分類集合間の類似度を用いた博物館横断検索

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，各館から提供される分類語彙を集積した分類集合間の類似度を用いて，横断検索を実現する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

現状では，ほとんどの館がそれぞれ独自のサービスを提供しているため，

－ 検索方法が館によって異なる
－ ある館の情報を他の館の情報と関連付けて見ることが難しい

などの問題点があり，横断検索を困難にしている。そこで横断検索のためには，各館の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，扱う収蔵品が多岐にわたる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，対象が図書のようなテキストである場合に用いられる全文検索等の手法も適用できない。オントロジ[6]や概念辞書[9]を用いた手法も検討されているが，個々の収蔵品に関して詳細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. 分類集合

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，収蔵品の分類に関する情報を用いて，これを統合する試みが報告されている[10]。

ある館が所蔵する収蔵品に関して，通常何らかの分類がなされ，分類されたそれぞれの集合毎にラベルとして分類語彙が付与されている状況を考える。この作業は博物館毎，あるいは博物館の学芸員毎，コレクション毎に独立してなされるため，たとえ，同じ分類語彙が用いられていたとしても，それらが同じ内容を指し示すことは保証できない。また，分類は各館が収蔵する物に対して行われるため，原則としてそこに存在しない物に対してはなされない。このような条件の下で，個別に作成された分類を統合することを考える。

このためにまず，各館毎に類概念をまとめあげた分類語彙の構造を利用して，分類語彙の集合間で共通に用いられている語彙を手がかりに，それらの類似度を定義する。

たとえば，分類語彙として a,b,c,d をもつ分類集合 L，同じく b,c,d,e をもつ分類集合 M，a,e,f,g をもつ分類集合 N を考える。L と M の間の共通語彙は b,c,d であるのに対し，L と N の間の共通語彙は a のみである。このとき，N よりも M のほうが L により類似しているとみなす。このような分類集合対毎の類似度を

$$\frac{X\text{と}Y\text{の共通語彙数}\times 2}{X\text{の語彙数}+Y\text{の語彙数}}$$

で表す。先の例では L と M の間の類似度が 75%，L と N の間の類似度が 25%となる。

4. 横断検索における分類集合の利用

分類集合を用いた横断検索について検討するために，[10]と同じく国立歴史民俗博物館の収蔵品データから衣類に関する 15 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などをもとに，コレクション毎に分類語彙を設定した上で，これを分類集合とみなし，それらの間の類似度を求めた。使用した分類集合（コレクション）のリストを図 1 に示す。

```
<collection id="H-35" name="野村正治郎衣裳コレクション"/>
<collection id="H-36" name="徳川（有栖川宮）実枝子所用衣裳"/>
<collection id="H-1182" name="海軍礼服"/>
<collection id="H-1314" name="柳沢家伝来服飾資料"/>
<collection id="H-1359" name="幕末・明治初期服飾資料"/>
<collection id="H-1395" name="大正末～昭和初期着物類"/>
<collection id="H-1694" name="近・現代女性洋装衣裳"/>
<collection id="F-7" name="青森県の衣生活用具"/>
<collection id="F-8" name="青森県南地方の仕事着コレクション"/>
<collection id="F-148" name="婚礼衣裳・婚礼用具及び生活用具"/>
<collection id="F-150" name="秋田・青森・岩手の仕事着及びこぎん・ひしざし模様見本"/>
<collection id="F-211" name="秋田の衣服"/>
<collection id="F-220" name="津軽・南部の衣服"/>
<collection id="F-337" name="東京近郊農家の衣生活資料"/>
<collection id="F-338" name="加藤家衣生活資料"/>
```

図 1　分類集合のリスト

次に，例として「野村正治郎衣裳コレクション」と他の分類集合との間の類似度を図 2 に示す。この結果から「野村正治郎衣裳コレクション」は「徳川（有栖川宮）実枝子所用衣裳」との類似度が最も高いことがわかる。このとき，「野村正治郎衣裳コレクション」で用いられている分類語彙は「衣裳」「小袖」「振袖」「帷子」「被衣」「襦袢」「腰巻」であり，「徳川（有栖川宮）実枝子所用衣裳」で用いられている分類語彙は「衣裳」「袿」「帯」「綴織帯」「振袖」であった。

一方，「幕末・明治初期服飾資料」との類似度は18%であった。こちらで用いられている分類語彙は「服飾資料」「羽織」「下着」「間着」「錦帯」「形染」「肩衣」「言袴」「合羽」「小袴」「小袖」「振袖」「染帯」「染裂」「前垂」「袖無」「打掛」「丹前」「長袴」「長着」「半袴」「腹掛」「紋裂」「浴衣」「帷子」「襴帯」と多岐にわたっている。

```
<collection id="H-36" name="徳川（有栖川宮）実枝子所用衣裳" similarity="33%" />
<collection id="H-1182" name="海軍礼服" similarity="0%" />
<collection id="H-1314" name="柳沢家伝来服飾資料" similarity="0%" />
<collection id="H-1359" name="幕末・明治初期服飾資料" similarity="18%" />
<collection id="H-1395" name="大正末～昭和初期着物類" similarity="0%" />
<collection id="H-1694" name="近・現代女性洋装衣裳" similarity="9%" />
<collection id="F-7" name="青森県の衣生活用具" similarity="8%" />
<collection id="F-8" name="青森県南地方の仕事着コレクション" similarity="0%" />
<collection id="F-148" name="婚礼衣裳・婚礼用具及び生活用具" similarity="10%" />
<collection id="F-150" name="秋田・青森・岩手の仕事着及びこぎん・ひしざし模様見本" similarity="0%" />
<collection id="F-211" name="秋田の衣服" similarity="9%" />
<collection id="F-220" name="津軽・南部の衣服" similarity="0%" />
<collection id="F-337" name="東京近郊農家の衣生活資料" similarity="0%" />
<collection id="F-338" name="加藤家衣生活資料" similarity="5%" />
```

図 2　分類集合間の類似度

このとき，「野村正治郎衣裳コレクション」を見ていて，類似する他の収蔵品を見たい場合に，最も類似度の高い「徳川（有栖川宮）実枝子所用衣裳」を検索するといった利用法が考えられる。たとえば，「徳川（有栖川宮）実枝子所用衣裳」は「小袖」という分類語彙を含んでいないが，「野村正治郎衣裳コレクション」の「小袖」を見ているときに，分類集合間の類似度が最も高いものを検索することにより，「小袖」を含まない「徳川（有栖川宮）実枝子所用衣裳」を取り出すことができるようになる。

このように特定の分類集合を軸として，類概念による検索範囲の拡張を行う場合に，分類集合対の類似度は有用であると考えられる。

一方で，特定の分類集合ではなく，検索語彙として「小袖」を指定した場合には，これを含む分類集合としては「野村正治郎衣裳コレクション」と「幕末・明治初期服飾資料」の二つが出てくる。しかし，両者はともに「小袖」は含んでいるものの，類概念として列挙されている他の分類語彙のカバーする範囲が大きく異なる。次にこの違いを表現する方法について考察する。

5. 分類集合の範囲の違い

同じ分類語彙を含む分類集合でも，分類の観点の違い等により，異なる類概念がまとまっていることがある。また，集合としてカバーする範囲も，比較的狭い範囲のものから広い範囲のものまでが混在する可能性がある。これらの差異を表現するための指標として，テキスト情報検索の分野で用いられている TF-IDF を利用することを検討する。

TF-IDF は文書中で重要とみなされるキーワードを抽出するための指標で，TF (Term Frequency) と IDF (Inverse Document Frequency) の積によって算出される。文書中のある単語 i の TF 値は，単純に文書 j の中での i の出現頻度 n_i で表されることもあるが，今回は j 自体の大きさに対する相対的な値として，文書 j の単語総数 $\sum_k n_k$ に対する i の出現頻度 n_i ではかることとする。また，IDF 値は総文書数 $|D|$ を単語 i を含む文書数 $|\{d : i \in d\}|$ で除算した値の対数値である。計算式を次に示す。

$$tfidf = tf \bullet idf = \frac{n_i}{\sum_k n_k} \log \frac{|D|}{|\{d : i \in d\}|}$$

この文書を分類集合と置き換えて，各分類集合の分類語彙毎に TF-IDF 値を求める。

TF 値により，ある分類集合の大きさを表すことになる。ここで，多くの分類語彙を含む分類集合ほど，より一般的であるとみなす。一方，IDF 値はある分類語彙について，分類集合に関わりなく一定であり，多くの分類集合に出現する語彙ほど一般的であるとみなすことで，その語彙自体がどの程度一般的かという指標となる。

これを先の例に当てはめると，「野村正治郎衣裳コレクション」における「小袖」の TF 値が 1/7，IDF 値が log(15/2)，TF-IDF 値が 0.125 であるのに対し，「幕末・明治初期服飾資料」の「小袖」の TF 値は 1/26，IDF 値が同じく log(15/2)，TF-IDF 値が 0.034 となる。これより，「野村正治郎衣裳コレクション」の「小袖」のほうが，より一般性が低い，言い換えれば代表性が高いということができる。

一方，「幕末・明治初期服飾資料」における「長着」については，「長着」という分類語が 8 つの分類集合に現れるため，TF 値が 1/26，IDF 値が log(15/8)，TF-IDF 値が 0.011 となり，同じ分類集合中の「小袖」よりもさらに一般性が高い（代表制が低い）ということができる。

これらの指標を先の類似度と組み合わせることで，特定の分類集合ではなく，任意の検索語彙から出発した場合の横断検索の手順は次のようになる。

(1) 入力された検索語彙（たとえば「小袖」）を分類語彙として持つ分類集合を検索する。
(2) 得られた分類集合（先の例では「野村正治郎衣裳コレクション」と「幕末・明治初期服飾資料」）のうち，TF-IDF 値が最も大きいもの（「野村正治郎衣裳コレクション」の「小袖」）から順に代表的な検索結果として返す。
(3) 次に，類概念に拡張した検索結果として，最も大きな TF-IDF 値をもつ分類語彙を含む分類集合（先の例では「野村正治郎衣裳コレクション」）に含まれるものを返す。
(4) さらに拡張する場合には，(3)の分類集合に対して類似度の高い分類集合（先の例では「徳川（有栖川宮）実枝子所用衣裳」）に含まれるものから順に返す。

この一連の動作をインタラクティブに行うことで，利用者の意図した横断検索を実現する。基本的な考え方としては，TF-IDF 値によって表現される語彙の代表性によって検索の絞り込みを，逆に分類集合間の類似度によって拡張を行うというものである。

本稿では，範囲を限定して，非常に小さなサンプルを対象に考察を行ったが，本手法をより広範なデータに対して適用した場合に，望ましい結果が得られるか否かについては実証的な実験が必要である。

また，本手法では，同じ分類集合内で類概念としてあげられている語彙に関する情報（共起情報）を用いていないため，同じ語が異なる文脈で用いられていた場合に，それらの区別が十分にできない。これについては，実例に基づき検討していく必要があると考えている。

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，博物館から提供可能な分類語彙を集積した分類集合を用いて，記述内容に基づく横断検索を行う方法について検討を行った。

個々の館から提供される分類語彙がカバーする範囲は部分的であるが，共通語彙を手がかりにして，分類集合間の類似度を計算することによって，他の館の収蔵品と関連づけた横断検索に利用することができる。

また，各語彙の代表性を TF-IDF 値を用いて計算することにより，カバーする範囲の異なる分類集合の中から，より代表的なものから選択的に示すことが可能となる。

今後，実際の博物館で用いられている広範な分類語彙を対象にした実証的な実験が必要であると考えてい

る。また，代表性以外の指標を用いた検索範囲の絞り込みや，種々の文脈情報の利用，利用者毎の個別化とチューニングは，今後の検討課題である。

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：博物館情報横断検索における分散オントロジの検討，画像ミュージアム研究会，2006-03.

[8] 山田篤，他：博物館情報検索のためのオントロジ・ユースケースの検討，画電年次大会，2006-06.

[9] 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.

[10] 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.