1P-02　ディジタル展覧会システムにおけるキーワード検出機能の検討

河野　聡子　　斉藤　典明
{s.kouno, saito.n}@rdc.east.ntt.co.jp
NTT 東日本　研究開発センタ

1．はじめに

電子美術館システムによる新たな美術教育への取り組みが始まっている。例えば、所蔵作品をディジタル化し、作品のタイトルや作者名、解説といったテキストデータと合わせて、ネットワーク上で公開する試みが、国内、海外を問わず多くの電子美術館で行われている。

電子美術館において、所蔵作品を紹介するためのコンテンツ（以下、絵画コンテンツ）に含まれる解説には、他作品との相違点や関わりについて述べられているものも多く、閲覧中のコンテンツから、それら比較対象も関連情報として参照できる仕組みが求められている。これに対して、絵画コンテンツのハイパーテキスト化は有力な手段の一つといえる。しかし現状では、その関連付けは人手によるもので、大量の情報が蓄積されたデータベースに対応しきれないという問題がある。

本稿では、キーワードによる絵画コンテンツの自動ハイパーテキスト化を目的とし、関連するコンテンツ間においてハイパーリンクを張るべき対象キーワードの抽出方法について検討する。

2．作品解説文における特徴と関連

著者らはディジタル化した絵画や写真をテキストデータ（タイトル、作者名、解説）と合わせて、ネットワーク上で閲覧できるディジタル展覧会システム[1]を開発している。その美術作品に関する解説内容（300 字から 500 字程度）は、主に次のようなものがあげられる。

・他作品との比較
・作品に描かれた土地や人物などに関する解説
・時代的背景や作者の経歴に関する紹介

絵画に関する解説は上記のような特徴を持つことから、文章中には作者名や作品のタイトルが頻繁に出現し、これらのキーワードはコンテンツ間の関連を見つけ出す上で重要な手がかりといえる。

ディジタル展覧会システムでは、これら絵画コンテンツの特徴を用いて、以下 4 つのパターンに該当するものに対し、キーワードによる絵画コンテンツのハイパーテキスト化を目的としている。

① 閲覧中のコンテンツが任意のコンテンツを一意に表すキーワードを持つ場合
② 任意のコンテンツが閲覧中のコンテンツを一意に表すキーワードを持つ場合
③ ①、②が同時に成立する場合
④ 両コンテンツがその特徴を表すキーワードを共有する場合

①～③は片方あるいは両方のコンテンツが相手を確実に示す名前や番号等を引用している場合が考えられる。④は解説される人物や土地、年代等が共通しているコンテンツ間の結びつきを示す。

Keyword Extraction Method at Digital Museum System
Satoko KONO　Noriaki SAITO
Research and Development Center NTT EAST
3-9-11 Midori-Cho Musashino-Shi Tokyo 180-8585 Japan

3．システム構成

3．1 ディジタル展覧会システム

ディジタル展覧会システムでは、予めコンテンツから抽出したキーワード（タイトル、作者名、関連キーワード）に対してハイパーリンクを生成する。図 1 に示すように、閲覧者がコンテンツを参照する際、閲覧コンテンツのテキスト中に他のコンテンツのキーワードが含まれていれば、自動的にハイパーリンクを生成し、キーワードを介して閲覧者に関連コンテンツを提供する。この予め抽出されるキーワードのうちタイトルと作者名はコンテンツのメタデータより抽出され、主に①～③の関連コンテンツを結び付ける。一方、関連キーワードは各コンテンツに含まれる解説文から、④のような関連コンテンツの接点となる単語を抽出しておく必要がある。以下にその関連キーワードの抽出方法を検討する。

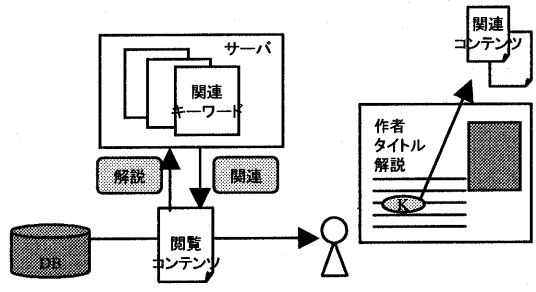

図 1　システム概要

3．2 関連キーワードの抽出

本方式では、形態素解析により抽出された単語を、コンテンツに現れる人物名（作者名）の類似性と、tfidf 法[2]により定義される重要度によって絞り込み、関連キーワードとして抽出する。図 2 を例にとるとまず各コンテンツに対して、コンテンツに出現する人物名（作者名）を要素とした人物ベクトル P_d と、コンテンツに含まれる解説文中の単語を要素とした

キーワードベクトル T_d を作成する。人物ベクトル P_d によりコンテンツを分類し、図2の斜線部のように、登場人物が類似するコンテンツのキーワードベクトル T_d の和から、重要度の高い単語を関連キーワードとして抽出する。以下に、その詳細をステップ1～3で説明する。

ステップ1：人物ベクトル P_d の作成

まず、データベース中に蓄積された任意のコンテンツ d の人物ベクトル P_d を作成する。

$$P_d = (c(d, 1), c(d, 2), \ldots, c(d, i)) \quad (1)$$

c(d,i)はコンテンツ d のメタデータもしくは解説文中に出現する人物名（作者名）i の出現頻度である。

ステップ2：キーワードベクトル T_d の作成

（キーワード候補 t の切り出し）

次に、コンテンツ d のキーワード候補 t として、形態素解析により、コンテンツに含まれる解説文中から名詞と未定義語を抽出する。また、一部複合語に関しては、括弧や記号で強調された部分の抜き出しや、形態素解析済みの単語列パタンマッチによって行う。

（tfidf 法による重要度 w(d,t)の算出）

キーワード候補 t の重要度 w(d,t)を tfidf 法により算出する。

$$w(d, t) = TF(d, t) \times IDF(t) \quad (2)$$

$$IDF(t) = \log \frac{\text{データベース中の総コンテンツ数}}{\text{キーワード候補t が現れるコンテンツ数}} \quad (3)$$

ただし TF(d,t)は、コンテンツ d におけるキーワード候補 t の出現頻度を、コンテンツ d が持つ形態素数(延べ数)で割った値である。

（キーワードベクトル T_d の作成）

前段階で切り出したキーワード候補 t を要素、その重要度 w(d,t)を要素値としたコンテンツ d のキーワードベクトルを T_d で表す。

$$T_d = (w(d, 1), w(d, 2), \ldots, w(d, t)) \quad (4)$$

ステップ3：関連キーワード t'の抽出

最後に、ステップ1で作成した人物ベクトル P_d を用いて、以下の式から類似度を求め、ある一定のしきい値により、類似した登場人物をもつ集合にコンテンツをクラスタリングする。

$$\text{類似度} = \frac{\text{2つの作者ベクトルの共通の要素の値の和}}{\text{2つの作者ベクトルの要素の値の和}} \quad (5)$$

人物ベクトルP_d				コンテンツd	キーワードベクトルT_d
・・ モネ	マネ	・・	ルノワール		
・・ 0	1	・・	0	レモン	{(静物画, 2.3)、(果物, 2.1) … }
・・ 0	0	・・	1	ぶらんこ	{(モンマルトル, 2.5)、(木漏れ日, 1.9)、..}
・・ 1	1	・・	0	草上の昼食	{(サロン, 3.1)、(オルセー美術館, 1.5) ..}
・・ 1	1	・・	0	船遊び	{(アルジャントゥイユ, 2.8)、(セーヌ川, 2.2) ..}
・・・	・・・	・・・			・・・ ・・・ ・・・

図2　作者ベクトルとキーワードベクトル例

ここで、各クラスタのキーワードベクトル T_d を足し合わせたとき、重要度 w'(d,t)の高いキーワード候補 t を関連キーワード t'として抽出する。

3．3 実施例

以上の手順に従い「小学館ウィークリーブック 週刊 美術館」（発行；小学館）を一部電子化し、関連キーワードを抽出した。その結果例を図3に示す。人手によって検証したところ、図3の「カミーユ」のように、全コンテンツ対象では低頻度の単語でも、特定のコンテンツ間で重要と思われる単語が切り出されており、また、この関連キーワードによってハイパーリンクが張られたコンテンツの中に、その内容から明らかに関連がないと思われるようなコンテンツは含まれていなかった。

【タイトル】アトリエ船で描くモネ 【作者】マネ

【解説】アトリエに仕立てた船で製作中のモネ。モネは良く<u>セーヌ川</u>に浮かべたアトリエ船を利用していた。奥に見えるのは<u>カミーユ</u>。

関連キーワード:カミーユ

■【タイトル】庭のモネ一家 【作者】マネ
マネのすばやい…モネの「カミーユ（緑衣の女性）」がサロンに入選し…、

■【タイトル】庭の女たち 【作者】モネ
1893年、サロンに落選し…夏の庭にたたずむ女性のモデルはカミーユ。

図3　絵画コンテンツの解説文と関連キーワード（例）

4．まとめ

本稿では、キーワードによってハイパーリンクを自動生成するディジタル展覧会システムにおいて、ハイパーリンクを張る関連キーワードの抽出方法について検討した。

本方式では、絵画コンテンツの解説文中に、人物名（作者名）が頻繁に出現し、それがコンテンツ間の関連を発見する大きな手がかりとなることに着目した。そこで、解説文に現れる作者名によりコンテンツを分類し、その中で重要度の高い単語を関連キーワードとして抽出する手法を考案した。今後は、関連キーワードによって結び付けられるコンテンツの妥当性と適用領域、および計算コストについて検証していく予定である。

参考文献

[1]関良明，日高哲雄，斉藤典明："コンテンツ連携技術を応用したディジタル展覧会システム"，電子情報通信学会，2000.5

[2]G.Salton and C.Buckley, "Term weighting approaches in automatic text retrieval", Information Processing and Management, Vol.24, pp.513-523, 1988