳細は 3.1 節で述べる．また，4.1 節および 4.2 節で述べるように，本研究における同一性および相補性は要素数が 2 以上のオブジェクト集合でのみ定義されるため，入力として与えるオブジェクト集合の要素数は 2 以上とする．

提案手法の流れを**図 1** と以下に示す．まず，提案手法の一般的な流れは以下のようになる．

(1) ユーザから入力として（ドメイン名，オブジェクト集合の要素数，我々の提案する同一性または相補性）の 3 つ組を受け取る．
(2) 入力として与えられたドメイン名から，そのドメインに属するオブジェクトを取得する．
(3) 手順 (2) で得られた各オブジェクトの属性値を取得する．
(4) 全オブジェクトの全属性値に対してクラスタリングを行う．このとき，各クラスタが 1 つの観点となる．
(5) 入力として与えられた要素数を持ち，かつ入力として与えられた同一性または相補性を満たすオブジェクト

[*1] http://alaginrc.nict.go.jp/hyponymy/index.html

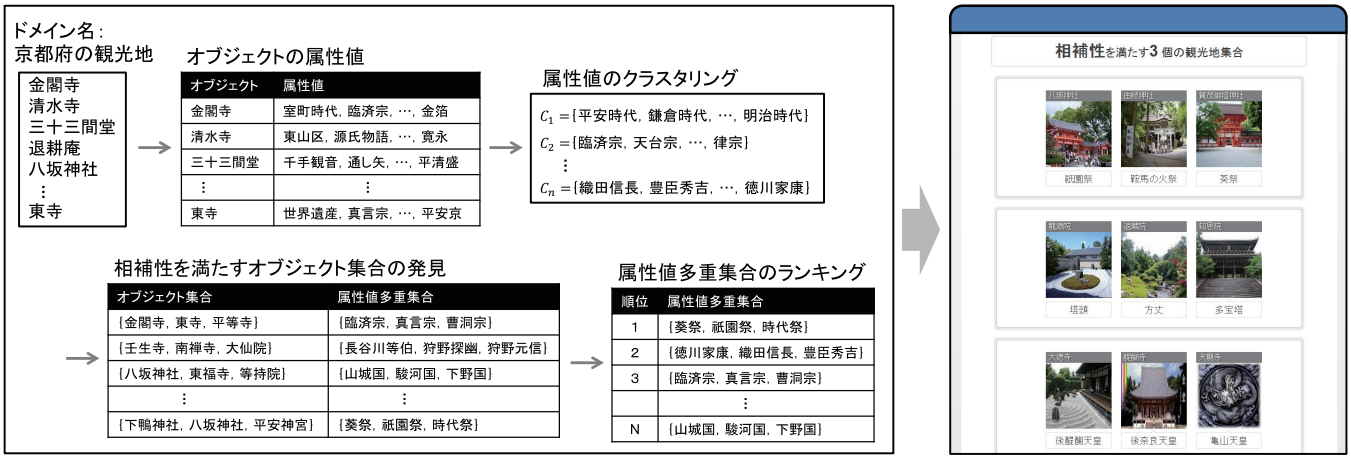

図 1　提案手法の流れとアプリケーション例
Fig. 1　Overview of the proposed methods and an application example.

集合および属性値多重集合をクラスタから抽出する．(6) 手順 (5) で抽出された属性値多重集合をその有用度に応じてランキングを行う．

次に，京都府の観光地集合を例として上記の各手順の説明を行う．ここでは，手順 (1) において，ユーザから入力として（京都府の観光地，3，相補性）の 3 つ組を受け取ったとする．手順 (2) では，京都府内の観光地というドメインに属するオブジェクト（観光地）として “金閣寺” や “京都タワー”，“平安神宮” などを取得する．手順 (3) では，たとえば “金閣寺” の属性値として “臨済宗” や “室町時代”，“金箔” などを取得する．手順 (4) で得られるクラスタの例として，“時代” という観点を表す {平安時代，鎌倉時代，室町時代，江戸時代，明治時代} というクラスタがあげられる．手順 (5) では，入力として与えられた要素数 3 を持ち，かつ入力として与えられた相補性を満たすオブジェクト集合および属性値多重集合をクラスタから抽出する．たとえばオブジェクト集合 {金閣寺，東寺，平等寺} は図 1 中のクラスタ C_2 の属性値として “金閣寺” は “臨済宗” を，“東寺” は “真言宗” を，“平等寺” は “曹洞宗” を持っているため，相補性を満たしている．手順 (6) では，たとえば，{葵祭，祇園祭，時代祭} という属性値多重集合は，京都府で訪れる観光地集合を決めるうえで最も有用な観点であれば有用度は 1 位となり，{山城国，駿河国，下野国} という属性値多重集合は，京都府で訪れる観光地集合を決めるうえで最も有用でない観点であれば有用度は最下位となる．

図 1 中の右図は，本稿の提案手法により実現されるアプリケーションの 1 つであり，たとえば（京都府の観光地，3，相補性）というクエリに対して，属性値集合の有用度が高い順にオブジェクト集合をユーザに提示する．これにより，ユーザは多様な観点からオブジェクト集合検索を行うことが可能となる．

実験では，“京都市内の観光地” を対象オブジェクトとし，クラスタの精度および，属性値多重集合の有用度に基づく分類とランキングの精度を評価した．実験の結果，属性値多重集合の有用度に基づく分類とランキングではドメイン名と属性値の関連度および，属性値の認知度が特に重要であることが明らかになった．

本稿の以降の構成は以下のとおりである．2 章では関連研究について述べる．3 章ではユーザから与えられたドメイン名に対して必要な前処理について述べる．4 章では同一性，相補性，およびそれぞれに基づくオブジェクト集合検索について定義する．5 章では提案手法について述べ，6 章では実験について述べる．7 章では提案手法に基づくアプリケーション例について述べ，8 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 [2], [4], [22], [26], [27]．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 [2] の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，田中ら [27] 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 [22] 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 “同一性” と “相補性” に近いものといえる．

2.2　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 [12], [13]．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値の候補を抽出する．その後各オブジェクトの属性値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値とする．本研究では，我々は Wikipedia [*2] の記事からオブジェクトの属性値を抽出するが，彼らの手法を用いてオブジェクトの属性値を抽出することも考えられる．Nie らは抽出した属性値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 [14]．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値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，その属性値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 [24], [25]．ページをオブジェクトと見なすと，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するできるだけ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同一性や相補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.　事前処理

3.1　ドメイン内のオブジェクトの取得

ユーザから入力として与えられたドメイン内のオブジェクトを取得するために，本研究では ALAGIN フォーラムから提供されている上位語階層データを用いる．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 個の下位語からなる．我々は，上位語階層データに含まれる上位語をドメイン名とし，そのドメイン名を上位語として持つすべての下位語をドメイン内のオブジェクトと見なす．たとえば，“日本の大学” という語をドメイン名として選択すると，“京都大学” や “早稲田大学” などがオブジェクトとして得られる．

本研究では上記の上位語階層データを用いるため，ユーザが入力するドメイン名は上位語階層データに含まれる上位語でなければならないという制約がある．

[*2] http://ja.wikipedia.org/

3.2　オブジェクトの属性値の取得

オブジェクトの属性値を取得するための情報源として，本研究では，オブジェクトが見出し語となっている Wikipedia の記事を用いる．属性値の情報源としては，オブジェクト名で Web 検索をした際の検索結果などもあげられるが，Web 検索の結果中にはクエリとは無関係な情報も含まれうるため，そのような文書集合から属性値を取得するとノイズとなる語も多く含まれてしまうと考えられる．一方，オブジェクトが見出し語となっている Wikipedia の記事では，そのオブジェクトと何らかの観点で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．さらに，オブジェクトとより関連のある語のみを抽出するために，記事中でリンクが張られている語のみを属性値として抽出する．

3.3　属性値のクラスタリング

ドメイン内の全オブジェクトの全属性値が得られたら，次はそれらのクラスタリングを行う．ドメイン D に属する全オブジェクトの全属性値の集合を A_D とし，まず各属性値 $t_i \in A_D$ の上位語を用いて TF-IDF [16] に基づき特徴ベクトル $\boldsymbol{v}_{t_i} = (h_1, h_2, \cdots, h_m)$ を作成する．ここで，m は A_D 内の全属性値の全上位語の数である．ある上位語 h_j に対する “TF” とは t_i がその上位語を持てば 1，持たなければ 0 となるバイナリ値であり，“DF” とは A_D の中でその上位語を持つ属性値の数を表す．属性値 t_1 と t_2 の距離は特徴ベクトル間のコサイン距離により求める．

我々は階層的クラスタリング手法を用いて属性値のクラスタリングを行う．階層的クラスタリング手法では，クラスタリングを停止するための条件が必要となる．最適な階層的クラスタリング手法および停止条件については 6.2 節で述べる．

4.　同一性と相補性に基づくオブジェクト集合検索

本章では，本研究における属性値の “同一性” と “相補性” の定義を述べ，それらに基づくオブジェクト集合検索の問題を定義する．

これらを定義するにあたり，3 章で得られた各データに対する記号を以下のように定義する．

- ドメイン D に属するオブジェクトの集合 O_D．たとえばドメイン名を “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．
- オブジェクト $o_i \in O_D$ の属性値集合 A_{o_i}．たとえば，オブジェクト “金閣寺” の属性値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 におけるクラスタ集合 C_D．$C_D^i \in C_D$ には共通の観点を持つ属性値が含まれる．たとえば，“京都市の観光地” というドメイン名に対して，{“臨済宗”，“天台宗”，“華厳宗”} は “宗派” という共通の観点を持つ属性値集合であり，C_D の要素となる．

4.1　同一性

あるオブジェクト集合 $S \subseteq O_D$ が同一性による関係を持つというのを，S が以下の式 (1) を満たし，かつ式 (2) および式 (3) を満たすクラスタ $C_D^i \in C_D$ が 1 つ以上存在することと定義する．

$$|S| \geq 2. \tag{1}$$

$$\forall o_k \in S, \left|C_D^i \cap A_{o_k}\right| = 1. \tag{2}$$

$$\left|\bigcup_{o_k \in S} C_D^i \cap A_{o_k}\right| = 1. \tag{3}$$

1 つ目の条件は，オブジェクト集合の要素数が 2 以上であることを表す．2 つ目の条件は S 内のいずれのオブジェクトも C_D^i の属性値を 1 つ持つことを表す．3 つ目の条件は，その属性値が 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．

4.2　相補性

あるオブジェクト集合 $S \subseteq O_D$ が相補性による関係を持つというのを，S が以下の式 (4) を満たし，かつ式 (5) および式 (6) を満たすクラスタ $C_D^i \in C_D$ が 1 つ以上存在することと定義する．

$$|S| \geq 2. \tag{4}$$

$$\forall o_k \in S, \left|C_D^i \cap A_{o_k}\right| = 1. \tag{5}$$

$$\left|\bigcup_{o_k \in S} C_D^i \cap A_{o_k}\right| = |S|. \tag{6}$$

1 つ目の条件は，オブジェクト集合の要素数が 2 以上であることを表す．2 つ目の条件は，S 内のいずれのオブジェクトも C_D^i の属性値を 1 つ持つことを表す．3 つ目の条件は，その属性値が S 内のすべてのオブジェクトで異なることを表す．

4.3　同一性と相補性に基づく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，入力の 1 つとしてオブジェクト集合のサイズ k（≥ 2）を受け取る．観光地集合の検索の場合，ユーザが訪れたい観光地の数に相当する．k が与えられたとき，同一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 (2) および式 (3) を満たすクラスタ $C_D^i \in C_D$ が 1 つ以上存在し，かつ $|S| = k$．また，相補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 (5) および式 (6) を満たすクラスタ $C_D^i \in C_D$ が 1 つ以上存在し，かつ $|S| = k$．

5.　属性値多重集合のランキング

本研究では，属性値の認知度，ドメイン名と属性値の関連度，属性値のドメイン内での生起確率の 3 つの特徴量を用いて属性値多重集合の有用度に基づくランキングを行う．属性値多重集合の有用度を求める際は，まず訓練データとして正例（人により有用であると評価された属性値多重集合）と負例（人により有用でないと評価された属性値多重集合）を用意する．その後，各訓練データに対する認知度，関連度，生起確率の 3 つの特徴量のうち，l 個（l は 1 以上 3 以下の整数）のスコアを基に，SVM を用いて分類器を作成する．最後に，有用度を求めたい属性値多重集合を，分類器を作成した際と同じ特徴量を用いて分類器にかけ，正例への所属確率を有用度とする．以下でそれぞれの特徴量におけるスコアの求め方を述べる．

5.1　ドメイン名と属性値の関連度

1 つ目の特徴量は，属性値多重集合内の各属性値のドメイン名との関連度である．たとえば “京都府の観光地” というドメイン名においてある 3 つの観光地がいずれも “祇園祭” という属性値で同一性を満たしており，またある 3 つの観光地がいずれも “伊達政宗” という属性値で同一性を満たしているとする．前者の属性値はドメイン名との関連度は高く，後者の関連度は低いといえる．相補性についても同様に，属性値の多重集合によってドメイン名との関連度は異なる．

属性値多重集合 $A_i = \{t_1, t_2, \cdots, t_N\}$ が同一性または相補性を満たすとき，ドメイン名との関連度に基づく A_i のスコア $f_{rel}(A_i)$ は次式により求められる．

$$f_{rel}(A_i) = \sum_{k=1}^{N} rel(d, t_k)/|A_i|. \tag{7}$$

ここで $rel(d, t_k)$ はドメイン名 d と属性値 t_k の関連度を表す．

本研究では，ドメイン名 d と属性値 t_k の関連度を求めるために WebPMI [1], [9] という指標を用いる．WebPMI では，ドメイン名 d と属性値 t_i の関連度は次式により求められる．

$$rel(d, t_i) = \begin{cases} 0 & \text{if } hits(d, t_i) \leq c \\ \log_2\left(\frac{hits(d \wedge t_i)/N}{(hits(d)/N) \times (hits(t_i)/N)}\right) & \text{otherwise.} \end{cases} \tag{8}$$

本研究では Bollegala ら [1] にならい，$c = 5$，$N = 10^{10}$ と

した．$hit(t_k)$ は属性値 t_k で Web 検索を行った際の検索結果数を表す．検索結果数を取得する方法としては，Yahoo! ウェブ検索 API[*3]がある．しかし，API により取得される検索結果数を見ると，$hits$(“日露戦争”) の 126,000 件に対して $hits$(“京都観光” “日露戦争”) が 215,000 件と，語を追加すると検索結果数が増加する場合が多かった．一方，Yahoo!JAPAN[*4]で検索をした際はそのような問題は起こらなかったため，本研究では Yahoo!JAPAN で検索を行い，検索結果数を人手で取得した．

検索エンジンの検索結果数をもと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 [1], [9]．

5.2　属性値の認知度

2 つ目の特徴量は，属性値多重集合内の各属性値の認知度である．“京都府の観光地” というドメイン名を例とし，ある 3 つの観光地がいずれも “織田信長” という属性値で同一性を満たしており，またある 3 つの観光地がいずれも “佐藤義宣” という属性値で同一性を満たしているとする．前者の属性値の認知度は高く，後者の属性値の認知度は低いといえる．相補性についても同様に，属性値の多重集合によって認知度は異なる．

属性値多重集合 $A_i = \{t_1, t_2, \cdots, t_N\}$ が同一性または相補性を満たすとき，認知度に基づく A_i のスコア $f_{pop}(A_i)$ を次式により求める．

$$f_{pop}(A_i) = \log \sum_{k=1}^{N} hits(t_k)/|A_i|. \tag{9}$$

つまり，Web 検索結果数の多い語ほど認知度が高いと考える．

5.3　属性値のドメイン内での生起確率

3 つ目の特徴量は，属性値多重集合内の各属性値のドメイン内における生起確率に基づくものである．例として “京都市の観光地” というドメイン名における属性値の生起確率について考える．たとえば，ある 3 つの観光地がいずれも “京都市” という属性値を持っており，同一性を満たしているとする．また，ある 3 つの観光地がいずれも “枯山水” という属性値を持っており，同一性を満たしているとする．いずれの属性値もドメインとの関連度は高く，認知度も高いと考えられるが，“3 つの観光地が京都市という属性値で同一性を満たしている” という情報は，このドメインにおいては明らかであり，ユーザが訪れる観光地集合を決めるうえで有用な情報にはなりにくいと考えられる．一方で “3 つの観光地が枯山水という属性値で同一性を満たしている” という情報は，このドメインにおいては自明でなく，ユーザが訪れる観光地集合を決めるうえで有用な情報になりうると考えられる．そこで，これら 2 つの属性値多重集合の有用度を区別するために，属性値の生起確率を考える．京都市に存在する観光地の数は多いため，3 つの観光地から構成される観光地集合が “京都市” という属性値で同一性を満たす確率は高い．一方，京都府の観光地の中で，枯山水に関連のある観光地は多くないため，3 つの観光地から構成される観光地集合が “枯山水” という属性値で同一性を満たす確率は低い．相補性についても同様に，属性値の多重集合によって相補性を満たす確率は異なる．

属性値多重集合 $A_i = \{t_1, t_2, \cdots, t_N\}$ が同一性または相補性を満たすとき，生起確率に基づく A_i のスコア $f_{prob}(A_i)$ は次式により求められる．

$$f_{prob}(A_i) = \log \prod_{k=1}^{N} p(t_k). \tag{10}$$

ここで，$p(t_k)$ は O_D の中で t_k を属性値として持つオブジェクトの割合である．

ただし，同一性と相補性いずれの場合も，$f_{prob}(A_i)$ の値が高いほど，あるいは低いほど有用度は高いといったものではなく，ドメイン名と属性値の関連度および属性値の認知度との組合せにより有用度が決まると考えられる．

6.　実験

提案手法の有用性を検証するために実験を行った．本実験では，京都市の観光地集合を検索するというタスクを想定した．次節以降では，まず本実験で用いるデータセットについて述べ，続いて属性値のクラスタリング結果について述べる．その後，ドメイン名と属性値の関連度および属性値の認知度の評価について述べ，最後に属性値多重集合の分類およびランキングに関する結果を述べる．

6.1　データセット

3.1 節で述べたように，本研究で用いるドメイン名は上位語階層データに含まれる上位語である．本実験では，京都の観光地を下位語として多く持つ “京都市の重要文化財” をドメイン名として用いた．“京都市の重要文化財” は 168 個の下位語を持つが，その中には “同志社大学” や “実隆公記” のように観光地として適切でない語も含まれていた．そこで，168 語の中で観光地として適切な語のみを人手で選択した結果，“京都市の重要文化財” 内のオブジェクトとして 157 語が得られた．本実験ではこの 157 語を対象オブジェクトとして用いた．

続いて，3.2 節で述べた手法により，各オブジェクトの属性値を抽出した．その結果，1 つのオブジェクトあたり

[*3] http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html
[*4] http://www.yahoo.co.jp/

平均で 51.2 個の属性値が，ドメイン内の全属性値数として 3,629 個の属性値が得られた．

6.2　属性値のクラスタリング結果

本節では，対象ドメインにおいて属性値のクラスタリングを行う際のクラスタリング手法とクラスタ数に関する評価を行う．

クラスタの精度を測るために，Wagstaff ら [20] によって提案された評価指標を用いた．彼らの評価指標では，“must-link” および “cannot-link” と呼ばれる 2 種類の制約を用いる．Wagstaff らの論文中で，must-link は “must-link constraints specify that two instances have to be in the same cluster [20]”，cannot-link は “cannot-link constraints specify that two instances cannot be in the same cluster [20]” とそれぞれ定義されている．これらの定義に従い，本実験では次のようにして正解セットを作成した．まず，クラスタリングを行う対象である 3,629 個の全属性値間の距離を計算する．次にその距離に基づいてすべての属性値の組を 20 個のグループに分割する．このとき，n 番目 $(1 \leq n \leq 20)$ のグループには距離が $(n-1) \cdot 0.05$ より大きく $n \cdot 0.05$ 以下の属性値の組が含まれる．最後に，各グループからランダムに 10 組ずつ，計 200 組選択し，各属性値の組が同じクラスタに属するべきか（must-link），異なるクラスタに属するべきか（cannot-link）について第 1 著者がラベル付けを行った．この結果，99 個の must-link と 101 個の cannot-link が得られた．

本実験ではクラスタリング手法として，最短距離法 [18]，最長距離法，重心法，メディアン法 [11]，群平均法 [17]，ウォード法 [21] の 6 つを用い，クラスタリングの停止条件としてクラスタ数に基づく閾値を用いる．最適なクラスタ数を求めるために，クラスタ数を 10 個から 3,000 個まで 10 個ずつ増やし，各手法の各クラスタ数における精度を次のように評価する．まず，クラスタリングの結果に対して must-link の精度（$p_{must\text{-}link}$）と cannot-link の精度（$p_{cannot\text{-}link}$）をそれぞれ計算する．$p_{must\text{-}link}$ は，must-link とラベル付けされた 99 個の属性値の組のうち，実際に同じクラスタに含まれる組の割合を表し，$p_{cannot\text{-}link}$ は cannot-link とラベル付けされた 101 個の属性値の組のうち，実際に異なるクラスタに含まれる組の割合を表す．これらをもとに，クラスタリングの精度（$accuracy$）を次式により求める．

$$accuracy = \frac{2 \cdot p_{must\text{-}link} \cdot p_{cannot\text{-}link}}{p_{must\text{-}link} + p_{cannot\text{-}link}}. \qquad (11)$$

そして，精度が最大となるクラスタリング手法とクラスタ数を用いた結果をこれ以降の実験で用いる．このように，must-link と cannot-link による精度に基づいて最適なクラスタを求める方法は Daniels ら [3] も用いている．

図 2 にクラスタリングの精度の結果を示す．群平均法

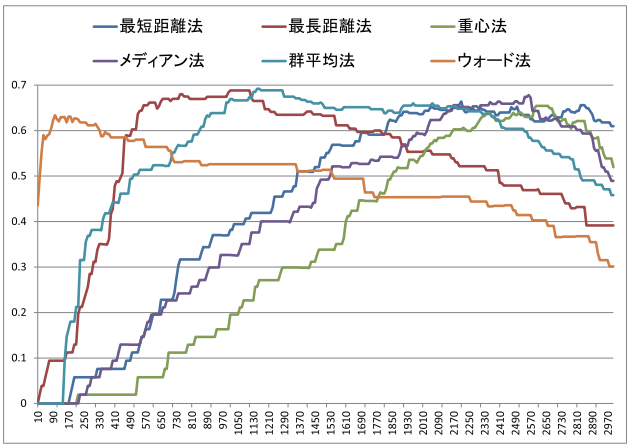

図 2　6 種類のクラスタリング手法の精度
Fig. 2　Clustering performances of six methods.

表 1　6 種類の各手法で精度が最大となったときのクラスタ数，精度，$p_{must\text{-}link}$，$p_{cannot\text{-}link}$
Table 1　Clustering performances of six methods. Only the results of clustering methods with the highest accuracy are shown.

手法	クラスタ数	精度	$p_{must\text{-}link}$	$p_{cannot\text{-}link}$
群平均法	1,150	0.692	0.848	0.584
最長距離法	1,010	0.688	0.657	0.723
メディアン法	2,560	0.678	0.646	0.713
最短距離法	2,210	0.664	0.828	0.554
重心法	2,600	0.654	0.636	0.673
ウォード法	100	0.634	0.667	0.604

を用いてクラスタ数を 1,150 個としたときに精度は最大値 0.692 となった．次に，**表 1** に各クラスタリング手法で精度が最大となったときのクラスタ数，$p_{must\text{-}link}$ および $p_{cannot\text{-}link}$ を示す．6 種類の手法は精度の高い順に並べられている．この結果から，群平均法は must-link をより重視した手法であることが分かる．

以上の結果から，これ以降では群平均法を用いてクラスタ数を 1,150 個としたときの結果を用いる．ただし，本実験ではクラスタサイズが 3 以上のもののみ用いることにした．この結果，最終的に得られたクラスタ数は 353 個であり，353 個のクラスタに含まれる属性値数の合計は 2,603 個であった．

次に，上記の手法によって生成されるクラスタの観点について評価を行った．そのためにまず，上述の 353 個のクラスタをクラスタサイズが大きい順にソートした．続いて，n を 0 以上 88 以下の自然数としたときに，クラスタサイズが $4n+1$ 位から $4n+4$ 位の中からランダムに 1 つずつ，計 89 個のクラスタを評価用にサンプリングした．ただし，$n = 88$ のときは 353 位のクラスタを選択した．89 個のクラスタに対して，第 1 著者と第 3 著者がクラスタに観点を付与した．その際，(1) クラスタ内の全属性値に共通する観点の中で粒度が最も小さいものを観点とする，(2) “人”

表 2　観点が存在すると判断されたクラスタの例
Table 2　Example of cluster which has a viewpoint.

観点	クラスタサイズ	属性値の例
仏教学者	9	那須政隆，坪井俊映，望月信亨，伊藤唯真，中村康隆
茶道の流派	7	武者小路千家，速水流，裏千家，表千家，庸軒流
建築様式	6	祇園造，寝殿造，春日造，入母屋造，権現造，流造
平安時代の事件	3	薬子の変，天慶の乱，承和の変

表 3　観点が存在しないと判断されたクラスタの例
Table 3　Example of cluster which does not have a viewpoint.

クラスタサイズ	属性値の例
4	京狩野，狩野派，花柳流，篠塚流
3	裳階，耐震，鉄筋コンクリート
3	仏師，九相図，仏画
3	堀尾金助，赤松則村，旗本

や “地物” のように他の多くのクラスタにもあてはまる観点は付与しない，という条件のもとで観点を付与した．観点が存在しないと判断した際は “観点なし” のラベルを付与するようにした．観点の付与を終えた後，両者の観点をクラスタごとに照らし合わせ，少なくとも一方が “観点なし” のラベルを付与しているクラスタおよび，両者の付与した観点が大きく異なるクラスタは観点が存在しないものとした．

評価の結果，89 個のクラスタのうち 72 個（80.9%）のクラスタでは観点が存在し，17 個（19.1%）のクラスタでは観点が存在しないという結果が得られた．観点が存在したクラスタと観点の例を**表 2** に，観点が存在しなかったクラスタの例を**表 3** に示す．観点が存在しないと判断されたクラスタの特徴の 1 つ目として，クラスタ内のすべての属性値に共通する適切な粒度の語が存在しないという点があげられる．たとえば，{京狩野，狩野派，花柳流，篠塚流} というクラスタについて考えると，“京狩野” と “狩野派” は絵画に関する流派であり，“花柳流” と “篠塚流” は舞踊に関する流派であるため，同一のクラスタに属するべきではないと考えられる．しかし，これら 4 つの属性値に共通の上位語として “流派” という抽象度の高い語が存在していたため，同一のクラスタに属するという結果になっていた．また，{裳階，耐震，鉄筋コンクリート} というクラスタについて考えると，“裳階” は建物の一部の構造であり，“耐震” と “鉄筋コンクリート” は建物の全体的な構造であるため，同一のクラスタに属するべきではないと考えられる．しかし，これら 3 つの属性値に共通の上位語として “建築構造” という抽象度の高い語が存在したため，同一のクラスタに属するという結果になっていた．

観点が存在しないと判断されたクラスタの特徴の 2 つ目として，本研究で使用した上位語階層データに上位下位関係が適切でない語が含まれているという点があげられる．たとえば，{仏師，九相図，仏画} というクラスタについて考えると，“仏師” という称号と “九相図” という絵画が混在しているため，適切な観点は存在しないと判断された．このクラスタに含まれるいずれの語も “仏教美術” という語を上位語として持っていたため同一のクラスタに属していたが，“仏教美術” は “仏師” の上位語として適切ではなく，関連語と考えた方が適切である．また，{堀尾金助，赤松則村，旗本} というクラスタについて考えると，“旗本” という身分の名称と “赤松則村” という人物名が混在しているため，観点が存在しないと判断された．このクラスタに含まれるいずれの語も “武士” という語を上位語として持っていたため同一のクラスタに属していたが，“武士” は “旗本” の上位語として適切ではなく，関連語と考えた方が適切である．

正解データを作成する際に属性値の組の選び方を工夫したり [8]，特徴空間をゆがめたり [15] することでクラスタの精度を高めることが今後の課題の 1 つとしてあげられる．

6.3　ドメイン名と属性値の関連度および属性値の認知度の評価

本節では，属性値多重集合の分類およびランキングの評価を行う前に，5.1 節と 5.2 節で述べた方法により得られる，ドメイン名と属性値の関連度および属性値の認知度の妥当性について評価を行う．5.3 節の属性値のドメイン内での生起確率はデータセットに依存して決まる値であるため，評価の対象外とした．

以下ではそれぞれの評価用のデータの抽出方法について述べる．属性値の認知度の場合，5.2 節の方法により 2,603 個の全属性値の認知度を求め，認知度の値に基づいて全属性値を 10 分割する．このとき，1 番目のグループには認知度の値の高さが 1 位から 261 位の属性値が含まれ，2 番目のグループには認知度の値の高さが 262 位から 522 位の属性値が含まれる，というように分割する．最後に，各グループから属性値をランダムに 5 個ずつ，合計 50 個の属性値を取得し，ランダムに並べ替えたものを評価用のデータとした．ドメイン名と属性値の関連度の場合も同様にして合計 50 個の属性値を取得し，評価用のデータとした．ただし，6.1 節で述べたように，本実験ではドメイン名として “京都市の重要文化財” を選択したが，観光地集合を検

表 4　ドメイン名と属性値の関連度および属性値の認知度の評価結果
Table 4　Evaluation results for the degree of association between a domain and an attribute value and the popularity of an attribute value.

評価内容	ピアソンの相関係数	κ 係数
ドメイン名と属性値の関連度	0.684	0.579
属性値の認知度	0.391	0.640

索するというタスクにおいてドメイン名と属性値の関連度を測るうえでこのドメイン名は適切でないため，ドメイン名と属性値の関連度を測る際はドメイン名を “京都観光” とした．

評価は 20 代の男性 2 名が行った．ドメイン名と属性値の関連度の評価では関連度を 5 段階で評価してもらい，属性値の認知度についても 5 段階で評価をしてもらった．その際，順序効果を考慮したうえで評価を行った．

結果を**表 4** に示す．評価者間の評価値の一致度を表す quadratic weight による κ 係数 [5] はいずれの場合も 1%の有意水準で一致していた．50 個の属性値の各指標での値と，評価者による評価の相関をピアソンの相関係数を用いて測った結果，ドメイン名と属性値の関連度については 0.684 と高い相関が，属性値の認知度については 0.391 と中程度の相関が得られた．いずれの場合も，ピアソンの相関係数は 1%水準で有意であった．

6.4　属性値多重集合の有用度に基づく分類とランキング

本節では 5 章で述べた 3 つの特徴量から，属性値多重集合の有用度の推定に関する評価を行う．実験の目的は，3 つの各特徴量が属性値多重集合の有用度に与える影響について調べることである．そのために，まず属性値多重集合を “有用である” と “有用でない” の 2 クラスへ分類する精度の評価を行い，次に属性値多重集合の有用度に基づくランキング精度の評価を行う．本実験では，属性値多重集合のサイズを 3 とした．

6.4.1　属性値多重集合のデータセット

本項では評価実験に用いたデータセットの作成方法について述べる．

同一性を満たす属性値多重集合の評価用データセットは以下のようにして作成する．まず，ドメイン内のオブジェクトの 3 つ組について，同一性を満たす属性値多重集合をすべて求める．これにより，たとえば {“三十三間堂”, “大覚寺”, “大徳寺”} は {“入母屋造”, “入母屋造”, “入母屋造”} という属性値多重集合で同一性を満たすといった結果が得られる．本実験で用いたドメインでは，全部で 227 種類のユニークな属性値多重集合が得られた．これを $A_{identity} = \{A_1, A_2 \cdots A_{227}\}$ とする．4.1 節の定義より，$A_{identity}$ の要素はいずれも，C_D 内のいずれかのクラスタで同一性を満たしている．ここで，$A_{identity}$ に含まれる属性値多重集合の中で，$C_D^i \in C_D$ で同一性を満たすものが l 個あるとする．この l 個の属性値多重集合の中から，f_{prob}，f_{pop}，f_{rel} のそれぞれについて最大値，最小値をとるものを評価用の属性値多重集合として抽出した．C_D 内のすべてのクラスタから同様にして抽出した結果，197 個の評価用の属性値多重集合が得られた．

表 5　同一性を満たす属性値多重集合における評価者間の κ 係数
Table 5　Kappa agreement between assessors regarding attribute value multisets which satisfy identity.

	評価者 1	評価者 2	評価者 3
評価者 2	0.388		
評価者 3	0.792	0.460	
評価者 4	0.636	0.404	0.673

表 6　相補性を満たす属性値多重集合における評価者間の κ 係数
Table 6　Kappa agreement between assessors regarding attribute value multisets which satisfy exclusivity.

	評価者 1	評価者 2	評価者 3
評価者 2	0.377		
評価者 3	0.521	0.562	
評価者 4	0.585	0.590	0.656

相補性の評価用データセットについても同様に，ドメイン内のオブジェクトの 3 つ組について，相補性を満たす属性値多重集合をすべて求めた結果，全部で 13,529 種類のユニークな属性値多重集合が得られた．これを $A_{exclusivity} = \{A_1, A_2 \cdots A_{13,529}\}$ とする．同一性の場合に比べて，相補性を満たす属性値多重集合の数が多いため，本実験ではクラスタサイズが 5 以上の各クラスタ $C_D^i \in C_D$ から 1 つずつ属性値多重集合を評価用データとして抽出した．その際，C_D 内のクラスタをサイズの大きい順にソートし，n を自然数としたとき，$6 \cdot (n-1)$ 番目のクラスタからは f_{prob} が最大のものを，$6 \cdot (n-1)+1$ 番目のクラスタからは f_{prob} が最小のもの，$6 \cdot (n-1)+2$ 番目のクラスタからは f_{pop} が最大のもの，$6 \cdot (n-1)+3$ 番目のクラスタからは f_{pop} が最小のもの，$6 \cdot (n-1)+4$ 番目のクラスタからは f_{rel} が最大のもの，$6 \cdot (n-1)+5$ 番目のクラスタからは f_{rel} が最小のものをそれぞれ選択した．この結果，141 個の評価用の属性値多重集合が得られた．

6.4.2　手法

5 章で述べたように，我々は属性値多重集合の有用度を測るために 3 つの特徴量を提案した．属性値のドメイン内での生起確率を **prob**，ドメイン名と属性値の関連度を **rel**，属性値の認知度を **pop** とし，特徴量として 3 つすべてを用いる手法（**prob+rel+pop** 手法），2 つを用いる手法（**prob+rel** 手法，**prob+pop** 手法，**rel+pop** 手法），1 つのみ用いる手法（**prob** 手法，**rel** 手法，**pop** 手法）の計 7 手法を用意した．

表 7　同一性，相補性を満たす属性値多重集合の中で評価者の平均評価値が高かった上位 10 件
Table 7　Top 10 attribute value multisets which satisfy identity and exclusivity.

	同一性		相補性	
順位	属性値多重集合	スコア	属性値多重集合	スコア
1	{織田信長，織田信長，織田信長}	5.0	{祇園祭，葵祭，鞍馬の火祭}	5.0
2	{徳川家光，徳川家光，徳川家光}	5.0	{日蓮，空海，親鸞}	5.0
3	{豊臣秀吉，豊臣秀吉，豊臣秀吉}	5.0	{徳川家康，織田信長，豊臣秀吉}	4.75
4	{源頼朝，源頼朝，源頼朝}	5.0	{大奥，新選組，柳生一族の陰謀}	4.5
5	{聖徳太子，聖徳太子，聖徳太子}	5.0	{徳川家斉，徳川家綱，徳川綱吉}	4.5
6	{古都京都の文化財，古都京都の文化財，古都京都の文化財}	5.0	{日蓮宗，浄土宗，臨済宗}	4.5
7	{枯山水，枯山水，枯山水}	4.75	{清少納言，紫式部，西行}	4.25
8	{特別名勝，特別名勝，特別名勝}	4.75	{八幡大菩薩，弥勒如来，閻魔王}	4.25
9	{新選組，新選組，新選組}	4.75	{入母屋造，春日造，流造}	4.0
10	{百人一首，百人一首，百人一首}	4.75	{在原業平，桓武天皇，神武天皇}	4.0

6.4.3　実験方法

4 名の評価者を用いて実験を行った．4 名の評価者のうち，3 名は 20 代の男性であり 1 名は 20 代の女性である．4 名のうち 2 名は 6.3 節の評価者と同じである．

同一性に関する評価では，“あなたは現在，京都市内のお寺を 3 カ所巡る計画を立てています．その際，3 つのお寺がいずれも以下のものと関連があると知った場合，その情報はあなたがその 3 つのお寺巡りをすると決めるうえでどれほど有用であるか 5 段階で評価してください．(1：有用でない～5：有用である)” という文章を最初に見せ，6.4.1 項で得られた 197 個の各属性値多重集合の有用度を評価してもらった．相補性に関する評価では，“あなたは現在，京都市内のお寺を 3 カ所巡る計画を立てています．その際，3 つのお寺がそれぞれ以下のものと関連があると知った場合，その情報はあなたがその 3 つのお寺巡りをすると決めるうえでどれほど有用であるか 5 段階で評価してください．(1：有用でない～5：有用である)” という文章を最初に見せ，6.4.1 項で得られた 141 個の各属性値多重集合の有用度を評価してもらった．実験は順序効果を考慮したうえで行った．

分類器の構築には，サポートベクタマシン（SVM）を用いた．実際の分類器の構築には，SVM のライブラリである LIBSVM [*5] を使用し，カーネルとして RBF カーネルを用いた．その際，パラメータは LIBSVM の初期設定値である $C = 1$，$\gamma = \frac{1}{k}$（k は入力ベクトルの次元数）を用いた [6] [*6]．

分類性能の評価では，まず属性値多重集合に対する 4 名の評価者の平均評価値が 3 以上のものを有用なクラス，つまり正例とし，3 未満のものを有用でないクラス，つまり負例とした．続いて 5 分割交差検定を行い，各分割に対して適合率を求め，適合率のマクロ平均を求めた．実際の検索においては，有用度の高い属性値多重集合を大量に発見することよりも，有用度の高いものを精度高く発見することが重要であると考えたため，本実験では適合率による評価を行った．

表 8　同一性・相補性それぞれにおける正例と負例の数
Table 8　Number of positive and negative examples for identity and exclusivity.

	正例	負例
同一性	79	118
相補性	45	96

属性値多重集合のランキングの評価では，Mean Average Precision（MAP）および Normalized Discounted Cumulated Gain（nDCG）[7] を用いた．LIBSVM では，クラス分類を行う際に各クラスへの所属確率を求めることができる [23] ため，有用なクラスへの所属確率の高い順に属性値多重集合をランキングし，各指標を求めた．MAP を求める際は，5 分割交差検定の各分割ごとに平均適合率を求め，その平均値を MAP とした．nDCG の場合も同様に，5 分割交差検定の各分割ごとに nDCG を求め，その平均値を 1 つの手法の nDCG とした．

6.4.4　実験結果

まず，評価者間の評価値の一致度を表す quadratic weight による κ 係数 [5] について，同一性に関する結果を**表 5** に，相補性に関する結果を**表 6** に示す．同一性，相補性ともにいずれの評価者間でも 1%の有意水準で評価は一致しており，κ 係数の平均値は同一性で 0.559，相補性で 0.549 と中程度の一致を示した．同一性，相補性を満たす属性値多重集合の中で評価者の平均評価値が高かった上位 10 件は**表 7** のとおりであった．評価者の評価により得られた正例と負例の数を**表 8** に示す．同一性では評価に用いた属性値多重集合のうち 40.1%が，相補性では 31.9%が正例であった．これ以降の実験では，正例と負例の数を合わせるために，同一性の場合は 79 個の負例を，相補性の場合は 45 個の負例をランダムに選択して使用した．

*5　http://www.csie.ntu.edu.tw/~cjlin/libsvm/
*6　パラメータのチューニングを行った場合の結果も求めたが，過学習により精度が低下したため本稿では LIBSVM の初期設定値のパラメータを用いた際の結果を掲載する．

表 9　7 手法の分類の適合率．太字で表された数値は手法間での最大値を表している

Table 9　Classification precision for the seven methods. Highest score between methods is shown in bold.

手法	同一性	相補性
prob+rel+pop	**0.591**	0.624
prob+rel	0.575	0.600
prob+pop	0.429	0.596
rel+pop	0.580	**0.645**
prob	0.138	0.494
rel	0.530	0.605
pop	0.573	0.511

表 10　7 手法の MAP．太字で表された数値は手法間での最大値を表している

Table 10　MAP results for the seven methods. Highest score between methods is shown in bold.

手法	同一性	相補性
prob+rel+pop	0.600	0.599
prob+rel	0.580	0.614
prob+pop	0.570	0.563
rel+pop	0.607	**0.646**
prob	0.505	0.493
rel	0.601	0.608
pop	**0.690**	0.620

次に，6.4.2 項で述べた各手法での分類の適合率を**表 9** に示す．同一性を満たす属性値多重集合の分類では **prob+rel+pop** 手法が最も高い適合率となった．**prob** 手法が他の手法と比べて適合率が低いこと，また **rel+pop** 手法の適合率が全手法の中で 2 番目に高いことから，評価者が同一性を満たす属性値多重集合の有用度を判定する際にはドメイン名と属性値の関連度および，属性値の認知度を重視していることが分かる．相補性を満たす属性値多重集合の分類では **rel+pop** 手法が最も高い適合率となった．同一性の場合と同様に，**prob** 手法の適合率が低いことから，属性値多重集合の有用度にはドメイン名と属性値の関連度および，属性値の認知度の影響が大きいといえる．

最後に，属性値多重集合のランキングの結果について述べる．同一性および相補性に関する MAP の値を**表 10** に示す．また，同一性に関する nDCG の値を**表 11** に，相補性に関する nDCG の値を**表 12** に示す．nDCG は上位 3 件，5 件，10 件までを見たときの値をそれぞれ求めた．同一性に関しては，MAP，nDCG ともに **pop** 手法が最も高い値となり，分類の適合率が最も高かった **prob+rel+pop** 手法は **pop** 手法と比べると MAP，nDCG の値はいずれも低かった．このことから，**pop** 手法では，「有用である」と分類された属性値多重集合のうち，評価者の評価で特に有用度が高かったものを上位にランキングできているといえる．相補性に関しては，nDCG@3 のみ **pop** 手法が最も高い値をとり，その他の指標では **rel+pop** 手法が最も高い値となった．**rel+pop** 手法は相補性を満たす属性値多重集合の分類においても最も精度高く分類できていた手法であり，このことはドメイン名と属性値の関連度および，属性値の認知度の 2 つの特徴量の重要性を表している．

表 11　同一性を満たす属性値多重集合のランキング結果に対する 7 手法の nDCG．太字で表された数値は手法間での最大値を表している

Table 11　nDCG results for the seven methods regarding ranking of attribute value multisets which satisfy identity. Highest scores among methods is shown in bold.

手法	nDCG@3	nDCG@5	nDCG@10
prob+pop+rel	0.628	0.587	0.609
prob+rel	0.607	0.610	0.613
prob+pop	0.594	0.593	0.601
rel+pop	0.628	0.602	0.610
prob	0.568	0.538	0.556
rel	0.654	0.652	0.624
pop	**0.682**	**0.682**	**0.650**

表 12　相補性を満たす属性値多重集合のランキング結果に対する 7 手法の nDCG．太字で表された数値は手法間での最大値を表している

Table 12　nDCG results for the seven methods regarding ranking of attribute value multisets which satisfy exclusivity. Highest score among methods is shown in bold.

手法	nDCG@3	nDCG@5	nDCG@10
prob+pop+rel	0.561	0.595	0.640
prob+rel	0.593	0.629	0.651
prob+pop	0.602	0.557	0.596
rel+pop	0.605	**0.633**	**0.667**
prob	0.550	0.541	0.551
rel	0.592	0.610	0.641
pop	**0.623**	0.630	0.646

7. アプリケーション

本章では，本稿で提案した属性値多重集合の同一性および相補性を考慮することで可能となるオブジェクト集合検索のアプリケーション例として，京都市の観光地集合検索について述べる．アプリケーションを実装する際は 6.4.1 項で述べたデータセットと同じものを用いた．

本アプリケーションでは，ドメイン名として “京都市の重要文化財” がすでに与えられており，ユーザが入力として与えるのは京都市内で訪れたい観光地の数と，我々の提案する同一性または相補性のいずれかである（**図 3**）．4.3 節で述べた定義では，検索するオブジェクト集合のサイズは 2 以上であれば制限はないが，同一性または相補性を満たす属性値多重集合の発見に要する計算量がオブジェクト集合のサイズとともに指数的に増加するため，本アプリケーションでは入力として与えるオブジェクト集合のサイズを

図 3　検索条件の設定例
Fig. 3　Configuration example of retrieval conditions.

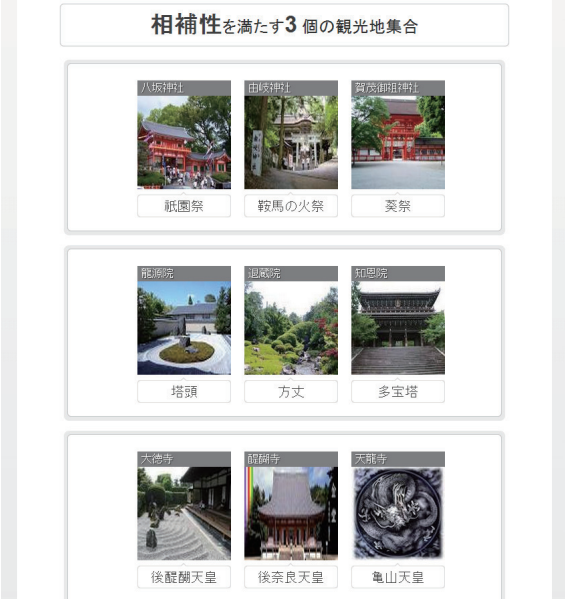

図 4　検索結果の例
Fig. 4　Example of search results.

2 または 3 としている．計算量を抑えるための工夫が今後の課題の 1 つとしてあげられる．

観光地の数を “3”，提示する属性値多重集合の関係を “相補性” としたときの結果を**図 4** に示す．属性値多重集合のランキングをする際は，6.4.4 項の結果に基づき，**rel+pop** 手法を用いている．また，相補性を満たすある属性値多重集合に対応する観光地集合が複数ある場合は，ランダムに 1 つの観光地集合を選択して提示するようにしている．図 4 のように，提案手法を用いることでユーザは多様な観点から観光地集合の検索を行えると考えられる．

8.　まとめと今後の課題

本稿ではオブジェクト集合の検索を実現するために，オブジェクト間の属性値の組合せの “同一性” および “相補性” という考えを導入し，その有用度について属性値多重集合のランキングを行う手法を提案した．実験の結果，同一性を満たす属性値多重集合の有用度を測る際には属性値の認知度が重要であり，相補性を満たす属性値多重集合の有用度を測る際にはドメイン名と属性値の関連度および，属性値の認知度が重要であることが明らかになった．また，属性値多重集合の有用度を測ることで可能となるアプリケーション例として，京都市の観光地集合を検索するシステムを実装した．

今後の主な課題として，次の 2 つがあげられる．1 つ目は，提案手法の他ドメインへの適用である．本稿で提案した手法はドメインに非依存の手法であるため，他ドメインにおいて属性値多重集合の有用度を測るうえで重要な特徴量の違いなどを調べる予定である．2 つ目は，属性値多重集合の有用度を測る際の特徴量を増やすことである．本稿では 3 つの特徴量に着目したが，ランキングの精度を高めるために，アンケートなどを行い，人が属性値多重集合の有用度を測る際に重視する点を明らかにすることで，より有用な特徴量を提案することを考えている．

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）および平成 25 年度研究拠点形成費等補助金若手研究者養成費（卓越した大学院拠点形成支援補助金）によるものです．ここに記して謝意を表します．

参考文献

[1] Bollegala, D., Matsuo, Y. and Ishizuka, M.: A Web Search Engine-Based Approach to Measure Semantic Similarity between Words, *IEEE Trans. Knowledge and Data Engineering*, Vol.23, No.7, pp.977–990 (2011).
[2] Byrne, D.E.: *The Attraction Paradigm*, Academic Press (1971).
[3] Daniels, K. and Giraud-Carrier, C.: Learning the Threshold in Hierarchical Agglomerative Clustering, *Proc. 5th International Conference on Machine Learning and Applications*, *ICMLA '06*, pp.270–278 (2006).
[4] Endou, K., Yamane, I. and Hori, H.：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*, Vol.12, pp.85–91 (1990).
[5] Fleiss, J.L. and Cohen, J.: The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability, *Educational and Psychological Measurement*, Vol.33, pp.613–619 (1973).
[6] Hsu, C.-W., Chang, C.-C. and Lin, C.-J.: A Practical Guide to Support Vector Classification (2003) (online), available from ⟨http://www.csie.ntu.edu.tw/˜cjlin/papers/guide/guide.pdf⟩.
[7] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (2002).
[8] Klein, D., Kamvar, S.D. and Manning, C.D.: From Instance-level Constraints to Space-Level Constraints: Making the Most of Prior Knowledge in Data Clustering, *Proc. 19th International Conference on Machine Learning*, *ICML '02*, pp.307–314 (2002).
[9] Lu, G., Huang, P., He, L., Cu, C. and Li, X.: A New Semantic Similarity Measuring Method Based on Web Search Engines, *W. Trans. Comp.*, Vol.9, No.1, pp.1–10 (2010).
[10] Lucchese, C., Orlando, S., Perego, R., Silvestri, F. and Tolomei, G.: Identifying Task-based Sessions in Search Engine Query Logs, *Proc. 4th ACM International Conference on Web Search and Data Mining*, *WSDM '11*, pp.277–286 (2011).
[11] Mosier, C.T.: An Experiment Investigating the Application of Clustering Procedures and Similarity Coefficients

to the GT Machine Cell Formation Problem, *International Journal of Production Research*, Vol.27, No.10, pp.1811–1835 (1989).
[12] Nie, Z., Ma, Y., Shi, S., Wen, J.-R. and Ma, W.-Y.: Web Object Retrieval, *Proc. 16th International Conference on World Wide Web*, *WWW'07*, pp.81–90 (2007).
[13] Nie, Z., Wu2, F., Wen, J.-R. and Ma, W.-Y.: Extracting Objects from the Web, *Proc. 22nd International Conference on Data Engineering*, *ICDE '06*, pp.123–125 (2006).
[14] Nie, Z., Zhang, Y., Wen, J.-R. and Ma, W.-Y.: Object-level Ranking: Bringing Order to Web Objects, *Proc. 14th International Conference on World Wide Web*, *WWW '05*, pp.567–574 (2005).
[15] Rangapuram, S. and Hein, M.: Constrained 1-Spectral Clustering, *Proc. 15th International Conference on Artificial Intelligence and Statistics*, *AISTATS '12*, pp.1143–1151 (2012).
[16] Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA (1986).
[17] Seifoddini, H. and Wolfe, P.M.: Application of the Similarity Coefficient Method in Group Technology, *IIE Transactions*, Vol.18, No.3, pp.271–277 (1986).
[18] Sneath, P.H.A. and Sokal, R.R.: *Numerical Taxonomy*, Freeman (1973).
[19] Spink, A., Park, M., Jansen, B.J. and Pedersen, J.: Multitasking during Web Search Sessions, *Inf. Process. Manage.*, Vol.42, No.1, pp.264–275 (2006).
[20] Wagstaff, K. and Cardie, C.: Clustering with Instance-level Constraints, *Proc. 17th International Conference on Machine Learning*, pp.1103–1110 (2000).
[21] Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Vol.58, No.301, pp.236–244 (1963).
[22] Winch, R.F.: *Mate-selection; a Study of Complementary Needs*, Harper (1958).
[23] Wu, T.-F., Lin, C.-J. and Weng, R.C.: Probability Estimates for Multi-class Classification by Pairwise Coupling, *J. Mach. Learn. Res.*, Vol.5, pp.975–1005 (2004).
[24] Yumoto, T. and Tanaka, K.: Finding Pertinent Page-pairs from Web Search Results, *Proc. 8th International Conference on Asian Digital Libraries: Implementing Strategies and Sharing Experiences*, *ICADL'05*, pp.301–310 (2005).
[25] Yumoto, T. and Tanaka, K.: Page Sets as Web Search Answers, *Proc. 9th international conference on Asian Digital Libraries: Achievements, Challenges and Opportunities*, *ICADL'06*, pp.244–253 (2006).
[26] 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*, Vol.31, No.1, p.109 (2005).
[27] 田中国夫，中里浩明：人格類似性と対人魅力—向性と欲求の次元，心理学研究，Vol.46, pp.109–117 (1975).

佃 洸摂　（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後期課程在学中．電子情報通信学会学生会員．

大島 裕明　（正会員）

京都大学大学院情報学研究科社会情報学専攻特定准教授．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）．主に情報検索，ウェブマイニング，デザインの研究に従事．電子情報通信学会，日本データベース学会，ACM 各会員．

加藤 誠　（正会員）

京都大学大学院情報学研究科社会情報学専攻特定助教．2012 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）．主に情報検索の研究に従事．日本データベース学会，人工知能学会，ACM 各会員．

田中 克己　（正会員）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院博士前期課程修了．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，ウェブ検索の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会各会員．

（担当編集委員　奥 健太）