社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－CH－61 (3)
2004／1／23

Nグラムと文字データベースによる漢字仏教文献の分析

師　茂　樹†

近年、漢字仏教文献をNグラムモデルによって分析する方法が成果をあげつつある。しかしながらこの方法は、文字の同定などの本質的な部分において文字コードに依存しているのが課題であった。本稿では、筆者が開発に参加しているCHISEプロジェクトの文字データベースと文字処理システムを用いて、文字コードに依存しないNグラム処理の試みについて報告する。

Analysis of Chinese Buddhist Texts with N-gram Model and Character Knowledge Database

SHIGEKI MORO†

In recent years, statistical analysis of Chinese Buddhist classics using N-gram model is getting a result. However, it has not been solved that N-gram processing, such as identification of a character, is essentially dependent on a character code sets. In this paper, I would like to report a test implementation of N-gram processing independent of any character code sets, using the character database and character processing system of CHISE project which I have participated in development.

1. 問題の所在

1.1 作業の支援から発想の支援へ

近年、漢字文献データベースが急速に充実してきており、漢字仏教文献についても『大正新脩大蔵経』などの叢書レベルでの入力、公開が盛んに行われている。しかし、このようなデータベースが構築されたことによって、文献学的な研究の便利になったか?と問われれば、ある一面に限れば首肯しうるものの、全体としてはむしろ文献学者に今まで以上の労苦を強いることになりそうだ、と答えざるを得ない。従来の紙ベースの研究では、研究対象の範囲が人間の能力や伝統、研究史・出版史などによって制限されているため、そのような制限のある中でデータベースを使えば確かに効率は劇的に上がるだろう。しかし、大規模なデータベースを横断検索が可能な環境では、そのような制限は意味をなさなくなる。データベースが充実すればするほど、研究者がコンピュータの前に座る時間は相対的に短くなり、逆に膨大な検索結果を検討する時間が増大していくであろうことは容易に想像できる。

一方、キーワードや正規表現などによる通常の検索はある程度の予見を必要とする（テキストの中身がわからなければ検索はできない）が、データベースが大規模化すると、それに比例して予見が立てづらくなる、という面もある。また、そもそも、文献研究において予見があること自体が疑問視されるべきことであろう☆。古典研究においては、研究者の先入見を生み出す研究者自身のコンテクストを如何にコントロールするかというのが、大きな方法論的課題となっているからである。

近藤泰弘氏が指摘するように[5]、コンピュータを利用したテキスト研究によってこれらの問題の解決に大きな前進があることが期待される。特に、氏が指摘する次の二点は非常に重要である。

（1）徹底的に網羅的な研究（すべての単語・すべての文字の単位にまで網羅性を及ぼすことが可能になる）。

（2）それによって現代人には通常認知できないデータの構造性や規則性を探り出す。それは、現代人の古典語に対する「内省」(introspection)（語感）の欠如を補うことができ、文学研究に貢献する。なぜなら、古典文学の正しい読みにとって、「内省」（文法的直観と言語

† 花園大学 Hanazono University
s-moro@hanazono.ac.jp

☆ 豊島正之氏が指摘するように、TEIに対してDTDの予見性に対する批判があったことは注目される[17]。

外知識など）の欠如は大きな障害のひとつだからである。

これは語学・文学研究をテーマにしたレジュメからの引用であるが、仏教学を含む古典テキストに関する多くの研究にも同様に当てはまる。コンピュータは人間のようにテキストの内容を「読む」ことはできないが、逆に人間のようにコンテクストに囚われることのない「徹底的に網羅的な研究」が可能であるため、コンテクストに束縛されがちな研究者の発想を支援することが期待できるのである。

1.2 なぜ N グラムなのか

ところで、このような確率統計的なテキスト分析は、従来、キーワード採取や形態素解析によるものが多かった。しかし、上記のような目的であることを鑑みれば、研究者の予見がより混入しやすいこれらの手段を使うことは積極的に避けたいところである。また、これらの作業は多大な労力を必要とするため、発想支援システムとしては欠点となろう。

もちろん、まったくテキストに反映されないことについては、どのようなモデルによっても探し出すことはできない。しかし、我々の気づかない思わぬところに、当時の常識の断片を見つけ出すことができるかもしれない。アンソニー・ケニー氏は次のように述べている[3]。

> 文体に指紋があるとすれば、それはどのようなものだろうか? それはおそらく、ある著者の文体的な特徴—例えば 'such as' の生起度数といった、まったく取るに足りないと言ってもよいような特徴を組み合わせたもの—であって、指紋と同様にその人に特有のものであろう。文体上些細で取るに足りぬ特徴だからといって、文体分析に利用しない理由にはならない。指先にある渦巻や輪が我々の容姿においては大切でも目につくわけでもないが、指紋が一生涯変わらないように、そういったものこそが著者の叙述において変化することのない特徴となるはずであり、他の書き手には見られないその人だけのものとなるはずであろう。(p. 24)

このようなテキストの「指紋」を見つけ出すのにコンピュータを使えないだろうか、というのが、N グラム・モデルを選択する大きな理由である。N グラム・モデルについてはこれまで多くの欠点が指摘されているが、従来の文献学的な研究と合致する部分も多く、他の確率的言語モデルと比較して実装が容易などの理由から簡便な発想支援システムとして有用であると考えられる。

1.3 文字コードに依存した文字処理の克服

しかしながら現時点では、単純な検索であれ確率統計的な分析であれ、文字の同定など、文字処理の本質的な部分において文字コードに依存している。Unicode[18] の *character* の定義（p. 15）に見られるように、文字コードの多くは、漢字学で言われる形・音・義に代表される文字に関する様々な要素や、文字が/に属しているコンテクストを捨象した抽象的な "文字" に対して番号をふるため、文字に対する多角的な知識が必要とされるテキスト分析などにおいては、文字コードのみで処理することは事実上不可能であると言ってよい[12]。また、文字コードの制定が古典研究を指向したものではない。したがって、成立時期や地域、あるいはエディションの異なるテキスト間の比較分析においては、目的別、分野別に文字の包摂などの（人力による）前処理が不可避的に必要になってくる。これは、作業量が増えるという点だけでなく、テキスト分析が文字コードの持つコンテクストに依存してしまうという大きな問題をはらむため、簡便な発想支援システムを目指す我々の方法にとって大きなマイナス点である。

筆者はこれまで、これらの問題を克服すべく、筆者が参加している CHISE プロジェクトの文字知識データベースと文字処理システムを用いて、文字コードに依存しない正規表現システムなどを開発、公開してきた[11]。本稿でも同様に、文字知識データベースを用いた文字単位での N グラム処理について試験的に実装を試みたことについて報告を行いたい。

2. CHISE プロジェクトについて

2.1 概　　要

ここで、CHISE プロジェクト☆[7] についてごく簡単に紹介したい。

CHISE（CHaracter Information Service Environment; 知世）プロジェクト☆☆では、文字コードに依存しない文字処理環境の実現を目標に、守岡知彦氏によって提唱された Chaon モデル☆☆☆に基づく様々な実装が開発されている。現在進行中のサブプロジェクトは以下の通りである。

- XEmacs/CHISE

☆ http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise/, http://cvs.m17n.org/chise/, http://mousai.as.wakwak.ne.jp/projects/chise/

☆☆ 以前は UTF-2000 プロジェクトと称していた。

☆☆☆ 以前は UTF-2000 モデルと称していた。

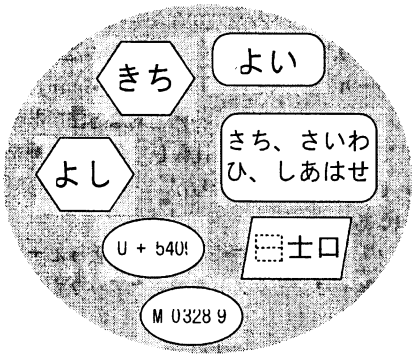

図 1 Chaon モデルにおけるある文字
Fig. 1 a character based on Chaon model

- 文字知識データベース
- libchise
- TopicMaps による
- Perl/CHISE
- Ruby/CHISE
- Ω/CHISE
- KAGE
- 文字情報の可視化

CHISE プロジェクトはメーリングリストにおける議論やオープンソースによる開発が活動の主体となっており、成果物はほとんどすべてフリーソフトウェアとして公開されている。

2.2 Chaon モデル

CHISE における文字のモデルのひとつとして、Chaon モデルがある。このモデルでは、字形・発音・意味や、既存の文字コードのコードポイントなど、様々な文字の知識（素性 feature）の集合によって文字を表現する。大雑把に図示すると、図 1 のようになる。

Chaon モデルにおいては、素性の集合☆によって文字が表現されるため、文字間の比較は図 2 のような集合演算になる。

現時点では、Chaon モデルによって素性のコンテクスト依存性を表現することはできないため、図 2 のような単純な集合演算になるが、将来的には TopicMaps☆☆等によって素性間の関係などが記述できるようになるよう、研究・開発が進められている。

☆ CHISE プロジェクト内では、これを文字オブジェクトと呼んでいる。ただし、オブジェクト指向モデルにおけるオブジェクトとは概念が異なる[12]。

☆☆ http://www.topicmaps.org/

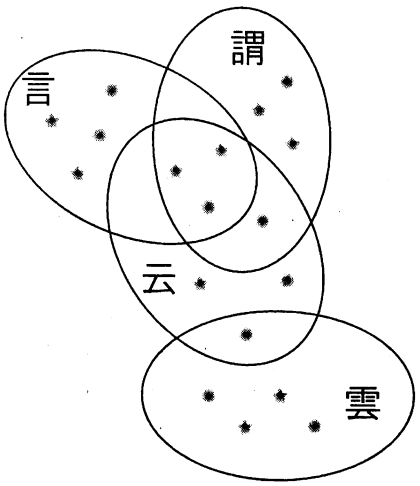

図 2 文字どうしの集合演算
Fig. 2 set operation between characters

3. 文字素性による N グラム分析の試み: 音韻分析を例に

3.1 目　　的

今回は玄奘訳『般若心経』を対象に、音韻に基づいた N グラム分析を試験的に行った。プロトタイピングが目的であり、『般若心経』自体の音韻の検討は行わない（短くて便利という理由のみでテキストを選択している）。

なぜ音韻なのかと言えば、近年、齊藤隆信氏によって進められている、漢字仏教文献中の音韻に注目した一連の研究[13)14)15)16)] を念頭においていることを述べておきたい。氏の研究においては、経典の散文（長行）中にある韻文（偈頌）的要素を抽出する作業を通じて、仏教経典における散文とは何か、韻文とは何かを考察し、また時代や版本によって散文が韻文化したりその逆が起きたりすることを検討している。なぜこのような研究がなされているかと言えば、経典解釈における一般論として、韻文が古く長行が新しい場合が多いという説があり、散文であるか韻文であるかがテキストの解釈上極めて重要な意味を持つ場合があるからである。

3.2 方　　法

3.2.1 全体の流れ

今回試みた実装においては、次のような作業の流れとなっている。

(1) 入力
(2) 前処理
- 音韻データベースの読み込み

- 句読点などの除去

(3) 文脈依存の処理

- 単語ベース

(4) N グラム処理

(5) 出力

3.2.2 音韻データベースの作成と扱い

今回は実装試験ということもあり、Unicode Consortium が公開している Unihan データベース☆の音韻に関する部分を CHISE 文字知識データベースに取り込んで使用した。原則として、Hugh M. Stimson. *T'ang Poetic Vocabulary.* Far Eastern Publications, Yale Univ. 1976. に基づく唐代音（kTang）を用いたが、そこに収録されていない文字についてはピンイン（kMandarin）で代用した。

しかしながら、本格的な文献分析においては、音韻データの底本の選択を慎重に行わなければならないのは言うまでもない☆☆。併せて、ひとつのデータベースでカバーできない場合にどうするのかについても検討せねばなるまい。

3.2.3 文脈依存音の解決方法

漢字には、前後の文脈に応じて音韻が変化するものが少なくない。それを文字レベルもしくは現状の Chaon モデルで解決することは難しいため☆☆☆、テキストレベルでの処理が必要である。現時点で考え得る処理には以下のものがあるが、今回はこのうち単語ベースでの処理のみを行っている。

(1) 単語ベース

(2) ルールベース

(3) マークアップ

単語ベースとは、下記のような単語と発音とのテーブルを用意する外延的な方法である。

般若 bo1/re3

波羅蜜多 bo1/luo2/mi4/duo1

這儿 zher

那儿 nar

利点としては作業が単純であることがあげられるが、反面、考え得る用例をあらかじめすべて用意しなければならず、また例外の例外が発生する可能性もある、という欠点がある。今回はこの方法のみを用いた。

次にルールベースとは、「$na + er \rightarrow nar$」のような変化☆4のルールを記述する内包的な方法である。利点としては、単語ベースと比較して全ての用例を準備しないですむ点をあげられるが、一般化が難しいなどの問題がある。

マークアップによる方法とは、下のようにテキスト自体に音韻やコンテクストをマークアップする方法である。

```
<tag phone="nar">那儿</tag>
<tag lang="skt">掲帝掲帝…</tag>
```

実際の作業において、すべての音韻データをテキストに混入させることは作業効率上よい方法だとは言えないだろうが、例外を記述するための方法としては有用であろう。

3.3 結果と評価

『般若心経』を音韻の 3 グラムによって分析した際の出力結果の一部を下に示す。

```
 ban1 bo1 pan2 ban3/luo2 luo1 luo5/jie1
qi4  3   1
    般羅揭  1
ban1 bo1 pan2 ban3/luo2 luo1 luo5/seng1
3   1
    般羅僧  1
bhiêt4/qiêi1/bo1    3   1
    佛依般  1
bi2/she2 gua1/shen1 yuan2 juan1 3   1
    鼻舌身  1
biêt4/gou4/biêt4    3   1
    不垢不  1
biêt4/jian3/shi4    3   1
    不減是  1
biêt4/jing4/biêt4   3   1
    不淨不  1
biêt4/mie4/biêt4    3   1
    不滅不  1
biêt4/sheng1/biêt4  3   1
    不生不  1
biêt4/xu1/gu4 gu3   3   1
    不虛故   1
biêt4/yi4/kong1 kong4 kong3 3   1
    不異空  1
```

☆ http://www.unicode.org/Public/4.0-Update1/Unihan-4.0.1d3b.txt

☆☆ 12 月 6 日に開催された漢字文献情報処理研究会第 6 回大会の場において、大阪外国語大学の山崎直樹氏より、伝統的な音韻分類である韻鏡を採用してはどうかというアドバイスを頂いた。今後、是非取り組みたいと考えている。山崎氏には記して感謝申し上げたい。

☆☆☆ Haralambous 夫妻が提唱する Probability Character Model は、文脈依存性を文字レベルで解決しようという試みである[1]。

☆4 「那儿」などははこのように変化するが、その際変化後の発音を 2 グラムとするか 1 グラムとするかは大きな問題であろう。今後、検討していきたい。

```
biêt4/yi4/se4 shai3 3   1
    不異色  1
biêt4/zeng1/biêt4   3   1
    不增不  1
bo1/luo2/mi4    3   7
    波羅蜜  7
bo1/re3/bo1 3   7
    般若波  7
```

この結果は2つの部分からなっており、最初に音韻による結果があって、それに該当する漢字の結果が続く。上の例で言えば、1行目の `ban1 bo1 pan2 ban3/luo2 luo1 luo5/jie1 qi4` が音韻の列（各文字の発音は「/」で区切られている）、次の3がグラム数、最後の1が頻度を表している。ほとんどがピンインであるが、`biêt4` などというピンインにはない表記は唐代音である。また、一番上の例のように「般」という文字に対して `ban1 bo1 pan2 ban3` などとなっているのは、文脈によって変化する発音が複数あることを示す。逆に一番下の例における「般」は、「般若」という単語においては `bo1` という発音が使われることを前もって指定しておいたため、発音がひとつにしぼられていることがわかる。

上の例では、1つの音韻列に1つの漢字列しか対応していないが、音韻データの抽象度を高める（例えば子音のみ、母音のみ、近い子音によるグループ化など）ことにより、1つの音韻列に複数の漢字列が対応する結果を得ることもできよう。それによって、テキストの内部に隠れている音韻パターンを発見するための手がかりを得ることが期待できる。

4. まとめ

以上、雑駁な報告であったが、発想支援システムとしての文字知識データベースによるNグラム分析について、その有用性が示せたのではないかと思われる。

今後の課題としては、以下のことを予定している。

- 汎用化
  - 音韻だけでなく、任意の文字素性によるNグラム処理を可能に。
- データベースの改良と充実
  - 韻鏡のデータベース化。
  - その他
- Nグラム処理の改良
- スムージング処理[4]の実装。
- 複数テキストの比較
  - NGSM[2)8)]への応用。
  - クラスター分析への応用[4)9)10)]。

謝辞　本報告は、科学研究費補助金・若手研究B「Nグラムモデルを用いたクラスタ分析による大規模漢字文献分析の基礎的研究」（課題番号15700215、研究代表者: 師茂樹）および科学研究費補助金・基盤研究C「次世代中国古典文献データベース構築の基礎的研究」（課題番号14510494、研究代表者: 村越貴代美慶応大学助教授）による成果の一部である。

参考文献

1) Haralambous, Tereza and Haralambous, Yannis. “Characters, Glyphs and Beyond.” (書体・組版ワークショップ〔京都大学21世紀COE東アジア世界の人文情報学研究教育拠点〕予稿集、近刊予定)
2) 石井公成 “Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus”. 『電子佛典』3、東國大學校、2001.
3) Kenny, Anthony. *Computation of Style: An Introduction to Statistics for Students of Literature and Humanities.* 1982; 吉岡健一訳『文章の計量 文学研究のための計量文体学入門』(南雲堂、1996) ☆
4) 北研二「確率的言語モデルに基づく多言語コーパスからの言語系統樹の再構築」(『自然言語処理』Vol. 4, No. 3、1997).
5) 近藤泰弘「コンピュータによる文学語学研究にできること—古典語の「内省」を求めて—」(全国大学国語国文学会夏季大会シンポジウム「情報技術は文学研究をいかに変えるか」要旨、2001、`http://klab.ri.aoyama.ac.jp/public/paper/20010602.pdf`)
6) 近藤泰弘・近藤みゆき「N-gramの手法による言語テキストの分析方法—現代語対話表現の自動抽出に及ぶ—」(『漢字文献情報処理研究』2、2001)
7) 守岡知彦、江渡浩一郎、苫米地等流、宮崎泉、師茂樹「CHISE Project」(『漢字文献情報処理研究』4、2003)
8) 師茂樹「XMLとNGSMによるテキスト内部の比較分析実験—『守護国界章』研究の一環として」(『漢字文献情報処理研究』2、2001)
9) 師茂樹「Nグラムモデルとクラスター分析を用いた漢?古典テキストの比較研究—『般若心経』の異本の比較を例に」(京都大学大型計算機センター第69回研究セミナー「東洋学へのコンピュータ利用」予稿集、2002)
10) 師茂樹「Nグラムによる比較結果からの用例自動抽出—禅宗系の偽経を題材に」(東洋学へのコンピュータ利用第14回研究セミナー予稿集、2003)
11) 師茂樹「Perl/CHISEによる正規表現の拡張の試み—文字素性による後方参照の実装実験と

☆ 本稿での引用ページは邦訳に基づくが、訳文は適宜独自のものに改めている。

課題—」(Linux Conference 抄録集: 第 1 巻 (2003年)、ISSN 1348-7876、http://lc.linux.or.jp/paper/lc2003/)
12) 師茂樹 "Surface or Essence: Beyond Coded Character Set Model." (書体・組版ワークショップ〔京都大学 21 世紀 COE 東アジア世界の人文情報学研究教育拠点〕予稿集、近刊予定)
13) 齊藤隆信「漢訳仏典中偈頌的韻律与《演道俗業経》」(『法源』18、2000)
14) 齊藤隆信「支謙と鳩摩羅什訳仏典における偈の詩律」(『仏教史学研究』43-1、2001)
15) 齊藤隆信「漢語仏典における偈の研究 —有韻の偈—」(『香川孝雄博士古稀記念論集 仏教学浄土学研究』、永田文昌堂、2001)
16) 齊藤隆信「支謙所訳経典中偈頌的研究 —四部経典中偈頌的漢訳者」(『法源』19、2001)
17) 豊島正之「TEI から見た SGML の話」(http://jcs.aa.tufs.ac.jp/mtoyo/TEI/JALLC-12-TEI.pdf、1992/1996/2000)
18) Unicode Consortium. *The Unicode Standard, Version 4.0.* Boston: Addison-Wesley. 2003.