目録データベースの高次化によるデータマイニングを可能とするために

—複数種のオントロジ辞書の利用・接合により検索効率の向上を試みる—

相田 満
人間文化研究機構国文学研究資料館文学形成研究系

In order to make data mining possible by enriching the contents of table databases

--The trial in which the retrieval accuracy is improved by the junction of the Ontology-dictionary--

AIDA,Mitsuru
National Institute of Japanese Literature

オントロジは、情報リソースから独立した上位層(メタレイア)に位置付けられ、情報を意味的に組織化、検索、ナビゲートするための新しいパラダイムとして注目を集めている。では、この理念の実効性はどれほどのものなのだろうか。このことを検証するために、少ない情報量しか持たない文献目録データベースに対して、複数のオントロジ辞書を組み合わせたり接合することによって、どれだけ中味を充実させることができるかを試みる。

Ontology is positioned by the higher rank layer which became independent of an information resource.And it systematizes, searches and navigates informations semantically from there.This attracts attention as a new paradigm for information retrieval.However, does this idea have efficiency nature truly?I want to verify this. In order to verify this, I'll try which to be able to enrich contents by combining two or more ontology dictionaries, or joining to a reference table database only with the few amount of information.

1. はじめに

「オントロジ[Ontology]」は、いわゆる「標準化」された世界の対極的に位置し、専門性に特化し、限定された分野の知識体系から積み上げられた、知識概念木で構築される。これは現在、情報リソースから独立した上位層(メタレイア)に据えて運用することにより、情報検索やナビゲートに援用することによって、情報爆発の被害を免れた、確度の高い情報を導き出すための、新しい「パラダイム（概念的枠組み）」として急速に注目を集めている。

これをあえて「パラダイム」と評する所以は、情報学と人文科学とが融合した形で提示されたこの考え方が、本義の「存在論」にも接近するかのような、哲学的命題を包み込み、さらには人文科学ではこれまで表現し得なかった思考の枠組みを提供してくれているからである。

たとえば、「ＡとＢは関連性がある」という説明における関連性の内容は、自由な定義が許されており、その内容がいかなるものであるかということを説明する概念体系が不明なままでも、それを規定する枠組みが、一貫した思考体系で貫かれている限りにおいては、それを「オントロジ」として扱うことが許される。その意味において、それまで比較的厳密な語彙関係の定義が要求される「シソーラス」とは全く異なる観点による知識セットの構築が可能になる。

この発想は、人間がものごとを関連づける営みを続けてきた行為自体を文化資源としてとらえ直す取り組みや[1]、「概念を操作する思考技法」としての「オントロジ・アルゴリズム」の提言[2]など、伝統的人文科学の思考の枠組みを、別の次元からとらえ直す契機にもなっているように、人文科学を含むさまざまな分野に広がりを見せ始めている。

2. 研究の目的

人の考えは様々であり、またそれに対する見

方も一様でない。情報検索における「オントロジ[Ontology]」では、検索者の個性が反映されることがうたわれる。専門性の高い知識体系である「オントロジ」を組み合わせることで、検索者の要望や考えに近い情報が引き出されることが期待されるものとなっているのである。
オントロジというものが構想され、提言された事情の背景には、そうした汎用的なシソーラスが構想され、挫折を繰り返していった反省もあった。

用語体系としてのオントロジは、知識、語彙、概念などと、それらの間の関係を明確にした辞書をいうが、言葉の関係を示した辞書で、単語と単語の関係を定義を厳密に定義するシソーラスよりも、関係の明示性はさらにフレキシブルな設定が可能であるという特徴がある。

そして、すでに存在する知識体系を分類・集約し、その成果をコンピュータを利用することによって人間科学に反映させるという前提に立つものとなっている。その意味で、オントロジの志向するところは、グローバルな意味での標準化とは対極的な所にあるといえよう。

その意味で、比較的厳密に定義された概念体系は、適切なサイズの領域内で蓄積・整備された知識の集合体となる。そして、そうした知識の積み上げにより、大局的には実世界の知識を扱うための基礎理論・分類体系から技術までを再構築することを目指すものとなっているのである。

では、こうしたビジョンが実現する可能性はどうだろうか。

そのことを、文献目録のデータを例とりあげて考えたい。すなわち、文献目録中に記載される情報を、別に用意される辞書データとのマッチングによって、どのような広がりを見せるか検証することによって、いわゆるオントロジ接合の実効性を試みる。

3.対象となる文献目録データについて

3.1.国文学論文目録データベース

文献目録は個人、機関の作成を問わず、もっとも多く作成されるデータベースのひとつである。しかし、それらのデータに対して、キーワード付け作業や、シソーラス管理など、十全な検索効率の向上に配慮したメンテナンスを恒常的に行えるところは多くない。

この事情は、国文学研究資料館より公開される「国文学論文目録データベース」でも同様で、データ自体の作成よりも、その検索系を支える基盤データの構築の方に、より大きな負担がかかることが、その理由であった。

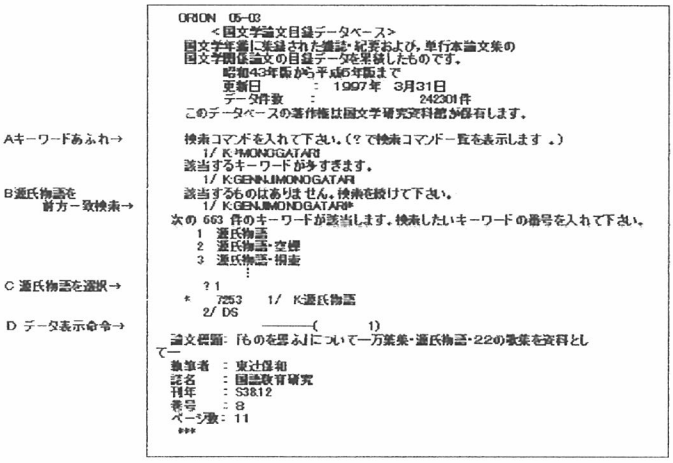
このデータベースは、国文学研究資料館（国文研）が、毎年の研究文献を集積した『国文学年鑑』を刊行したデータの中から雑誌・紀要・および単行本論文集から取材された研究論文目録をデータベース化、オンライン検索に供したものである。

現在公開されるデータは約 40 万件。およそ 1924 年（一部）からおよそ 80 年間にわたる国文学研究に関する研究論文を収める。

その電子化への取り組みの歴史は古く、国文学研究資料館の設立にかかる 1972 年にまでさかのぼる。しかも、電子化に際して抽出された文字データ群は、1978 年に JIS C 6226 として制定された、第１・第２水準漢字の基礎資料にも供されたほどの歴史をもつ。

1992 年度からオンラインサービスが開始されたが、当時は、汎用機を使用してのデータ処理・公開がなされ、検索もコマンドプリフィクスと検索語を入力した後に検索結果を出力させる方式であった（図①）。

このシステムは 2000 年度第 6 期システムからワークステーションによる Web に対応したシステム（OpenText 使用）へダウンサイジングされ、HTML 使用による GUI ナビゲーションによるインターフェイスへと置きかえられた。



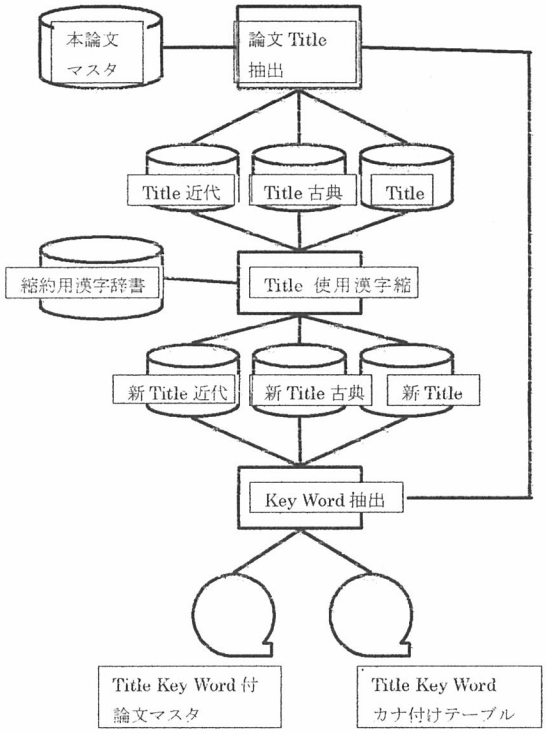
図①　国文学論文目録データベース(旧システム)検索画面

項目名	項目名	項目名
冊子年	分野５	開始頁
入力者	分野５よみ	終了頁
入力日	分野６	**総頁**
更新１	分野６No.	和暦年
更新２	分野７	月
更新３	分野７よみ	日
漢字	頭書	**西暦年**
メモ	主題	**翻複１**
連番	副題	**翻複２**
旧連番	**題名**	**作品名**
時代分類	英文題	**作者名**
時代分類No.	**執筆１**	概念
分野１	執筆２	要約
分野１No.	英文執筆	キー
分野２	種別	作成年
分野２No.	請求	**作品名よみ**
分野３	**誌著名**	**作者名よみ**
分野３よみ	英文誌	**M（和暦年月次／必須）**
分野４	**巻号**	**月桁**
分野４No.	通巻	

図② 国文学研究文献目録データベースのデータ項目（第6期システム）
（太字ゴシックの部分の項目が受け継がれている）

項目名（解説）	項目名（解説）
￥N（一連番号／必須）	￥M（和暦年月次／必須）
￥０（時代分類／必須）	￥X（西暦年／必須）
￥１（ジャンル／必須）	￥R（翻刻・複製）
￥２（頭書）	￥Y（翻刻・複製のヨミ）
￥３（主題）	￥S（作品名）
￥４（副題）	￥J（作者名）
￥T（タイトル／必須）	￥C（共通事項）
￥A（執筆者）	￥D（詳細事項）
￥Z（雑誌名／必須）	￥K（キーワード）
￥G（巻号）	￥Q（論文本文）
￥P（ページ数）	

図① 国文学研究文献目録データベースのデータ項目（移行前）



図④　キーワード付加処理の流れ

3.2.前システムのキーワード処理

当該データベースに搭載されるデータは、予約される情報項目は豊潤に予約されてはいるものの、実際は情報ソースとなっている冊子体『国文学年鑑』を編集する際に必要な分類と書籍内における配列作業のために付与された関連作品・関連作者名と、冊子の配列のために分類名が与えられるのみで、情報検索への便宜をはかるために、特別にキーワードを付与したり、検索システムにシソーラスを搭載することなどは行われていない。しかし、それでも汎用機が使用された前システムでは、

完全一致・前方一致・後方一致

と AND,OR 程度の演算子を使っての検索に制限されていた。システム処理によるキーワード「自動」切り出し処理と、ヨミ情報の付加処理が行われてきた（図④）[3]。

前システムで行われたキーワードメンテナンスで特徴的なことは、古典を扱う論文と、近代以降を扱う論文とでは学術用語の体系が異なることを予想して、古典用・近代用と辞書をそれぞれ作成し、切り分けて処理を行っていたことである。当時は階層的な用語管理はなされなかったが、その発想は異なるオントロジ辞書を組み合わる処理のプロトタイプともいってよい。

しかし、いくら「自動」が標榜されるとはいえ、実際に切り出された語句は、そのまま辞書に再登録されるべきものとは程遠かった。

たとえば、

「和歌の」「和歌の様」「和歌の様式」

のように、付属語との組み合わせの切り分けが不十分なものなどが頻出するため、機械的に切り出された語句が使用可能なレベルまで鍛え上げるためには、予めデータの分かち書きを施すなどの前処理が必要であった。

しかも、論文タイトルに頻出する語句につい

ては、あまりにヒット数が多すぎる検索の用をなさなくなってしまうという問題もあった。

当時の検索システムでは、ヒットする件数が1,500 件を超えると「件数が多すぎます」というメッセージとともに、検索語の再設定を要求する仕掛けになっていた。しかし、「和歌」「物語」などの主要ターム、さらには「源氏物語」「宮沢賢治」に代表されるような、人気の高い研究テーマになればなるほど、ヒット数が増加しすぎてしまい、検索が不可能となるという矛盾を抱え込むこととなり、それを防ぐためのメンテナンスに要する負担も膨大なことが予想された。

その後、現行第６期システムでは、中間一致検索が可能となり、キーワードの切り出し処理が不要となった。そこで、前述の自動処理によるキーワード埋め込み用のデータ項目は予約されるものの、上記処理のために蓄積されたデータは使われなくなった。

しかし、将来的にはシソーラスやオントロジ辞書を独立させて運用することにより、キーワードによるナビゲーション機能を付加することは期待された。そこで、データ構築が続けられることとなり、現在に至っているのである。

4.使用辞書について

今回検証に使用するオントロジ辞書は、上記経緯で作られた３種類である。

いずれも日本文学に関する研究用語辞書ではあるが、それぞれ異なる作成責任者の監修のもとに辞書がデザインされ、取材源、取材方法もまったく異なるものとなっている。

内訳は以下の通り。作成責任者が付けた名称や、その特徴を尊重して、それぞれＡ：連想辞書、Ｂ：翻訳辞書、Ｃ：知識支援辞書と呼んでおき、次にその特徴を略述するとともに、各辞書のフォーマットを構成の把握に配慮して、イメージで示す。

A:連想辞書(国文学研究語彙辞書)(図⑤)

約 35,000 語。基準語句から連想される語句を、複数の日本文学研究者の協力を得て追加、さらに増補を加えたもの。異表記・同義語・分野別連想語・関係時代・ローマ字表記・英語（一部データ）などの付属情報を持つシソーラス辞書[4]。

B:翻訳辞書(日本文学研究用語翻訳辞書)(図⑥)

約 7,000 語。日英対訳辞書。小西甚一著『日本文芸史』の英訳本[5]索引から取材[6]。

Ｃ：知識支援辞書（図⑦）

見出し語 17,000 語、約総語数約 4,000 語（今回の実験に使用するもののみ）。辞書の見出し項目と解説文から取材された辞書。辞書の見出に記述される語句中より、ふりがなが付与された要語を抽出。当該語句に、人名、作品名、地名…などの属性を付与。[7]

基準かな	基準漢字	異表記よみ	ローマ字(ヘボ	ローマ字(訓	英語	時代	同義語1	同義語2	連想1(ジャンル)	連想2(作品・作者・関連人物)	連想3(その他関連事	予備
サンカンショウ	山間商家		SANKANSHO	SANKANSY		D	石川山店		俳諧	石川山店/北部/松尾芭蕉	山店/蕉門/江戸	
ヤマギハ	山木派		YAMAGIHA	YAMAGIHA		D			演劇/芸能/箏	山田流箏曲	歌沢節/歌沢芝金	
ヤマモトバン	山本版		YAMAMOTO	YAMAMOTC		D		山本九	演劇/芸能	浄瑠璃本/山本九兵衛/大	西沢九右衛門/正	
サンガクコ	散楽戸		SANGAKUKO	SANGAKUKO		ABC		楽所/			雅楽/うたづかさ/う	
サンゴウ	散更	サルゴウ/	SANGOU	SANGOU		ABC		中楽/	演劇/芸能		能/大和猿楽/能	
ヨモダムカン	四方太無		YOMODAMU	YOMODAMU		E					無意味/白紙文字	
ヨモレン	四方連		YOMOREN	YOMOREN		D			詩歌		狂歌/蜀山人/四	
シキヒョウリン	史記評林		SHIKIHYOURI	SIKIHYOURIN		D			演劇/芸能		評判記/役者評判	
サシコガミ	指神子					CD	さすの巫				卜占/陰陽子/ト	
ウジサトキ	氏郷記		UJISATOKI	UZISATOKI		CD			歴史書/演劇/	蒲生氏郷/名古屋山三郎/	お国歌舞伎/名古屋	
	氏別女					A					采女/独身/後宮	
シノフダ	詩の札		SHINOFUDA	SINOHUDA		ABC			漢詩文		白楽天/詩板/詩	
シハン	詩板		SHIHAN	SIHAN		ABC			漢詩文		白楽天/詩牌/詩	
ジツイ	字対		JITSUI	ZITUI		ABC			漢詩文		漢詩/修辞/謡曲の	

図② A:連想辞書(国文学研究語彙辞書)

用語番号	原資料の巻号	用語(英語)	補足説明	備考	構成単語数	構成単語数(カッコあり)	カッコの有無(0括弧なし・1括弧あり)	言語コード	和語の是非	固有名	訳語
1	#43				1	1	0	J	1	0	だあらあに
2	#43				2	2	0	JE	0	0	
3	#43				4	4	0	EEEE	0	0	
4	#43				4	4	0	JJEE	0	0	
5	#14	Akinaka	FF		1	2	1	J	1	1	
6	#14	Akisue	FF		1	2	1	J	1	1	
7	#14	Akisuke	FF		1	2	1	J	1	1	
8	#14	Arie	FF		1	2	1	J	1	1	有家
9	#14	Ietaka	FF kanyu		1	3	1	J	1	1	家隆
10	#14	Ietsune	FF	poem	1	2	1	J	1	1	
11	#14	Ieyoshi	FF		1	2	1	J	1	1	
12	#14	Kimikage	FF	poems	1	2	1	J	1	1	

図③　B:翻訳辞書(日本文学研究用語翻訳辞書)

FFFFFFFF	¥テスト	#検索テスト用辞書									
EEEEEEEE	レコードID	K1 ID	K1キーワード	K1 異表記	K1 代表読み	K1 副読み	K1 メモ	関係属性名	K2キーワード	K2 代表読み	K2 副読み	K2メモ
1	5684	間狂言		あいきょうげん			その他／その他	間	あい			
2	5684	間狂言		あいきょうげん			その他／その他	俳	おかし			
3	5684	間狂言		あいきょうげん			その他／その他	会釈間	あしらいあい			
4	5684	間狂言		あいきょうげん			その他／その他	口開間	くちあけあい			
5	5684	間狂言		あいきょうげん			その他／その他	習間	ならいあい			
6	5684	間狂言		あいきょうげん			その他／その他	一役間	いちやくあい			
7	5684	間狂言		あいきょうげん			その他／その他	仕科	しぐさ			
8	5684	間狂言		あいきょうげん			その他／その他	台白	せりふ			
9	5684	間狂言		あいきょうげん			その他／その他	能力	のうりき			

図⑦　C：知識支援辞書

5. それぞれの辞書の包含関係

まず、それぞれの辞書に収録された語句の関係を図⑤に示す。階層関係にある辞書には重出語句を含むが、ここに示したものは重出語を単一化したもので示している。

各辞書は人手によって抽出された語句に関する類縁語が定義されているかどうかという点で、A：連想辞書（国文学研究語彙辞書）とその他という関係で大別される。

双方の辞書間に共通する語句は、全体でそれぞれ25％程度だが、内訳が異なる。

6.『源氏物語』論文ナビゲーションの可能性

次に、これらの辞書が、論文目録のデータに対して、どのようなパフォーマンスを発揮するか、検証を試みた。

なお、実験を行うに際しては、古典を扱う論文中で、最多の出現数を誇る『源氏物語』についてどれだけの一致率を見ることとした。

この試みから、分かることに、

①国文学における『源氏物語』に関する知識体系の占める位置

②類似する情報群から個別情報へのナビゲーションがどれだけ可能か

ということである。

前述のように『源氏物語』や『宮沢賢治』のような、個人の能力を超える膨大な量の論文・研究書・一般書が執筆・生産される研究分野では、データベースによる検索・評価・引用、新たな知見の提示といった、通常採られる研究手法が機能しない状況にある。そのため、研究者は、人脈による精選・選別化という方法で研究動向を把握することがなされることが多いのが現状である。

¥Ｔ源氏物語のモデル,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ源氏物語のモデル（承前）,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ源氏物語著作の時期,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ源氏物語論の考察,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ源氏物語に表はれた物の気に就て,,¥Ｓ源氏物語,¥Ｊ紫式部

〈中略〉

¥Ｔ源氏物語の鑑賞,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ源氏物語を学ぶ人の為めに,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ壷前栽（一）（源氏物語）,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ源氏物語の性質,,¥Ｓ源氏物語,¥Ｊ紫式部

〈中略〉

¥Ｔ壷前栽（二）（源氏物語）,,¥Ｓ源氏物語,¥Ｊ紫式部

¥Ｔ陽明文庫本源氏物語青表紙系本文の仮名遣い―「お」と「を」の仮名遣い,,¥Ｓ三宝絵詞，色葉字類抄，大般若経音義，源氏物語，下官集，明月記,¥Ｊ藤原定家，紫式部

¥Ｔ〈翻〉　国冬本源氏物語　１（翻刻　桐壷・帚木・空蝉）,¥Ｒ源氏物語（国冬本）,¥Ｓ源氏物語,¥Ｊ紫式部

図⑧　『源氏物語』に関する論文例

	Ａ：語彙辞書	Ｂ：翻訳辞書	Ｃ：知識辞書	源氏の論文数
一致語彙	2,941	201	879	5,792
全語数	35,582	1,876	17,084	
比率（一致率／全語数）	8%	11%	5%	
一致延べ頁（175頁中）	2,026	2,740	5,893	

図⑨　『源氏物語』に関するヒット数

確かに、従来専門分野というものは、そのような性質を持つものであった。しかし、近年は、学問の国際化・脱領域化の流れの中で、旧来型の文学研究の講座を持つところは減少の一途をたどっている。

しかしながら、国文学研究論文目録に収載されるデータ件数は毎年１万２～４千件で推移している。

このことは、文学研究の担い手が文学研究という場から拡散しつつある証左でもある。そのような状況を踏まえか考えるに、初学者やいわゆる専門外の研究者への足がかりとして機能するデータベースデザインは、今後ますます重要なものとなるのである。

もとより、図⑧にも示したように、きわめて限られた情報しか提示されない論文標題を便りに、そこからデータマイニングに導くことには限界がある。

しかし、データベース検索における対する信頼性を高めることは、『源氏物語』のみに特化した研究者にとっても、関連する他の学際領域との交流をうながす契機ともなり、そのことがひいては学界の活性化をうながすことにもなろう。

7. 文学研究における『源氏物語』研究の位置

作成した辞書をもとに、『源氏物語』に関する研究論文がどれだけヒットしたかを示したものが図⑨である。

本データで使用した辞書は、国文学研究にお

ける古典分野全体を通観する辞書より取材されたもので、見方を変えれば、『源氏物語』に関連する知識体系の占める位置を、研究語彙で図るならば、およそ５％から 11 ％の間に位置するということになるが、どうだろうか。

8. まとめにかえて

検索結果の有効性が、辞書の特質の差によって、どのように反映したかということを表現・説明するかということは意外に難しい。

今回の検証を通して、確かにヒット率は向上したものの、その感触をどのように表現するか、さらにはそれぞれの辞書による検索効率の差は、具体的にどのように違うかをどうやって説明するかなどといった、評価方法にかかわる新たな課題も発生した。

このことは、辞書データを形成するに際しては、少なからぬ時間と労力とが必要となるにもかかわらず、その成果・重要性に対する周囲の理解が十分でないという社会的状況とも密接に関わる問題といってよかろう。

また、特定の研究主題が、学問全体におけるどのような位置づけ、関連性を持つかということについて、全体の研究用語に占める関連概念との照らし合わせによって提示できるのではないかという可能性が見えてきた。

語彙の豊富さは、それに関する文化の厚みを象徴するということは一般に諒解される所である。

その意味において、学問動向の変化をとらえるひとつの指標として、本論で試みた方法は有効である。

注・参考文献

[1]相田満:和漢古典学のオントロジ(1)，2004,科研報告書,国文学研究資料館ほか
[2]相田満:「記録・情報・知識」の世界―オントロジ・アルゴリズムの研究,2004,中央大学出版部
[3]㈱平和情報の Hapiness データベース使用。
[4]作成責任者：新井栄蔵、松村雄次。
[5]by Jin'ichi Konishi ; translated by Aileen Gatten and Nicholas Teele ; edited by Earl Miner:A history of Japanese literature,rinceton,N.J Princeton University Press, c1984-1991
[6]作成責任者：藤原鎮男
[7]作成責任者：相田 満