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† 安達 文夫‡ 小町 祐史§

Atsushi YAMADA† Fumio ADACHI‡ Yushi KOMACHI§

†京都高度技術研究所 †ASTEM RI/Kyoto
‡国立歴史民俗博物館 ‡National Museum of Japanese History
§大阪工業大学 §Osaka Institute of Technology
E-mail: †yamada@astem.or.jp, ‡adachi@rekihaku.ac.jp, §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関連する語を抽出する方法と，そこでの資料名称中の主要語選定の効果について述べる。

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。

たとえば，資料群 A は{a,a,a,b,b}，B は{b,c,c,c,d}という資料名称がそれぞれグループ化されていたとすると，資料名称 b は資料群 A では a と共起し，B では c,d と共起していることになる。これを行列で表すと，

	a	**b**	**c**	**d**
A	3	2	0	0
B	0	1	3	1

という頻度表が作成できる。資料群が多くなればなるほど，様々な資料名称が出現し，この行列は大きくなるとともに，一つの資料群について見た場合は出現しない資料名称も増え，全体として非常に疎な行列になる。これは文書検索における頻度付き索引データと類似であるため，[10]では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。このとき，資料名称を単位とすると，資料群によっては非常に長い名称が存在し，共起計算に利用しにくいため，形態素解析を行うことと，修飾語との共起を排除するために主要語を選定することが，非常に小規模なデータについて述べられている。本報告ではより大規模なデータセットについて，これらの効果を検討した。

4. 実データに基づく検討

国立歴史民俗博物館の収蔵品データから，考古資料 21,935 点の資料名称をもとに頻度表を作成した。資料名称の異なり総数は 4,492，資料群は A-1 から A-633 まであった。資料名称全体で頻度上位のものを図 1 に示す。

1252	石鏃
575	打製有茎石鏃
537	深鉢形縄文式土器
451	石片類
339	石匙
303	文字平瓦
254	石匕
244	磨製石斧
233	出土地不明打製有茎石鏃
225	平瓦
225	壷形土器
217	白磁碗

図 1　高頻度の資料名称

これを見ると，頻度第 1 位の「石鏃」と第 2 位の「打製有茎石鏃」，第 9 位の「出土地不明打製有茎石鏃」は互いに関連おり，これらは適切に形態素解析を行うことによって，「石鏃」の部分を共通の語として取り出すことで関連づけることができる。「文字平瓦」と「平瓦」の関係も同様である。

そこで，形態素解析器として MeCab，形態素解析用電子化辞書として UniDic[9]を用いて，語（短単位）への分割を行った。形態素解析結果における高頻度語を図 2 に示す。

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

図 2　形態素解析結果における高頻度語

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り語数は 2,917 であった。

UniDic の特徴として，短単位という斉一な単位で語が分割されるため，資料名称の解析においてはもう少し長い単位が欲しいといった要求もあり得る。また，未知語や形態素解析誤りによって，一文字素片が出力されている場合もあり，形態素解析用辞書の整備も欠かせない課題であることがわかる。

次に，形態素解析結果のすべての語を用いて，資料群毎の頻度表を作成し，汎用連想計算エンジン GETA*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，入力語に対して，関連する資料群を計算し，さらにそこから関連する語を計算することで，入力語の関連語を計算するという手法を採った。類似度の計算には GETA の WT_SMART の設定をそのまま使用した。「土器」という入力語に対する関連語の計算結果を図 3 に示す。形態素解析の結果得られた短単位語が並んでいる。上位に並んでいる「鉢形」「広口」「深鉢」などは「鉢形土器」「広口土器」「深鉢土器」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，この結果では上位に現れてこないが，「土器」を含む資料群において出現している語である。

これらを区別するために，一つの資料名称から一語のみを取り出して，同一資料群中における共起を用いた計算を行ってみた。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，これらの異なり数は 662 となった。

```
Word: 土器
[Result]
 1: 鉢形        4.72997096261726
 2: 広口        4.46729327630274
 3: 深鉢        4.46364161276952
 4: 土器        4.00571543518008
 5: 人面        3.42199224791764
 6: 小形        3.31716407412078
 7: 浅  3.29019933184202
 8: 縄文        3.26822661319976
 9: 小型        3.18929685168478
10: 口  3.1677677992656
11: 注  3.12597234842501
12: 把手        3.10503859070239
13: 絵画        3.05664322675671
14: 石偶        2.92339935886476
15: 軽石        2.89399611164759
16: イモ        2.84135364522323
17: 手  2.79654309553799
18: づく        2.73325762056586
19: 形  2.63269532995118
20: 墨書        2.6277729663247
```

図 3 全形態素解析結果を用いた連想検索結果

そして，同一資料群におけるこれらの語の頻度表を作成し，GETA を用いて，先と同様の実験を行った。「土器」という入力語に対する関連語の計算結果を図 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」と同じ資料群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。先の全形態素解析結果を用いた場合と比べて，意味的に関連性の高い語が並んでいると思われる。

* 「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。

Word: 土器
[Result]
1: 土器 7.33459505175907
2: 石鏃 5.89435861293338
3: 敲石 5.24628118897332
4: 石斧 5.19186644425113
5: 石 4.78793785692116
6: 破片 4.61088641189617
7: 土偶 4.60107250865646
8: 石器 4.57767866245767
9: 匕 4.5745460902211
10: 錘 4.52815606013407
11: 錐 4.34824637748181
12: 礫 4.26112634243512
13: 器 4.04099856102012
14: 箆 3.8761675810929
15: 石槍 3.80598195306418
16: 品 3.7590165673881
17: 片 3.54669643009943
18: スクレイパー 3.4573573622418
19: 軽石 3.36423234019861
20: 石皿 3.34132646401116

図 4 主要語のみを対象とした連想検索結果

5. 考察

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も，関連語を取り出すことができた。

課題としては，いかに適切な主要語を選定するかという点がある。今回対象にした資料名称では「金属製雛形品　鏃　複製」のように，最後尾の語が「複製」で終わるものが見受けられた。この場合，「複製」を主要語として取り出すべきかどうかという問題がある。

また，資料名称の末尾に括弧書きで注釈等が書かれる場合があり，これについては末尾の括弧内の記述は取り除いて選定を行った。

ほかにも末尾が接尾辞で終わる場合は，接尾辞を主要語とするよりは，その前の自立語を選定すべきであったり，並記されている場合は両方を主要語とすべきであったりと，主要語の選定方法についてはさらなる検討が必要である。

6. 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関する情報を用いて，関連語を計算する際に，主要語を選定することの効果について述べた。

博物館分野の形態素解析辞書の充実ならびに主要語の選定手法についてはさらなる検討が今後必要ではあるが，適切にグループ化された資料群から，関連語が抽出できることを示した。

考古資料以外の様々な資料についても同様の手法が適用できるか，またそれらが混在した場合に適切な関連語が抽出できるかについても，今後検討を要する。

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] 伝康晴，他：「コーパス日本語学のための言語資源-形態素解析用電子化辞書の開発とその応用-」日本語科学 22 pp.101-123, 2007.

[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，画像ミュージアム研究会，2009-03.