情報処理学会論文誌　データベース　Vol. 1　No. 3　96–107 (Dec. 2008)

聖教書誌情報全文検索システムの構築

村川猛彦[†1]　宇都宮啓吾[†2]　中川優[†1]

寺院に所蔵されている様々な文書や典籍（聖教）において，その奥書には書写・伝授などの由来が記録されており，人文研究者にとっては，聖教の形成や当時の人間関係など様々な歴史的事象を知る手がかりとなる．そのような研究を支援するため，全文検索エンジン Hyper Estraier を用いた書誌情報検索システムを試作した．登録した聖教情報は，7 種類の出典目録から合わせて 44,135 件である．検索語入力の手間や入力ミスといった問題を解消するため，検索結果においてキーワードを強調表示し，そのキーワードで検索できるようにした．そのためにまず，聖教情報に含まれるキーワード 6,690 個の機械的抽出を行った．年代情報および人物名は，候補となる文字列を生成・抽出して全文検索に問い合わせ，出現するものをキーワードとした．寺院情報および地名に対しては，「寺」「国」などの接尾辞に着目してテキストマイニングを行った．キーワードは関係データベースに登録し，任意のテキストデータに対して該当するキーワードの位置を効率良く求められるようにした．120 件の聖教情報に対してキーワード検出を試みたところ，適合率・再現率とも 90%を超え，十分な実用性を確認した．

Full-text Search System for Bibliographic Data of Ancient Documents

Takehiko Murakawa,[†1] Keigo Utsunomiya[†2] and Masaru Nakagawa[†1]

Ancient documents and scriptures in the temples provide a clue to the clarification of various historical phenomena such as their background and the human relationships, due to the colophons and other bibliographic elements. We constructed a document retrieval system powered by the full-text search engine Hyper Estraier, where 44,135 records are registered. To discharge the problems of time-consuming or error-prone input of search terms, we improved the search interface so that the users can see the resulting documents with the keywords highlighted for further search. For that purpose, we found 6,690 keywords in bibliographic data. The words about eras and persons were produced or extracted and subsequently filtered through the full-text search. Those about temples and places were obtained by a text-mining approach aiming at the suffix characters. The keywords are stored in a relational database apart from the full-text index. Using the database, we implemented the feature of calculating the position of the keywords for a given string efficiently. The experiment showed that the precision and the recall of the detected keywords on 120 bibliographic documents were greater than 90%.

1. はじめに

近年，コンピュータは急速に発展し，研究に欠かせない重要な構成要素になりつつある．これは，人文系・芸術系分野においても例外ではない[1)]．そして，史料の持つテキスト情報を翻刻したり，テキストや年代の情報を付与したりしてサーバに格納し，検索に提供されるようになりつつある．総合的な検索に関しては，PORTA（国立国会図書館デジタルアーカイブポータル）[⋆1]，早稲田大学の古典籍総合データベース[2),⋆2]，東京大学史料編纂所[⋆3]のデータベース検索などが知られている．また漢訳仏典では，大正新脩大蔵経のテキスト化が多くの人の手によってなされており，全文検索システムも公開されている[⋆4]．

人文系の中でも，国文学や歴史学の分野では，古代・中世の仏教ならびに仏教界をめぐる種々の問題を，当時の文献を多数活用して，解明しようとしている．筆者らの 1 人はこれまで，仏教活動の成果である聖教の形成・継承の問題や聖教をめぐる僧侶の活動について調査，検討してきた[3)–5)]．ここで聖教とは，「寺院社会内で教義・行法に関して記録したもので僧尼の修学や宗教活動の実践に際して活用され，かつ師弟間における原本授受または書写伝授によって法脈継承を根拠づける文献」[6)] であり，経典はその代表的なものである．

そこで，寺院に所蔵されている聖教の全体把握と個別の聖教の精査が活発に行われている．多くの聖教は「奥書」と呼ばれる書誌情報を有し，そこには「いつ誰が命じて誰が書写をしたか」「経典がどこの寺院に所蔵されたか」といった情報が記載されている．したがっ

†1 和歌山大学
Wakayama University
†2 大阪大谷大学
Osaka Ohtani University
⋆1 http://www.dap.ndl.go.jp/
⋆2 http://www.wul.waseda.ac.jp/kotenseki/
⋆3 http://www.hi.u-tokyo.ac.jp/index-j.html
⋆4 CBETA 中華電子佛典協會，http://www.cbeta.org/．広済寺 仏教典籍検索，http://www.kosaiji.org/~kyoten/

て奥書は，聖教自体の形成・継承の問題や聖教をめぐる僧侶の活動を把握するのに重要な情報といえる．

ここで，聖教奥書を参照した研究活動を行う際の課題について述べる．中心となるのは平安・鎌倉時代の古典籍であり，それらは文化財であるため，寺院や博物館などが厳重に保存しており，普段は原本を閲覧できない．しかし，編著者らの方針により聖教が整理された「聖教目録」がいくつか出版されており，これを活用することが可能である．とはいえ，掲載聖教数が膨大で索引が整備されていない場合には，特定の典籍名を確認するのも多大な負荷となる．また，人物索引は作られないため，ある僧侶が関与したとされる聖教を見つけるには，時期からページを限定して，奥書を 1 つ 1 つ見ていくしかない．それには大変な苦労を要し，見落としも起こりやすい．私的使用の範囲内で，書誌情報をテキスト化し，Microsoft Excel や Access で表形式にして検索できるようにすれば，見落としの可能性が減るが，複数の目録から探す場合には，別々に検索しなければならない．さらに，「22 年」「廿二年」といった表記の揺らぎにより，単純な検索語句では，期待する情報が取得できないおそれもある．なお，テキスト化における情報の内容の正当性や，1 回の検索に要する時間も課題にあげられるが，本研究ではそれらは対象外としている．

本研究では，全文検索システム Hyper Estraier を用いた Web アプリケーションを構築し，聖教書誌情報を効率良く検索できるようにした[7),8)]．さらに検索の利便性を高めるため，次の工夫も行った．(1) 年の検索では，西暦年となる整数値を指定すれば，漢数字による西暦年や，元号と年数で表現される和暦についても検索するようにした．全文検索を行うので，聖教の作成時期だけでなく，奥書や備考に年の情報が書かれていても，取得可能である．(2) 字体変換テーブルを内部に持たせている．これにより，利用者は単一の検索語を指定しても，内部で字体変換を行い複数の検索語を生成し，その OR 条件として検索することになる．(3) 検索結果では，検索語を強調表示するとともに，年代・寺院・地名・人物にリンクを付与し，その語句を含む検索ができるようにして，毎回検索語を入力しなくても，あるキーワードをともに含む聖教を容易に見つけられるようにしている．

年代・寺院・地名・人物にリンクを付与するためには，あらかじめそれらの情報を取得しなければならない．またその語句の数が膨大になっても，検索結果の聖教情報に対して効率良くリンクの位置を求めなければならない．前者については，構築した全文検索エンジンと，外部の Yahoo!検索サービス，および既存のデータセットを用いて抽出を試みた．後者については，キーワードをトライ構造で表現し，そこからリンク位置を効率良く求めるものとした．リンク付与の妥当性についても評価を行い，120 件の聖教書誌情報において，適合率・再現率とも 90%を超え，手作業による修正を加えればさらに向上することを確認した．

本論文の構成は以下のとおりである．2 章では，全文検索エンジン Hyper Estraier および対象とする聖教情報などを概説したのち，構築した聖教書誌情報全文検索システムを紹介する．3 章では，キーワードの抽出方法を，年代，人物名，寺院名および地名に分けて説明し，4 章でキーワードデータベースと検索インタフェースの改良を述べる．5 章では，検出キーワードを用いたリンク付与の評価実験とその評価，および著作権に関する検討を述べ，6 章では，自他の既存研究と比較して本研究の特色を明らかにする．7 章はまとめと今後の展望である．

2. 聖教書誌情報全文検索システム

2.1 準　　備

本研究で使用する Hyper Estraier[9)] について紹介する．

フリーソフトウェアの全文検索エンジンであり，内部で転置インデックスを用い，前処理に時間を要するが，インデックスができあがれば検索はきわめて高速である．インデックス構築のための文書の切り出し方法として，形態素解析と N-gram が利用可能となっており，本研究では N-gram のみを使用する．文字コードは UTF-8 を使用しており，日本語に限定せず登録，検索できるのは，通常漢文により記述される奥書と，日本語で表記される備考の情報から一括して検索でき，都合が良い．C，Java，Ruby の各言語からアクセスするためのインタフェースが整備されており，我々はもっぱら Ruby を用いて，Web アプリケーションや，Web を介さない各種処理プログラムを作成した．

さらに，文書には任意の属性を付加することができる．この特徴を用いて本システムでは「属性検索」も実装している．

2.2 対象とする聖教情報とその登録

対象とした聖教情報の出典および件数を**表 1** に示す．「基本聖教目録」については，書籍や Web で公開されている情報を収集しデータ化したものである．「基本聖教目録」も 1 つの出典として，各出典のことを本論文では「出典目録」と呼ぶ．

出典目録ごとに，表形式のファイル（Excel 形式または Access 形式）で管理されている聖教目録ファイルを作成したが，テキスト化の時期や用途が異なるため，表の互換性はない．これをテキストファイルとして保存し，全文検索システムに登録しても，検索は可能であるが，表示が煩雑になるという問題点がある．そこで本研究では，聖教奥書の検索と表示に適した，統一したフォーマット（列構成）を定義し，各文書をこの列構成に変換して，登

表 1　出典目録
Table 1　Catalogs of references.

出典目録名	件数	出典等
東寺	31,418	東寺観智院金剛蔵聖教目録[10]
大覚寺	5,389	大覚寺聖教目録[11]
築島	2,168	平安時代訓点本論考[12]
古点本	850	古点本の国語学的研究[13]
来迎院	598	来迎院如来蔵聖教文書類目録[14]
宗性奥書	480	宗性・凝然写本目録[15]
基本聖教目録	3,232	その他の書籍，Web など
合計	44,135	

表 2　聖教情報の列構成
Table 2　Attributes of bibliographic data.

列名	属性名	値の例
書名	@syomei	弘賛法華伝，三観義
巻数	@kannsuu	二冊，一帖
出典目録	@mokuroku	基本聖教目録，来迎院
所蔵	@syozou	奈良東大寺，来迎院
書写時期	@syosyajiki	保安元年（一一二〇），平安後期
装幀	@soutei	巻子本，綴葉装
訓点	@kunntenn	池上阿闍梨点，円堂点
奥書	@okugaki	大日本国保安元年（略）羊僧覚樹記之
備考	@bikou	本書は平安時代末期の写本と思われるが（略）

録することにした．

定義した列構成を**表 2** に示す．ここで装幀にはその聖教史料の製本体裁に関する情報を，また訓点には経典などの漢文を訓読する際に付記する記号の種類を格納する．これらの属性は，奥書と異なりそれ単体で聖教の出自を示すものではないが，聖教作成の時期，場所や寺院，宗派などを限定し，聖教の正真性を補強するものとなる．なお，いくつかの聖教情報では，装幀，訓点，奥書の分割が容易ではなく，備考にまとめて記載している．

変換作業について述べる．まず，各聖教目録ファイルを，テキスト処理のしやすい CSV（Comma Separated Values）形式に変換する．この CSV ファイルに対して，登録に適した形に項目（列）を変更すると同時に，アラビア数字による年数を漢数字に置き換える．たとえば「弘仁 13 年」「(822)」はそれぞれ「弘仁十三年」「(八二二)」となる．

編集された CSV ファイルを，登録ページを介してアップロードすると，サーバ側ではそ

図 1　聖教書誌情報の登録
Fig. 1　Registration of bibliographic data of ancient documents.

```
@bikou=（奥）保延五年九月二日巳時奉書□　念仏宗僧運覚　願以書写功　必為往生因
　法界衆生　生西方浄刹
@digest=a1d7adfe7c3d027d4d0593df0b46e0a0
@id=2022
@kannsuu=一巻
@kunntenn=→
@line=2022
@mokuroku=基本聖教目録
@okugaki=→
@soutei=→
@syomei=無量清浄平等覚経巻下
@syosyajiki=保延五年（一一三九）
@syozou=河内金剛寺
@uri=file:///var/stext/kihon-2022.txt

無量清浄平等覚経巻下, 一巻, 河内金剛寺, 保延五年（一一三九）, →, →, →,（奥）保延
五年九日巳時奉書□ 念仏宗僧運覚 願以書写功 必為往生因 普法界衆生 生西方浄刹
```

図 2　文書ファイルの例
Fig. 2　Example of document.

れを行ごとに分割し，属性を付加して文書ファイルを構成して，Hyper Estraier のインデックスに登録する（**図 1**）．文書ファイルの例を**図 2** に示す．「@」から始まる各行が属性検索用の情報であり，空行を置いて，コンマ区切りによる全文検索用の情報を記載している．@bikou などの属性名は表 2 に記載の各項目に対応する．@digest，@id，@uri の各属性は元の文書にはなく，インデックス登録時に付与される．Hyper Estraier の提供する属性検索には，完全一致，前方一致，数値型としての検索などがあるが，本システムでは文字列型の部分一致検索のみを使用している．

これまで，表 1 に示したように，7 種類の出典目録より 44,135 件の聖教情報を実験的に

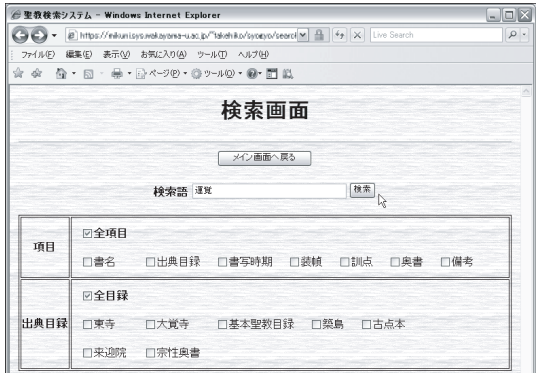

図 3　検索画面
Fig. 3　Screenshot of search form.

図 4　検索結果画面
Fig. 4　Screenshot of search result.

登録した．この著作権については，5.3 節で述べる．

この全文検索システムを稼働させたサーバの仕様は，Intel Q6600（2.40 GHz のクアッドコアであるが，マルチコアを意識した処理は行っていない），2 GB メモリ，OS は Debian GNU/Linux である．ソフトウェアについては，Hyper Estraier 1.4.13 のほか，3 章以降で述べるキーワード管理用に PostgreSQL 8.1 を，各種処理用に Ruby 1.8 で動作する自作プログラムを使用している．Ruby から，Hyper Estraier，PostgreSQL，および 3.3 節で述べる Yahoo!検索 Web サービスに接続するためのライブラリを別途導入している．44,135 件，総バイト数 15,758,432 バイト（文字コードは UTF-8）の聖教文書ファイルを Hyper Estraier で検索できるよう登録するのに要した時間は 99 秒であり，インデックスのファイルサイズは約 34.6 MB，インデックスが持つ語数は 107,736 となった．

2.3　検　索　例

検索画面の例を**図 3** に，また文献 5) で着目した僧侶「運覚」を検索語として，得られる結果画面の例を**図 4** に示す．検索可能な項目には，書名，出典目録，書写時期，装幀，訓点，奥書，備考がある．また検索対象の目録を限定することもできる．結果では，指定可能な各項目のほかに巻数も表示される．

1 語から数語であれば検索は瞬時に行われ，CGI プログラムを起動する時間，および HTML 生成とブラウザでの表示の時間を合わせても，1 秒以下である．

2.4　異体字変換を用いた検索改善の試み

ここまで述べた検索システムを用いれば，期待する検索語を持つ聖教情報を瞬時に獲得でき，また同一の奥書が複数の文献目録に記載されている場合には，見比べることで，内容の精度を高めることが可能となる．

しかし，検索語が，対象とする聖教情報から 1 文字でも異なっていれば，その情報は検出できない．毎回検索語を入力するのも手間がかかり，打ち間違いによるタイムロスも起こりやすい．検索に苦労を要する例として，「弘賛法華伝」があげられる．これを検索語とすると 1 件しか見つからないが，「弘贊法華伝」とすれば 3 件あり，「弘賛法花伝」なら 1 件が見つかる．Hyper Estraier の OR 検索を用いて「弘賛法華伝 | 弘贊法華伝 | 弘賛法花伝」という検索語を与えれば該当データをすべて取得できるが，このような指定を毎回利用者が行うのは非現実的である．そこで，異体字変換を取り入れ，検索語指定の省力化を図った．

異体字変換のテーブルは，筆者らが漢訳仏典の検索のために構築したもの[16] と，「異体字処理 付表」[17] のうち「B → A」の形式の漢字の組を取り出したものを組み合わせて構築した．変換ルールは 643 組であるが，変換対象の文字列を Ruby の正規表現で見つけて変換するようにしており，変換処理は瞬時に行われる．

登録では，異体字変換を行い，全文検索インデックスに登録する．もとのテキストファイルは残しておき，結果表示の際にはこの変換前のデータを取り出し利用者に提示する．検索時にも，検索語の持つ異体字を変換してから，検索を行うようにする．これにより，利用者は，たとえば「弘賛法華伝」のみで，「弘賛法華伝」「弘贊法華伝」「弘賛法花伝」を検索で

き，それぞれ原文のままで閲覧できる．

3. キーワードの抽出

検索インタフェース改善のための次の取り組みとして，検索結果のどこに検索語が出現しているかを強調表示するとともに，さらなる検索のために，古典籍でよく見かけるキーワードがあればそれも（検索語とは別に）強調表示し，それをクリックするだけで検索できるようにしたい．この章では，そのためのキーワードの抽出方法について述べる．

文献 7) では，年代・寺院・地名・人物の 4 種類について，聖教情報を支援するデータベース化を課題にあげたが，それぞれについてキーワード獲得を試みた．キーワードの獲得方法には，外部の情報をもとにキーワード候補を抽出あるいは生成し，前章で述べた全文検索システムに問い合わせて出現すればキーワードとする方法と，聖教情報からキーワード候補を抜き出して外部の検索システムに問い合わせて適切に出現していればキーワードとする方法の 2 種類に大別できる（**図 5**）．外部の検索システムには，Google や Yahoo!などの世界規模の全文検索サイトに限らず，古典籍に関する検索サイト（必ずしも全文検索を使用していなくてもよい）も利用可能である．

キーワードは，全文検索エンジンとは別のキーワードデータベースに格納するものとし，かつ，聖教データベースに出現しないものは排除することとした．これは，データベースの容量を減らすだけでなく，聖教検索システムで，提示するキーワードで検索して「0 件」という検索結果を出さないようにするためである．聖教情報からキーワード候補を抜き出す際には，2.4 節で述べた異体字変換を行っているが，それ以外の字句の修正は行っていない．さらに，全文検索システムは Hyper Estraier および N-gram に基づき漏れのない方式を採用していること，聖教情報は異体字変換を行ったものを登録していることから，この方法で求めたキーワードについては，検索すれば必ず 1 件以上見つかることが保証される．

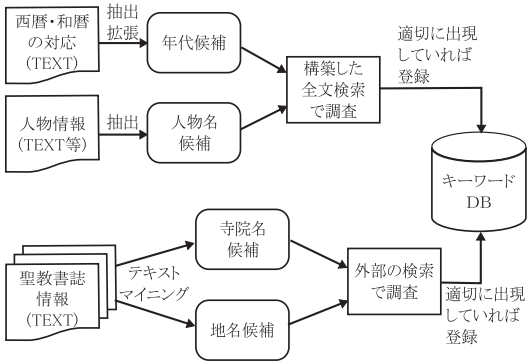

図 5　キーワード抽出の流れ
Fig. 5　Overview of keyword extraction.

3.1 年代情報

本研究では，「平成二十年」のような，元号および漢数字による年数を組み合わせた表記（和暦）と，「(二〇〇八)」の形で表記される括弧付き西暦表記が，聖教書誌情報に多く出現することに着目し，これらの情報の検出を試みた．

構築した検索システムにいくつか和暦を与えて調べたところ，年数については，改元の年はほぼ「元年」であるが，「一年」という表記も見られた．また，「一〇年」，「廿二年」，「卅年」といった表記も存在する．これらの年数表記の揺れに注意して，和暦を機械的に生成し，全文検索により 1 件以上見つかったものを，キーワードとした．

和暦生成では，西暦—元号変換用辞書『元号法』[⋆1] よりダウンロードできる書庫ファイルに収録の「元号 (年).dic」から，各西暦年に対応する元号および年数を抽出する．たとえば西暦 749 年については，「749(天平感宝 1 年/天平勝宝 1) 年」という表記から，正規表現で「天平感宝」とその直後の「1」を取り出し，ここから「天平感宝元年」「天平感宝一年」を生成する．同様に，「天平感宝」とその直後の「1」から，「天平勝宝元年」「天平勝宝一年」が得られる．さらに，「1 年」であることから改元の年であることを意味し，前年の和暦を 1 年増やしたものも，和暦候補とする．西暦 798 年については，「748(天平 20) 年」となっていることから，正規表現で「天平」とその直後の「20」を見つけ，「天平二十一年」「天平二一年」「天平廿一年」を生成する．

括弧付き西暦表記は，西暦の整数値を文字列に変換し，各算用数字を漢数字に置き換え，前後に括弧を付けることで得られる．西暦 749 年を例にすると，「(七四九)」である．

以上の 8 つが，西暦 749 年のキーワード候補である．そのそれぞれについて，聖教情報に 1 件以上の出現があるか調べたところ，「天平二十一年」「天平廿一年」「(七四九)」の 3 つが該当することが分かった．したがって，これらを西暦 749 年のキーワードとして，キーワードデータベースに登録している．

3.2 人物名

人物名については，『天台血脈』[3]，『諸嗣宗脈紀』[4]，および『野澤血脈集』[18] の電子デー

⋆1 http://www.sal.tohoku.ac.jp/~kirihara/gengo.html

表 3　人物キーワード
Table 3　Keywords about person.

	天台血脈	諸嗣宗脈紀	野澤血脈集
候補総数	946	2,545	8,761
非出現語数	545	1,737	5,417
出現語数	401	807	3,344
非頻出語数	394	784	3,138
頻出語数	7	23	206
頻出キーワード数	3	5	14
合計キーワード数	397	789	3,152

タを活用することとした．聖教データベースには西暦 661 年から 1992 年までの年代情報が記載されているが，聖教研究の中心となるのは平安・鎌倉時代であり，上記の電子データはこの時代の僧侶情報を概ね網羅するものである．

人物名に相当する位置の文字列を人物名候補とし，自作 Ruby プログラムを用いて各候補文字列を検索システムに問い合わせ，出現件数を求めた．出典ごとのキーワード数を**表 3** に示す．なお，合計 12,252 個に対する総検索時間は 1.37 秒であった．1 個あたり 0.1 ミリ秒強の検索時間は，ユーザによる検索とは別に，本システムの有用性を示すものとなる．

キーワード選定の方法は，次のとおりとした．まず，出現件数が 0 のもの（非出現語）は，前述のとおりキーワードとしない．次に，出現件数が過大なものについても，適切かどうか個別に確認した．『天台血脈』の人物データについては，出現件数の多かったものを上位から並べると，「大日」(779 件)，「不空」(246 件)，「承信」(183 件)，「弘法大師」(156 件) であるが，「大日」による検索結果を見たところ，「大日本国」「大日如来」「大日経」の一部として出現することが分かり，キーワードにすべきではないと判断した．「不空」も同様であった．「承信」「弘法大師」は僧侶名として記載されていたのを確認し，ともに人物キーワードとした．

『諸嗣宗脈紀』『野澤血脈集』の人物データについても，同様に確認した．しかし，『野澤血脈集』から抽出した人物名候補は 3,344 にのぼり，そのすべての検証は困難である．ここで，出現件数が多いものを誤抽出または抽出漏れにしてしまうと影響が大きいことを考慮し，出現件数に上限を設けた．具体的には 60 件以上出現する人物名候補は頻出語として個別に検査し，1 件以上 60 件未満のものは非頻出語として，検査せずに人物キーワードと見なした．これらを結合して，重複を取り除き，最終的に 3,373 個を人物キーワードとした．

3.3　寺院名および地名

接尾辞に着目したキーワード獲得の方法について述べる．方針としては，聖教書誌情報のテキストデータを全探索して「寺」などで終わる部分文字列を取得し，そこから適切なキーワードとなるよう文字列処理を行っている．「寺」を接尾辞とするキーワード（寺号）の獲得は，以下の手順で行った．

(1)　異体字変換していない聖教テキストデータから，「『寺』で終わる 6 文字以内の漢字の並び」を，正規表現を用いてすべて抽出する．

(2)　(1) で抽出した文字列に対して，除去リストに適合する文字があれば，文字列の先頭から該当箇所までを除去する．これにより，空文字列または「寺」1 文字になった場合，もしくはすでに登録しているキーワードになれば，その文字列に対する処理を終える．

(3)　(2) で獲得した文字列に対して，異体字変換の対象文字が含まれていれば，変換後の文字列を求め，変換前後の文字列をともにキーワード候補とする．異体字変換の対象文字がなければ，その文字列がキーワード候補となる．

(4)　(3) で獲得したキーワード候補に対して，Yahoo!検索 Web サービス[*1]を用いてその出現状況を調べる．結果には，該当件数だけでなく，検索語に関する上位 10 件（該当件数が 10 件未満なら，その件数）までの Web ページのタイトルと要約が含まれている．そこで，キーワード候補がいずれかのタイトルまたは要約の中に部分文字列として（空白や記号類を間に含むことなく）出現していれば，キーワードとする．

「除去リスト」は，初めは空とし，除去すべき語句を増やしていった．最終的な除去リストには，以下の文字列が含まれている．実際の処理では各文字列を区別することなく除去リストに格納し，それらの OR 条件となるような単一の正規表現を生成して，除去のためのパターンマッチングを行っている．

奥書の読解において意味を持つ文字：月，日，以，於，廿，在，并，但，州，郡，町，村，領，語，市，国，街，県，都，読，字，山，宗．

頻出する地名：京都，近江，奈良，河内，東京，長安，大沢，今津，大和，土佐，名古屋，紀伊．

寺号に含めるべきでない寺院情報：専寺，之寺，当寺，本寺，入寺，同寺，造寺．

ただし上述の手順と除去リストを適用して，機械的にキーワードを求めるのでは，「石山

⋆1 http://developer.yahoo.co.jp/search/

寺」などの有名なものも漏れてしまう．そのような寺号については，別途，目視で見つけた．具体的には以下のものがあげられる．

除去文字を含む寺院名：月輪寺，日輪寺，大日寺，興国寺，国分寺，高山寺，香山寺，山階寺，光明山寺，仲山寺，金峯山寺[⋆1]，大山寺，霊山寺，前山寺，立本寺，円宗寺．

Yahoo!検索 Web サービスを用いた検索に関して，該当件数を基準としてキーワードの妥当性を判断しなかったのは，個別に結果を見たところ，不適切なものが多数含まれていたためである．たとえば「栗野寺」(2 件）を検索すると「夏栗　野寺」のように，検索語が分割された形で出現しており，キーワードとして適切でないと判断した．同様の理由で「境内寺」(506 件）はキーワードとせず，一方，「神鳳律寺」(1 件）は適切に出現していたのでキーワードとした．

次に，「院」「房」「坊」のいずれかで終わる，寺院関係のキーワードを抽出した．なお，『妙法蓮華経』(築島）の書誌情報には「保安二年（一一二一）九月廿一日於宿院双厳房移点畢」という奥書と，「双厳房は天台宗延暦寺谷流の頼昭」という備考があるように，房名が人物名を指すことがある．院号や坊名についても存在しうる．しかしここではそのような内容に立ち入らず，寺院のキーワードとして獲得を試みた．

キーワードの求め方は，寺号と同様であるが，「奥書の読解において意味を持つ文字」に，寺，院，房，坊の 4 文字を加え，かつこれらの文字については，処理中の文字列の末尾と一致しているものは除去対象としないというルールを設けた．これにより，上述の奥書から，「宿院」と「双厳房」が分離してキーワード候補となる．ただし「双厳房」については，Yahoo!検索 Web サービスの結果では部分文字列として出現していなかったため，キーワードとしていない．この方法で取得できたキーワードには，「三千院」「明法房」「小池坊」が含まれている．

処理時間は，Yahoo!検索 Web サービスへの通信時間や同サービス内部での処理時間のため，検索語あたり約 6 秒，「寺」「院」「房」「坊」のいずれかで終わる 2,017 個の全キーワード候補に対して，問合せを終えるのに 3 時間強を要した．

地名には，「下野輪王寺」の「下野」のように，寺号の直前につくものも見られるが，ここでは，寺号などと同じように，接尾辞に着目した取得を試みた．着目した文字は「国」「郡」「州」である．院号などと同様の方法でキーワード候補を求める．さらに，キーワード候補から接尾辞を除いたものが 1 文字でなければ，これもキーワード候補とする．たとえば「山城国」に対して，「国」を除去した「山城」も候補とし，Yahoo!検索 Web サービスにより出現を確認したので，登録している．

⋆1 キーワードとしては，異体字変換により，「金峰山寺」となる．

表 4　抽出したキーワード
Table 4　Results of extracted keywords.

	種類	個数	キーワード例
年代	和暦	1,151	斉明七年，平成四年
	括弧付き西暦	468	(七九四)，(一〇九九)
人物		3,373	弘法大師，俊源
寺院	寺号	463	浄瑠璃寺，石山寺
	院号	505	三千院，金剛乗院
	房名	335	竹林房，明法房
	坊名	94	西坊，小池坊
地名	国名	66	大日本国，山城国
	郡名	58	賀茂郡，吉野郡
	州名	51	大和州，伊州
	その他	126	山城，隠岐
	合計	6,690	

3.4　キーワード数

前章の各手法により抽出したキーワードの個数と例を，**表 4** に示す．

4.　キーワードデータベースとそれを用いた検索システムの改良

4.1　キーワードデータベースの構築および聖教情報におけるキーワードの検出

聖教検索システムの検索結果において，検索語を強調表示するとともに，前節までのキーワードが含まれていれば，それをリンクとして，クリックするだけで当該キーワードで検索できるようにしたい．しかし，現時点でキーワードの総数が数千あり，手作業でもキーワードを登録できることを念頭に置いたシステムにおいて，リンク化処理においてキーワード数に比例する回数のテキストマッチング処理を行うのは効率的とはいえない．

そこで，聖教情報ごとに，含まれるキーワードの位置と長さをすべて求め，配列（「キーワード出現配列」と呼ぶ）として格納しておき，結果表示時にはこの配列と検索語を用いて，強調表示とリンク付与を高速に行えるようにした．

聖教情報に含まれるキーワードを求めるのは検索時ではなく前処理になるが，少しでも効率良くするため，登録されているキーワードを各文字をノードとするトライ[19]で表現し，データベースに格納した．「三乃」「三井」「三井寺」「三井郡」「三会院」の 5 つの文字列に対するトライの構成例を**図 6** に示す．

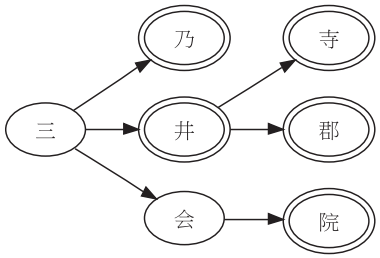

図 6　トライによるキーワードの構成例
Fig. 6　Construction of keywords using trie.

キーワードおよび聖教情報のキーワード出現配列を格納するため，全文検索システムとは別のデータベースを PostgreSQL で構築した．現時点で登録されているキーワード数は 6,690，文字ノード数は 10,134 である．

キーワードの先頭文字からなる集合（頭字集合）を別途構成しておく．マッチング作業は，与えられたテキストデータに対して異体字変換を施してから，頭字集合に属する文字の位置をすべて求め，次にその各位置から，トライ構造を参照して最長一致のキーワードを見つける．そのような位置と文字列長を要素とする配列を構成する．この配列は文字列化してデータベースに登録し，検索結果表示時に使用する．キーワードの登録もしくは削除があれば，そのキーワードをトライに配置もしくは除去してから，それを含む聖教情報を全文検索により求め，それぞれについてキーワード出現配列を再計算すればよい．

『弘賛法華伝』（基本聖教目録）の書誌情報を例に，処理の流れを述べる．「大日本国保安元年七月八日於太宰府勧俊源法師書写畢、宋人蘇景自高麗国奉渡聖教之中、有此法華伝、仍為留多本令書写也、羊僧覚樹記之、」という奥書に対して適用するとき，頭字集合にマッチするのは，大日国保安元年七月日太俊源法景高国聖教中有法華伝多令覚の 27 カ所の文字であり，それぞれを先頭として最長一致のキーワードを見つけると，「大日本国」「日本国」「保安元年」「太宰」「俊源」「高麗国」「覚樹」が得られる．このうち，「日本国」は「大日本国」に完全に含まれるが，この段階ではこの位置と長さも，キーワード出現配列の要素に加える．また，「太宰」については，奥書の意味上では「太宰府」がより適切な単語であるが，これはキーワードとして機械的に取得できていないためであり，別途登録することで改善される．

検索結果を表示する際には，対象となるテキストデータを取り出し，異体字変換を行ってから，検索語が出現していればその各箇所について，位置と長さを求める（「検索語出現配列」と呼ぶ）．複数の検索語による AND 検索，OR 検索については，あらかじめ単語ごとに分割しておく．一般に検索語数は少ないので，このマッチング作業は単語ごとの単純な文字列照合をとっている．検索語出現配列の要素間に重複があれば，最左最長のものを残し他は除去する．次に，書誌情報のキーワード出現配列の各要素（位置と長さ）について，それが検索語出現配列のいずれかの要素と重なっていれば除去する．最後に，キーワード出現配列の中でも，重複があれば最左最長のものを残して他を除去する．この時点でのキーワード出現配列をもとに得られる，テキストデータの部分文字列を「検出キーワード」，それを求めることを「キーワードの検出」と呼び，3 章における，聖教テキストデータ全体を対象とした「キーワードの抽出」と区別する．

あとは，対象となるテキストデータの異体字変換前のものに対して，求めた検索語出現配列およびキーワード出現配列の各要素に応じて，強調表示もしくはリンク付与を行う．たとえば上述の奥書で，「法華伝」を検索語とした場合，実際に「法華伝」が出現するが，異体字変換前は「法花伝」であり，これを強調表示する．キーワードに関しては，前述の「日本国」のみ除去し，残りについてリンク付与することで，「大日本国 保安元年 七月八日於太宰 府勧 俊源 法師書写畢、宋人蘇景自 高麗国 奉渡聖教之中、有此＜法花伝＞、仍為留多本令書写也、羊僧 覚樹 記之、」が得られる．＜…＞が強調表示部であり，下線部はリンクとなる．

登録した各聖教情報に対して，キーワード出現配列を求め，検索語なしとして検出キーワードを求めたところ，キーワード数の平均は 1.81 個であり，聖教情報の字数に対する検出キーワード文字数の割合の平均は 4.75%であった．これは，出典目録のうち東寺の情報で，キーワードが出現しないものが多かった（31,418 件中 16,182 件が該当した）ことが影響しており，よく参照される基本聖教目録に限れば，それぞれ 6.48 個，10.45%となっている．1 つの聖教情報で最もキーワード出現数の多かったのは『成唯識論』（築島）で，1,263 文字の中に 91 個が検出された．

4.2　検索システムの改良

キーワードデータベースを用いて，検索システムの改良を試みた．具体的には，検索語の変換と，結果の強調表示である．

サーバは検索語を受け取ると，異体字変換を行うとともに，年代に関して検索語の展開を行う．たとえば「1000」が検索語として与えられると，西暦 1000 年と見なして，全文検索に先立ち，キーワードデータベースにこの年と結び付けられているキーワードを検索する．ここでは「長保二年」のみであり，これを全文検索用の検索語として，検索を行う．

図 7　検索語の変換例
Fig. 7　Example of conversion of search term.

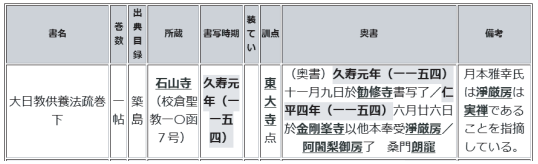

図 8　検索結果例
Fig. 8　Example of search result.

「1234-1235」とすると，指定した西暦年の区間の年代で，キーワードデータベースに登録されている，「嘉禎元年」「天福二年」「文暦二年」「文暦元年」「(一二三五)」の OR 検索を試みる（**図 7**）．「平成 4」は，「平成四年」の検索となる．

「1154」を検索語に与えて得られる，聖教情報の画面の一部を**図 8** に示す．全文検索エンジンに渡される検索語は，仁平四年，久寿元年，（一一五四）の 3 つであり，この例ではそれらが反転表示されている．それと別に，登録されているキーワードは強調表示され，マウスでクリックすれば，その語句で検索できる．「金剛峯寺」のように，異体字変換前のものもキーワード表示されているのは，前節で述べたとおり，テキストデータごとに，キーワードの位置と長さを格納しているためである．この「金剛峯寺」をクリックすると，この文字列を新たな検索語として検索を試みるが，検索システム内部では「金剛峰寺」を探すことになる．

5. 評価および考察

5.1 検出キーワードの評価実験

4.1 節の「検出キーワード」について，広く用いられている情報検索システムの検索性能の評価尺度と同様の考え方で，その適合率および再現率を求めた．検索語は与えないものとして，文書の持つ検出キーワードの個数を W，その中で適切なものの個数を R，文書中のキーワードとすべきもの（重複があれば最左最長のものを選ぶ）の個数を C としたとき，適合率を R/W，再現率を R/C と定義する．これまで例に用いてきた「大日本国 保安元年

表 5　検出キーワードの適合率および再現率
Table 5　Precision and recall for detecting keywords.

	W（個）	R（個）	C（個）	R/W（%）適合率	R/C（%）再現率
A 群	377	354	392	93.90	90.31
B 群	315	286	315	90.80	90.80
B’ 群	314	291	315	92.68	92.38

七月八日於 太宰 府勧 俊源 法師書写畢、宋人蘇景自 高麗国 奉渡聖教之中、有此法華伝、仍為留多本令書写也、羊僧 覚樹 記之、」については，$W = 6$ であり，その中で「太宰」だけが間違っているので，$R = W - 1 = 5$ である．「太宰府」と「蘇景」がキーワードとすべき語なので，$C = R + 2 = 7$ である．したがって適合率は 5/6 すなわち 83.33%，再現率は 5/7 すなわち 71.43%となる．

登録した聖教ファイルのうち，次の 3 群について，適合率と再現率を求めた．なお，文書（聖教ファイル）によって検出キーワードの個数が大きく異なるため，適合率・再現率は文書単位の平均ではなく，群ごとに W，R，C の値を集計してから算出している．A 群と B 群を合わせて，検出キーワードの基本性能を見ることができる．また，B 群と B’ 群を比較することで，手作業によるキーワードの変更による性能向上を知る手がかりとなる．

A 群：『弘賛法華伝』（基本聖教目録）およびランダムに選んだ 59 件を合わせた 60 件の聖教情報を対象とし，4 章までに構築したデータベースによりキーワードを検出する．

B 群：A 群とは別に，ランダムに選んだ 60 件の聖教情報を対象とし，4 章までに構築したデータベースによりキーワードを検出する．

B’ 群：聖教ファイルは B 群と同じものとし，4 章までに構築したデータベースに，A 群の適合率・再現率がともに 100%になるよう最小限のキーワードを登録・削除してから，キーワードを検出する．

結果を**表 5** に示す．不適切な検出キーワードの例として，「京都国」「極楽」などがあげられる．前者は「京都国立博物館」から地名として誤抽出されキーワードデータベースに登録されたもので，後者は人物キーワードとなっていた．検出に失敗したキーワードには，「円惣房」「良諄」「大原」などがあった．

5.2 考　　察

前節の結果は，次のように説明できる．すなわちまず，適合率・再現率ともに 90%を超えていることから，10 個のキーワード検出があればそのうち 1 個弱は誤りであり，かつそ

の間に 1 個程度の検出されていないキーワードがあるということである．そして，B 群と B' 群の差は，上記の「10 個」が「13 個」に変わるものになり，その分，検出ミスや検出漏れの度合いが減ることを意味する．最後に，寺院，人物や地名といった明確な命名規則を持たないキーワードに関して，また奥書（漢文表記）に特有の知識を多く与えることなく，自他の検索エンジンと外部の人物データを活用しながら多くを自動処理によって，書誌情報中に出現するその大部分を適切に取得できたことは，本手法の有用性，また他の古典籍テキストに対する特徴情報取得の可能性を示唆するものである．

ここで，精度向上や他文書群への適用を念頭に置き，人手が不可欠な作業を整理する．まず，対象とするテキストデータ（本論文では聖教テキストファイル）が適切に整備されていなければならない．2.2 節で列構成を定め，属性を付与したが，これは必須ではなく，検索語の存否を効率良く知るためだけであれば，属性なしで各文書を登録すればよい．次に不可欠な作業は，どのような種類の情報を抽出したいかを決定することである．そして，その情報のデータセットがすでに存在する（人物）か，生成可能である（年代）ならば，そのデータセットの各データについて，その存否を全文検索で調べることにより，キーワードの判別を行う．そのようなデータセットが存在しなければ（寺院と地名），対象とするテキストデータに対して，正規表現を用いて特徴的な語句を拾い出すか，N-gram 法に代表される統計的手法により「もっともらしい」もの，具体的には頻度の高いものを見つけることになる．ここで，どのような正規表現を与えればよいか，頻度の高い中で何をキーワードとすべきかは，対象物に対する事前の知識，あるいはその分野の専門家の経験によって決定されるものとなる．なお，聖教テキストファイルに対して unigram から 6-gram までによる頻度調査を試みたが，高頻度のものからキーワードになるものを見つけることはできなかった．

本論文の手法，もしくはより簡便な方法により，質は必ずしも高くなくてもキーワード候補を機械的にデータベースへ登録し，その後，実例にあてはめてキーワードを追加・削除するというアプローチも考えられる．この場合は，キーワード管理のためのインタフェースがその成否を握る．また，本システムのように，キーワードに年代・寺院・地名・人物の種別を設定している場合には，種別ごとに，キーワード強調表示の可否をその場で変更できるようにしておくと，キーワード管理だけでなく，利用者の関心のあるトピックで文書を見つけやすくなる（聖教研究を例にとると，人物名に着目して聖教の関わりを把握する）ことにも役立つと考えられる．

リンク付与の利用に関して，いくつか指摘したい．まず，年代に関して年のみに着目し，月日や時代区分（平安など），元号についてはキーワード化していない．これは，年の検索が基本となるためである．年代に関する検索で強く期待される検索方法の 1 つは，ある西暦年を基準として，その前後数年で，着目する僧侶がどのような聖教の書写や加点（訓点を付けること）に関わったかを知ることである．この用途に，4.2 節で示した西暦年の範囲検索はきわめて有用である．ただし，時代や元号に関するキーワード抽出により，文献間の連携を強める効果も考えられる．月日のみでの検索は想定しにくいが，年と組み合わせた検索のニーズはありうる．本システムでは，元号のみや年月（日）の検索については，キーワードからのリンクではなく，従来型の検索語入力によりできるようになっている．

キーワードの登録・削除を，キーワードデータベースに対してではなく，文書単位で行えるようにできれば，リンク付与をより適切なものにできる．そのような例として，人物名としても形容動詞の語幹としても出現する「忠実」があげられる．

5.3　著作権への配慮

Web サイトの情報を収集（クロール）して検索用インデックスを構築し，利用者の与える検索語に応じて検索結果を返す検索エンジンサービスについて，法制上の課題の検討がなされている[20]．諸外国の状況と比較しつつ，日本国内では，クロール作業が著作物の複製に，また検索結果の表示は，著作物の送信可能化および自動公衆送信に該当し，原著作物の著作権に配慮しなければならないことが指摘されている．

本検索システムでは，クロール作業に代えて書籍の電子化作業を行ったという違いがあるが，元の書籍は著作物[⋆1]であり，現時点では公開して利用に供することは（各著作者の許諾を得ない限り）困難であるが，私的使用の範囲内で活用することや，引用として検索例を提示することは問題ないと考える．

なお，本論文で転載している聖教やキーワードの例は，本システムの使用方法を示すための必要最小限にとどめている．

6. 関 連 研 究

古典籍の全文検索システムの構築に関して，木村ら[21] は『兵範記』のテキストデータに関するデータベース化を行い，単純な検索に加えて，現代語を入力して古語を推測して検索するという検索方法を示している．ただしその検索システムでは，キーワードに関するデータベース検索が行われており，その検索精度は，事前に登録するキーワードに強く依存

⋆1 査読者の 1 人より，編集著作物ではないかという助言を得た．感謝するとともに，本システムおよび古典籍データベースの開発・公開において，参考にしたい．

する．本研究のシステムのように，全文検索エンジンを用いれば，事前登録のキーワードの多寡にかかわらず，検索語を与えればそれが出現する情報を知ることができる．さらに，検索語の前後に文字列を付加することで，件数を絞り込める効果も期待できる．たとえば，本検索システムでは「東寺」を検索語とすると 32,409 件が該当するが，「於東寺」とすれば 14 件となり，容易にそのすべてを読むことができる．

古典籍に含まれるキーワードを分析し，記録と年代を結び付けたり，情報どうしの関連性を知る手がかりとしたりすることについて，これまで多くの取り組みがなされている．年代の取得や表現については，たとえば相田[22] は，紀元前 660 年から 2082 年までの各日に関する暦日データベースを作成し，改元，閏月，旧暦と太陽暦の違いなどに配慮している．金田[23] は，百科事典の検索・表示のための年代データについて検討しており，「年」の後ろにつく接尾語も利用して，年代の機械的な推定を行っている．人物名の処理については，Yahoo!検索 Web サービスに問い合わせ，Web 上の情報を対象に著名人の呼称を抽出する試みがある[24]．同名異人あるいは異名同人の対処に関する検討もなされている[25),26)]．

本研究では，古典籍テキストデータに出現する年代・寺院・地名・人物の文字列を満遍なく取得し，キーワードデータベースに登録しており，これによりキーワードを連携させて多方面に検索の幅を広げ，古典籍検索を支援している．キーワード抽出手法に関しても，高速に漏れなく検索する自前の全文検索エンジンと，Web 上で広く収集している検索サービスとを組み合わせ，それぞれの特性を生かして「もっともらしい」キーワードを発見しようとする試みは，これまでにないものである．

筆者らは，人物とりわけ僧侶に関する血脈（師弟関係）や親子・兄弟関係の情報のデータベース化を試み，系図の作成・表示の方法について検討してきた[7),27)]．本検索システムと，ユーザインタフェースおよびデータ構造の両面で連携を進め，一体として利用可能な検索閲覧システムを提供することで，文献 5) ほかのような，時代・聖教・人物をまたいだ歴史的な課題について，その発見，解明，検証が効率良く行えるようになると考える．

7. おわりに

本研究では，聖教書誌情報約 44,000 件を対象とした全文検索システムを構築し，検索語を含む聖教書誌情報を瞬時に獲得できることを確認した．そして，毎回検索語を入力する際の手間や打ち間違いを減らすため，キーワードを抽出してデータベースで管理し，検索時にリンクとして提示するほか，年代検索に関するより柔軟な検索方法を実装した．

抽出した年代・寺院・地名・人物のキーワードは，検索作業の省力化だけでなく，キーワードの統計分析など，聖教研究の分野に対してこれまでにない活用が期待できる．古典籍に関するシソーラスや，漢文の形態素解析[28] における辞書などにも利用できるよう，精度を高めるとともに素性を付与していき，公開できる水準にしたい．

謝辞　本研究は科研費（17520310，20300067）の助成を受けたものである．

参考文献

1) 八村広三郎：人文科学とデータベース，情報処理学会誌，Vol.38, No.5, pp.377–382 (1997).
2) 松下眞也：古典籍総合データベースの構築と展開，早稲田大学図書館紀要，No.53, pp.1–24 (2006).
3) 宇都宮啓吾：東寺観智院金剛蔵『天台血脈』について，大谷女子大学紀要，Vol.38, pp.44–99 (2004).
4) 宇都宮啓吾：血脈資料『諸嗣宗脈紀』について—龍谷大学本を手懸かりとしたその成立とデータ公開を巡る問題，大阪大谷国文，Vol.1, pp.45–60 (2007).
5) 宇都宮啓吾：聖教調査におけるデータの整備を巡る問題—『念仏宗僧運覚』・書誌情報を中心として，基盤研究（A）(1)「金剛寺一切経の総合的研究と金剛寺聖教の基礎的研究」報告書 (2007).
6) 上川通夫：中世聖教史料論の試み，史林，Vol.79, No.3, pp.108–129 (1996).
7) 朴　明哲，森本雅史，立花純児，村川猛彦，宇都宮啓吾，中川　優：人文研究を支援するデータベースシステム—聖教検索および系図表示，情報知識学会誌，Vol.17, No.2, pp.105–110 (2007).
8) 村川猛彦，山中克真，宇都宮啓吾，中川　優：古典籍書誌情報におけるキーワード抽出手法，情報知識学会誌，Vol.18, No.2, pp.87–92 (2008).
9) Hyper Estraier: a full-text search system for communities. http://hyperestraier.sourceforge.net/index.html
10) 京都府教育委員会：東寺観智院金剛蔵聖教目録 (1977).
11) 嵯峨美術短期大学綜合美術研究所：大覚寺聖教目録 (1992).
12) 築島　裕：平安時代訓点本論考 (1996).
13) 中田祝夫：古点本の国語学的研究 総論篇，講談社 (1954).
14) 文化庁文化財保護部美術工芸課：来迎院如来蔵聖教文書類目録 (1972).
15) 東大寺図書館：宗性・凝然写本目録 (1959).
16) 村川猛彦，丁　敏，中川　優：仏典全文検索システムの構築と評価，人文科学とコンピュータシンポジウム論文集，情報処理学会シンポジウムシリーズ，Vol.2007, No.15, pp.221–228 (2007).
17) 異体字処理　付表．http://www.arc.ritsumei.ac.jp/lib/geino/hyoban/doc/fuhyou.htm
18) 僧名露視　稿（二）．http://home.hiroshima-u.ac.jp/mmatumo/sites/

soumei(2).html
19) 青江順一：トライとその応用，情報処理，Vol.34, No.2, pp.244–251 (1993).
20) 文化審議会著作権分科会法制問題小委員会平成 19 年度・中間まとめ，pp.45–61 (2007). http://www.bunka.go.jp/chosakuken/singikai/pdf/housei_chuukan_1910.pdf
21) 木村文則，小牟礼雅之，前田　亮，佐古愛己，杉橋隆夫：古典史料データベース検索システムの提案，情報処理学会研究報告 人文科学とコンピュータ，2008-CH-78, pp.45–52 (2008).
22) 相田　満：暦象オントロジの構築—日本旧暦時代の文献分析支援のために，情報処理学会研究報告 人文科学とコンピュータ，2007-CH-76, pp.25–32 (2007).
23) 金田　泰：百科事典から動的に年表を生成するテキスト検索法のための年代情報の抽出法と表現法，情報処理学会研究報告 情報学基礎，1999-FI-055, pp.81–88 (1999).
24) Hokama, T. and Kitagawa, H.: Extracting Mnemonic Names of People from the Web, *Proc. 9th International Conference on Asian Digital Libraries* (*ICADL 2006*), LNCS 4312, pp.121–130 (2006).
25) 小野真吾，佐藤一誠，吉田　稔，中川裕志：重要語抽出を用いた Web 文書上の同姓同名の曖昧さ解消，電子情報通信学会第 19 回データ工学ワークショップ（DEWS2008），No.A7.3 (2008).
26) 研谷紀夫：デジタルネットワークにおける歴史的人名・組織情報の現状とその利活用モデル，情報知識学会誌，Vol.18, No.2, pp.93–98 (2008).
27) 田中猛彦，冨金原賢次，宇都宮啓吾，中川　優：平安・鎌倉時代を対象とした僧侶データベースシステム，情報知識学会誌，Vol.13, No.2, pp.18–31 (2003).
28) 守岡智彦：MeCab を用いた古典中国語の形態素解析の試み，情報処理学会研究報告 人文科学とコンピュータ，2008-CH-79, pp.17–22 (2008).

(平成 20 年 6 月 20 日受付)
(平成 20 年 10 月 8 日採録)

（担当編集委員　高久 雅生）

村川　猛彦（正会員）

昭和 46 年生．平成 10 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．奈良先端科学技術大学院大学情報科学研究科助手，和歌山大学システム工学部助手を経て，平成 15 年より和歌山大学システム工学部講師，現在に至る．Web アプリケーション，ディジタルアーカイブに関する研究に従事．

宇都宮啓吾

昭和 41 年生．平成 5 年広島大学大学院文学研究科博士課程後期国語学国文学専攻単位取得のうえ退学．修士（文学）．大谷女子大学専任講師を経て，平成 18 年より大阪大谷大学教授，現在に至る．国語学・文献学（特に訓点資料を中心とした研究）に関心を持つ．日本語学会，訓点語学会，戒律文化研究会，智山勧学会各会員．

中川　　優（正会員）

昭和 45 年大阪大学基礎工学部卒業，昭和 47 年同大学院修士課程修了．同年日本電信電話公社武蔵野通信研究所，平成 6 年近畿大学生物理工学部教授，平成 9 年和歌山大学システム工学部教授．工学博士（大阪大学）．IEEE，情報知識学会，人工知能学会各会員．