

# 博物館情報探索における関連語の順位を考慮した到達容易性の評価

画像電子学会 画像ミュージアム研究会 博物館・美術館 DTD-SG

山田 篤<sup>†</sup> 安達 文夫<sup>‡</sup> 小町 祐史<sup>§</sup>

Atsushi YAMADA<sup>†</sup> Fumio ADACHI<sup>‡</sup> Yushi KOMACHI<sup>§</sup>

<sup>†</sup> 京都高度技術研究所

<sup>†</sup> ASTEM RI/Kyoto

<sup>‡</sup> 国立歴史民俗博物館

<sup>‡</sup> National Museum of Japanese History

<sup>§</sup> 大阪工業大学

<sup>§</sup> Osaka Institute of Technology

E-mail: <sup>†</sup> yamada@astem.or.jp, <sup>‡</sup> adachi@rekihaku.ac.jp, <sup>§</sup> komachi@y-adagio.com

## 1. はじめに

博物館・美術館情報の電子化が進み、ネットワークを通じて個々の館の収蔵品に関する情報の提供サービス、検索サービスなどが開始されている。博物館情報の利用者にとっては、どの館にアクセスするかを意識せず、各館の差異を意識せずにシームレスに検索ができること、つまり横断検索できることが望ましい[1]。

本稿では、利用者の検索要求から関連する収蔵品を見つけ出すために、個々の収蔵品に対して付与されている資料名称をもとに計算される関連語を利用した探索を行う際の到達容易性の評価方法について述べる。

## 2. 横断検索のためのフレームワーク

収蔵品の横断検索のためには、個々の館が持つ情報を統合する仕組みが必要となる[2]。画一的な共通フォーマットを用いずに、多様性を許容する情報構造として、次の3レベルに階層化される情報共有のフレームワークが提案されている[1]。

(1) 情報記述構造レベル

(2) 情報記述内容レベル

(3) 情報ナビゲーションレベル

このうち、(1)については、様々な取り組みがなされており、記述スキームの共有や標準化といった試みもある[3][4]。

これに対して、内容レベルの情報共有については、様々な収蔵品を扱うという事情から困難な問題として残っている。対象物が基本的に物であることから、図書のような全文検索といった手法も適用できない。オントロジ[6]や概念辞書[7]を用いた手法も検討されているが、収蔵品に関して様々なメタデータを付与することは博物館にとっても大きな負担となる。

このため、個々の博物館がなるべく簡単に用意できる情報をもとにして、内容レベルの情報共有を行い、横断検索を可能にすることが望まれる。

## 3. 資料群と資料名称

収蔵品に関して、博物館になるべく負担をかけずに収集可能なメタデータとして、資料群と資料名称を考える。

博物館では、様々な理由により収蔵品を群として管理していることがある。このようにグループ化された資料の集まりを資料群と呼ぶことにする。資料群はさらに多階層に階層化されて管理されていることもある。[8]では資料群を分類と見なしているが、厳密な分類ではなくとも、グループ化され同じ資料群に属する収蔵品の間には何らかの関連があると考えられる。現状で、資料群の構成方法について、統一された基準は見あたらないが、大量の資料群を集めてくることにより、互いに関連する可能性の高い収蔵品を見いだす手がかりが得られるのではないかと考えている。

一方、個々の収蔵品には名称（資料名称）が付与されている。これと先の資料群を組み合わせると、資料群によってグループ化された資料名称の集まりができる。

[9]では資料群と資料名称との関係を、文書検索と同様の手法を用いて取り扱うことを提案している。また、[10]では、この考え方にに基づき、実データに対して、資料名称の形態素解析を行い、資料名称中の主要語を抽出して、関連語の計算を行っている。

このとき、たとえ関連語が抽出できたとしても、その数が極めて多い場合に、より関連性の高い語から順にみていくとすると、たとえ何らかの関連があったとしても、低い関連性をもつ語にはなかなか到達しないという問題がある。

そこで、本報告では、主要語に基づく関連語が計算された場合に、それらを辿っていくことで、特定の語から出発して別の語に到達することが可能な場合に、語によってどれだけ到達し易さが変わるか、また同じ語の組み合わせに対して複数の経路が存在した場合にそれらの間での到達し易さの違いを評価する方法について検討を行った。

#### 4. 関連語に基づく到達可能性

準備として、[10]と同じく、国立歴史民俗博物館の収蔵品データから、考古資料 21,935 点の資料名称を対象として、関連語の計算を行った。

関連語の計算の元となる各資料名称に対する主要語としては、より精度を高めるために、形態素解析結果に基づくものではなく、人手によって付与したものを用いた。また、一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。主要語の異なり語彙数は 751 であった。関連語の計算方法は[10]と同様である。

このとき、たとえば「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 位までのものを図 1 に示す。ここで第 3 列に示している数値は、GETA によって計算された関連度を表す数値で、値が大きいほど関連性が高いことを表している。

すべての主要語について、関連語を計算した結果、43403 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」という経路を辿ることによって「土器」から「ハンドアックス」に到達することができる。このように関連語を順次辿ることによって、「土器」に対しては、665 の語へ到達可能となった。この一部を図 2 に示す。

土器	鉢	8.03913828624174
土器	深鉢	7.86422961961796
土器	石鉢	7.83571978935358
土器	縄文式土器	7.47299868019516
土器	敲石	7.38722016253306
土器	磨石	6.91231106632793
土器	石錘	6.60617903218393
土器	石斧	6.53347584629854
土器	壺	6.38438041261366
土器	尖頭器	6.26528183115589
土器	浅鉢	6.23886951547771
土器	石器	6.00940483766365
土器	礫	5.97652008832945
土器	石匕	5.88919695337085
土器	石錐	5.73358667002854
土器	剥片	5.66514528536452
土器	石皿	5.57998733397973
土器	斧	5.56782684783571
土器	石槍	5.36499175059727

図 1 「土器」の関連語の例

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

図 2 「土器」から到達可能な語の例

ただし、この中には到達が容易なものとそうでないものがある。たとえば、多数の関連語を、関連

性が高いものから順に提示した場合に、関連性が低いものは高いものに比べて到達することがより困難であると考えられる。

また、到達可能な語に対して到達する経路も一つとは限らない。この結果、到達可能な語であっても、経路毎の到達容易性に差異が生じる。

## 5. 単純なコストを用いた評価

はじめに、ある語から出発して、到達可能な語にたどり着く経路に関する到達容易性に関して検討を行った。

このために、関連性が高い語ほど、到達が容易であると考え、図 1 に示した関連度の逆数がある語からみた別の語に到達するためのコストとみなし、経路全体のコストをこれらの総和として計算した。

博物館情報の探索という観点からは、一般的な語彙から出発して、より専門的な語彙に到達することに興味がある。そこで、探索を始める語として「土器」、到達目標の語として「鉄滓」を設定し、3つの関連語を辿って到達可能な全経路と、そのコストを計算した。経路の総数は7008であった。計算結果の一部を表1に示す。

この表は経路全体のコストの低いものから順に並べている。この結果、「土器」から「鉄滓」に至る直接の経路を辿った場合が最もコストが低いという結果になっている。しかしながら、「土器」から直接到達可能な353個の語彙の中では「鉄滓」は187番目であり、相対的には選ばれにくいと考えられる。

コスト	経路		
0.961047	鉄滓		
1.041381	杯	鉄滓	
1.099758	土器	鉄滓	
1.12785	蓋	鉄滓	
1.146301	須恵器	鉄滓	
1.180091	土器	杯	鉄滓
1.210528	甕	鉄滓	
1.234522	壺	蓋	鉄滓
1.244233	斧	杯	鉄滓
1.252283	斧	鉄滓	
1.266561	土器	蓋	鉄滓
1.275575	壺	杯	鉄滓
1.285011	土器	須恵器	鉄滓
1.286826	斧	須恵器	鉄滓
1.296337	石斧	鉄滓	
1.301158	土師器	鉄滓	
1.31954	鉢	蓋	鉄滓
1.345758	鉢	甕	鉄滓
1.349238	土器	甕	鉄滓
1.356295	壺	甕	鉄滓

表 1 「土器」から「鉄滓」に至る経路のコストによる評価

一方、全体の経路長が長いということは、多くの関連語を辿らねばならず、経路長が短い方が到達し易いということが言える。よって、到達容易性の評価においては、

- 1) 同時に提示される関連語の中での順位
- 2) 経路全体の長さ

の2点を考慮する必要がある。

## 6. 関連語の順位を用いた評価

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る。たとえば、「土器」から出発した場合、「土器」の関連語として、図1に示した語をこの順で提示する。このとき、先頭の語からはじめて、一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。「土器」の例の場合は、

鉢— $p$ →深鉢— $p$ →石鏃— $p$ →縄文式土器— $p$ →敲石— $p$ →磨石— $p$ →石錘— $p$ →石斧— $p$ →…

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」に至る各経路の確率を計算することができる。今、仮に  $p=0.89$ ,  $q=p^{20}$  として計算してみた場合に、上位にくる経路を表2に示す。

確率	経路		
0.000919	杯	木炭	鉄滓
0.000457	石槍	化石	鉄滓
0.000322	香炉	大甕	鉄滓
0.000287	杯	拓本	鉄滓
0.000255	杯	須恵器	鉄滓
0.000202	剥片	化石	鉄滓
0.000202	須恵器	木炭	鉄滓
0.00016	石斧	化石	鉄滓
0.000127	杯	古銭	鉄滓
0.000127	蓋	須恵器	鉄滓
0.000127	須恵器	拓本	鉄滓
0.000113	杯	鉄滓	
8.94E-05	杯	瓦経	鉄滓
8.94E-05	杯	瓦譜	鉄滓
8.94E-05	杯	仏	鉄滓
8.94E-05	杯	すり石	鉄滓
8.94E-05	杯	カード	鉄滓
8.94E-05	杯	漆喰	鉄滓
8.94E-05	杯	泥塔	鉄滓
8.94E-05	杯	鰐口	鉄滓

表2 「土器」から「鉄滓」に至る経路の関連語の順位を元にした確率による評価

この例では、関連語リストの 20 語めを見る確率と、次の関連語を辿る確率を同等としている。この場合、先のコストによる評価とは相当異なる結果が得られる。コストによる評価では 1 位であった「鉄滓」に直接至る経路は 1161 位に後退した。

別の設定として、 $p=0.95$ ,  $q=p^{60}$  として計算してみた場合に、上位にくるものを表 3 に示す。これは関連語リストの 60 語めを見る確率と、次の関連語を辿る確率を同等とした場合である。

確率	経路		
0.04607	杯	木炭	鉄滓
0.033866	石槍	化石	鉄滓
0.029035	香炉	大甕	鉄滓
0.027584	杯	拓本	鉄滓
0.026205	杯	須恵器	鉄滓
0.02365	剥片	化石	鉄滓
0.02365	須恵器	木炭	鉄滓
0.021344	石斧	化石	鉄滓
0.019263	杯	古銭	鉄滓
0.019263	蓋	須恵器	鉄滓
0.019263	須恵器	拓本	鉄滓
0.0183	杯	鉄滓	
0.016515	杯	瓦経	鉄滓
0.016515	杯	瓦譜	鉄滓
0.016515	杯	仏	鉄滓
0.016515	杯	すり石	鉄滓
0.016515	杯	カード	鉄滓
0.016515	杯	漆喰	鉄滓
0.016515	杯	泥塔	鉄滓
0.016515	杯	鰐口	鉄滓

表 3 「土器」から「鉄滓」に至る経路の関連語の順位を元にした確率による評価（2）

先の設定と比較すると、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、順位自体の変動は見られなかった。

## 7. おわりに

本稿では、博物館、美術館の収蔵品の横断検索において、資料名称と資料群に基づき計算された関連語を用いて、ある語から別の語へ辿っていく場合に、関連語の順位と、全体の経路長を元にした経路の選択確率を定義し、到達容易性を評価する方法について述べた。ある語から別の語に至るすべての可能な経路の確率の総和を比較することで、語の対毎の到達容易性の比較も行うことができると考えている。

## 文 献

- [1] 山田篤, 他: 博物館情報の知的横断検索のためのフレームワーク, 画電年次大会, 2002-06.
- [2] 山本泰則, 中川隆: 博物館資料情報共有の試み, 画電年次大会, 2004-06.
- [3] 文化財情報システムフォーラム (<http://www.tnm.go.jp/bnca/>).
- [4] The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) (<http://www.cidoc.icom.org/>).
- [5] 山田篤, 他: 博物館情報の分類マッピングを用いた横断検索, 画電年次大会, 2004-06.
- [6] 山田篤, 他: 博物館情報の横断検索におけるオントロジ利用の試み, 画像ミュージアム研究会, 2005-03.
- [7] 山田篤, 他: 博物館横断検索に向けた概念辞書の枠組みの検討, 画像ミュージアム研究会, 2007-03.
- [8] 山田篤, 他: 部分的分類知識の統合による博物館情報の横断検索の提案, 画像ミュージアム研究会, 2008-02.
- [9] 山田篤, 他: 博物館資料群中の語の共起関係を用いた関連語抽出, 画像ミュージアム研究会, 2009-03.
- [10] 山田篤, 他: 博物館資料群中の語の共起関係を用いた関連語抽出における主要語選定の効果, 画電年次大会, 2009-05.