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，各館が個々の収蔵品に対して付与する資料名称をもとに，関連する収蔵品を見つけ出すための手法について述べる。

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。資料群はさらに階層化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には何らかの関連があると考えられる。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として用いることを考える。

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，文書検索と同様の手法を用いて取り扱うことを考えてみる。

4. 予備実験

国立歴史民俗博物館の収蔵品データから「亀ヶ岡遺跡出土縄文時代遺物」,「槻の木遺跡出土品」,「青森県南地方の仕事着コレクション」,「婚礼衣裳・婚礼用具及び生活用具」,「錦絵コレクション」,「野村正治郎衣裳コレクション」「水木家資料」の 7 つのコレクションを選び，それぞれのコレクションに含まれる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 6,468，資料群の総数は 72 であった。全体で頻度上位のものを図 1 に示す。

978	石鏃
451	石片類
247	石匙
165	掻器類
163	壷形土器
148	石玉及び石材
110	鉢形土器
99	石製具
97	石斧
88	石刃類
80	注口土器
73	台付鉢形土器
66	石棒
64	裁着
62	着物

図 1　高頻度の資料名称

このようにして作成した頻度表に対して，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する。具体的には，特定の資料名称を入力して，関連する資料群を計算し，さらにそこから関連する資料名称を計算することで，入力した資料名称に関連した資料名称を計算した。類似度の計算には GETA の WT_SMART の設定をそのまま使用した。「着物」という入力に対する結果を図 2 に示す。

```
Word: 着物
[Result]
 1: 肌着          9.03448696044758
 2: 帯　8.87234209980558
 3: 前掛          8.40845927199848
 4: 着物          8.07526877826644
 5: 羽織          7.8912101725996
 6: 袴　7.8912101725996
 7: 襦袢          7.8912101725996
 8: 男締め        7.8912101725996
 9: コート        7.8912101725996
10: 腰巻          7.8912101725996
11: 紐　7.8912101725996
12: ズボン        7.8912101725996
13: チョッキ      7.8912101725996
14: モーニング    7.8912101725996
15: 懐剣          7.8912101725996
16: 角巻          7.8912101725996
17: 上着          7.8912101725996
18: 帯締め        7.8912101725996
19: 二重廻し      7.8912101725996
20: 浴衣          7.8912101725996
```

図 2　連想資料名称検索の例

この結果を見てみると，対象とした資料群には遺跡からの出土品や，錦絵など様々な収蔵品を集めたものが混在していたが,「着物」という入力に対しては，着物に関連したものだけがうまく抽出されていることがわかる。

---

* 「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。

5. 問題の所在と手法の検討

先の例で比較的うまくいくのは，資料名称が短く，少数の語からなる場合である。たとえば，実験に用いた資料群の中には「染分綸子地藤草花模様絞縫箔小袖」といった資料名称も存在したが，これをひとかたまりの語として取り扱うと，出現頻度が低いため，同じ資料群に現れるものしか得られない。また，得られる結果も同様に非常に長い資料名称のリストになり，関連するものが適切に抽出できたとは言い難い。

そこで，資料名称をいったん形態素解析し，語を小さな単位に分割した上で，これらの頻度表を作成して同様の実験を行うことにした。現在入手可能な形態素解析用辞書は，必ずしも博物館の資料名称の解析に適しているとは言い難く，実際に随所に解析誤りが見られたが，辞書の構築は今後の課題として，語の斉一性が保証される既存の辞書として UniDic[9]を用いて，語（短単位）への分割を行った。たとえば，先の例は UniDic では「染/分/綸子/地/藤/草花/模様/絞/縫箔/小袖」と解析される。

そして，あらかじめ衣裳関連の 15 個のコレクションを選び，これらに含まれる資料名称の形態素解析を行い，解析結果から資料群毎の語の出現頻度をカウントした頻度表を作成し，同様の実験を行った。「振袖」という入力語に対する結果を図 3 に示す。

```
Word: 振袖
[Result]
 1: 鼠 1.7473409920529
 2: 菊花　　1.67894497490603
 3: 散らし　　1.6161026238565
 4: 紬 1.6161026238565
 5: 島 1.6161026238565
 6: 杢 1.6161026238565
 7: 鶴松　　1.59590548351114
 8: 流水　　1.57961458639608
 9: 藍 1.56403142155295
10: 色 1.56324472248727
11: 振袖　　1.56281420200995
12: 縮緬　　1.56191746818541
13: 縹 1.53504314324275
14: 下着　　1.53504314324275
15: 松葉　　1.53504314324275
16: 松樹　　1.53504314324275
17: 菊枝　　1.53504314324275
18: 若草　　1.53504314324275
19: 小花　　1.53504314324275
20: 折 1.53504314324275
```

図 3 形態素解析結果を用いた連想検索の例

形態素解析結果の小さな語の単位で，関連する語が並んでいることがわかる。また，これらの語から，それを含む元の資料名称をたどることは容易である。しかしながら，この結果を見ると，「振袖」にどう関連しているのかが不明な語が多く出現している。これは資料名称を構成する語をすべて同じ重みで計算に用いたためである。これにより，「小袖」と「振袖」が同じ資料群に出現するのと同じ重みで，「鼠」や「菊花」がカウントされた結果，これらが関連語として取り出されている。先に例に挙げた「染/分/綸子/地/藤/草花/模様/絞/縫箔/小袖」で，もっとも重要な語は最後尾の「小袖」で，それ以外はある構造をもって「小袖」を修飾しているのであり，これらを同列に扱うべきではない。

そこで，各資料名称の形態素解析結果の中からもっとも重要な主要語のみを取り出し，それらで頻度表を再構成して計算を行うことにした。ただ

し，主要語の選定手法については今後の課題とし，今回は簡易的に最後尾の語をもって主要語と見なすこととした。この結果を図 4 に示す。

```
Word: 振袖
[Result]
 1: 下着        3.08320072945125
 2: 打掛        3.08320072945125
 3: 振袖        2.92038875155844
 4: 裂 2.81634341224024
 5: 小袖        2.74974398722883
 6: 帷子        2.57903211446141
 7: 染 2.55904828088159
 8: 浴衣        2.35944100473588
 9: 袖 2.29063496948367
10: 帯 2.13843690655955
11: 上下        1.94386515113386
12: 模様        1.85512250709309
13: 肩衣        1.82501048684413
14: 前垂        1.82501048684413
15: 無 1.82501048684413
16: 間着        1.82501048684413
17: 丹前        1.82501048684413
18: 長袴        1.82501048684413
19: 腹掛        1.82501048684413
20: 筥迫        1.81623367373224
```

図 4 主要語を対象とした連想検索の例

必ずしも最後尾の語が主要語とは限らないため，少しノイズは残っているが，今回は比較的入力語に関連すると思われる語が並んでいる。ここから，これらの語を主要語として含む資料名称にたどり着くことは容易である。

6. 考察

資料群中の資料名称を構成する語の共起関係を用いた関連語抽出の利用方法について考えてみる。

まず，一般の利用者は，専門的な語彙がわからないか思いつかないということが考えられる。たとえば「着物」は思いついても，「被衣」を検索語として思いつくことはないだろう。そこで，連想検索の結果，誰でもが思いつく一般的な語から，より専門的な語が提示されることにより，それらを選んで検索を進めていくという使い方が考えられる。一度の連想検索では出現しなくとも，連想フィードバックを繰り返すことで，より専門的な語にたどり着く可能性がある。先の例において，「着物」から出発し，関連語として「浴衣」，「振袖」とたどって「被衣」に到達できることを確認している。

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関する情報を用いて，関連するものを検索する方法について提案した。

本手法の実現には，博物館分野の形態素解析辞書の充実と，主要語の選定手法が必要となるが，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。

現実の博物館の大規模なデータに対する本手法の有効性の検証は，今後の課題である。

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.