人文科学とコンピュータ 37－2
（1998. 1. 31）

インド学仏教学論文データベース INBUDS を用いた術語間関係の大きさの推定について

相場徹, 生出恭治
{aiba,k-oide}@vacia.is.tohoku.ac.jp
東北大学大学院 情報科学研究科

概要

本稿で我々はインド学仏教学論文データベース INBUDS の中に出現している術語間の 関係の大きさ (語彙的な近さではない) を計算する方法について述べる。

INBUDS が論文書誌データベースであること、またインド学・仏教学の分野ではまだ十分な電子辞書がないことから、我々は INBUDS の中で 同じ論文の中に出現している術語間の関係に着目し、これらの術語間の相互情報量を用いた術語間関係の推定をおこなった。しかし直接的な術語間の関係を用いただけでは十分な情報量が得られなかったため、我々は「チェイン」を通じた 間接的な術語間の関係も利用して実験をおこなった。

また我々はこの結果を INBUDS の検索エンジンに取り込み、ユーザの検索支援に利用している。

Calculation of the Relational Value between Technical Terms in Indian and Buddhist Studies Treatise Database (INBUDS)

Tooru AIBA, Kyoji OIDE
{aiba,k-oide}@vacia.is.tohoku.ac.jp
Graduate School of Information Science, Tohoku University

Abstract

In this paper, we propose a method to calculate the relational value (but not lexical association) between technical terms appeared in the INBUDS, i.e. the bibliographical database for Indology and Buddhist Studies in Japan.

Because of the dearth of electronic lexicons, we tried to develop a method to estimate the mutuality of words which appear in the same article in INBUDS. Since the estimate, in terms of simple, direct word correspondence was not sufficiently reliable, we have also incorporated indirect word correspondence by employing the idea of "chain" as well.

Finally, we describe the incorporation into the INBUDS search engine of our calculations.

1 はじめに

計算機が一般に普及してくるとともに、人文科学の諸分野、とくに文献学の領域においても計算機を利用した研究に関心が持たれるようになってきた。文献学は、コーパスさえ用意すればコーパス中の単語を検索するなどのことは簡単にできることもあり、研究に計算機を利用したとき その効果が出やすい分野であると考えられる。それゆえ大規模コーパス構築のプロジェクトが世界各地でおこなわれている。一方で 木越 [4] のようなコーパスからの単語の検索方法について簡単に説明した論文も求められている。

このようにコーパスの構築・利用の動きが盛んなのに対し、コーパスの校正作業あるいは電子的な辞書の構築などの作業は後回しにされる傾向がある。このうちコーパスの校正については、インド学など研究者の数が限られ、また対象言語がサンスクリット語・チベット語など特殊な言語にならざるを得ない分野では 人間の手間を最小限に抑えることができるデータ校正支援環境が必要である。計算機を用いたコーパス校正支援環境を構築するためには データを機械で扱うための辞書、また計算機用の知識を取得するために用いる 信頼性の高いコーパス等の存在は必須である。

かような状況において公開されたのが「インド学仏教学論文データベース (INBUDS)」 [1] である。これは国内で発表されたインド学仏教学関連の学術論文を網羅的に集めた書誌データベースであり、約 16000 件の論文データが収録されている大規模なものである。INBUDS は論文書誌データベースなので術語・固有名詞が多量に収録されている。すなわち INBUDS は複数の言語に対応した術語・固有名詞辞書として応用できる可能性がある。またこのデータベースは無償配布されているものゆえ、我々の成果がそのまま一般に公開され、誰もが利用可能である。そこで我々は INBUDS を用いて以下のことを行なうことにした。

- 検索システムおよび検索支援環境の構築
- データ入力および校正支援環境の構築

我々はこれら一連の作業の手始めとして、検索支援環境の構築についての実験および評価をおこなった。我々は検索支援の具体的な内容として「ユーザが入力した術語と関連している術語を推定し提示すること」 を目的としたが、これにより INBUDS が抱えるデータの記述あるいは構造に関する問題点が明らかになり、それが次のステップへの足がかりになると考えたからである。

2 INBUDS 論文データベース

INBUDS 論文データベースについて、データの構造および問題点などについて簡単に述べる。

2.1 データの構造

掲載誌	
番号	0-00040-00001
名	東北大学文学部研究年報
号	1
年	19510300
ISSN	ISSN 0563-6566
論文	
番号	0-00040-00001-001
著者	金倉 / 円照 // かなくら / えんしょう
題目	タルカバーシャ / The Tarkabhāṣā Keśava Miśra
副題	印度の論証法 / A Treatise of Indian Dialectics
頁	pp.67 - 133
地域	インド
時代	中世
分野	インド哲学 / ニャーヤ学派
文献	タルカ・バーシャー
人物	ケーシャヴァ・ミシュラ
術語	tarka, 六句義

表 1: INBUDS 所収の論文データの例

INBUDS に収録されている論文データの例を表 1 に示す。INBUDS には表 1 のような構造を持った論文データが約 16000 件収録されている。データの特徴は以下のようなものである。

- すべての論文データには「論文番号」がつけられており、この論文番号によって一意に論文が特定されること。
- 地域・時代・分野・文献・(他の) 術語の各項目に記述されているのはデータ入力者[1]が論文から選び出した術語 (個数は自由) であること。

2.2 データの問題点

INBUDS が抱える問題点として、データ表記の不統一性、および記述ミスの多さに代表される信頼性の低さがあげられる。

このうち「データ表記の不統一性」については、INBUDS データベース構築の際に どの術語をどう

[1]著者自身の場合もあるが、それは一般的ではない。

収録するかをデータ入力者に完全に任せてしまったため、たとえば「brahman」「ブラフマン」「梵」など同じ概念であるはずのものが異なった術語・表記として入力されている例がかなりある。

また INBUDS 構築プロジェクトはかなり大規模なものであり、以下のような状況で構築されているものである。

- インド学仏教学に関係した全国の「協力機関」(37 機関) が入力作業を分担しておこなったものであること
- 入力作業は現在まで 10 年ちかくの歳月をかけておこなわれ、今後も継続しておこなわれる予定であること

このように物理的また時間的にかなり分散した環境で構築されるため、データ表記の不統一は将来的にも INBUDS データベースが抱える問題として残るのではないかと思われる。

記述ミスの多さ については、データをチェックする機関として「データベースセンター」という機関を設立してデータを配布・訂正するという取り組みがなされているが、データの分量が大量なのに比してデータ訂正作業をおこなう人員の数が限られてしまうため、訂正作業が十分になされる状況ではないという問題もある。

2.3 対処

このような INBUDS が抱える問題点をふまえると、与えられたデータをもとにした検索システムを作成するだけでは作業としては不十分であり、データ訂正・整理を支援する環境の作成までも視野に入れた取り組みが必要であるといえる。

そこで我々はまず検索支援を目的にして「関係が大きい術語のよせあつめ」をおこなうこととした。ある術語が検索されたとき、その術語と関係が大きいと思われる術語集合を示すことができればユーザにとって有益なのは勿論のこと、データの構築作業をおこなう人たちにとっても有益であると考えたからである。

3 術語間関係の大きさの推定

本稿では術語間関係の大きさを推定するため、術語間関係の大きさをスコアという数値的な形で算出し、そのスコアをもとにして、それぞれの術語と関係が大きい術語の推定をおこなう。そこで問題となるのは いかにして術語間関係の大きさを数値で表現するか、ということになる。

論文データベースという性質上、INBUDS に載せられたデータはほとんどが特殊な学術用語であること、またインド学関連の電子的なシソーラス・術語辞書が存在していないことを考慮に入れると、現存する INBUDS の中にある記述を利用して術語間関係の推定をおこなうしか方法はないと思われる。

そこで我々は、同じ論文に出現する データ間の「共起関係」を利用して術語間の関係の大きさを計算し、その値を利用して術語間関連の推定をおこなうこととした。

3.1 方法

「同じ論文の中で一緒に出現しやすい傾向にある術語どうしの関係は大きい」という我々の直感にもとづき、術語間関連の大きさを計算する。 たとえば表 1 の例であるが、この「(他の) 術語」の中に出現している “tarka” と “六句義” という術語のあいだには何らかの関係があると推定できる。このように共起しやすい傾向にある術語同士ほど関係が大きいとするのである。

そこで我々は Hindle [2] が用いた名詞のクラスタリングの手法を流用した。Hindle はコーパス中における主動詞と主語名詞・目的語名詞のあいだの共起の分布を利用したが、我々は INBUDS の術語の分布を利用する。ただし、それぞれの術語の語彙的な類似を推定するには INBUDS だけでは情報量が不足であると考え、とりあえずは術語間の関係の大きさを推定するだけにとどめた。

まず INBUDS における単語の出現 / 共起について以下のような定義をおこなう。

- [出現] それぞれの語がキーワードとして採用された数を数え、これを **出現回数** とする。
- [共起] ある、ひとつの論文のキーワードとして術語 A と術語 B が存在しているとき、**術語 A と術語 B が共起している** と定義する。

  また、ある術語と術語が「共起」して出現している回数を **共起回数** とする。

そして、このように定義した上で、単語間の関連の強さを求めるため Hindle [2] が提案した以下の式を用いて単語間の距離を計算することにする。

$$I(a,b) = \log \frac{P(a,b)}{P(a)P(b)} \qquad (1)$$

ここで $P(a)$ は術語 a が出現する確率、また $P(a,b)$ は術語 a と術語 b が共起する確率を示している。また左辺の $I(a,b)$ は術語 a,b のあいだの「相互情報

量」(mutual information) を示しているが、この値が大きくなればなるほど 術語 a,b の関係は大きい と言うことができる。

この式は以下のように展開される。

$$
\begin{aligned}
I(a,b) &= \log \frac{P(a,b)}{P(a)P(b)} \\
&= \log \frac{\frac{f(a,b)}{N}}{\frac{f(a)}{N}\frac{f(b)}{N}} = \log \frac{f(a,b)\times N}{f(a)f(b)} \\
&= \log \frac{f(a,b)}{f(a)f(b)} + N'(N'\text{は定数}) \quad (2)
\end{aligned}
$$

ここで $f(a,b)$ は単語が共起している共起回数、$f(a)$ は単語 a の出現回数をあらわしており、N, N' はともに定数で、N は術語の全体の個数である。[2]

3.2 実験

上述の式を用いて術語間関係の大きさを計算し、またその結果求められた値についての評価をおこなう。

3.2.1 出現回数の数えあげ

論文 A	智度論, 中論, 般若経
論文 B	般若灯論, 般若経, 維摩経, 中論

表 2: 記載されているデータの例

表 2 の例で説明する。この場合、「中論」「般若経」は論文 A,B でそれぞれ出現しており、合わせて 2 度出現しているので出現回数は 2 回、それ以外の術語はすべて 1 回、といった具合に INBUDS に記述されているすべての術語について出現回数をカウントする。

3.2.2 共起回数の数えあげ

表 2 の論文 A につけられたキーワード群をもとにして「『智度論』と『中論』の共起回数が 1 回」「『智度論』と『般若経』の共起回数が 1 回」「『般若経』と『中論』の共起回数が 1 回」といった具合にカウントをおこなう。また 論文 B をはじめとした、他のすべての論文についても同様にカウントをおこない、すべての術語間関係について共起回数のカウントをおこなう。

[2]我々は、それぞれの術語間での $I(a,b)$ の値の大小関係を比較することを目的としているため、定数である N' の値は無視できる。

3.2.3 評価の基準

このようにカウントした出現・共起回数のデータを用いてそれぞれの術語間関係の大きさの値の計算をおこなう。その結果得られた値の妥当性を確認するため、我々は以下のような仮説をたてた。

- 同じ、あるいは類似した分野に属する術語間の関係は、そうでないもの同士の関係よりも一般に強くなるのではないか。

たとえば初期仏教系の術語なら、それと同じ初期仏教系の術語との関係が大きくなりやすく、一方で同じ仏教でもインド仏教の最後期に属する タントラ密教系の術語とのあいだの関係は小さくなるのではないか、ということである。

そこで我々は 金倉圓照 著『インド哲学史』 [3] を用いて、そこから術語、今回はとくに文献名を手作業によって抽出した。我々が 金倉 [3] を用いた理由は以下のとおりである。

- この本は初心者向けの概説書なので、各章が「ヴェーダの宗教と思想」「ブラーフマナの世界観」など分野ごとに分けられていること
- 上と同じ理由により、各分野において代表的・重要と思われる術語が抽出できること

上述の仮定に従えば、金倉 [3] の中の同じ章に出現している術語どうしの関係は大きくなる、すなわち術語間の関係の大きさを示す数値の平均は大きくなるはずであり、また同様に異なった章で出現している術語間の関係の大きさを示す数値の平均は小さくなるはずである。それゆえ実験の結果がこの仮定のとおりになっているかを調べることにより、我々がおこなった術語間関係の推定のしかたが妥当であったかどうかが確認できる。

3.2.4 評価の方法

前節で 金倉 [3] から抽出した書誌データを用いて、以下のような手順で実験をおこなった。

- INBUDS に対応する術語がないデータは除く (例: *nyāyāvatāra* は [3] にあるが INBUDS には記述がない。このようなデータは除く)
- 術語を、それぞれの術語が出現している章[3]ごとにグループ化する。
  (例: *mīmāṃsāsūtra* は「11. ミーマーンサー学派」で説明されているので、「11 章」のグループに入れる)

[3]2-4,7-20 の計 17 章。

- それぞれの章ごとに、それと同じ章に含まれる術語とのあいだの共起スコアの平均値を求める。同時に、他のそれぞれの章に含まれる術語とのあいだの共起スコアの平均値も求める。

  その結果、同じ章に含まれる術語とのあいだの共起スコアの平均値が最大になったときは ○ (正解) とし、それ以外のときは × (間違い) とする。

- 可能なすべての術語の組み合わせのうち、どの程度の組み合わせから共起スコアを取ることができたかについての調査をおこなう。これを今回の手法による「カバレジ」として扱う。

3.3 結果および評価

章数	○	×
17	1 (5.9 %)	16

表 3: 実験結果 (精度)

	総組数	スコア組数	平均スコア
同じ章	1326	225 (17.0%)	0.833
別の章	15145	466 (3.1%)	0.830
全体	16471	691 (4.2%)	0.831

表 4: 実験結果 (カバレジ)

実験の結果を表 3 および表 4 に示す。表 4 における**総組数** は金倉 [3] から作成することができる可能な術語の組み合わせ数であり**スコア組数** は **総組数** のうちスコアが実際にとれた組の数および割合である。これらの結果から、以下のことがいえる。

- 表 4 から、カバレジを「同じ章」と「別の章」とで比較してみた場合、「同じ章」では約 17 % の術語間関係のスコアがとれたのに対し、「別の章」では約 3 % 程度の術語間関係しか取り出すことができなかった。

- 表 3 の結果をみると、正解は 1 組 (5.9 %) のみ、とかなり悪い

このように結果はかなり悪いものになってしまった。この原因について考えてみる。表 4 によると、「同じ章」におけるカバレジが総術語組合せの 17 % ほどであるのに対し、「別の章」におけるカバレジは 3 % 程度にしかならなかった。この数字から、今回の計算では統計的に信頼できるほどの共起データが取れなかったこと、すなわちデータのスパースネスが原因で精度が悪くなっていることが容易に推測できる。ただし「同じ章」のカバレジがそれ以外のものよりかなり大きいことから、論文データにおける共起情報を利用して術語間の距離を計算する我々の方法は、方針としては間違っていない、といえそうである。

4 チェインを用いた拡張

前節の実験の結果、術語間の共起スコアを用いて術語間の関係の大きさを推定することが妥当であることが推定できた。しかしカバレジが全体的に低すぎることが問題になった。

そこで我々は「チェイン」という概念を用いて、術語間の関係のとりかたを拡張することにした。本節ではこの拡張の内容および評価について述べる。

4.1 方法

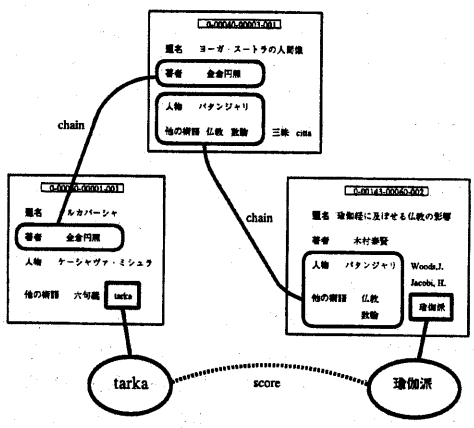

図 1: 'tarka' と「瑜伽派」の間接的な関係の取得

我々は 複数の論文で採用されている術語を利用して、間接的な術語間関係も考慮に入れた計算をおこなうことにより、データのスパースネスを補うことを考えた。たとえば 図 1 のように、複数の論文中に登場している術語を利用することにより、直接には共起していない **tarka** と **瑜伽派** のあいだの関係の値をとることを可能にする。我々は、ここでの 「著者: 金倉円照」のような、複数の論文データに出現している術語を結び付けるものを「チェイン」と名付けた。

このような「チェイン」を用いて単語間の関連の強さを計算する場合、関連のスコアをつける要素になるのは以下のものになると思われる。

- 関係を調査したい各術語と、経路となっている各ノードのあいだのそれぞれの単語間関係の大きさの総体

すなわち 図 1 の例では

1. 術語:tarka と著者: 金倉円照 の関係の大きさ
2. 著者: 金倉円照 と人物: パタンジャリ (or 術語: 仏教 ...) の関係の強さ
3. 人物: パタンジャリ (or 術語: 仏教 ...) と術語: 瑜伽派 の関係の強さ

これらの 3 つの関係の値を合わせることによって、最初の 術語:tarka と最後の術語: 瑜伽派 のあいだの関係の値が計算できると考えた。

4.2 実験

同じ論文中にある各チェインのノードになっている術語、たとえば図 1 の上にある「ヨーガ・スートラの人間像」という論文における 著者: 金倉円照 と 人物: パタンジャリなどとの関係の値のとりかたは前節で用いた Hindle の手法をそのまま用いた。

また「間にあるチェインの数が多くなればなるほど術語間の関係は遠いものになる」という我々の直感を反映させるため、チェインのノード間の共起スコアを含むすべての術語間の関係値を調整し 1 以下の正数にする[4]。

こうしておいて複数のチェインを要する術語間関係を計算するときに、各チェインについた共起スコアの値の積をとり、その結果を術語関係の関係のスコアにすれば、チェインの数が多くなるほど術語間関係は小さくなるという我々の直感が活きることになる。

さらに我々はスコア計算をおこなう際に以下の制約を設けた。

- チェインの最大長に制限をかける。(今回は 3 以下)
- 同じ種類のノードをたどらない。たとえば一度「著者」でチェインをたどった後は「著者」をチェインのノードとすることができない。
- チェインのノードとして用いるのは「文献」「人物」「術語」「著者」のカテゴリに限定する。

これらの制約を加えることによって計算量をなるべく抑えるようにした。この最後の項目について簡単に理由を説明する。

	地域	時代	分野
総数	15196	6565	27924
パターン数	580	1155	2209
平均	26.2	5.683	12.64
分散	106458	1276	17352
(1) 数	5768	893	4266
術語	日本	鎌倉時代	日本仏教
割合 (%)	37.95	13.60	15.27
(2) 数	4688	549	3023
術語	インド	現代	インド仏教
割合 (%)	30.85	8.362	10.82
(3) 数	2586	392	2085
術語	中国	鎌倉	中国仏教
割合 (%)	17.01	5.971	7.466

表 5: おもな項目でのデータの分布 (1)

	文献	(他の) 術語	人物
総数	30150	43304	23524
パターン数	12333	23519	7515
平均	2.444	1.841	3.130
分散	116.5	15.19	529.0
(1) 数	816	181	1649
術語	教行信証	念仏	親鸞
割合 (%)	2.706	0.417	7.009
(2) 数	308	138	534
術語	法華経	縁起	法然
割合 (%)	1.021	0.318	2.270
(3) 数	191	134	374
術語	選択集	戒律	善導
割合 (%)	0.633	0.309	1.589

表 6: おもな項目でのデータの分布 (2)

INBUDS における、おもな項目での術語データの出現分布を表 5 および表 6 に示した。これらの表からデータの記述内容の分布を項目ごとに見てみると、表 5 にあげた「地域」「時代」「分野」と表 6 にあげた「文献」「人物」「術語」とでは術語の記述の分布にかなりの違いがあることがわかる。[5]

[4] しかし関係の値の大小関係を変更してしまうと問題があるので、その点には注意を払ってある。

[5] 「著者」は総数 16429, パターン数 3616, 平均 4.54, 分散 74.22 と 「文献」などのほうに近い構成である。

たとえば表 5 の「分野」について見てみると、記述されている術語は全部で 27924 個 (2209 パターン) であるが、そのうちの 4266 個 (15.2 %) が「日本仏教」である。このように記述内容が一部の術語に極端に偏ってしまっている項目をチェインに利用してもあまり意味がないと思われたため、今回はチェインの候補とはしないこととした。

ところで図 1 の例では「パタンジャリ」「仏教」「数論」といった複数の術語がチェインのノードの候補になっている。このような場合、それぞれを別のチェインとしてスコア計算をおこない、最終的に術語間の関係値が最大になったもののみを採用するようにして実験をおこなった。

4.3 結果および評価

チェインによる拡張を加えて術語間関係の大きさを計算した結果を表 7 および表 8 に示した。このうち表 7 はカバレジの変化を、表 8 は精度の変化を示している。どちらも **plain** がチェインを用いないときの結果、**chain** がチェインを用いた拡張の結果を示している。また表 7 での **max** は最大チェイン長を、**総組数** は金倉 [3] から作成できる可能な術語間の組み合わせ数を示していて、そのうち術語間関係のスコアがとれた組の数および割合を示しているのが **スコア組数** である。

type	max		総組数	スコア組数	
plain	0	同じ章	1326	225	(16.97 %)
		違う章	15145	466	(3.08 %)
		全 体	16471	691	(4.20 %)
chain	3	同じ章	1326	674	(50.83 %)
		違う章	15145	4575	(30.21 %)
		全 体	16471	5249	(31.87 %)

表 7: 拡張の結果 (カバレジ)

type	章数	○	
plain	17	1	(5.9 %)
chain	17	12	(70.6 %)

表 8: 拡張の結果 (精度)

表 7 から、拡張によって「同じ章」で 3 倍以上、「違う章」では 10 倍ものカバレジの向上が見られたことがわかる。また表 8 から、チェインを用いた拡張の結果、実験をおこなった 17 章のうち 12 章 (70.6 %) で「同じ章」とのスコア値の平均値が最大 (すなわち正解) となり、拡張する以前の結果と比較して かなりの精度の向上が見られたことがわかる。

chain	スコア組数	
0	456	8.69%
1	505	9.62%
2	1751	33.36%
3	2537	48.33%

表 9: chain 長ごとのカバレジの内訳

さらに我々はチェインの長さとカバレジの関係を調査するため、スコアがとれた組における、チェインの長さごとの内訳を調査した。その結果を表 9 に示す。この表によって、チェインの長さを長くすればするほどカバレジが向上していくであろうことが確認できる。

5 考察

チェインを用いた拡張を加えた実験の結果、以下のことが言える。

- カバレジに関しては、チェインをつくることによって、結果が大幅に向上した。またチェインを長くすればするほどカバレジが向上することも予想できる。
- チェインを用いることにより 精度が飛躍的に向上した。この原因について調査をおこなった。

	平均		chain 長ごとの平均スコア			
type	長さ	スコア	0	1	2	3
違う章	2.30	0.49	0.87	0.66	0.52	0.40
同じ章	1.63	0.58	0.86	0.66	0.51	0.40

表 10: chain 長ごとの平均スコア値

表 10 に「同じ章」「違う章」それぞれのチェインの長さ・スコアの平均値、またチェインの長さごとのスコアの平均値を示す。チェインの長さごとの平均値を比較してみると、チェインの長さが 0-3 のいずれの場合も「同じ章」と「違う章」の両方の値がほぼ同じであることがわかる。このことから「同じ章」と「違う章」のあいだのスコア平均値の差は、チェ

インの長さの平均の違いによるものであることがわかる。

この結果は「『同じ章』のほうが『違う章』よりも単語間の関係が近いので要するチェインは少なくて済むはず」という我々の仮定に沿うものであることから、今回我々がとった術語間関係の大きさの計算の方法、およびチェインを用いた拡張の手法が妥当であったことを示していると考えられる。

現在の状態ではまだ「同じ章」のほうが「違う章」よりカバレジが大きい。このことから、可能なチェインの最大長をさらに大きくして計算をおこなうと、「同じ章」よりも「違う章」のカバレジの向上に効果があると考えられる。術語間の関係のスコア値は、チェインの長さが長くなるほど小さくなる傾向にあるので、可能なチェインの最大長を伸ばせば伸ばすほどカバレジが向上するだけでなく、精度もさらに向上することが見込まれる。しかしチェインの最大長を大きくすればするほど、計算すべき術語間の組み合わせ数が爆発的に増加していくため、可能な組み合わせ候補の数をどう絞り込んでいくかが課題となる。

6 おわりに

我々は INBUDS にある術語の直接的・間接的な共起関係を利用した術語間関係の推定をおこなった。その結果、ある程度の信頼性が持てるデータを得ることができた。そこでこのデータを利用して、我々が開発した INBUDS の論文検索システム「印仏検」に検索支援機能を追加した。ただし機能の内容としてはユーザが入力・選択した術語と関係が大きい術語をリストとして示す、という簡単なものである。「印仏検」は以下の URL で一般に公開している:

`http://www.vacia.is.tohoku.ac.jp/cgi-bin/ibsis`

このページには毎月 2,500-3,000 程度のアクセスがあるが、これはインド学という学問分野で計算機を用いたデータ検索・検索支援に対する需要がかなり高まっていること、そして我々のシステムがある程度までその需要に答えるものであること、の現れではないかと考えている。

今後の課題として以下のことがあげられる。

- 可能なチェインの最大長をさらに大きくしてスコア計算をしてみる。同時に、組み合わせ爆発を抑えるための工夫をさらに追加する。
- 現在のところ、チェインのノードとして「著者」「文献」など一部の項目しか利用していないが、他の項目も活用できないかどうかを検討する。
- 表音文字による記述を利用した、同一の術語の推定が可能かどうか調査をおこなう。たとえば ātman はこれ以外にも「アートマン」「我」として表記されている。このうち "ātman" と「アートマン」に関しては、ローマナイズをカタカナ表記に変換するなどの規則を用意すれば、同じ術語と推定することが可能であると思われるので、調査してみる。

最後に、本稿の冒頭で述べたとおり我々は INBUDS データベースの信頼性をあげることにより、将来的にはこのデータベースがインド学関係の術語・固有名詞辞書として活用できるものになると考えている。それゆえデータ校正支援環境の構築および INBUDS データ入力支援環境の構築についても今後の重要な課題として位置付けている。

謝辞

大規模データベースである INBUDS を企画・構築され、また無償公開された 江島惠教先生 をはじめといたします日本印度学仏教学会の先生がたに感謝いたします。また 先生がたへの連絡の窓口となってくださったり、我々の検索システムについて種々のコメントをくださいました東京大学東洋文化研究所の鈴木隆泰さんに感謝いたします。

参考文献

[1] 日本印度学仏教学会データベースセンター,『インド学仏教学データベース』,1996.

[2] D. Hindle, 'Noun Classification from Predicate Argument Structure.', ACL 90-6,pp.268-275,1990.

[3] 金倉圓照,『インド哲学史』, 平楽寺書店,1962.

[4] 木越治,「国語国文学徒のための情報処理講座(入門編)」, 情報処理語学文学研究会会報 19, 1996.