「本との出会いを支援するシステム」の開発

館野 紅理奈* 浦谷 則好† 杉原 厚吉*
明治大学大学院先端数理科学研究科* 東京工芸大学大学院工学研究科†

1 はじめに

著者らは Wikipedia のカテゴリ構造を利用した新しい書籍の推薦手法を提案し[1], この提案を基にした書籍推薦システムを開発している．提案したシステムの仕組みは，Wikipedia のカテゴリにおける上位下位概念構造を取得した上で「書籍の内容を示す特徴語」と「取得された Wikipedia のテーマ（カテゴリ）の内容を示す特徴語」のマッチングを取ることで書籍を Wikipedia のカテゴリに分類するというものである．ただし，推薦システムを開発する際には，取得した上位下位概念の全てのカテゴリを対象とするのではなく，その中から「学問の分野構造」に注目してそれを利用する．このようにすることで，『書籍同士の関係性』の見通しが良くなり，かつ，分野の詳細な階層構造は知識の全体構造を把握することにもなるために『各書籍の知識構造における所属位置』が分かり，そして『分野間の関係性』を通して書籍を推薦することができると考えられる．

2 章では過去に提案した推薦手法について再度振返り，Wikipedia からの学問分野取得方法に関しては 3 章にて詳細に述べる．4 章で自動取得した分類体系について考察し，5 章では，取得されたシステムで必要なデータを利用して，実際に開発されるシステムの特徴について述べる．そして最後に，まとめと今後の課題について考察する．

2 Wikipedia を利用した書籍分類法

従来の図書分類法では，新しい分野への対応の遅さ，多重にカテゴリ分類できないこと，また，カテゴリ間の親子関係が大まかであり図書分類表の分類体系や書籍の内容を熟知していないとその書籍がどのカテゴリに分類されているのか分かりづらいという問題点があった．人手による管理・出納を目的として定められた図書分類法のカテゴリだけでは，多様な内容を含む書籍に合わせた分類を行うことは困難である．

そこで我々は，集合知の概念が利用されている日本版 Wikipedia を利用して，書籍の内容によって自動で分類を行う手法を提案した．類似の研究としては，Reference Navigator [2]があるが，これは豊富な概念を持つ Wikipedia を探索の窓口として図書分類法のカテゴリ体系に誘導させる統合的なもので，図書のレファレンス作業を Wikipedia で拡張させたものであるために，上記した図書分類法の制約を受けざるをえない．それに対して我々は，体系構造も書籍の所属も全面的に Wikipedia のカテゴリを利用する．そして，内容による分類も自動で行う．

具体的には，Wikipedia のカテゴリの上位下位構造を Hyponymy extraction tool (Version1.0) [3]を利用して抽出することで，Wikipedia の階層構造の情報を得る．そして，「BOOK」データベース[4]と Wikipedia から

それぞれ「書籍名とその書籍の内容を表す特徴語」と「カテゴリ名とカテゴリを表す特徴語」のペアを作成し，連想検索エンジン GETA [5] を利用して書籍の Wikipedia カテゴリへの自動分類を行うことに成功している．

しかしながら，この手法をそのまま推薦システムに適用するには，Wikipedia のカテゴリは書籍のカテゴリとして相応しくないものも多数存在するという問題点が残ったままであった．そこで，Wikipedia の柔軟で密なカテゴリ間の関係をそのままに，新たな書籍の分類体系と成り得ると考えられる「学問分野の構造」に注目し，この構造を書籍分類のカテゴリとして適用する．

3 Wikipedia 学問分野構造の自動取得

著者らの求める学問分野の構造は，言い換えれば，”知を分類し階層化したラティス構造”である．学問分野の分類は，日本十進分類法や NDLC といった書籍によくある明確な分類法はなく，国や教育機関等によって差異がある[6]．今回取得したものは，情報源が日本の Wikipedia であるため，それに依存した分類法となる．

Wikipedia から取得される学問分野構造には，以下のような特徴がある．

1. ネットワーク構造（ループを含み複数の上位ノードを許す）になっている．
2. 各分野には下位分野が存在する．
3. 下位分野には人物名や学会,学術的な所属機関名も含まれる．
4. A > B > C が A > C となるような(ただし，A > B は B が A の下位概念であることを表す)省略パスが存在する場合がある．

これらは Wikipedia カテゴリの上位下位概念の構造に依存した特徴である．

このように Wikipedia の構造をそのまま利用するには構造が複雑であり，書籍の所属するカテゴリとして（もしくは知の分類として）ふさわしくないものが数多く含まれているため，学問分野抽出にはフィルタリングに関して工夫を施す必要がある．そのための手法は，テキストのパターンマッチを利用する．

まず始めに，カテゴリとして抽出されたものから「〇〇学」という表現が上位または下位に現れる上位下位の組を拾いだす．ただし，「テーマ別の~」や「〇〇に関する~」などもカテゴリ名として相応しくないと見なして，分野構造抽出の際にフィルタリングにかけておく．次にそれらの間のループ構造や省略パスを作る上位下位の組と人物名や機関名などを取り除く．ループ構造や省略パスを取り除くのは書籍推薦の為にカテゴリ間の最短経路を求める際に邪魔になるためである．パターンマッチを用いたフィルタリングで最大の困難が，人物名の除去であるが，「人物名」はより下位の概念を持たないことが多いとする特徴を利用して，下位概念を持たないカテゴリを取り除く．これにより，人物名の大半を除去することができる．”学”の付かない下位概念が,上位概念となるような上位下位のペアをフィルタリングにかけながら繰り返し取得することによって，Wikipedia から学問分野構造を自動取得する．これを学問分野構造の階層構造とみなす．

以上の手続きで学問分野構造を取得した結果，まず始めに取得された“学”のつく概念お

よびその上位概念(この集合を A とおく)が 350 種類，A に属する概念の下位概念で他のペアの上位概念となり，かつ語尾に”学”を含まないもの(この集合を B とおく)が 498 種類，また B に属する概念を上位概念としたときの下位概念(この集合を C とおく)が 1668 種類であった．さらに，C に属す概念の下位を上記手続きで辿って取得されたもの(この集合を D とおく)が 42120 種類あった．従って，だんだんと再検索される数が増えていき，取得される上位下位の組は A から B までの階層で 6850 組，A から C までの階層で 25600 組，A から D までの階層で 67294 組となった．しかし，D まで取得されたものには，学問構造から逸脱したものや，詳細すぎるためカテゴリらしくないものが数多く含まれている．従って，これ以上の数の取得は，C の各カテゴリから人の手でより詳細に取得すべき D の要素を探し出して抽出するほうが良いと考えられる．ただし，C までで想定する大半の学問分野の構造が精度よく取得できており，その必要は少ないと推察される．

4 本手法による分類体系と図書分類体系との比較

日本図書協会の基本件名標目標(Basic Subject Headings ;BSH) [7]のカテゴリ数が 11192 件 (2012 年現在)，また，国立国会図書館件名標目表 (National Diet Library Subject Headings ;NDLSH) [8]のカテゴリ数が 19485 件 (2012 年現在)ある．我々の分類体系をこれらと比べると，カテゴリ数が少なく，代わりにカテゴリ間のつながりは圧倒的に多いのは，カテゴリの内容を構造の関係で表しているためである．BSH や NDLSH にもなくて本手法で取得したカテゴリの例に「パタン・ランゲージ」などがある．ちなみに，「パタン・ランゲージ」の上位概念を辿っていくと【人文科学 > 地理学 > 人文地理学 > 都市地理学 > 都市計画 > パタン・ランゲージ】【人文科学 > 地理学 > 人文地理学 > 文化地理学 > コミュニティ > 都市計画 > パタン・ランゲージ】となっており，構造がラティス構造であることがわかる．

また，BSH になく NDLSH にはあるカテゴリの例として「自然言語処理」に注目してみる．NDLSH には上位概念に「情報処理」，下位概念に「漢字処理」「機械翻訳」「テキストマイニング」の 3 件のみしかなく関連語として「言語」「コーパス言語学」「オントロジー (情報科学)」がある構造になっている．それに対して本手法では，直前の上位概念に「言語学」「人工知能」，直後の下位概念は全 25 件あり，これを書き出すと，「CD 理論」「未知語」「構造木」「コーパス」「全文検索」「文書分類」「構文解析」「機械翻訳」「照応解析」「形態素解析」「計算言語学」「かな漢字変換」「固有表現抽出」「潜在意味解析」「自動要約生成」「自然言語理解」「自然言語生成」「シャローパーサ」「チャートパーサ」「確率文脈自由文法」「語彙の曖昧性解消」「ベイジアンフィルタ」「中国語入力システム」「日本語入力システム」「コンピュータ言語学」となった．本手法での「テキストマイニング」は「統計学」と「言語学」の下位概念に位置しており，「言語学」を通して「自然言語処理」と繋がっている．このように，本手法の分類体系と図書分類体系を比べると，同義語と下位概念が入れ替わっているものも多く見つかった．

5 推薦システムの特徴

本システムの利用用途には以下のようなものがあると考えられる．これは本システムの目指すところでもある．

- 嗜好本の学問分野大系における位置把握
- 分野間のつながりの発見
- ２冊の書籍間を繋ぐ分野と書籍の発見
- ユーザの興味の啓発

本システムではクエリとしてユーザの嗜好本を与える．その嗜好本から，既存手法[1]で登録しておいた「書名とその所属するカテゴリ群」のデータベースをみて所属するカテゴリを２つ取り出す．この２つのカテゴリがつくる２点の最短経路を計算し，出力された経路を Graph で表示させる．ここで問題は，最短経路が複数存在する場合の対策，最短経路が存在しない場合の対策である．さらに，類似度の高い順から２つのカテゴリを取得する方法だけでは，思いがけないつながりを見せるには概念が近すぎるものが取得されることがあるのも問題となる．今後，上記目標に近づけるためにこれらの課題に取り組む必要がある．

6 まとめ

本稿では，既存手法[1]での推薦システムを実現する際に問題となった，Wikipedia のカテゴリとその階層構造から書籍にふさわしいカテゴリのみを抽出する方法を述べた．その結果，本手法で書籍の目次部分に対応するほど詳細なカテゴリが自動取得できた．この分類体系は，一般書籍以外でも，専門性の高い学術論文や分野横断的なつながりを知りたいものには特に有用である． 次に，抽出した分類体系の構造を利用した書籍の推薦システムについて特徴を考察し，開発する際に発覚した問題点を述べた．当面の課題は，開発の際に生じた問題にとりかかることである．その後，人の思考にとってスムーズに情報を取得できる，より良い UI について Graph 表現を拡張させる方面で考えていきたい．

参考文献

[1]館野,浦谷,「本との出会い」を支援するシステム,言語処理学会第 17 回年次大会発表論文集,2011 年 3 月 7 日- 10 日, P1-8.

[2]清田,田村,中川,増田, Reference Navigator: 異種オントロジーの統合ブラウジングツール～図書館の分類体系と Wikipedia カテゴリの対応付け～, 言語処理学会第 13 回年次大会発表論文集,2007 年 3 月 17 日- 20 日, PD1-3.

[3] http://alaginrc.nict.go.jp/hyponymy/index.html#example,jyoui, 上位下位関係抽出 Version1.0:Hyponymy extraction tool

[4]株式会社紀伊国屋書店,株式会社トーハン, 株式会社日本出版販売,株式会社日外アソシエーツ，図書内容情報「BOOK」データベース.

[5] http://geta.ex.nii.ac.jp/geta.html,汎用連想検索エンジン GETA）

[6] http://ja.wikipedia.org/wiki/学問の一覧,学問の一覧.

[7]http://infos.net.cias.kyoto-u.ac.jp:8083/bsh1/word-list-all.jsp,基本件名標目表 BSH(Basic Subject headings)トピックマップ.

[8] http://id.ndl.go.jp/auth/ndla/, Web NDL Authorities 国立国会図書館典拠データ検索・提供サービス.