F-006

# 日本十進分類法と Wikipedia のカテゴリを用いた蔵書検索クエリの拡張 － 蔵書検索結果を用いた語彙構造化 －

本間 維† 永森 光晴† 杉本 重雄†
筑波大学大学院図書館情報メディア研究科‡

## 1. はじめに

現在, Web 上ではキーワードやタグといったメタデータを付与して公開される情報資源が多数存在している. 付与されるキーワードやタグは, 主にコンテンツの主題を表す語句が用いられ, キーワードを対象とした検索を行うことで, 同一主題の情報資源をまとめて取得することができる. CiNii[1]などの論文検索システムでは各論文にキーワードが付与され, はてなブックマーク[2]や Flickr[3]などのコンテンツ共有サイトではソーシャルタギングが行われている. しかし, 情報資源のキーワードとして付与される語句は, 多くの場合は各個人の頭に浮かんだ語句, あるいは特定のコミュニティで用いられる語彙から選択されることが多く, 共通の語彙は利用されていない. このため, 主題表現に用いられる語句は分散し, 結果としてキーワードによる同一主題コンテンツの検索が難しくなる. タグ付けに用いる共通語彙を作成する試みとして Common Tag[4]という活動もあるが, その利用はまだ進んでいない.

一方, 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, それらを連携させる Linked Open Data[5]と呼ばれる試みが盛んになっている. 増え続ける Web 情報資源による Linked Open Data を実現するためには, 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.

上記のような語彙に起因する問題を解決するため, 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, 蔵書検索結果から抽出される件名標目を利用した語彙構造化手法を提案する. これは, 各語彙を件名標目表の語彙に集約することで, 件名表目標をハブとした語彙横断利用を試みるものである. また本稿では, Wikipedia[6]の記事カテゴリや日本十進分類法(NDC)[7]による蔵書検索クエリの拡張について述べ, 提案手法の評価を行う.

## 2. 語彙構造化手法

多様な語彙の横断的利用を実現するためには語彙間の対応付けが求められるが, 論文のキーワードやフォークソノミーなどのタグ付けで用いられる語彙の多くは個々に独立しているため, 何らかの手法で対応付ける必要がある. 語彙間を対応付ける手法として各語彙をスター型で対応付けていく方法が考えられるが, 対応付ける語彙が増加した場合, 各語彙間で対応付けを行うコストが大きくなる. 多様な語彙を結びつけるには, ハブ型の対応付けの方が適しているが, その場合は任意の語彙をハブとなる語彙へ結びつける手法が問題になる.

ハブ型モデルを利用した語彙の構造化手法として, 我々は国立国会図書館件名標目表(NDLSH)[8]と NDC を利用した語彙構造化手法を提案した[9]. 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. その上で, NDC を用い, 適切な主題分野の件名標目に対応付けられるよう修正を行う.

### 2.1 国立国会図書館件名標目表(NDLSH)

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. 図 1 は件名標目の記述例である.

NDLSH の件名標目には, NDC 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. 図 1 の例では, 件名標目“意味論”が付与される資料は”007.1（情報理論）”や”801.2(語源学. 意味論)”などの NDC 分類記号が併せて付与される傾向にあることを表している.

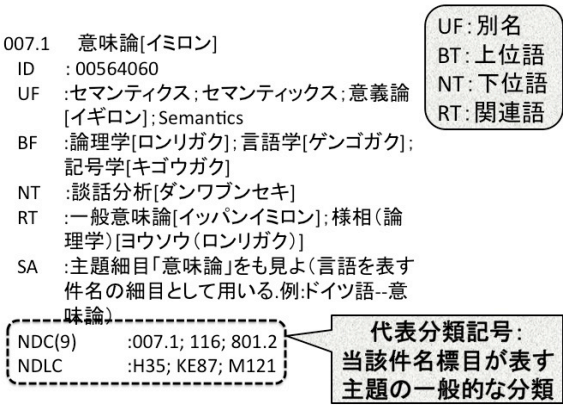


**図 1 NDLSH の記述内容**
**例：件名標目”意味論”**

“Improvement of Matching Functions by NDC and Wikipedia: Adding Structure to Unstructured Subject Vocabularies by Linking Terms using Retrieval Result of Book”
† Tsunagu Honma. (tsunagu.honma@a.slis.tsukuba.ac.jp). Mitsuharu Nagamori. Shigeo Sugimoto.
‡ Graduate School of Library, Information and Media Studies. U of Tsukuba.

表 1　NDC の第 1 次区分

| | | | |
|---|---|---|---|
| 0 | 総記 | 5 | 技術. 工学 |
| 1 | 哲学 | 6 | 産業 |
| 2 | 歴史 | 7 | 芸術. 美術 |
| 3 | 社会科学 | 8 | 言語 |
| 4 | 自然科学 | 9 | 文学 |

### 2.2 日本十進分類法(NDC)

NDC は, 日本図書館協会(JLA)が図書の分類を目的として考案した図書分類法であり, 日本の図書館における代表的な分類法として利用されている. NDC において, 分類記号はアラビア数字とピリオドで表現され, 各分類記号にはラベルが定義されている. 例えば”3”は「社会科学」(表 1 参照), ”33”は「経済」, ”337”は「貨幣, 通貨」と定義されていて, 経済全般に触れている資料には”33”という分類記号が付与される.

### 2.3 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 蔵書目録は, NDL が所蔵する資料に付与されているメタデータの集合である. これを用いて, 以下の手順で語と件名標目の対応付けを行う(図 2).

1. 任意の語をクエリとして, 蔵書検索を行う.
2. ヒットした資料には 0 から 3 つほどの件名標目が付与されているので, これを抽出・集計する.
3. ヒットした資料には 0 から 1 つの NDC 分類記号が付与されているので, これを抽出する.
4. 2 の集計結果をもとに, 出現頻度の高い件名標目をスコアの高い対応付け候補とする.
5. 3 で抽出された NDC 分類記号と同じ記号を NDC 代表分類記号として持つ件名標目は, そのスコアを上方修正する. 図 2 の例では, ヒットした資料から NDC”547.483”などが抽出されるため, これを NDC 代表分類記号として持つ件名標目”動画共有サイト”がスコア修正の対象となる.
6. 任意の語に対応する件名標目を, スコアの高い順に出力する.

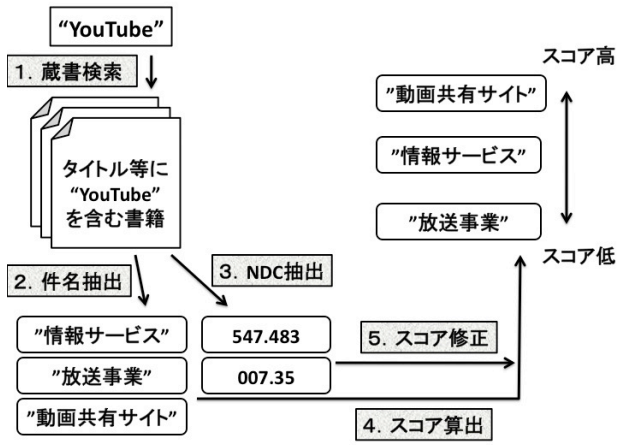


図 2　任意の語に対応する件名候補を提示する手順
例：”YouTube”に対応する件名候補を提示

上記のモデルを用いて語と件名標目の対応付けを見つけることで, 任意の語彙に含まれる語を NDLSH にマッピングする.

### 2.4 蔵書検索クエリの拡張

蔵書検索の結果が 0 件である, あるいはヒットした資料に件名標目が付与されていないなどの理由で, 対応付ける件名標目の候補が選出できない場合がある. また, 件名標目と対応付けたい語が多義語である場合にその文脈を特定できず, 適切な件名標目を提示できない場面も考えられる. 例えば, 一時記憶領域を表す”キャッシュ”に対して, ”キャッシュフロー”や”キャッシュカード”など経済に関する件名標目が提示される場合がある. こうした問題に対して, 本研究では, 1) 蔵書検索クエリに用いる語の置き換え, 2) NDC 第 1 次区分による文脈の付与という 2 つの手法で解決を図った.

蔵書検索クエリの置き換えには, Wikipedia のカテゴリを用いた. Wikipedia のカテゴリとは, 各記事を分野別にまとめるためのラベルであり, その語彙は統制されていない. 蔵書検索クエリに対して提示される件名標目が 0 件であれば, Wikipedia の記事から当該蔵書検索クエリをタイトルに含むものを検索し, ヒットした記事のカテゴリ名を新たな蔵書検索クエリとする. 例えば, ”フルハイビジョン”を蔵書検索クエリとしたときの検索結果が 0 件であれば, Wikipedia からタイトルに”フルハイビジョン”を含む記事を取得する. 取得した記事のカテゴリ名が”テレビ放送”と”デジタル方式”であれば, それらを新たな蔵書検索クエリとして再度件名標目の提示を試みる.

文脈の付与には, NDC 第 1 次区分(表 1 参照)を利用する. これは NDC 分類記号の 1 文字目に該当するものであり, 0 から 9 の 10 種類がある. この記号を蔵書検索クエリの背景にある文脈として指定し, 一致する代表分類記号を持つ件名標目をより適切な件名標目とする. 例えば, ”キャッシュ”という検索語に対しては”キャッシュフロー”など経済に関する件名標目が多く抽出されるが, NDC 第 1 次区分の”0”を文脈として与えることで, ”0”から始まる代表分類記号”007”を持つ”情報処理”などがより適切な件名標目の候補として提示される.

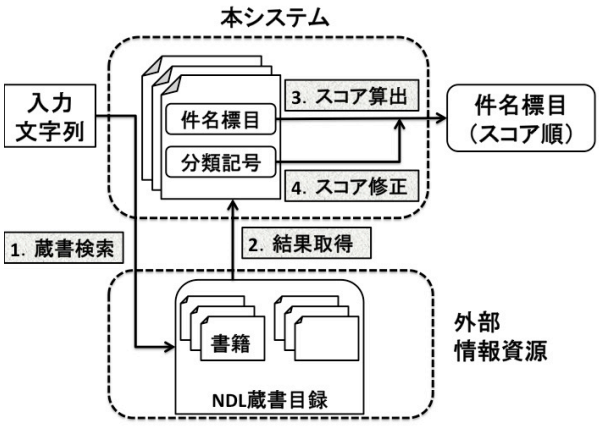


図 3　本システムにおける処理の流れ

## 3. 関連研究

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[10]. 上田らは任意の語と各件名標目をベクトル化し, その類似度によって任意の語に対応する件名標目を導いている. 任意の語をベクトル化する際は, その語を検索クエリに用いることで Wikipedia や Amazon などの Web 情報資源から関連語を取得し, 任意の語と関連語から成る合成ベクトルを作成する. 件名標目は, 当該件名標目とその下位標目から合成ベクトルを作成する.

本研究では上記のようなベクトルの類似度計算という手法ではなく, 任意の語と関係する書籍を探索し, 結果に含まれる件名標目を集計することで対応する件名標目を決定している.

## 4. 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 はシステムにおける処理の流れを表したものである. 本システムは, 入力された文字列から形態素解析器 MeCab[11]により抽出された名詞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する. Wikipedia カテゴリ名を用いたクエリ置き換えや, NDC 第 1 次区分を用いた文脈付与の機能も実装した. 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[12]で提供される API を利用した. NDLSH は 2008 年度版のデータを用い, Wikipedia は 2010 年 3 月 6 日のデータを用いている.

## 5. 評価実験

本システムが入力語に対して適切な件名標目を提示できるのかを確認するため, 評価実験を行った. 入力語には「IT 用語辞典 e-Words」[13]の 2009 年 11 月 30 日におけるアクセス数上位の用語 50 語と, 「やさしい経済用語の解説 基礎知識編」[14]の 50 語を用い, 各入力語に対して提示される件名標目上位 5 件に適切な件名標目が含まれているものを正答であると評価した. 各実験は 3 名の評価者により行った.

### 5.1 実験手法

**実験 1：** クエリ拡張をせずに用語をそのままシステムに与える.

**実験 2：** 実験 1 で件名標目を 1 つも提示できなかったクエリに対して, クエリ置き換えを行いシステムに与える. 置き換えに用いる Wikipedia カテゴリ名は, ヒットした記事群で最も出現頻度の高い 1 つのみとする.

**実験 3：** 実験 1 で件名標目を提示できたが評価者全員に不正解と見なされたクエリについて, NDC 第 1 次区分を与えて再度件名標目の提示を試みる. IT 用語に対しては"0", 経済用語に対しては"3"を文脈として与える.

### 5.2 実験結果

表 2 から表 4 は、各実験結果の一部を示したものであり, 表 5 は各手法における精度の一覧である. なお, 本システムで利用した NDL 蔵書検索は仕様上, 濁音や長音が無視される. クエリとして"アドオン"を与えると"アトオン"で検索が行われる. 被検索語も同様に濁音等が無視され, かつタイトル等は読みがカタカナ表記で記述されているため, "ジュリアード音楽院"は"シユリアトオンカクイン"として扱われている.

**実験 1：** 精度は, IT 用語 50 語において 0.52, 経済用語 50 語において 0.88 となった. このとき, 件名標目を 1 件も提示できなかったクエリが IT 用語で 9 件, 経済用語で 4 件発生した. また, 件名標目を提示できたが評価者全員が不正解と見なしたクエリは IT 用語で 9 件, 経済用語で 0 件となった.

**実験 2：** 精度は, IT 用語 9 語において 0.33, 経済用語 4 語において 0.25 となった.

**実験 3：** 精度は, IT 用語 9 語において 0.15 となった. なお, 実験 1 で件名標目を提示できたが評価者全員が不正解と見なした経済用語は 0 件であったため, 経済用語に対する文脈付与は行わなかった.

### 5.3 考察

評価実験から,1) 入力語の分野によって精度に違いが出る, 2) 件名標目を提示できない場面では, 蔵書検索クエリを置き換えて改善を図ることができる, 3) NDC 第 1 次区分で文脈を付与しても評価の変動は小さい, ということがわかった.

入力語の分野によって精度に違いが出る原因として, NDLSH の分野の偏りが挙げられる. "経済"を表す NDC 代表分類記号"33"以下の件名標目が 1500 件以上存在するのに対して, "情報処理"を表す NDC 代表分類記号"007"以下の件名標目は 100 件ほどである. 情報処理に関する件名標目が経済に関する件名標目と比較して細分化されていないと考えると, IT 用語はより上位の件名標目との対応付けを余儀なくされるため, 入力語と件名標目のギャップは大きくなり, 不正解と評価されることが多くなると考えられる.

Wikipedia のカテゴリによるクエリ置き換えについては, 件名標目を提示できなかったもののうち 3 割前後が, 妥当な件名標目を提示できるようになったことから, 蔵書再検索において Wikipedia のカテゴリがクエリ候補になりうると考えられる. また, 今回は Wikipedia の記事を検索する際に記事タイトルのみを検索対象とした結果, IT 用語 9 語中の 2 語は Wikipedia 記事がヒットせず, 拡張用のカテゴリ名を取得できなかった. 表 2 における入力語"フル HD"がその一つである. 記事全文を対象として検索を行い, 適合度の高い記事が持つカテゴリ名を優先的にクエリ置き換えに用いる等の処理が必要である. また、カテゴリ名が必ずしも記事の主題を表すものではないため、クエリ置き換えに用いないカテゴリ名をあらかじめ指定すべきである.

NDC 第 1 次区分により文脈を与える手法は, クエリ置き換えと比較してうまく機能していない. 文脈の付与により評価が不正解から正解へ転じうるのは, 件名標目の上位 5 件に正解が含まれず, 6 位以下に正解が含まれる場合であり, もともと 5 件以下の件名標目しか

提示できなかったクエリに対しては文脈を付与して件名標目の優先順位を変更しても評価は改善されない. 5 件以下の件名標目しか提示できなかったものが IT 用語 9 語中に 4 語存在していたことが, 評価改善の妨げとなったと考えられる. 表 3 における入力語”アドオン”のように, 不正解と評価され, かつ提示される件名標目が 5 件以下の場合は, 文脈付与ではなくクエリ置き換えを試みるべきであった.

## 6. 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. また, 件名標目の提示に失敗した場合の改善案として, Wikipedia カテゴリ名や日本十進分類を用いたクエリ拡張を提案した. その上で, 提案手法の効果を確かめるために評価実験を行った. 実験の結果, 1) 入力語の分野によって精度に違いが出る, 2) 件名標目を提示できない場面では, 蔵書検索クエリを Wikipedia のカテゴリに置き換えて改善を図ることができる, 3) NDC 第 1 次区分で文脈を付与しても評価の変動は小さい, ということがわかった.

本手法では文脈の付与に NDC 第 1 次区分を利用したが, これは多くの人にとってなじみの無いものであり, 適切な文脈付与は難しい. 入力語の背景にある文脈を自動的に判断し, 蔵書検索クエリに適切な文脈を反映する仕組みが必要である. また, 本研究の評価実験では, 提示される件名標目が不正解であると見なされたクエリに対してのみクエリ拡張を行ったが, 自動的にクエリ拡張を行おうとすると, レスポンスが不正解となるクエリを機械的に判別する必要がある. このため, 正解, 不正解の判別ルールを設定しなければならない.

今後は, 提示される件名標目を NDC 代表分類記号でクラスタリングすることで, 抽出される件名標目の傾向から適切な件名標目を求めるといった手法にも取り組んでいく.

## 参考文献

[1]CiNii. http://ci.nii.ac.jp/
[2]はてなブックマーク. http://b.hatena.ne.jp/
[3]Flickr. http://www.flickr.com/
[4]Common Tag. http://www.commontag.org/
[5]Linked Data. http://linkeddata.org/
[6]Wikipedia. http://ja.wikipedia.org/
[7]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.
[8]国立国会図書館件名標目表. http://www.ndl.go.jp/jp/library/data/ndl_ndlsh.html
[9]本間維ほか. 国立国会図書館の件名表目標と蔵書目録を利用した語彙の構造化. 情報処理学会第 72 回全国大会. 2010.
[10]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.
[11]MeCab. http://mecab.sourceforge.net/
[12]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/
[13]e-Words. http://e-words.jp/
[14]やさしい経済用語の解説 基礎知識編. http://www.nikkei4946.com/today/basic/index.html

**表 2　クエリ拡張無し(e-Words)**

| 入力語 | 件名標目 |
|---|---|
| Bluetooth | 無線通信, データ伝送, 通信網, … |
| SSD | 半導体記憶装置, ハードディスク, 宇宙工学, 平和, 平和運動 |
| IP アドレス | プロトコル, インターネット, 電子署名, 分散処理（コンピュータ） |
| プラグイン | プログラミング（コンピュータ）, データ伝送, 通信網, ハイブリッドカー, ブログ |
| インフラ | ワイン, 社会資本, 技術援助（日本）, 地域開発, 都市計画 |
| API | 人生訓, 日本, アメリカ合衆国, 成功法, 人間関係 |
| トロイの木馬 | 合弁会社, 国際投資, 外国語教育, 宗教社会学, 推理小説 |
| アドオン | 利子, ジュリアード音楽院 |
| eSATA | 該当なし |
| PPPoE | 該当なし |
| DTCP-IP | 該当なし |
| SDHC | 該当なし |
| ベリファイ | 該当なし |

**表 3 クエリ置き換え(e-Words)**

| 入力語 | 件名標目 |
|---|---|
| eSATA | コンピュータ, データ伝送, プログラミング（コンピュータ） |
| PPPoE | プロトコル, データ伝送, 通信網 |
| DTCP-IP | ウェブアプリケーション, データ処理, ハイパーテキスト |
| SDHC | 日本, クレジットカード, 中小企業 |
| ベリファイ | 該当なし |

**表 4 文脈付与(e-Words)**

| 入力語 | 件名標目 |
|---|---|
| プラグイン | プログラミング（コンピュータ）, コンピュータ・グラフィックス, システム開発, ソフトウェア工学, データ伝送 |
| インフラ | 情報産業, 情報化社会, 国際文化交流, データ管理（コンピュータ）, 情報処理 |
| API | システムエンジニア, システム開発, 目録法, 絵画, 人生訓 |
| トロイの木馬 | 合弁会社, 国際投資, 宗教社会学, 推理小説, 日本 |
| アドオン | 利子, ジュリアード音楽院 |

**表 5 各手法における精度**

| 入力語セット | 手法 | 精度 |
|---|---|---|
| e-Words | 拡張無し | 0.52 |
| | 置き換え | 0.33 |
| | 文脈付与 | 0.15 |
| 経済用語 | 拡張無し | 0.88 |
| | 置き換え | 0.25 |
| | 文脈付与 | |