「本との出会い」を支援するシステム

館野 紅理奈　浦谷 則好
東京工芸大学工学部コンピュータ応用学科

1 はじめに

情報爆発時代において，莫大な量のデジタルデータが蓄積され，また指数規模でのデータ増加が進む中で，如何にしてこれらの大規模データを有効活用するかは現代の大きな課題である．人は既知の情報からキーワードにより情報収集することはたやすいが，全く未知の情報からその存在を知ることは，何らかのきっかけがない限り難しい．近年，技術の進展に伴い，多様な分野同士が複雑に絡み合い，また専門家同士の知恵を融合することで新分野が誕生する事例もある．しかし，上記のような理由から，分野間の融合にはいくつか障害があるように感じられる．

そこで我々は，「専門同士を仲介する」ことを専門とする新しいシステムの形を提供するべく，本との出会いを支援するシステムを提案する．大規模データを扱うシステムを構築する上で枷となるのが，人間の情報処理能力への考慮である．それを踏まえて我々は，ユーザが直観で知的好奇心の対象を見つけ出せるように，可視化の手法に習って書籍の全体構造を把握できる検索システムを構築することを目指した．3では書籍を内容で分類するための手法について考察し，4では実際に分類システムを構築した．5では，4で構築した分類システムを利用して，我々が提案する書籍推薦手法を具体的に検討する．ここで発見性に対する考察も行う．最後に6で今後の課題を述べて本稿をしめる．

2 提案システムの概要

研究のテーマである「本との出会い支援」の目的は，ユーザがシステムを利用するうちに自然と思いがけない本を発見することである．ただし,Amazonや他の図書推薦と異なるのは，我々のシステムはユーザの他分野への興味拡張を目指しているという点である．つまり，書籍群が所属する分野とその構造を明瞭にし，ユーザの興味がある書籍が，分野全体の構造から俯瞰してどの位置に所属するか認識させることにより，繋がりのある他分野，そしてそのまた他分野へと視野を広げていく，そのための手助けをしようとするのである．具体的には，書籍と分野を繋ぐラティス構造を利用して，分野と分野，書籍と分野の繋がりを辿ることで，そこに所属する書籍を推薦する．

3 書籍分類手法の考察

書籍をカテゴリに分類することは書籍群の全体構造を把握するために必要不可欠である．我々は従来の図書分類法の弱点を克服するために，Wikipediaのカテゴリ概念を取り入れる．

3.1 従来の図書分類法

NDCはNippon Decimal classificationの略称であり，日本で最もよく利用されている分類表である．あらゆる知識を9区分して1～9の数字を割り当て，どれにも属さないものには総記として0を与える十進記号法で分類する．各区分の中を同様の作業で分割を繰り返すことで，中をより階層的に区分していくことが可能である．NDLCはNational Diet Library Classificationの略称で，日本の国立国会図書館で利用されている分類手法である．NDCが整数とピリオドのみを利用した純粋型記号法であるのに対し，NDLCは保管と出納を考慮したアルファベットと整数の混合による混合型非十進法により分類する．アルファベットを利用することで，NDCより高い区分能力が可能である．具体例を表1に示す．

三種類は共通してツリーで表せる階層構造であり，高い位の出現から順に細かな所属へと一意に判定できる構造になっている．従来の書籍分類法が抱える問題点には，図書館や販売側の物理的な書籍を管理・

表 1　従来の図書分類表による「非線形言語モデルによる自然言語処理-基礎と応用」の分類

図書分類表	表記	a>b:木構造の親と子の関係
C-CODE	3080	3(専門)＞0(単行本)＞80(語学表記)
NDC(9)	007.63	0(総記)＞0(総記)＞7(情報科学)＞67(ソフトウェア・システム・コンピュータ)
NDLC	VL47	U(学術一般)＞L(図書館情報センター)＞47(機械翻訳)

保管・出納することが目的であったことが根源としてあり，(1)分類分けがユーザ視点で行われていない．(2)ツリー構造の親子関係も分類がおおざっぱであるために，全て意味内容で親子関係があるとは限らない．そのため，ユーザが直観で求める書籍の親カテゴリを見極めるのは困難である．複数主題があった場合などは，各々の分類基準表に従って一つの主題に分類されてしまうため,排除された他の主題からその書籍を見つけることはほとんど不可能となる．

3.2 Wikipedia の分類法

Wikipedia の特徴として，記事の網羅性，即時性，密なリンク構造，質の高いリンクテキスト，多様なリンク構造などが知られている[7]．また，ページと概念が一対一で対応していること，ページの冒頭にある定義文には他の概念に対する is-a 関係が豊富であることも判明している[5]．さらに，カテゴリの多くは一つ以上の親カテゴリを持ち，上下関係が逆転することはないとする Wikipedia のカテゴリ方針から，Wikipedia の全体構造は一つ以上の親を持つ半順序関係を有するラティス構造の形をしていることがわかる．以上のことから我々は，Wikipedia の概念を書籍のカテゴリとして利用することで，書籍の内容による詳細な分類することを考えた．書籍に対して複数のカテゴリに所属することを許すことで，書籍を通して分野の繋がりを発見することが可能になる．このことは，我々が目指すシステムが構築可能であることを指す．

4 提案する書籍分類システムの構築

Wikipedia の概念の上下関係を利用して書籍を分類するシステムを構築した．利用した情報と構築手順について説明する．

4.1 書籍内容情報源

本研究で利用する書籍情報源として，図書内容情報「BOOK」データベースを利用する．「BOOK」データベースとは，㈱紀伊国屋書店，㈱トーハン，㈱日本出版販売，㈱日外アソシエーツの 4 社が 1986 年より共同構築している本格的なデータベースである．収録され利用可能な累積件数は約 118 万件(2010 年 2 月現在)である．年間では約 6.3 万件の増加量であり，これは収録を始めた 1986 年からみて約 2 倍以上の増加量であり，現在も年間約 1,000 件ずつ増加傾向にある．このデータベースはアクセスキーを通して必要とするデータを簡単に抽出することが可能であり，画像情報以外の基本的な図書内容情報は全てここで揃う．

4.2 分類情報源

本研究で利用する書籍をカテゴライズするための分類情報源として，日本語版 Wikipedia(以下，Wikipedia)を利用する．今回，Wikipedia を利用したデータベースを二つ用意した．カテゴリの上位下位関係を納めたデータベースと，カテゴリ名とその特徴語を納めたデータベースである．

4.3 構築手順

書籍分類システムのフローを図 1 に示す．

手順 1:「BOOK」データベース内の TAG を利用して，各書名 t1,・・,tn とその書名を特徴付ける特徴語 bi(i=1~k,k<0)の行列ファイルを作成する．図の n は書籍の全冊数，k は一冊あたりの特徴語の全個数を表しており，特徴語は書籍の書名・目次・要旨から自動抽出した固有名詞を利用する．出力は，書名とその特徴語の組(t1,bi)，(t2,bi),・・,(tn,bi),(i=1~k)である．

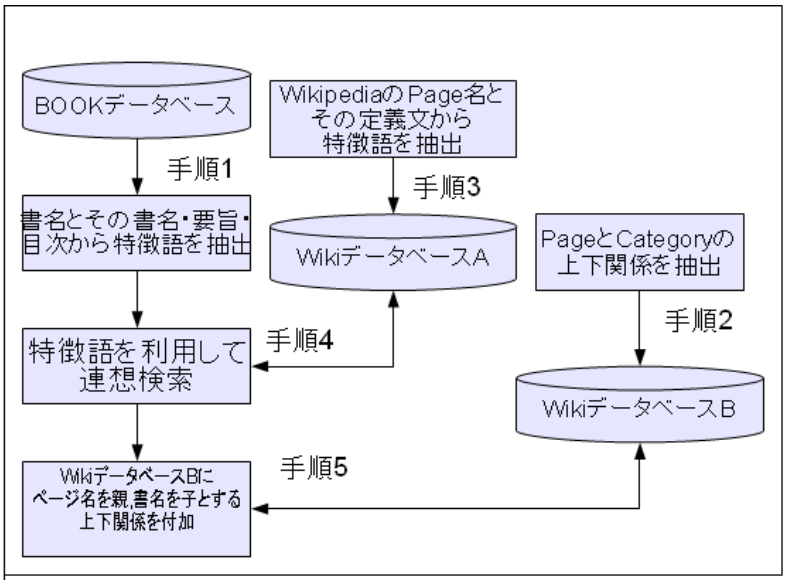

図 1　書籍分類システムのフロー

手順 2: Wikipedia から Page(記事)と Category, または Category と Category の上下位関係の組を抽出し, Wikipedia の上位下位関係データベース(以下, Wiki データベース B)を作成する. 上位下位関係の抽出には, 隅田ら[**4**]が開発した上位下位関係抽出ツール Hyponymy extraction tool(Version1.0)を利用した.

手順 3:Wikipedia から各 Page 名 p1,・・,pm とその Page を特徴付ける特徴語 w1j(j=1~r)または w2h(h=j+1~s)の行列ファイルを作成する. 図の m は Wikipedia の全 Page 数, r は 1Page あたりの特徴語の全個数を表しており, 特徴語は Page の最初の文(定義文)から形態素解析により抽出した固有名詞を利用する. また, w1j と w2h はそれぞれ Page と同名の Category が存在しない場合と存在した場合のものである. 存在した場合の w2h は, 同一名を XXX だと仮定すると,【カテゴリ:「XXX」にあるページ】の Page 名も特徴語に加える. つまり, 図の s は新たに加わった特徴語の全個数を表している. 出力は, 特徴語の組(p1,(w1j)U(w2h)),(p2,(w1j)U(w2h)),・・,(pm,(w1j)U(w2h)),(j=1~r,h=1~s)である.

手順 4:書名とその書籍の特徴語の組(t1,bi),(t2,bi),・・,(tn,bi),(i=1~k), Page 名とその特徴語の組(p1,(w1j)U(w2h)),(p2,(w1j)U(w2h)),・・,(pm,(w1j)U(w2h)),(j=1~r,h=1~s)に対して. 書名の特徴語 bi と Page 名の特徴語(w1j)U(w2h)のマッチングを取る. マッチングの量が多い上位の書名と Page 名の組(pb,ta),(0<a<n,0<b<m)を抽出する.

手順 5:組(pb,ta)は pb を親, ta を子とした上位下位関係の組とみなすことができる. この新しく作られた上位下位関係を, 手順 2 で作成した Wiki データベース B に付け加える. つまり, この書名 pb(＝書籍)は Page 名 pb に属しており, その Page 名(＝カテゴリ)にカテゴライズされる.

5 提案する書籍の推薦手法

前章で提案した我々の書籍カテゴライズ手法により, 書籍は複数のカテゴリに属すことが可能となった. この書籍とカテゴリの繋がりを利用して推薦手法では, どのような書籍が推薦可能になるか考察する.

クエリには, ユーザの嗜好に合った書籍を入れる. そうすることで, ユーザの嗜好による分野の関係が表示され, また関連する分野に所属する書籍を表示することで, 書籍を推薦する. 例として, 書名「非線形言語モデルによる自然言語処理－基礎と応用」(a)をクエリとした結果を図 2 に示す. ここで推薦される特徴的な書籍を(b)－(h)と置き, それぞれを考察した.

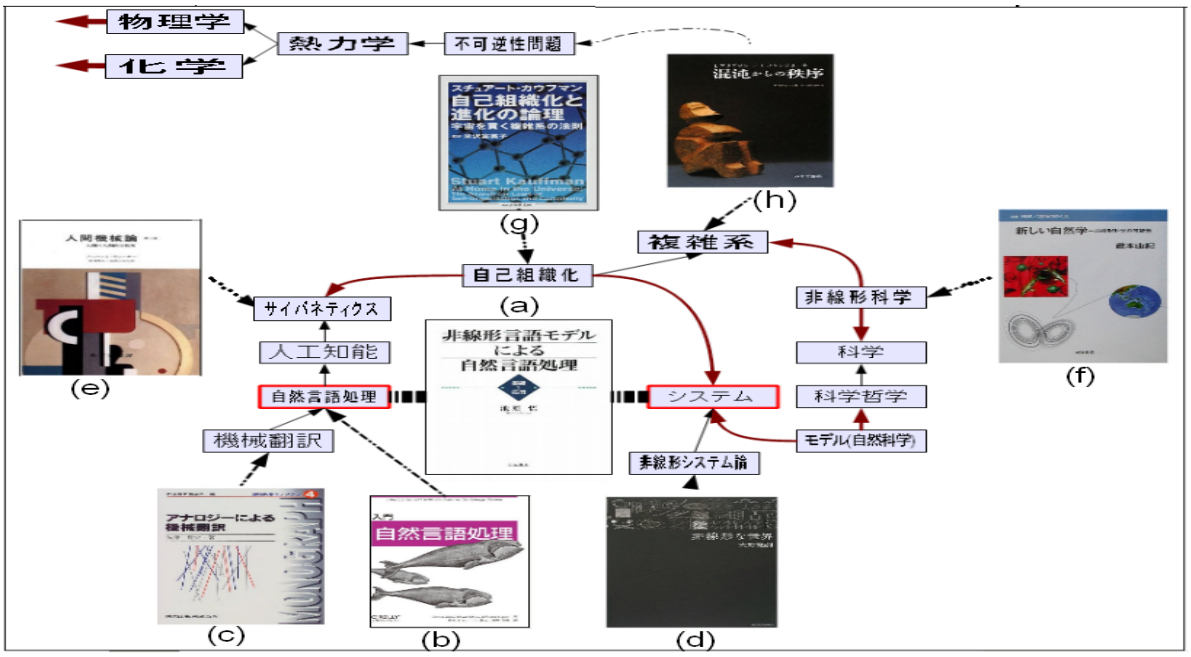

図 2　書籍推薦システムと推薦される書籍

図のように表示は有向グラフでその関係を示す．図の→は，「A→B:A が子，B が親とする親子関係」を表している．また，書籍gとそれが所属するカテゴリ G は便宜上 G∋gのように表すことにする．図 2 で(a)はカテゴリ[自然言語処理]と[システム]に属する書籍であることが分かる．[自然言語処理]をから生まれるノードを集合 M，[システム]から生まれるノードを集合 F とおくと，M={[自然言語処理],[人工知能],[サイバネティクス],[機械翻訳][自己組織化]･･･},F={[システム],[非線形システム論],[モデル(自然科学)],[科学哲学],[科学],[非線形科学],[複雑系]･･･}である．書籍(b)－(h)は(a)にとって以下の関係がある．(b):(a)と兄弟関係．(c), (d):(a)の甥姪関係．(e):A の祖父母関係．(g):M と F の両方を血縁関係にあるカテゴリ G に対して G∋g，この G は M 空間と F 空間をつなぐノードであり，(a)の振る舞いと近いと考えられる．(f):F 集合に属する関係で(a)から遠くの関係に見えるが「非線形」を扱う書籍として共通している．(h):(G→H∋h)，(a)から興味のある書籍(h)が見つかったと仮定すると，(h)は(a)と共通する集合 F と(a)とは別の集合 M'={[不可逆性問題],[熱力学],･･･}を生成する．つまりこの(h)から，新しい集合 M' へと興味を拡張させることが可能である．

6　今後の課題

5 で述べたような振る舞いを持つ(b)－(h)等の書籍に対して，どのような本がユーザにとって発見性が高いのか実験とデモを繰り返すことで評価を行い，改良を重ねる予定である．

参考文献

[1]日本十進分類法新訂 9 反分類基準
[2]国立国会図書館分類表
[3]「BOOK」データベースファイル仕様書
[4] 隅田飛鳥，吉永直樹，鳥澤健太郎，萬成健太郎．Wikipedia から大規模な上位下位関係取得．言語処理学会弟 14 回次大会発表論文集，2008.
[5] 中山浩太郎．Wikipedia マイニングによる大規模 Web オントロジの実現．The 22nd Annual Conference of the Japanese Society for Artificial Intelligence，2008.
[6]高野明彦，西岡信吾，今一修，岩山真，丹羽芳樹，久光徹，藤尾正和．汎用連想計算エンジンの開発と大規模文書分析への応用．2002.
[7] 白川真澄，中山浩太郎，原隆浩，西尾章治郎．Wikipedia のカテゴリ構造解析とクラスタリングによる概念ベクトルの生成．第 23 回人工知能学会全国大会，2009.