国立国会図書館の件名標目表と蔵書目録を利用した語彙の構造化

本間 維† 永森 光晴‡ 杉本 重雄‡
筑波大学図書館情報専門学群† 筑波大学図書館情報メディア研究科‡

1. はじめに

現在，RDFやOWLなどを用いて記述されたWeb情報資源が増加しており，それらによるWebオントロジーの構築が現実のものとなりつつある．増え続けるWeb情報資源からオントロジーを構築するためには，計算機処理により関連するWeb情報資源同士の結び付きを見つけることが必須となる．しかし，Web情報資源のメタデータ記述に用いられる語彙が多様であるため，計算機が情報のつながりを判断するには，語彙を横断的に理解し処理する必要がある．

本研究では，国立国会図書館が提供する件名標目表と蔵書目録に着目し，これらを利用した語彙構造化手法を提案する．多様な語彙を一つの体系として構造化し，SKOS[1]形式で記述することにより，各語彙を横断的に利用できるようにする．

2. 語彙構造化手法

フォークソノミーなどのタグ付けで用いられる語彙は，統制や構造化が行われていないため，語彙横断利用において障害となる．こうした語彙を構造化するため，本研究では国立国会図書館件名標目表(NDLSH)[2]を利用した語彙構造化手法を提案する．NDLSHは構造化された語彙であり，任意の語をNDLSHの適切な語(件名標目)に対応付けることで，語彙の横断的利用を可能にする．

2.1 国立国会図書館件名標目表

本研究で利用するNDLSHは，国立国会図書館(NDL)が各蔵書の主題を表すために用いている語彙である．NDLSHは，1)語彙が豊富である，2)定期的なメンテナンスが行われている，3)米国議会図書館件名標目表(LCSH)との対応付けが行われている，4)件名標目同士が上位語・下位語・関連語などの関係で構造化されている，等の特徴を持つ．2008年度版NDLSHの収録語数は17,953語で，参照形(各件名標目の別名)を含めると47,816語になる．

NDLSHの件名標目には，日本十進分類法(NDC)[3]による代表分類記号が付与されている．代表分類記号とは，件名標目が表す主題の一般的な分類を示すものである．NDLSHの各件名標目は，当該件名標目と併せて付与される傾向にあるNDC分類記号を代表分類記号として持つ．

2.2 蔵書目録を利用した語彙構造化

任意の語に対応する適切な件名標目をNDLSHから抽出するために，本研究ではNDL蔵書目録に注目した．NDL蔵書目録では，書籍の主題を表す件名標目が付与されている．同様に，書籍の検索対象項目であるタイトル等に含まれる語も書籍の主題を表す語である．よって，書籍検索を行った結果から抽出される件名標目は，検索に用いた任意の語と対応付ける件名っ標目の候補となる（図1）．対応付けの決定に用いる各件名標目のスコアは，抽出された全件名標目に占める割合を基本値とした．また，検索結果に含まれるNDC分類記号も抽出し，そのいずれかと同じ分類記号を代表分類記号として持つ件名標目はスコアを高くしている．

上記のモデルを用いて語と件名標目の対応付けを行うことで，任意の語彙に含まれる語をNDLSHの構造に組み込む．図2は，構造化されていない任意の語彙に含まれる語「Webサイト」「YouTube」「ニコニコ動画」をNDLSH中の件名標目に対応付けた例である．上位語である件名標目「ホームページ」に「Webサイト」を，下位語である件名標目「動画共有サイト」に「YouTube」と「ニコニコ動画」を対応付けることで，「Webサイト」の下位概念が「YouTube」と「ニコニコ動画」であるという構造を表現できる．

任意の語を他の語彙の適切な位置に対応付ける研究としては，上田らによるWikipedia等のWeb情報資源と基本件名標目表(BSH)を利用した手法がある[4]．上田らは任意の語に対する関連語をWeb情報資源から抽出し，それらの語から成る合成ベクトルを用いて類似する件名標目を決定しているのに対し，本研究では任意の語と関係する書籍を探索し，その主題傾向から対応する件名標目を決定している．

3. システムの作成

本手法に基づいて，任意の語と対応する件名標目を提示するシステムの作成を行った(図3)．本システムは，入力された語を書籍検索クエリとして使用し，結果に含まれる件名標目を集計・スコアリングして表示

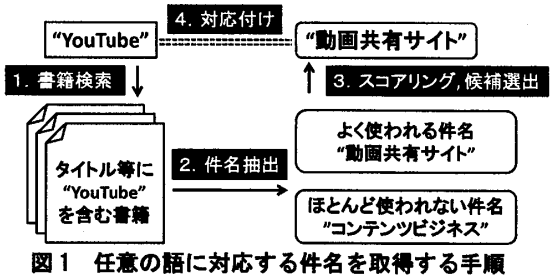

図1　任意の語に対応する件名を取得する手順
例：”YouTube”に対応する件名を取得する場合

"Adding Structure to Unstructured Subject Vocabularies by Linking Terms using NDL Subject Headings and Catalogs."
† Tsunagu Honma. (tsunagu.honma@a.slis.tsukuba.ac.jp)
† School of Library and Information Science. U of Tsukuba.
‡ Mitsuharu Nagamori. Shigeo Sugimoto.
‡ Graduate School of Library, Information and Media Studies. U of Tsukuba.

する．書籍の検索には，国立国会図書館デジタルアーカイブポータル[5]で提供されるAPIを利用した．NDLSHは，本研究室で開発したNDLSH閲覧システムHANAVI[6]で用いられているデータを利用した．

4. 評価実験

本システムを用いて，IT用語辞典e-Words[7]の新着用語50語，アクセス数上位の用語50語に対し，対応する件名標目の抽出を行った．表1と表2は件名標目抽出結果の一部を抜粋したものである．抽出された件名標目の妥当性を人手で確認した結果，新着で4割，アクセス数上位で6割ほどの件名標目が適当であった．

アクセス数上位の用語における正答率が新着用語の場合より高い理由として，注目を集めている語は書籍の主題となりやすい，新しい語が示す主題は書籍が出版されていないといったことが考えられる．

実験から，1)書籍の主題となりやすい語は，十分な数の書籍を検索結果として取得し，適切な件名標目をうまく提示できる，2)単独で書籍の主題となりにくい語(例：SSID)は書籍検索結果が0件あるいは非常に少ない数となるため件名標目を提示できない，ということが分かった．また，本システムでは用語の背景にある文脈や意図が書籍検索時に反映されないため，計算機の処理速度向上に用いる「キャッシュ」が，経済・経営分野の件名標目「キャッシュフロー計算書」に対応付けられてしまうといったように，多義語の対応付けに失敗する場面も見られた．

5. おわりに

本稿では，件名標目表と蔵書目録を利用した語彙構造化手法を提案し，任意の語に対応する件名標目を導くシステムについて述べた．本手法で提示される件名標目が分散する場面において，NDLSHの代表分類記号を用いて結果をクラスタリングすると，結果を2・3種類にまとめることができた．今後は，クラスタリングすべき条件の判別をどのように行うかを検討する．また，検索結果の不足により，件名標目を1つも取得できない場面もある．この場合，Wikipedia等から類似する語を取得し書籍再検索を行うことで結果を変化させ，件名標目の取得を可能にする必要がある．

この他，多義語と件名標目の対応付けを可能にするため，書籍検索クエリの作成あるいは抽出した件名標目のスコアリングにおいて，語の文脈や意図を付与する仕組みも考案する．

なお，本研究では，国立国会図書館が提供するNDLSHデータと蔵書目録データを利用した．

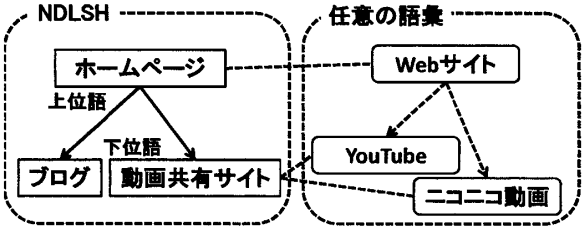

図2 NDLSHを用いた語彙構造化の例

表1 新着用語に対する件名標目抽出結果

入力	抽出された件名標目
タイムスタンプ	時刻認証
インタラクティブ	データ伝送
エンジン	検索エンジン
信頼性	信頼性(工学)，情報システム
DBA	データベース
カレント	糖尿病
保守	日本，アメリカ合衆国
エンハンス	エンハンスメント(医療)
トランク	倉庫業
ピア	PR
フィックス	コンピュータ・グラフィックス

表2 アクセス数上位の用語に対する件名標目抽出結果

入力	抽出された件名標目
SSD	半導体記憶装置
インタフェース	人間工学，コンピュータ
Java	プログラミング(コンピュータ)
HTML	ハイパーテキスト
Bluetooth	無線通信
SSL	ネットワークセキュリティ，暗号
RSS	ホームページ
Cookie	ビスケット
ルータ	インド
キャッシュ	キャッシュフロー計算書，情報処理
SSID	(該当なし)

参考文献

[1]SKOS. http://www.w3.org/TR/skos-reference/
[2]国立国会図書館件名標目表.
http://www.ndl.go.jp/jp/library/data/ndl_ndlsh.html
[3]もりきよし原編，日本図書館協会分類委員会改訂．日本十進分類法．新訂9版，日本図書館協会，1995.
[4]上田洋，村上晴美．蔵書検索のためのWeb情報源を用いた件名の提案．情報処理学会研究報告．2006.
[5]国立国会図書館デジタルアーカイブポータル.
http://porta.ndl.go.jp/
[6]HANAVI. http://raus.slis.tsukuba.ac.jp/
[7]e-Words. http://e-words.jp/

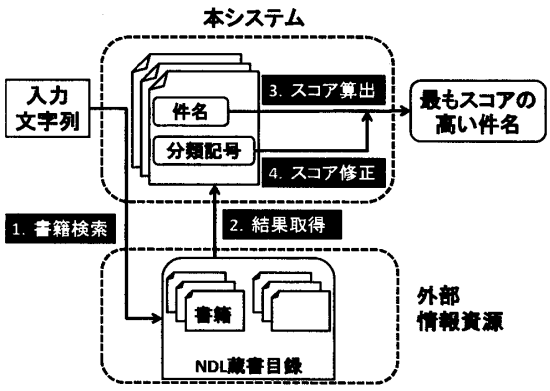

図3 本システムにおける処理の流れ