# 博物館資料情報の検索のための発見的検索手法

Heuristic Search for Information on Museum's Holdings

山田　篤・安達文夫・小町祐史

YAMADA Atsushi, ADACHI Fumio and KOMACHI Yushi



## ［論文要旨］

本稿では，博物館資料情報の高度な検索の一手法として，既存の資料名称から語を取り出し，それらの間の関連度を用いて，一般的な語彙から専門的な語彙にたどり着く発見的な検索手法を提案し有用性を検証する。

関連度の計算においては，一般の文書に含まれる語の間の関連度を測る方法を適用することができるが，資料名称に含まれるすべての語を対象とするのではなく，その中から主要な語を選定した方がより適切な関連語が得られる。

このようにして得られた関連語を関連度の高いものから順に提示し，関連語を順にたどっていくことで一般的な語彙から専門的な語彙にたどりつくことが可能になる。

関連語の計算においては，資料集合の設定が大きすぎると，何でも関連づけられてしまい，結果として望ましくない関連が発生し，資料集合を細かく分割し過ぎると，関連語としてふさわしい語までが排除されてしまうため，適切な資料集合の設定が課題となる。これに対して，コレクションの木構造を利用した資料集合の構成法を示し，その効果を検証している。

最後に，分野別のコレクションを併せて関連語を探す際には，全般的な傾向としては，語彙数の増加により，到達容易性は低くなるが，異なる分野において共通に出現する語を介して，分野間を横断した検索が可能になることもあることを示している。

【キーワード】資料名称，関連度，到達容易性，木構造

# ❶……はじめに

博物館には数多くの資料が収蔵されている。これを公開することは博物館の役割の 1 つである。この公開の手段として，その画像とともに目録情報をデータベースとして，Web で公開することが行われている。資料に関連する研究者だけではなく，一般の人の利用も可能となっている。

データベースとして公開される資料情報として，資料名称，製作年代，製作または使用された地域などが付与されている。この中で資料を特徴付け，全ての資料に共通して検索の手掛かりとなるのは，資料名称である。しかしながら，資料名称は専門的な語彙が用いられることが多い。このため，一般の人はもちろん，初学者やある資料の周辺分野の研究者では，探し出そうとする資料にたどり着けないことが生じる。一例をあげると，一般の人が「着物」を検索したいと思っても，実際の資料名称としては，たとえば国立歴史民俗博物館（以下，歴博と記す）の所蔵資料である「黒麻地蛇籠桜花文字模様絞染縫帷子」のような文字列が付与されており，帷子という語を知らない限り，この資料を探し出すことはできない。

このような問題に対処するため，普通に使われる語彙から，専門的な語彙で表現された資料にたどり着くことを補助する手法の実現が求められる。博物館資料の名称に関わる同義語や関連語のシソーラスを作り上げることが 1 つの方法であるが，幅広い対象に対する相当の知識と労力を必要とする。そこで，歴博の資料管理において，収集した際の資料のまとまりをコレクションとして扱うことを利用して，語と語の関連度を求め検索補助に適用する方法について検討を進めた。すなわち，1 つのコレクションを構成する資料の名称には互いに関連した語が含まれていることが期待される。この関係を利用して比較的平易な語と専門的な語との結び付きを得ようとするものである。これにより，例えば「着物」→「羽織」→「振袖」→「帷子」とたどり着く可能性が生まれる。

本稿では，博物館資料情報の高度な検索の一手法として，既存の資料名称から語を取り出し，それらの間の関連度を用いて，一般的な語彙から専門的な語彙にたどり着く発見的な検索手法を提案し有用性を検証する。なお，資料名称には，その資料が何であるかを記述するもの資料に多い「名称（name）」と，文献資料や錦絵などの絵画資料に記された内容に基づき付与される「題名（title）」がある。本稿は前者を対象としている。

以下，2 章で本検討の前提となる歴博の資料管理上の構造について記し，3 章で関連度を求める基本的な考え方を述べる。4 章で関連度を用いた発見的な検索手法，5 章で関連度の計算のもととなる資料集合の構成方法，6 章で分野別のコレクションを併せて関連語を探す際の効果と影響について考察する。

# ❷……歴博の資料管理の特徴と館蔵資料データベース

## （1）資料管理と資料番号

歴博は，所蔵する資料の目録を館蔵資料データベースとして公開している。この基本となる資料管理において，資料を歴史（History），考古（Archeology），民俗（Forclor）に区分するとともに，

資料を収集した際のまとまりをコレクションとして扱っている。資料番号として，先頭に H, A, F の記号を，次にコレクションの番号を与えている。個々の資料は，これに枝番を付けて管理する。コレクション中の資料点数が多く，図 1 のように資料を分類できる場合は，資料番号を階層的に与えて管理している。その例を図 2 に示す。

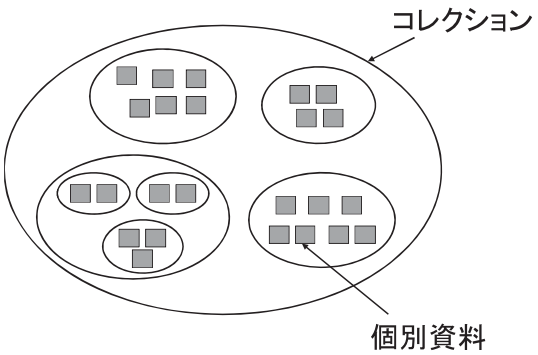



**図 1　歴博の資料管理体系**

H-3　紙本著色洛中洛外図屏風（歴博甲本）
- H-3-1 紙本著色洛中洛外図屏風（歴博甲本）右隻
- H-3-2 紙本著色洛中洛外図屏風（歴博甲本）左隻

（a）　洛中洛外図屏風

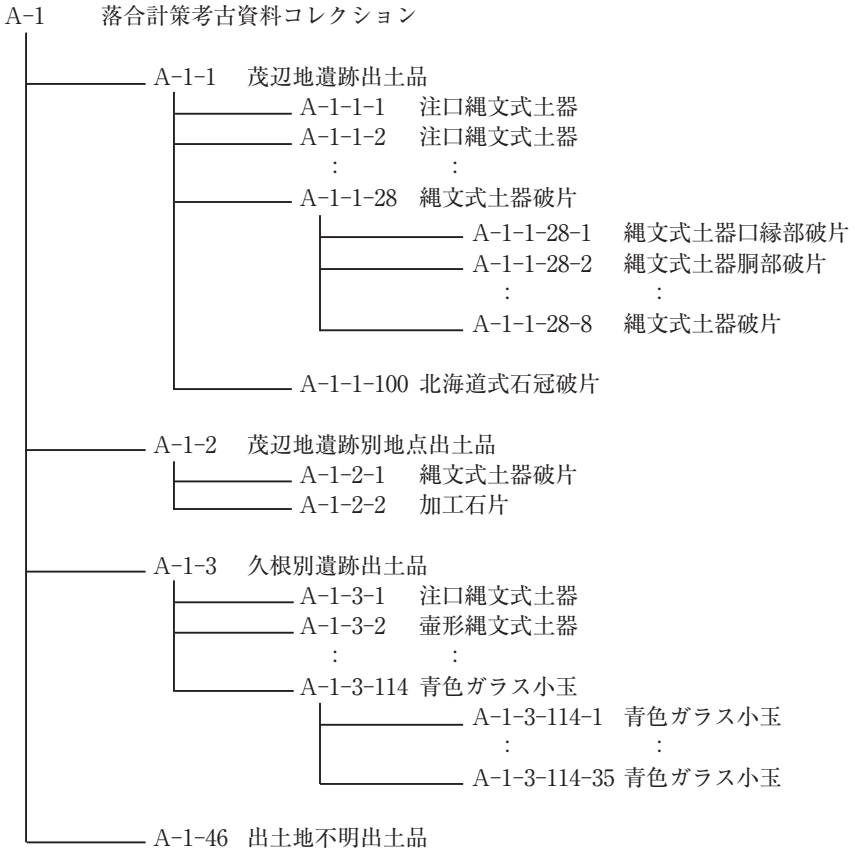
F-8　青森県南地方の仕事着コレクション
- F-8-1　麻地刺子仕事着
  - F-8-1-1　長着
  - F-8-1-53　前掛
  - F-8-1-67　こぎん布
- F-8-2　麻地仕事着
  - F-8-2-1　着物
  - F-8-2-2　短着
  - F-8-2-17　反物
- F-8-6　アイヌ模様衣類
  - F-8-6-1　長着
  - F-8-6-5　あつし
  - F-8-6-6　着物

（b）　仕事着コレクション

A-539 槻の木遺跡出土品
- A-539-1　土器類
  - A-539-1-1　注口土器
  - A-539-1-2　注口土器
  - ：
  - A-539-1-645　土器破片
- A-539-2　石器類
  - A-539-2-1　石篦
  - A-539-2-2　石匙
  - ：
  - A-539-2-2601　軽石製品
- A-539-3　その他
  - A-539-3-1　土製有孔円版
  - A-539-3-2　土偶
  - ：
  - A-539-3-85　土製耳栓

（c）　槻の木遺跡出土品

**図 2　館蔵資料の構管理造（1/2）**




（d）　落合計策考古資料コレクション

**図 2　館蔵資料の構管理造（2/2）**

図 2（a）の洛中洛外図屏風では，右隻と左隻が個々の資料として扱われ，両隻を併せて 1 つのコレクションとしている。同図（b）は仕事着のコレクションで，材質等を基に分類し，資料番号を付与している。図 2（c）では考古の出土品を土器類，石器類，その他で分類し，階層化した資料番号を付与している。（d）は（c）と同じ考古資料であるが，遺跡毎の分類となっている。

コレクションを構成する資料の名称の関連度において，コレクションを単位とするだけではなく，資料間の階層性を利用することも考えられる。

## （2）館蔵資料データベースのデータ項目

館蔵資料データベースのデータ項目とその例を表 1 に示す。多様な種類の資料を記述するため，比較的共通する 12 の項目により構成している。しかし，全ての資料に記載されているのは，資料番号，資料名称と数量，法量くらいである。製作年代や使用地は，検索の手掛かりとなり得るが，不明であったり記載に意味がない場合もあって，記述されないことも多い。このため，全資料に共通して検索の手掛かりとなるのは資料名称だけとなる。

表 1　館蔵資料データベースのデータ項目と実例

| データ項目 | 例 1 | 例 2 |
|---|---|---|
| 資料名称 | クロアサジジャカゴオウカモジモヨウシボリソメヌイカタビラ<br>黒麻地蛇籠桜花文字模様絞染縫帷子 | チュウコウドキ<br>注口土器 |
| コレクション名 | ノムラショウジロウイショウコレクション<br>野村正治郎衣裳コレクション | ツキノキイセキシュツドヒン<br>槻の木遺跡出土品 |
| 指定 | 指定 : 未指定 _ | 指定 : 未指定 _ |
| 数量 | 1 | 1 |
| 法量 | 縦 143.00 cm 横 61.00 cm | 高 10.00 cm |
| 材質 | _ | _ |
| 実物・模造 | 実物 | 実物 |
| 尺度 | _ | _ |
| 製作年代 | AD 時代 : 江戸時代 | BC 時代 : 縄文時代 |
| 使用地 | _ | 青森県東津軽郡平内町　槻の木遺跡 |
| 資料番号 | H-35-117 | A-539-1-1 |
| 備考 | 江戸中期 | 旧番号　土器－ 1　口縁部径 11.1cm<br>底部径 6.6cm　最大径 14.7cm<br>底部・口縁部一部欠損 |

## ❸……資料集合における関連度の計算

### （1）関連度の考え方

コレクションを構成する資料名称に含まれる語の間の関連度を見ることは，既に行われている一般の文書に含まれる語の間の関連度を測る方法を適用できると考えられる。

一般の文書として，たとえば，新聞の社会面，経済面，スポーツ面のそれぞれの記事の特徴を見てみると，どの分野の記事でも共通して用いられる語がある一方で，それぞれの分野で多用される語があり，逆にそれら特徴的な語が出現する記事は，その分野の記事である可能性が高い。これを一般化すると，

（A）　似たような語が出現する文書は，互いに関連する可能性が高い

（B）　関連する文書に出現する語は，互いに関連する可能性が高い

という仮説が成り立つ。

具体的には，文書毎に出現する語を計数して頻度を求め，この情報を基に，互いに関連する文書や関連する語を得ることができる。より具体的には，文書を行，語を列として，文書中の語の頻度付き索引データを行列として構成し，行間，列間の類似度を内積型のメジャーで計算することで，互いに関連する文書や関連する語を得る。

この考え方を博物館資料に適用することを考えると，先の仮説は，

（A’）　似たような語が資料名称に出現する資料のまとまりは，互いに関連する可能性が高い

（B’）　関連する資料のまとまりに出現する語は，互いに関連する可能性が高い

と書き換えることができる。歴博の館蔵資料の場合，コレクションという単位で資料が管理されており，これを資料のまとまり（資料集合と呼ぶ）として捉えることで関連度の計算を行うことが考えられる。

## （2）関連度の計算

具体的には，考古資料 21,935 点の資料名称を対象として，形態素解析器 MeCab，形態素解析用電子化辞書 UniDic を用いて，語への分割を行った。全資料名称を形態素解析した結果における高頻度語を表 2 に示す。

21,935 点の資料名称に対する形態素解析結果における異なる語の数は 2,917 であった。

次に，形態素解析結果のすべての語を用いて，コレクション毎の頻度表を作成し，汎用連想計算エンジン GETA[(1)] を用いて，行と行，列と列の間の類似度を内積型メジャーで計算した。具体的には，ある語に対して，関連するコレクションを計算し，さらにそこから関連する語を計算することで，元の語と関連度が高い語（以下，関連語と呼ぶ）を計算するという手法を採った。関連度の計算には GETA の WT_SMART の設定をそのまま使用した。「土器」という語に対する関連語の計算結果を表 3 に示す。上位に並んでいる「鉢形」「広口」「深鉢」は「鉢形土器」「広口土器」「深鉢土器」

**表 2　形態素解析結果に於ける高頻度語**

| 頻度 | 語 |
|---|---|
| 2935 | 土器 |
| 2537 | 石鏃 |
| 2507 | 複製 |
| 2375 | 瓦 |
| 2015 | 文 |
| 1870 | 石 |
| 1779 | 打製 |
| 1717 | 形 |
| 1511 | 縄文 |
| 1485 | 式 |
| 1323 | 平 |
| 1310 | 軒 |
| 1166 | 出土 |
| 1001 | 器 |
| 988 | 丸 |

**表 3　全ての語を用いた「土器」に対する関連語の計算結果**

| 順位 | 語 | スコア |
|---|---|---|
| 1 | 鉢形 | 4.73 |
| 2 | 広口 | 4.467 |
| 3 | 深鉢 | 4.464 |
| 4 | 土器 | 4.006 |
| 5 | 人面 | 3.422 |
| 6 | 小形 | 3.317 |
| 7 | 浅 | 3.29 |
| 8 | 縄文 | 3.268 |
| 9 | 小型 | 3.189 |
| 10 | 口 | 3.168 |
| 11 | 注 | 3.126 |
| 12 | 把手 | 3.105 |
| 13 | 絵画 | 3.057 |
| 14 | 石偶 | 2.923 |
| 15 | 軽石 | 2.894 |
| 16 | イモ | 2.841 |
| 17 | 手 | 2.797 |
| 18 | づく | 2.733 |
| 19 | 形 | 2.633 |
| 20 | 墨書 | 2.628 |

のような形で単一の資料名称内において「土器」との共起頻度が高いために現れていると考えられる。これらは,「土器」に関する説明を詳細化する目的で,修飾語として「土器」と共起している語群である。

他方,「土偶」や「石鏃」といった語は,「土器」を含むコレクションにおいて高頻度で出現している語であるが,この結果では上位に現れてこない。

## (3) 主要語の選定の効果

形態素解析結果のすべての語を用いた場合には,修飾語等の望ましくない関連語が得られるため,1つの資料名称から1語のみを取り出して,同一コレクション中における共起を用いた計算を再度行ってみた。ここで資料名称から取り出す語は,その資料名称中で最も重要な語(主要語)であることが望ましい。日本語においては原則として先行語が後続語にかかっていくため,最後尾の語を取り出して,実験を行った。先の21,935点の資料名称に対して,形態素解析結果から最後尾の語を取り出したところ,これらの異なる語の数は662となった。

そして,同一コレクションにおけるこれらの語の頻度表を作成し,GETAを用いて,先と同様の

**表4　形態素解析結果の最後尾語を用いた「土器」に対する関連語の計算結果**

| 順位 | 語 | スコア |
|---|---|---|
| 1 | 土器 | 7.335 |
| 2 | 石鏃 | 5.894 |
| 3 | 敲石 | 5.246 |
| 4 | 石斧 | 5.192 |
| 5 | 石 | 4.788 |
| 6 | 破片 | 4.611 |
| 7 | 土偶 | 4.601 |
| 8 | 石器 | 4.578 |
| 9 | 匕 | 4.575 |
| 10 | 錘 | 4.528 |
| 11 | 錐 | 4.348 |
| 12 | 礫 | 4.261 |
| 13 | 器 | 4.041 |
| 14 | 篦 | 3.876 |
| 15 | 石槍 | 3.806 |
| 16 | 品 | 3.759 |
| 17 | 片 | 3.547 |
| 18 | スクレイパー | 3.457 |
| 19 | 軽石 | 3.364 |
| 20 | 石皿 | 3.341 |

**表5　人手で付与した主要語を用いた「土器」に対する関連語の計算結果**

| 順位 | 語 | スコア |
|---|---|---|
| 1 | 鉢 | 8.039 |
| 2 | 深鉢 | 7.864 |
| 3 | 石鏃 | 7.836 |
| 4 | 縄文式土器 | 7.473 |
| 5 | 敲石 | 7.387 |
| 6 | 土器 | 7.209 |
| 7 | 磨石 | 6.912 |
| 8 | 石錘 | 6.606 |
| 9 | 石斧 | 6.533 |
| 10 | 壷 | 6.384 |
| 11 | 尖頭器 | 6.265 |
| 12 | 浅鉢 | 6.239 |
| 13 | 石器 | 6.009 |
| 14 | 礫 | 5.977 |
| 15 | 石匕 | 5.889 |
| 16 | 石錐 | 5.734 |
| 17 | 剥片 | 5.665 |
| 18 | 石皿 | 5.58 |
| 19 | 斧 | 5.568 |
| 20 | 石槍 | 5.365 |

実験を行った。この場合の「土器」という入力語に対する関連語の計算結果を表 4 に示す。

今回上位に現れている語は，「土器」と一緒に出土した等の理由で，「土器」と同じコレクションに含まれていたため，共起頻度が高かった資料名称の主要語である。先の形態素解析結果のすべての語を用いた場合と比べて，意味的に関連度の高い語が並んでいると思われる。

このように主要語を選定することにより，2 万点程度の資料名称を対象にして，意味があると考えられる関連語を取り出すことができた。ただし，現状では形態素解析辞書が対象分野の語彙を必ずしもすべてカバーしていないという問題がある。また，機械的に主要語を選定することが困難な場合もある。このため，考古資料について，人手によって付与した主要語を用いた関連度の計算を次に行った。また，1 つの資料名称に対して，複数の主要語を割り当てることを許した結果，主要語の異なり語彙数は 751 となった。

このとき，「土器」と関連があると計算された語彙は 353 あった。このうち上位 20 位までのものを表 5 に示す。

このように，現状では，資料名称からの主要語抽出は，機械処理よりも，意味を考えて人手で行うほうが，より正確な関連語計算が可能になると考えられる。

## ❹……関連語を用いた発見的検索

関連語の情報を用いることで，入力された検索語彙以外の語に到達する可能性が生まれる。具体的には，ある語と関連する語を候補として提示することで，検索者が興味をもった語に遷移することが考えられる。

そこで，関連語を用いて，どのような遷移が可能になるかを検討するため，考古資料について，3 章で述べた人手により選定した主要語を用いて，すべての主要語の関連語を計算した結果，43,403 組の 2 つ組が得られた。

この 2 つ組をつなぎあわせていくことで，ある語から出発して到達可能なすべての語を計算することができる。たとえば，「土器」の関連語の中に「ハンドアックス」は出現しなかったが，「ハンドアックス」は「土器」の関連語である「石器」の関連語であるので，「土器」→「石器」→「ハンドアックス」という経路が構成される。この結果，「土器」から出発した場合は，751 語の主要語中 665 語へ到達可能となった。これは，土器から出発してどうしても到達できない語が

表 6　「土器」から到達可能な語の例

| 語 | 経路 |
|---|---|
| 瓶子 | 土器 , 壷 , 瓶子 |
| 梵鐘 | 土器 , 土製品 , 梵鐘 |
| 下肢骨 | 土器 , 骨 , 下肢骨 |
| 盥 | 土器 , 皿 , 盥 |
| 硬玉 | 土器 , 管玉 , 硬玉 |
| 鍬先 | 土器 , 木製品 , 鍬先 |
| 短甲 | 土器 , 貝 , 短甲 |
| 足根部 | 土器 , 骨 , 足根部 |
| 尺骨 | 土器 , 骨 , 尺骨 |
| 高台 | 土器 , 皿 , 高台 |
| 寛骨 | 土器 , 骨 , 寛骨 |
| 短剣 | 土器 , 斧 , 剣 , 短剣 |
| 行灯皿 | 土器 , 皿 , 行灯皿 |
| 近世陶磁器 | 土器 , 皿 , 近世陶磁 |
| 燗瓶 | 土器 , 皿 , 燗瓶 |
| グレイバー | 土器 , スクレイパー , グレイバー |
| 扁額 | 土器 , 皿 , 扁額 |
| 膝蓋骨 | 土器 , 骨 , 膝蓋骨 |
| 天秤 | 土器 , 骨角器 , 天秤 |
| 丸碗 | 土器 , 皿 , 丸碗 |

86 語あることを示している。「土器」から到達可能な語の例を表 6 に示す。

また，到達可能な語については，複数の経路が存在することがある。関連語を関連度の高いものから順に提示することを考えると，経路によっても到達容易性に差が生じる。

そこで，ある語に対して，その関連語を関連度の高いものから順に提示することを考える。このとき，先頭の語からはじめて，1 つ下位の語をみる確率を p とし，n 番目の語に到達する確率を，$p^{n-1}$ で表すことにする。たとえば，表 5 に示した「土器」の例の場合は，

鉢 $\xrightarrow{p}$ 深鉢 $\xrightarrow{p}$ 石鏃 $\xrightarrow{p}$ 縄文式土器 $\xrightarrow{p}$ 敲石 $\xrightarrow{p}$ 磨石 $\xrightarrow{p}$ 石錘 $\xrightarrow{p}$ 石斧 $\xrightarrow{p}$ …

となる。

次に，ある語の関連語を見に行く確率を q とする。たとえば，「土器」の関連語について，先頭から「石鏃」まで見たときに，「石鏃」の関連語を見に行く場合，その確率は $p^2q$ となる。ここに p+q<1 とする。この定義により，「土器」から「鉄滓」に至る各経路のスコアを計算することができる。今，リストの 20 番目の語を見る確率と，別の関連語に遷移する確率が同程度という仮定を置くと q=p^20 とすることができる。このとき，p+q=p+p^20<1 を満たす p の値として p=0.89 として計算してみた場合に，上位にくる経路を表 7 に示す。

**表 7 「土器」から「鉄滓」に至る経路の関連語の順位を元にしたスコアによる評価（1）**

| スコア | 経路 | |
|---|---|---|
| 0.000919 | 土器 | 杯 |
| 0.000457 | 土器 | 石槍 |
| 0.000322 | 土器 | 香炉 |
| 0.000287 | 土器 | 杯 |
| 0.000255 | 土器 | 杯 |
| 0.000202 | 土器 | 剥片 |
| 0.000202 | 土器 | 須恵器 |
| 0.00016 | 土器 | 石斧 |
| 0.000127 | 土器 | 杯 |
| 0.000127 | 土器 | 蓋 |
| 0.000127 | 土器 | 須恵器 |
| 0.000113 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |
| 0.0000894 | 土器 | 杯 |

**表 8 「土器」から「鉄滓」に至る経路の関連語の順位を元にしたスコアによる評価（2）**

| スコア | 経路 | |
|---|---|---|
| 0.0461 | 土器 | 杯 |
| 0.0339 | 土器 | 石槍 |
| 0.0290 | 土器 | 香炉 |
| 0.0276 | 土器 | 杯 |
| 0.0262 | 土器 | 杯 |
| 0.0237 | 土器 | 剥片 |
| 0.0237 | 土器 | 須恵器 |
| 0.0213 | 土器 | 石斧 |
| 0.0193 | 土器 | 杯 |
| 0.0193 | 土器 | 蓋 |
| 0.0193 | 土器 | 須恵器 |
| 0.0183 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |
| 0.0165 | 土器 | 杯 |

別の設定として，リストの 60 番目の語を見る確率と，次の関連語を辿る確率を同等と仮定した場合についても同様に，p+q=p+p^60<1 を満たす p の値として p=0.95 として計算してみた場合に，上位にくるものを表 8 に示す。

p=0.95 という設定は，p=0.89 という設定と比べて，順位がより低い語を見に行くことが優位になる設定であるが，上位 20 位までの経路に関しては，順位の変動は見られなかった。このことから，p,q の値の違いが経路の順位に与える影響はそれほど大きくはないと言える。

## ❺……資料集合の構成法

### （1）資料集合による到達容易性の違い

ここまでの実験では，関連語を計算する資料集合の単位をコレクションとしてきたが，実際にはどのように資料集合を設定するかが問題となる。

歴博の館蔵品データはコレクション以下が階層的に構成されており，コレクションはこの中で最上位の単位である。

一般に，資料集合の設定が大きすぎると，何でも関連づけられてしまい，結果として望ましくない関連が発生することがわかっている。逆に，資料集合を細かく分割し過ぎると，関連語としてふさわしい語までが排除されてしまうことになる。

そこで，考古資料を例に取り，21,935 点の資料名称を対象として，資料群の構成方法の違いにより，到達容易性がどのように変化するかについての実験を行った。

どのような結果が望ましいかについては，「石器」,「石斧」,「土器」,「土偶」のそれぞれの語の関連語群を専門家に提示し，それぞれが関連語としてふさわしいか否かを判定してもらった。このうち，専門家が出発語との関連がないと判定した語について，到達容易性が低くなることが望ましいものであるとして評価を行う。

具体的には，専門家によって「土器」の関連語としてはふさわしくないと判定された「石鏃」と，ふさわしいと判定された「高杯」を選び，それぞれについて「土器」からの到達容易性を計算した。

到達容易性の計算方法は 4 章で示したとおり，ある語を入力した場合に，その関連語を関連度の高いものから順に提示することを考え，先頭の語からはじめて，1 つ下位の語をみる確率を p とし，n 番目の語に到達する確率を $p^{n-1}$，ある語の関連語を見に行く確率を q（ただし，p+q<1）として，それぞれの資料集合設定における到達容易性の計算を行った。具体的には p=0.89，q=p^20 として，関連語は第 60 位まで，別の関連語を辿る回数は 4 回までに制限した。

資料集合の構成方法としては，3 章の計算と同じくコレクションをそのまま資料集合とするもの（構成 1），階層構造毎に分割して資料集合とするもの（構成 2），階層構造の中で分割位置を人手で決めて資料集合とするもの（構成 3），階層構造の部分木毎に資料集合とするもの（構成 4）を作成して実験を行った。これらを図 3 に示す。このとき，構成 4 については，階層構造の上位でまとめた資料集合には望ましくない関連が含まれることがあるため，その深さに応じて下位の部分木ほど優遇するようにした。

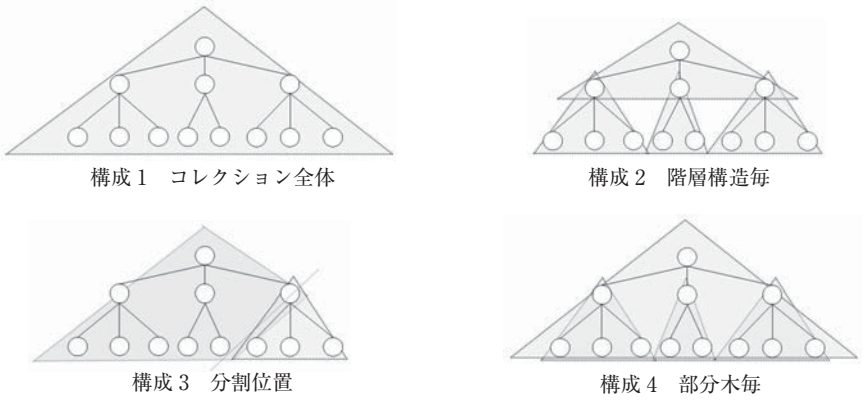



**図 3　資料集合の構成法**

**表 9　資料集合の構成法の比較**

| 構成 | スコア |
|---|---|
| 構成 1 | 0.169 |
| 構成 2 | 0.0471 |
| 構成 3 | 0.0744 |
| 構成 4 | 0.0594 |

(1)「土器」から「石鏃」に至る到達容易性

| 構成 | スコア |
|---|---|
| 構成 1 | 0.000498 |
| 構成 2 | 0.00142 |
| 構成 3 | 0.00203 |
| 構成 4 | 0.00425 |

(2)「土器」から「高坏」に至る到達容易性

こうして,「土器」から「石鏃」に至る到達可能な全経路の計算を行い，全経路のスコアの和を求めたところ，表 9（1）に示す結果となった。この場合，到達容易性の値が小さい方が望ましいため，望ましい順に，

構成 2 > 構成 4 > 構成 3 > 構成 1

となる。

次に，「土器」から「高坏」についても同様の計算を行うと，表 9（2）に示す結果となった。この場合は到達容易性の値が大きい方が望ましいため，望ましい順に，

構成 4 > 構成 3 > 構成 2 > 構成 1

となる。

これらの結果から，資料集合の構成時に階層構造を考慮することについて，一定の効果はあると考えられる。ただし，構成 4 では下位の部分木ほど一律に優遇しているため，浅い階層よりも深い階層が常に重視されることになる。これは階層の深さが一定であれば，それほど問題とはならないが，深い木と浅い木が混在する状況では，浅い木での関連が不利となるため，木構造の深さの違いを考慮した構成法を次に考える。

## (2) 木構造を反映した資料集合の構成法

木構造の深さを反映した資料群の構成法として，木の中間ノード間の枝の数に着目した方法を考える。同一の中間ノードを親に持つ終端ノードの集まりをパス長 0 の集合とする。次に，親となる中間ノード間の枝の数が 1 の終端ノードの集まりをパス長 1 の集合とする。以下，同様に親となる中間ノード間の枝の数に従い，パス長 2 の集合，パス長 3 の集合といったように定める。これらの関係を図示すると，図 4 のようになる。

今, 中間ノードを N, 終端ノードを T, カレントノードの子供の終端ノードを CT とし, makeGroup

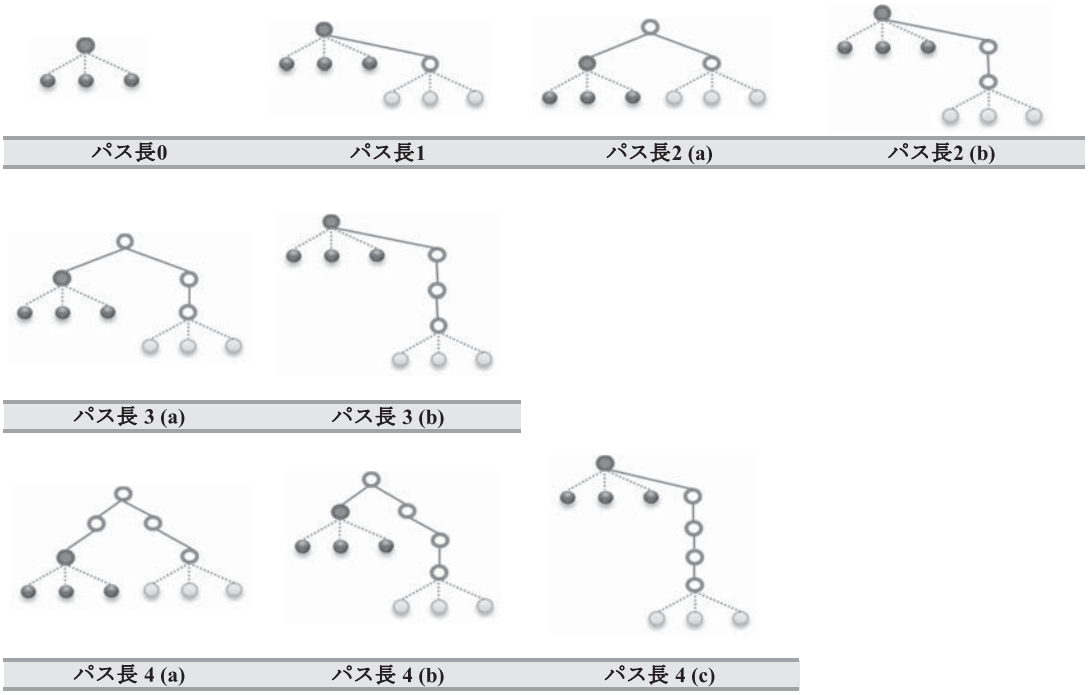



図 4　木構造を反映した資料集合の構成

表 10　木構造を反映した資料集合の構成法

| パス長 | 構成法 |
|---|---|
| 0 | makeGroup（CT） |
| 1 | for-each child::N<br>makeGroup（CT｜child::T） |
| 2 | for-each following-sibling::N<br>makeGroup（CT｜child::T） |
| | for-each child::N/child::N<br>makeGroup（CT｜child::T） |
| 3 | for-each（preceding-sibling::N｜following-sibling::N）/child::N<br>makeGroup（CT｜child::T） |
| | for-each child::N/child::N/child::N<br>makeGroup（CT｜child::T） |
| 4 | for-each parent::N/following-sibling::N/child::N<br>makeGroup（CT｜child::T） |
| | for-each（preceding-sibling::N｜following-sibling::N）/child::N/child::N<br>makeGroup（CT｜child::T） |
| | for-each child::N/child::N/child::N/child::N<br>makeGroup（CT｜child::T） |

**表 11　木構造を反映した資料集合の構成法の比較**

| 構成 | スコア |
|---|---|
| 構成 5（階層構造：パス長 0 のみ） | 0.044 |
| 構成 6（階層構造：パス長 1 まで） | 0.0705 |
| 構成 7（階層構造：パス長 2 まで） | 0.0231 |
| 構成 8（階層構造：パス長 3 まで） | 0.0134 |
| 構成 9（階層構造：パス長 4 まで） | 0.01 |

（1）「土器」から「石鏃」に至る到達容易性

| 構成 | スコア |
|---|---|
| 構成 5（階層構造：パス長 0 のみ） | 0.00155 |
| 構成 6（階層構造：パス長 1 まで） | 0.0000874 |
| 構成 7（階層構造：パス長 2 まで） | 0.000377 |
| 構成 8（階層構造：パス長 3 まで） | 0.0018 |
| 構成 9（階層構造：パス長 4 まで） | 0.000786 |

（2）「土器」から「高坏」に至る到達容易性

を，その引数のノードで資料群を構成する関数とすると，それぞれのパス長の資料群は XPath の記法を用いて表 10 に示す方法で構成することができる。

ルートノードから順に，中間ノードに対してこれらの処理を再帰的に適用することでそれぞれのパス長の資料集合を得ることができる。関連語の計算においては，より近い関係の資料集合を優遇するために，パス長 n で打ち切る場合，（n- パス長）の重み付けを行う。

このようにして資料集合を機械的に設定し，構成 5 から構成 9 をそれぞれパス長 0 からパス長 4 までの資料集合によって構成した上で，考古資料について，先と同様の計算を行ったところ，「土器」から「石鏃」に至る到達容易性は，表 11（1）のようになり，表 9（1）の結果とあわせると望ましい順に，

構成 9> 構成 8> 構成 7> 構成 5> 構成 2> 構成 4> 構成 3> 構成 1> 構成 6

となる。

次に，「土器」から「高坏」についても同様の計算を行うと，表 11（2）のようになり，表 9（2）の結果とあわせると，望ましい順に，

構成 4> 構成 3> 構成 8> 構成 5> 構成 2> 構成 9> 構成 1> 構成 7> 構成 6

となる。

次に，「石器」，「石斧」，「土器」，「土偶」のそれぞれについて，関連語の中で専門家によってどれか 1 つだけに関連すると判定された語に対して，到達容易性を計算してみた。たとえば，「のみ」は「石斧」以外の語の関連語としてはふさわしくないとされている。そこで，「石器」，「石斧」，「土器」，「土偶」のそれぞれの語を出発語とした場合の「のみ」への到達容易性を計算し，それらの値を比較したところ，表 12（1）のようになった。パス長 4 までを取り入れても「石斧」を出発語とした場合の到達容易性がそれほど上がっていないが，これは，元のコレクションの階層化において，「のみ」と「石斧」を関連づけるような区分がなされていなかったためであると考えられる。このように，対象の機能に着目するといった全く別の観点からの関連度を取り入れるためには，そのような観点からの階層化の事例が必要となる。

他にも「石器」からのみ×がつけられていなかった「石皿」，「土偶」からのみ×がつけられていなかった「冠」，「土器」からのみ×がつけられていなかった「杯」についても同様の計算を行った結果を表 12（2）（3）（4）に示す。

実際には，関連する語といっても，コンテキストの違いでどのように関連するかは異なるため，すべての観点をうまくカバーするような階層化がなされていない限り，対象によって，より強い共起関係をもつものが上位に現れている。

**表 12　木構造を反映した構成法に対する到達容易性の比較**

(1)「のみ」に至る到達容易性の比較

| 構成 | 到達容易性が高い順 |
|---|---|
| 構成 5（階層構造：パス長 0 のみ） | 土偶 > 土器 > 石器 > 石斧 |
| 構成 6（階層構造：パス長 1 まで） | 土偶 > 石器 > 器 > 石斧 |
| 構成 7（階層構造：パス長 2 まで） | 土偶 > 土器 > 石器 > 石斧 |
| 構成 8（階層構造：パス長 3 まで） | 土偶 > 石器 > 土器 > 石斧 |
| 構成 9（階層構造：パス長 4 まで） | 土偶 > 石器 > 石斧 > 土器 |

(2)「石皿」に至る到達容易性の比較

| 構成 | 到達容易性が高い順 |
|---|---|
| 構成 5（階層構造：パス長 0 のみ） | 石斧 > 石器 > 土器 > 土偶 |
| 構成 6（階層構造：パス長 1 まで） | 石斧 > 石器 > 土器 > 土偶 |
| 構成 7（階層構造：パス長 2 まで） | 石斧 > 石器 > 土偶 > 土器 |
| 構成 8（階層構造：パス長 3 まで） | 土偶 > 石器 > 石斧 > 土器 |
| 構成 9（階層構造：パス長 4 まで） | 土偶 > 石器 > 石斧 > 土器 |

(3)「冠」に至る到達容易性の比較

| 構成 | 到達容易性が高い順 |
|---|---|
| 構成 5（階層構造：パス長 0 のみ） | 石器 > 土偶 > 石斧 > 土器 |
| 構成 6（階層構造：パス長 1 まで） | 土偶 > 石斧 > 石器 > 土器 |
| 構成 7（階層構造：パス長 2 まで） | 石器 > 土偶 > 石斧 > 土器 |
| 構成 8（階層構造：パス長 3 まで） | 石器 > 土偶 > 石斧 > 土器 |
| 構成 9（階層構造：パス長 4 まで） | 石器 > 土偶 > 石斧 > 土器 |

(4)「杯」に至る到達容易性の比較

| 構成 | 到達容易性が高い順 |
|---|---|
| 構成 5（階層構造：パス長 0 のみ） | 土器 > 石器 > 土偶 > 石斧 |
| 構成 6（階層構造：パス長 1 まで） | 土器 > 土偶 > 石器 > 石斧 |
| 構成 7（階層構造：パス長 2 まで） | 土器 > 石器 > 土偶 > 石斧 |
| 構成 8（階層構造：パス長 3 まで） | 土器 > 土偶 > 石器 > 石斧 |
| 構成 9（階層構造：パス長 4 まで） | 土器 > 土偶 > 石器 > 石斧 |

## (3) 到達容易性に対するパス長の影響

到達容易性に対するパス長の影響を更に見るために，次の実験を行った。考古資料 21,935 点の資料名称を用い，「石器」，「土器」を出発語とし，「スポール」「円礫」「岩偶」「高杯」「高坏」「甑」「須恵器」「垂飾」「錐器」「石各」「石核」「石偶」「石錘」「石篦」「石匕」「石篦」「石鏃」「掻器」「台石」「彫器」「土師器」「独鈷石」「剥石」「仏飯器」「穂摘具」「坩」「敲石」「焙烙」「礫器」の 30 語を対象語として，対象語が出発語からたどり着く先の語としてふさわしいかどうかを専門家に判定し

てもらった。この結果を表 13 に示す。○が付与された語は，専門家にとっては出発語との関連があると判定された語であるため，到達容易性が高くなることが望ましいものであるとして評価を行う。

到達容易性をはかる尺度としては，3 章で述べた方法を用いる。ある語を入力した場合に，その関連語を関連度の高いものから順に提示することを考え，先頭の語からはじめて，1 つ下位の語をみる確率を p とし，n 番目の語に到達する確率を $p^{n-1}$，ある語の関連語を見に行く確率を q（ただし，p+q<1）として計算する。具体的なパラメータとしては p=0.89, q=p^20, 関連語は第 60 位まで，別の関連語を辿る回数は 4 回までに制限して，出発語から対象語に至る到達可能な全経路の計算を行い，それらの到達容易性の和を計算した。そして，専門家によって○が付与された語について，これらの和を計算して，この値が大きい方が，専門家の知見により合致しているとみなす。たとえば，表 13 において，石器を出発語とした場合に，「円礫」「岩偶」「錐器」「石核」「石偶」「石錘」「石篦」「石匕」「石篦」「石鏃」「掻器」「台石」「彫器」「独鈷石」「敲石」「礫器」への到達容易性の和を求め，その値を比較する。

評価対象となる資料集合の構成法としては，コレクションをそのまま資料集合とする方法と，木の中間ノード間の枝の数に着目した方法であるパス長 0 からパス長 4 までの構成法を対象とした。表 14 に石器を出発語とした場合の計算結果を示す。

表 14 について，値の大きい順に資料群の構成法を並べると，

パス長 0 > コレクション単位 > パス長 1 > パス長 2 > パス長 4 > パス長 3

となる。パス長 0 の同一ノードを親とするノードで資料群を構成する方法が最も値が高くなっている。また，パス長を延ばしていくと合計の値が低くなっていっている。

土器を出発語とした場合についても，同様の評価を行った。表 15 に到達容易性による評価結果を示す。

表 15 について値の大きい順に資料群の構成法を並べると，

パス長 4 > パス長 3 > パス長 2 > パス長 0 > コレクション単位 > パス長 1

となる。土器を出発語とした場合は，コレクション単位とパス長 0，パス長 1 の場合に値が極端に低く，パス長が延びるに従って値が大きくなる傾向にある。

このように，石器を出発語とした場合と土器を出発語とした場合で，異なる評価結果となった。

**表 13　石器，土器の関連語に対する専門家による判定結果**

| | 石器 | 土器 |
|---|---|---|
| スポール | | |
| 円礫 | ○ | |
| 岩偶 | ○ | |
| 高杯 | | ○ |
| 高坏 | | ○ |
| 甑 | | ○ |
| 須恵器 | | |
| 垂飾 | | |
| 錐器 | ○ | |
| 石各 | | |
| 石核 | ○ | |
| 石偶 | ○ | |
| 石錘 | ○ | |
| 石篦 | ○ | |
| 石匕 | ○ | |
| 石篦 | ○ | |
| 石鏃 | ○ | |
| 掻器 | ○ | |
| 台石 | ○ | |
| 彫器 | ○ | |
| 土師器 | | ○ |
| 独鈷石 | ○ | |
| 剥石 | | |
| 仏飯器 | | |
| 穂摘具 | | |
| 坩 | | ○ |
| 敲石 | ○ | |
| 焙烙 | | ○ |
| 礫器 | ○ | |
| 鏃 | | |

表 14　石器を出発語とした場合の到達容易性の比較

| | コレクション単位 | パス長 0 | パス長 1 | パス長 2 | パス長 3 | パス長 4 |
|---|---|---|---|---|---|---|
| スポール | 0.000305 | 0.00000448 | 0.0000421 | 0.00032 | 0.00051 | 0.000904 |
| 円礫 | 0.00473 | 0.0118 | 0.000761 | 0.0000797 | 0.0000495 | 0.0000624 |
| 岩偶 | 0.00517 | 0.00399 | 0.00157 | 0.00935 | 0.0169 | 0.012 |
| 高杯 | 0.000192 | 0.000148 | 0.000143 | 0.00114 | 0.0019 | 0.0026 |
| 高坏 | 0.000194 | 0.000368 | 0.0000404 | 0.0000862 | 0.000717 | 0.00028 |
| 甑 | 0.000127 | 0.000155 | 0.0000393 | 0.0000749 | 0.000237 | 0.000206 |
| 須恵器 | 0.000146 | 0.000387 | 0.000406 | 0.00226 | 0.0123 | 0.0173 |
| 垂飾 | 0.0082 | 0.0403 | 0.0464 | 0.,0604 | 0.0163 | 0.0145 |
| 錐器 | 0.00324 | 0.0000876 | 0.000102 | 0.000368 | 0.00306 | 0.0038 |
| 石各 | 0.000328 | 0.00367 | 0.000531 | 0.0132 | 0.00666 | 0.00313 |
| 石核 | 0.126 | 0.0328 | 0.00986 | 0.00523 | 0.0217 | 0.0286 |
| 石偶 | 0.000529 | 0.000116 | 0.0000149 | 0.0000542 | 0.000214 | 0.000212 |
| 石錘 | 0.126 | 0.171 | 0.0943 | 0.0491 | 0.0154 | 0.0178 |
| 石篦 | 0.0542 | 0.152 | 0.107 | 0.133 | 0.1 | 0.115 |
| 石匕 | 0.0452 | 0.0996 | 0.0732 | 0.11 | 0.0537 | 0.0685 |
| 石篦 | 0.00166 | 0.000373 | 0.0002 | 0.000176 | 0.00262 | 0.00543 |
| 石鏃 | 0.164 | 0.106 | 0.134 | 0.104 | 0.0433 | 0.0372 |
| 掻器 | 0.0135 | 0.0189 | 0.00557 | 0.00137 | 0.00727 | 0.00574 |
| 台石 | 0.00241 | 0.00392 | 0.0003 | 0.000109 | 0.000216 | 0.000171 |
| 彫器 | 0.00324 | 0.000356 | 0.000154 | 0.00103 | 0.00454 | 0.00559 |
| 土師器 | 0.000109 | 0.00118 | 0.000635 | 0.00761 | 0.0152 | 0.0211 |
| 独鈷石 | 0.0000118 | 0.0000228 | 0.0000116 | 0.0000164 | 0.000000207 | 0.000000399 |
| 剥石 | 0.000465 | 0.000177 | 0.0000844 | 0.0000042 | 0.00000293 | 0 |
| 仏飯器 | 0.00000525 | 0.0000348 | 0.000000446 | 0.0000022 | 0.00000854 | 0.00000572 |
| 穂摘具 | 0.0000061 | 0.000201 | 0.0000236 | 0.000674 | 0.000643 | 0.0000674 |
| 坩 | 0.000198 | 0.000143 | 0.000102 | 0.000487 | 0.000488 | 0.000406 |
| 敲石 | 0.121 | 0.0967 | 0.0417 | 0.006 | 0.0133 | 0.0135 |
| 焙烙 | 0.0000231 | 0.0000777 | 0.00000251 | 0.0000086 | 0.0000289 | 0.0000307 |
| 礫器 | 0.0213 | 0.00945 | 0.00078 | 0.00298 | 0.0182 | 0.0199 |
| 鏃 | 0.0336 | 0.0208 | 0.0171 | 0.0194 | 0.0159 | 0.0131 |
| 計 | 0.693 | 0.707 | 0.469 | 0.423 | 0.301 | 0.333 |

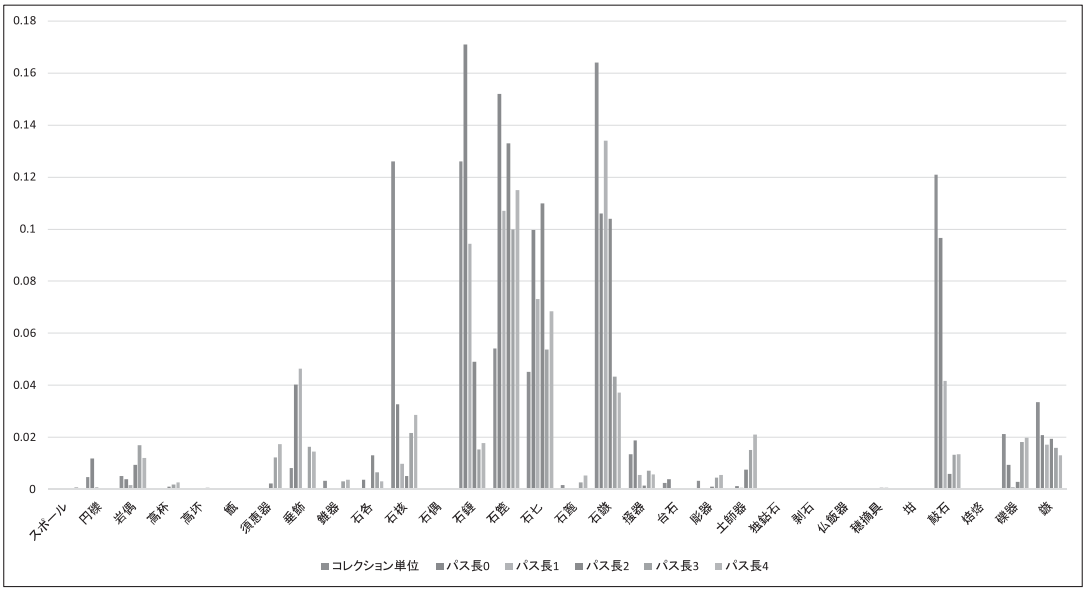
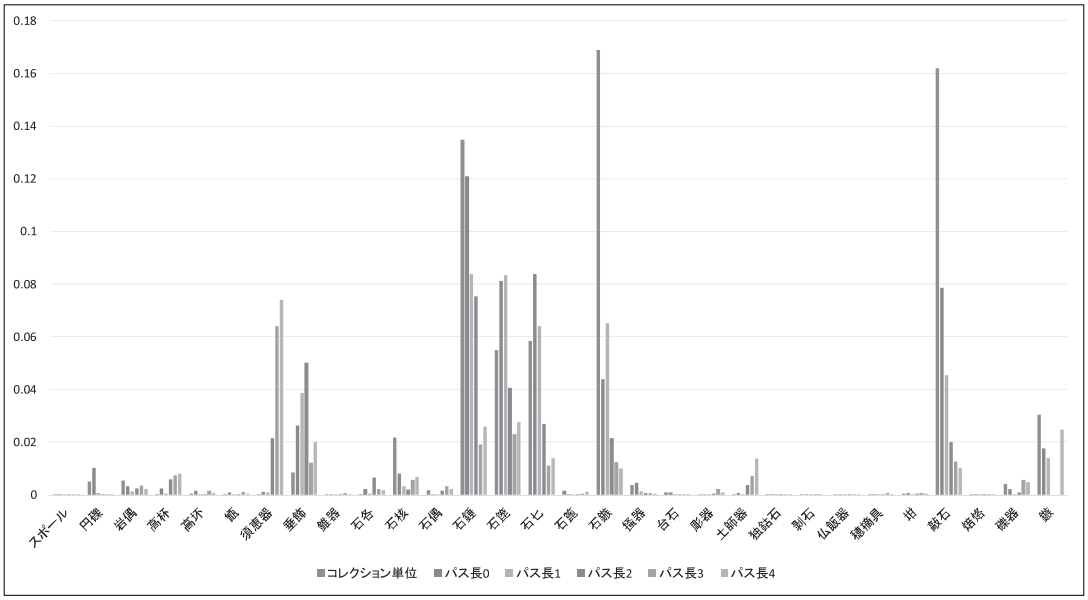


表 15　土器を出発語とした場合の到達容易性による比較

| | コレクション単位 | パス長 0 | パス長 1 | パス長 2 | パス長 3 | パス長 4 |
|---|---|---|---|---|---|---|
| スポール | 0.0000737 | 0.000000107 | 0.0000177 | 0.0000834 | 0.000171 | 0.000188 |
| 円礫 | 0.00504 | 0.0104 | 0.000669 | 0.000133 | 0.000226 | 0.000195 |
| 岩偶 | 0.00553 | 0.00335 | 0.00131 | 0.00257 | 0.00348 | 0.00236 |
| 高杯 | 0.000408 | 0.00242 | 0.000569 | 0.00598 | 0.00748 | 0.00819 |
| 高坏 | 0.000498 | 0.00155 | 0.0000813 | 0.000352 | 0.00168 | 0.000786 |
| 甑 | 0.000342 | 0.000858 | 0.0000632 | 0.000258 | 0.00107 | 0.000546 |
| 須恵器 | 0.000265 | 0.00113 | 0.00093 | 0.0215 | 0.0642 | 0.0742 |
| 垂飾 | 0.00864 | 0.0264 | 0.0387 | 0.0503 | 0.0123 | 0.02 |
| 錐器 | 0.000283 | 0.0000191 | 0.0000212 | 0.000163 | 0.000836 | 0.000345 |
| 石各 | 0.000299 | 0.00219 | 0.000417 | 0.00656 | 0.00221 | 0.00179 |
| 石核 | 0.0217 | 0.00807 | 0.00333 | 0.00209 | 0.00568 | 0.00684 |
| 石偶 | 0.00188 | 0.000176 | 0.0000193 | 0.00162 | 0.00342 | 0.00235 |
| 石錘 | 0.135 | 0.121 | 0.0839 | 0.0754 | 0.0193 | 0.026 |
| 石篦 | 0.0551 | 0.0814 | 0.0834 | 0.0408 | 0.0232 | 0.0276 |
| 石匕 | 0.0585 | 0.0839 | 0.0642 | 0.027 | 0.0112 | 0.0139 |
| 石篦 | 0.00162 | 0.000204 | 0.000183 | 0.00033 | 0.000614 | 0.0012 |
| 石鏃 | 0.169 | 0.044 | 0.0652 | 0.0216 | 0.0125 | 0.00999 |
| 掻器 | 0.00375 | 0.00464 | 0.00132 | 0.000695 | 0.000812 | 0.000516 |
| 台石 | 0.000986 | 0.000927 | 0.0000731 | 0.0000443 | 0.0000442 | 0.0000349 |
| 彫器 | 0.000283 | 0.0000169 | 0.000018 | 0.000539 | 0.00223 | 0.00101 |
| 土師器 | 0.000384 | 0.000815 | 0.000412 | 0.00376 | 0.00735 | 0.0137 |
| 独鈷石 | 0.0000223 | 0.0000321 | 0.0000284 | 0.00000768 | 0.0000041 | 0.00000291 |
| 剥石 | 0.0000904 | 0.0000607 | 0.0000284 | 0.00000165 | 0.000000617 | 0 |
| 仏飯器 | 0.0000208 | 0.000123 | 0.00000145 | 0.0000164 | 0.0000769 | 0.0000266 |
| 穂摘具 | 0.0000034 | 0.000161 | 0.0000168 | 0.00022 | 0.000634 | 0.000192 |
| 坩 | 0.000454 | 0.000745 | 0.000165 | 0.000497 | 0.000752 | 0.000596 |
| 敲石 | 0.162 | 0.0786 | 0.0455 | 0.02 | 0.0127 | 0.0103 |
| 焙烙 | 0.0000818 | 0.000364 | 0.00000718 | 0.0000574 | 0.000203 | 0.000117 |
| 礫器 | 0.00422 | 0.00234 | 0.000237 | 0.00106 | 0.00565 | 0.00496 |
| 鏃 | 0.0304 | 0.0177 | 0.014 | 0l.0105 | 0l.0191 | 0.0248 |
| 計 | 0.00217 | 0.00675 | 0.0013 | 0.0109 | 0.0185 | 0.0239 |



これは資料集合を構成するもととなっている歴博のコレクションの木構造の構成にも依存するところが大きいが，石器のようにパス長を延ばすとスコアが低くなるものと，逆に土器のように高くなるものがあることがわかった。

石器の場合と土器の場合で計算されたスコアの値を比較してみると，オーダが異なることがわかる。石器の場合は，比較的上位の語に望ましいとされた語が含まれているのに対し，土器の場合はそうではなかった。上位に現れる語は，同一資料集合の中に多く現れる語（直接の関連語）であることが多いため，そのようなものの中に望ましい語がある場合は，パス長を延ばしてもあまり効果はないということができる。逆に，直接共起する語の中に望ましい語がない場合は，パス長をのばすことにより，それらへの到達容易性の向上が期待できる。

## ❻……異なる分野の混在の効果

ここまでの実験は考古資料という単一の分野（ジャンル）の資料名称を対象として行ってきた。発見的検索において，探したい資料が属する分野が特定できる場合には，特定分野を対象として発見的検索を行うことができるが，そうでない場合には異なる分野の資料を対象とすることが考えられる。そこで，同時に異なるジャンルの資料名称を対象として関連語検索を行った場合に，単一のジャンルを対象とした場合と比べてどうなるかを明らかにするために，実験対象として，考古資料 21,935 点と民俗資料 16,335 点の資料名称を用いた実験を行った。それぞれのコレクション数は，考古資料が 633，民俗資料が 392 であり，今回はこれらのコレクションを資料集合構成の単位とした。

**表 16　考古資料と民俗資料の共通の主要語の例**

| 主要語 | 考古資料での出現数 | 民俗資料での出現数 |
|---|---|---|
| 土器 | 2847 | 3 |
| 鉢 | 1640 | 62 |
| 皿 | 836 | 769 |
| 斧 | 605 | 15 |
| 匙 | 402 | 19 |
| 甕 | 299 | 2 |
| 鏡 | 253 | 1 |
| 杯 | 102 | 84 |
| 蓋 | 76 | 3 |
| 小皿 | 76 | 8 |
| 剣 | 55 | 2 |
| 擂鉢 | 53 | 2 |
| 高坏 | 47 | 9 |
| 刃 | 47 | 8 |
| 徳利 | 32 | 9 |
| 香炉 | 30 | 2 |
| 人形 | 26 | 77 |
| 高杯 | 23 | 2 |
| 釣針 | 21 | 2 |
| 槍 | 21 | 2 |
| 鎌 | 21 | 89 |
| 茶碗 | 20 | 7 |
| 中皿 | 19 | 3 |
| 鍬 | 17 | 360 |

はじめにそれぞれの資料名称から，人手で主要語を抽出した結果，主要語の異なり語は，考古資料で 697 語，民俗資料で 4020 語であった。考古資料においては同一の資料名称が多数出現し，資料数に比べて主要語の異なり数が少なかったのに対し，民俗資料は名称に様々なバリエーションがあり，結果として主要語の異なり数が増えた。このうち，99 語が両者に共通の語で，考古資料と民俗資料をあわせた主要語の異なり語は 4618 語となった。よって，全体の 2% が共通であり，共通語彙は少ないと言える。考古資料と民俗資料で共通に現れた主要語の例を表 16 に示す。考古資料と民俗資料を対象にして同時に関連語検索を行った場合には，これらの共通の語を介して，両ジャンルの語が関連づけられる可能性がある。仮

に各ジャンルの主要語間で，全く共通する語がなかった場合には，それらのジャンルを統合して関連語検索を行ったとしても，一方のジャンルの語から異なるジャンルの語へ到達することはできない。

次に，すべての主要語について関連語を計算した結果，考古資料において 43,333 対，民俗資料において 907,207 対の主要語–関連語対が得られた。さらに両者を統合した場合の主要語–関連語対は 947,983 対となった。

異なる分野のコレクションを統合して関連語検索を行った場合の影響を調べるために，出発語と到達語として表 17 に示す語を選定し，それらのすべての組み合わせについて，考古資料単独で検索した場合，民俗資料単独で検索した場合，両者を統合して検索した場合を，到達容易性によって比較した。到達容易性をはかる尺度としては，3 章と同様，ある語を入力した場合に，その関連語を関連度の高いものから順に提示することを考え，先頭の語からはじめて，1 つ下位の語をみる確率を p とし，n 番目の語に到達する確率を $p^{n-1}$，ある語の関連語を見に行く確率を q（ただし，p+q<1）として計算する。具体的なパラメータとしては p=0.89, q=p^20，関連語は第 60 位まで，別の関連語を辿る回数は 3 回までに制限して，出発語から到達語に至る到達可能な全経路の計算を行い，それらの到達容易性の和を計算した。これらの全ての組み合わせについての計算結果を表 18 に示す。

考古資料のみに出現する語を出発語とし，考古資料のみに出現する語に到達しようとする場合，検索対象として民俗資料を対象としたときには，出発語，到達語の語彙が含まれていないため，到達できないのは当然である。考古資料のみを対象とした場合と，考古資料，民俗資料をあわせて対象とした場合とでは，語彙数が増えることと，関連語の順位の変動等の影響により，後者のほうが到達容易性は低くなる傾向がある。

考古資料のみに出現する語を出発語とし，民俗資料のみに出現する語に到達しようとする場合，

**表 17　出発語と到達語**

| | 出発語 | 到達語 |
|---|---|---|
| 考古資料のみに出現 | 石器，土偶 | 石鏃，石匕，土師器 |
| 民俗資料のみに出現 | 樽，膳 | 飯櫃，直綴，竪杵，切鍬 |
| 両者に出現 | 皿，杯，鎌 | 高杯，把手，鋤 |

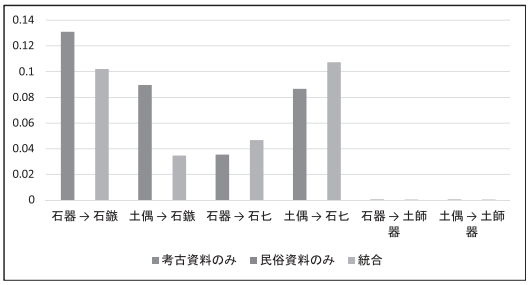
検索対象を考古資料のみ，民俗資料のみにしたときは，出発語ないし到達語が含まれていないため，到達できない。今回の実験では両者を統合した際にも，到達できなかったが，これは両者の共通語彙が少なく，それらが今回設定した出発語からたどることのできる関連語の上位に現れなかったためと考えられる。

考古資料のみに出現する語を出発語とし，両者に出現する語に到達しようとする場合，民俗資料のみを検索対象としたときには，出発語が含まれていないため，やはり到達できない。考古資料のみを対象とした場合と，考古資料，民俗資料をあわせて対象とした場合とでは，先と同様の傾向が見られる。一方，この実験では，考古資料のみを対象とした際には到達できなかった語への到達容易性が，統合した場合に計算されている例が見られる。これは統合の効果によるものと考えられる。

## 表 18　分野を統合した到達容易性の計算結果（1）

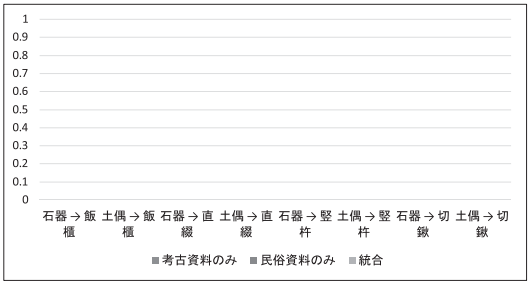
（1）考古資料のみに出現する語→考古資料のみに出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 石器 → 石鏃 | 0.131 | 0 | 0.102 |
| 土偶 → 石鏃 | 0.0897 | 0 | 0.0348 |
| 石器 → 石匕 | 0.0354 | 0 | 0.0469 |
| 土偶 → 石匕 | 0.0865 | 0 | 0.107 |
| 石器 → 土師器 | 0.0000466 | 0 | 0.0000425 |
| 土偶 → 土師器 | 0.000117 | 0 | 0.00000563 |



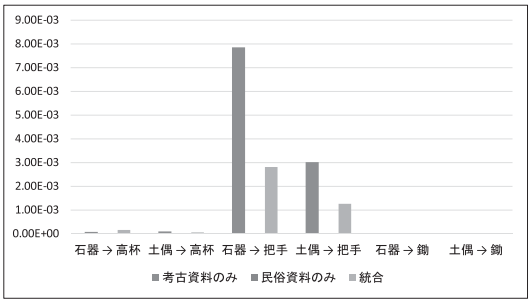

（2）考古資料のみに出現する語→民俗資料のみに出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 石器 → 飯櫃 | 0 | 0 | 0 |
| 土偶 → 飯櫃 | 0 | 0 | 0 |
| 石器 → 直綴 | 0 | 0 | 0 |
| 土偶 → 直綴 | 0 | 0 | 0 |
| 石器 → 竪杵 | 0 | 0 | 0 |
| 土偶 → 竪杵 | 0 | 0 | 0 |
| 石器 → 切鍬 | 0 | 0 | 0 |
| 土偶 → 切鍬 | 0 | 0 | 0 |



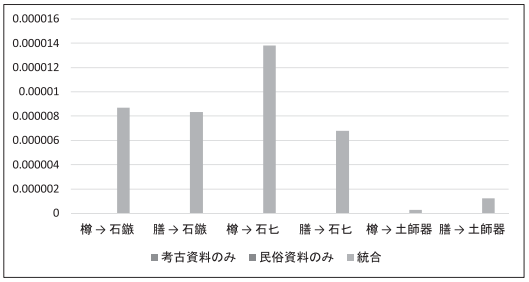

（3）考古資料のみに出現する語→両者に出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 石器 → 高杯 | 0.0000686 | 0 | 0.000164 |
| 土偶 → 高杯 | 0.000103 | 0 | 0.0000576 |
| 石器 → 把手 | 0.00785 | 0 | 0.00280 |
| 土偶 → 把手 | 0.00302 | 0 | 0.00125 |
| 石器 → 鋤 | 0 | 0 | 0.0000000161 |
| 土偶 → 鋤 | 0.0000000127 | 0 | 0.000000000632 |



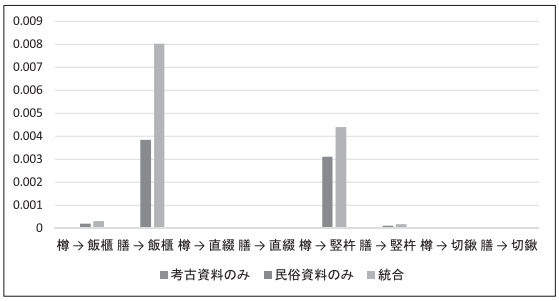

（4）民俗資料のみに出現する語→考古資料のみに出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 樽 → 石鏃 | 0 | 0 | 0.00000871 |
| 膳 → 石鏃 | 0 | 0 | 0.00000835 |
| 樽 → 石匕 | 0 | 0 | 0.0000138 |
| 膳 → 石匕 | 0 | 0 | 0.00000680 |
| 樽 → 土師器 | 0 | 0 | 0.000000296 |
| 膳 → 土師器 | 0 | 0 | 0.00000122 |



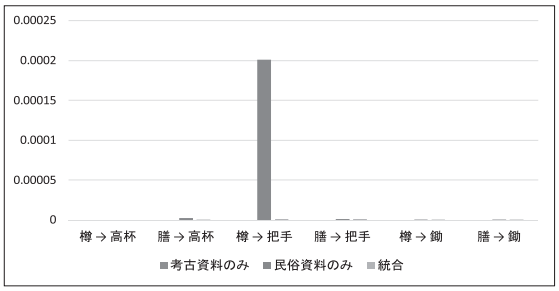
（5）民俗資料のみに出現する語→民俗資料のみに出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 樽 → 飯櫃 | 0 | 0.000200 | 0.000301 |
| 膳 → 飯櫃 | 0 | 0.00385 | 0.00801 |
| 樽 → 直綴 | 0 | 0.0000000143 | 0.0000000143 |
| 膳 → 直綴 | 0 | 0.0000000143 | 0.000000434 |
| 樽 → 竪杵 | 0 | 0.00312 | 0.00440 |
| 膳 → 竪杵 | 0 | 0.000122 | 0.000183 |
| 樽 → 切鍬 | 0 | 0 | 0 |
| 膳 → 切鍬 | 0 | 0 | 0 |



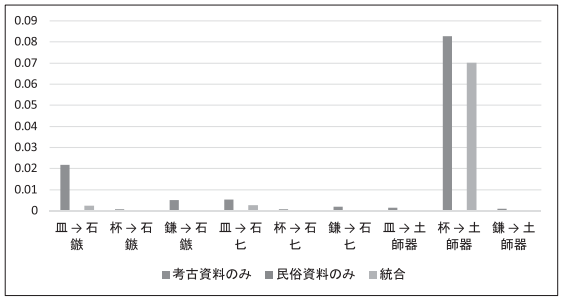

（6）民俗資料のみに出現する語→両者に出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 樽 → 高杯 | 0 | 0 | 0 |
| 膳 → 高杯 | 0 | 0.00000227 | 0.000000104 |
| 樽 → 把手 | 0 | 0.000201 | 0.00000119 |
| 膳 → 把手 | 0 | 0.000000866 | 0.00000104 |
| 樽 → 鋤 | 0 | 0.000000420 | 0.000000129 |
| 膳 → 鋤 | 0 | 0.000000264 | 0.000000114 |



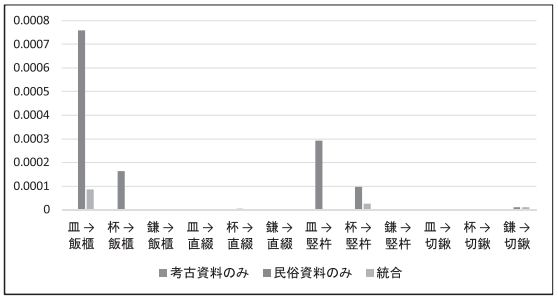

（7）両者に出現する語→考古資料のみに出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 皿 → 石鏃 | 0.0219 | 0 | 0.00241 |
| 杯 → 石鏃 | 0.000701 | 0 | 0.0000296 |
| 鎌 → 石鏃 | 0.00501 | 0 | 0.0000114 |
| 皿 → 石匕 | 0.00532 | 0 | 0.00258 |
| 杯 → 石匕 | 0.000744 | 0 | 0.0000271 |
| 鎌 → 石匕 | 0.00195 | 0 | 0.00000749 |
| 皿 → 土師器 | 0.00149 | 0 | 0.0000427 |
| 杯 → 土師器 | 0.0827 | 0 | 0.0702 |
| 鎌 → 土師器 | 0.00104 | 0 | 0.000000967 |



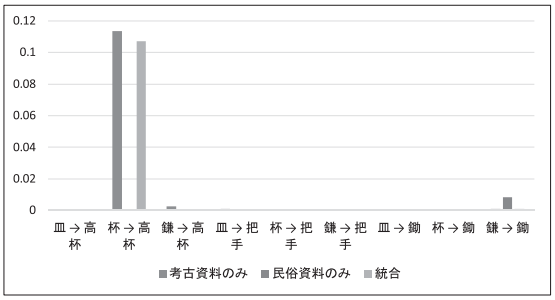

（8）両者に出現する語→民俗資料のみに出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 皿 → 飯櫃 | 0 | 0.000760 | 0.0000864 |
| 杯 → 飯櫃 | 0 | 0.000163 | 0.000000845 |
| 鎌 → 飯櫃 | 0 | 0.00000224 | 0.000000211 |
| 皿 → 直綴 | 0 | 0.000000058 | 0 |
| 杯 → 直綴 | 0 | 0.00000487 | 0 |
| 鎌 → 直綴 | 0 | 0 | 0 |
| 皿 → 竪杵 | 0 | 0.000293 | 0.00000299 |
| 杯 → 竪杵 | 0 | 0.0000959 | 0.0000266 |
| 鎌 → 竪杵 | 0 | 0.0000000113 | 0 |
| 皿 → 切鍬 | 0 | 0 | 0 |
| 杯 → 切鍬 | 0 | 0 | 0 |
| 鎌 → 切鍬 | 0 | 0.0000107 | 0.0000101 |



(9) 両者に出現する語→両者に出現する語

| | 考古資料のみ | 民俗資料のみ | 統合 |
|---|---|---|---|
| 皿 → 高杯 | 0.000598 | 0 | 0.0000494 |
| 杯 → 高杯 | 0.113 | 0 | 0.107 |
| 鎌 → 高杯 | 0.00258 | 0 | 0.0000299 |
| 皿 → 把手 | 0.00112 | 0.000000267 | 0.000171 |
| 杯 → 把手 | 0.000191 | 0.000186 | 0.00000130 |
| 鎌 → 把手 | 0.000265 | 0.000235 | 0.0000000288 |
| 皿 → 鋤 | 0 | 0.000000950 | 0.00000000104 |
| 杯 → 鋤 | 0.00000136 | 0 | 0.0000320 |
| 鎌 → 鋤 | 0.00116 | 0.00836 | 0.000923 |




次に，民俗資料のみに出現する語を出発語とし，考古資料のみに出現する語に到達しようとする場合，検索対象を考古資料のみ，民俗資料のみにしたときは，出発語ないし到達語が含まれていないため，やはり到達できない。しかしながら，両者を統合した場合には，到達容易性の値が計算されている。これは考古資料のみに含まれている語彙から出発して，両者に共通の語彙を経由し，民俗資料のみに含まれている語彙への経路が存在したためと考えられる。

民俗資料のみに出現する語を出発語とし，民俗資料のみに出現する語に到達しようとする場合，考古資料のみを検索対象としたときには，出発語，到達語ともに含まれていないため，やはり到達できない。民俗資料のみを対象にした場合と，考古資料，民俗資料をあわせて対象とした場合とでは，際だった違いはない。民俗資料は主要語数が多く，その中でも到達できないものが多くあるが，統合してもそれらの到達容易性が上がることはなかった。

民俗資料のみに出現する語を出発語とし，両者に出現する語に到達しようとする場合，考古資料のみを検索対象としたときには，出発語が含まれていないため，やはり到達できない。民俗資料のみを対象にした場合と，考古資料，民俗資料をあわせて対象とした場合とでは，際だった違いはない。

次に，両者に出現する語を出発語とし，考古資料のみに出現する語に到達しようとする場合，検索対象を民俗資料のみにしたときは，到達語が含まれていないため，到達できない。考古資料のみを対象とした場合と，考古資料，民俗資料をあわせて対象とした場合とでは，語彙数が増えることと，関連語の順位の変動等の影響により，後者のほうが到達容易性は低くなる傾向がある。

両者に出現する語を出発語とし，民俗資料のみに出現する語に到達しようとする場合，検索対象を考古資料のみにしたときは，到達語が含まれていないため，到達できない。民俗資料のみを対象とした場合と，考古資料，民俗資料をあわせて対象とした場合とでは，後者のほうが到達容易性は低くなる傾向がある。また，いくつかの例で，対象が民俗資料のみの場合に到達容易性の値が計算されていたものが，統合した場合に到達できなくなっている。これは，関連語の順位の変動により，もともと順位が下位にあったものが，更に下がって打ち切り範囲から落ちたためと考えられる。

最後に，両者に出現する語を出発語とし，両者に出現する語に到達しようとする場合，検索対象がどちらか一方のみでは到達できなかったものが，統合すれば到達できている。

異なる分野の統合の影響をさらに確認するために，民俗資料と歴史資料の間で同様の実験を行った。歴史資料の主要語の異なり語は，2296 語，民俗資料との共通語彙は 155 語であった。民俗資料との共通語彙数は考古資料よりも多い。出発語として，共通語彙の中から「着物」「衣裳」「羽織」

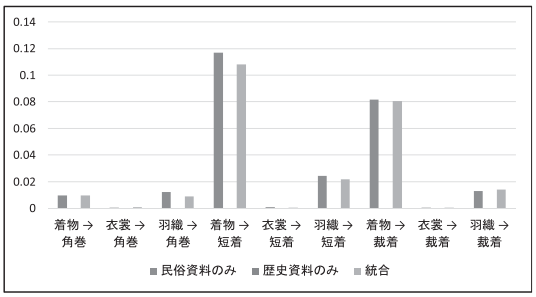
を選び，ここから民俗資料のみに存在する「角巻」「短着」「裁着」，歴史資料のみに存在する「小袖」「帷子」「被衣」，両者に共通に存在する「単衣」「長衣」「合羽」への到達容易性の和を計算した。これらの全ての組み合わせについての計算結果を表 19 に示す。結果としては，考古資料と民俗資料の統合の場合と同様の傾向が出ている。

以上をまとめると，複数の分野を統合して関連語検索を行うことにより，全般的な傾向としては，語彙数の増加により，到達容易性は低くなる。ただし，異なる分野において共通に出現する語を介して，分野間を横断した検索が可能になることもあることが確かめられた。これを一般化すると，

**表 19　分野を統合した到達容易性の計算結果（2）**

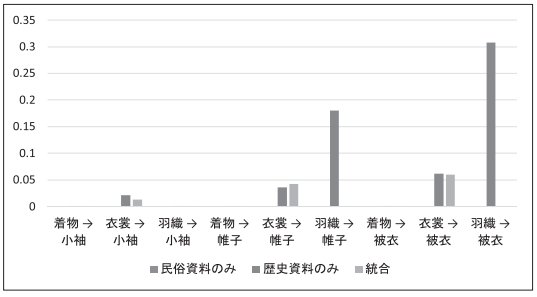
（1）両者に出現する語→民俗資料のみに出現する語

| | 民俗資料のみ | 歴史資料のみ | 統合 |
|---|---|---|---|
| 着物 → 角巻 | 0.00964 | 0 | 0.00952 |
| 衣裳 → 角巻 | 0.000242 | 0 | 0.000948 |
| 羽織 → 角巻 | 0.0121 | 0 | 0.009 |
| 着物 → 短着 | 0.117 | 0 | 0.108 |
| 衣裳 → 短着 | 0.000899 | 0 | 0.000469 |
| 羽織 → 短着 | 0.0243 | 0 | 0.0218 |
| 着物 → 裁着 | 0.0815 | 0 | 0.0806 |
| 衣裳 → 裁着 | 0.000382 | 0 | 0.000264 |
| 羽織 → 裁着 | 0.0129 | 0 | 0.0141 |



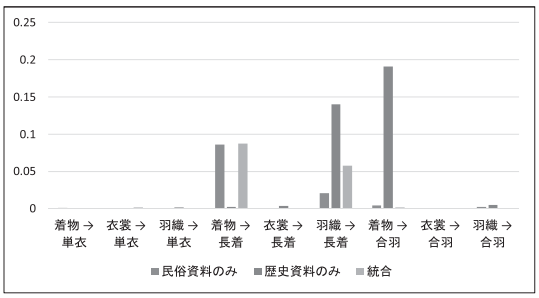

（2）両者に出現する語→歴史資料のみに出現する語

| | 民俗資料のみ | 歴史資料のみ | 統合 |
|---|---|---|---|
| 着物 → 小袖 | 0 | 0.00000896 | 0.00000517 |
| 衣裳 → 小袖 | 0 | 0.0218 | 0.0136 |
| 羽織 → 小袖 | 0 | 0.000981 | 0.00000914 |
| 着物 → 帷子 | 0 | 0.00033 | 0.000107 |
| 衣裳 → 帷子 | 0 | 0.0357 | 0.0421 |
| 羽織 → 帷子 | 0 | 0.18 | 0.000292 |
| 着物 → 被衣 | 0 | 0.000579 | 0.000168 |
| 衣裳 → 被衣 | 0 | 0.0619 | 0.0596 |
| 羽織 → 被衣 | 0 | 0.308 | 0.000445 |




（3）両者に出現する語→両者に出現する語

| | 民俗資料のみ | 歴史資料のみ | 統合 |
|---|---|---|---|
| 着物 → 単衣 | 0.00128 | 0.00000216 | 0.000375 |
| 衣裳 → 単衣 | 0 | 0.000609 | 0.00147 |
| 羽織 → 単衣 | 0.000595 | 0.00175 | 0.000589 |
| 着物 → 長着 | 0.0858 | 0.00235 | 0.0873 |
| 衣裳 → 長着 | 0.000694 | 0.00391 | 0.000551 |
| 羽織 → 長着 | 0.0205 | 0.14 | 0.0576 |
| 着物 → 合羽 | 0.00448 | 0.191 | 0.00172 |
| 衣裳 → 合羽 | 0 | 0.00000825 | 0.0000164 |
| 羽織 → 合羽 | 0.00219 | 0.00488 | 0.000134 |



検索対象がある程度確定していて，どの分野に存在するかわかっている場合は，検索範囲をその分野に絞って検索を行ったほうが到達容易性は高くなる。一方，分野を横断した検索を行いたい場合は，複数の分野を統合して関連語検索を行うことで，異なる分野の語に到達できる可能性が生まれる。考古資料と民俗資料の間では，分野間で共通な語の数が少なかったため，この効果は限定的であった。

## ❼……まとめ

本稿では，博物館資料の検索において，普通に使われる語彙から，専門的な語彙で表現された資料にたどり着くことを補助する手法の実現を目指して，既存の資料名称から語を取り出し，それらの間の関連度を用いて，一般的な語彙から専門的な語彙にたどり着く発見的な検索手法を提案し有用性を検証した。

関連度の計算においては，一般の文書に含まれる語の間の関連度を測る方法を適用することができるが，資料名称に含まれるすべての語を対象とするのではなく，その中から主要な語を選定した方がより適切な関連語が得られることが確認された。

次に，ある語に対して，その関連語を関連度の高いものから順に提示し，関連語を順にたどっていくことで一般的な語彙から専門的な語彙にたどりつくことが可能なことを示した。

関連語の計算においては，資料集合の設定が大きすぎると，何でも関連づけられてしまい，結果として望ましくない関連が発生し，資料集合を細かく分割し過ぎると，関連語としてふさわしい語までが排除されてしまうため，適切な資料集合の設定が課題となった。これに対して，コレクションの木構造を利用した資料集合の構成法を示し，その効果を検証した。その結果，同一資料集合の中に多く現れる語（直接の関連語）以外の語について，木構造におけるパス長をのばして計算することにより，それらへの到達容易性の向上が見られた。

最後に，分野別のコレクションを併せて関連語を探す際には，全般的な傾向としては，語彙数の増加により，到達容易性は低くなるが，異なる分野において共通に出現する語を介して，分野間を横断した検索が可能になることもあることが確かめられた。これを一般化すると，検索対象がある程度確定していて，どの分野に存在するかわかっている場合は，検索範囲をその分野に絞って検索を行ったほうがよいが，分野を横断した検索を行いたい場合は，複数の分野を統合して関連語検索を行うことで，異なる分野の語に到達できる可能性が生まれることが確認できた。

**註**

（1）――「汎用連想計算エンジン（GETA）」は，情報処理振興事業協会（IPA）が実施した「独創的情報技術育成事業」の研究成果である。

山田　篤　（公益財団法人京都高度技術研究所，国立歴史民俗博物館共同研究員）
安達文夫　（国立歴史民俗博物館研究部）
小町祐史　（国士舘大学大学院総合知的財産法学研究科，国立歴史民俗博物館共同研究員）

（2014 年 1 月 7 日受付，2014 年 5 月 26 月審査終了）

Bulletin of the National Museum of Japanese History
Vol. 189 January 2015

# Heuristic Search for Information on Museum's Holdings

YAMADA Atsushi, ADACHI Fumio and KOMACHI Yushi

This article suggests a heuristic search method as an advanced approach to searching information from museums. In this method, a search engine takes terms out of the names of the existing materials and analyzes the degree of relevance between them to extract technical terms from general terms. The effectiveness of the method is also verified in this paper.

The degree of relevance can be calculated by applying the method to measure the degree of relevance between terms in general documents. In this case, one can obtain more appropriate related terms by selecting keywords rather than using all the terms in the names of materials.

One can find technical terms in the list of general terms by sorting the related terms obtained through the above process in the order of relevance and searching one by one in the order from highest to lowest.

When measuring the relevance between related terms, it is important to set up a proper categorization of materials. For example, when the materials are divided into too few clusters, any terms can be related to each other, resulting in that irrelevant associations may be generated. On the contrary, when the materials are divided into too many clusters, related terms may be excluded. Therefore, this article suggests a categorization of materials based on the tree structure of collections while verifying the effectiveness of this set up.

In the end, this article indicates how to enable cross-disciplinary search. In general, it is more difficult to find adequate terms when searching related terms from collections in different fields due to a larger vocabulary; however, it is possible to find cross-sectional information by focusing on terms common in different fields.

Key words: material name, degree of relevance, easiness of search, tree structure