2006－CH－69（5）
2006／1／27

仏典データベースのためのテキスト処理について

田中 猛彦(1)　　仁野 洋平(2)　　中川 優(1)
(1)和歌山大学 システム工学部
(2)和歌山大学システム工学研究科

仏典撮影画像を対象としたディジタルアーカイブに検索性を持たせるためには，テキストデータの活用が不可欠である．本研究では，大正新脩大蔵経テキストデータと Hyper Estraier を用いた全文検索システムを構築し，9,000 を超えるテキストファイルを登録した．文字列の指定方法として，従来の文字順での検索のほか，縦書きの経典に対して横方向に連続する数文字を指定しても検索できる．経典画像とテキストデータを行単位で対応付けて参照するための，テキストファイルの加工方法も紹介する．

Text Processing for Buddhist Canon Database

Takehiko TANAKA(1)　　Yohei NINO(2)　　Masaru NAKAGAWA(1)
(1)Faculty of Systems Engineering, Wakayama University
(2)Graduate School of Systems Engineering, Wakayama University

Text data are indispensable for retrieving the image among the digital archive of enormous files taken with digital cameras. We constructed a full-text search system of over 9,000 text files of Taisho Tripitaka, using Hyper Estraier as the search engine. To identify the text file, one can specify not only the phrases in a traditional way but the characters positioned hosizontally in a Buddhist canon written vertically. Another approach of the text processing is introduced, which helps the user to view the Buddhist canon images and the line-by-line corresponding text data at the same time.

1 まえがき

図 1 は，金剛寺一切経のうち，放光般若經巻第一の冒頭である．この画像に対して，この内容を含むテキストデータを取得し，画像とそのテキストデータとを対応付けて閲覧できるようにするには，画像処理[1]だけでなく，対象となるテキストファイル，あるいはテキストファイル群の検討が不可欠である．

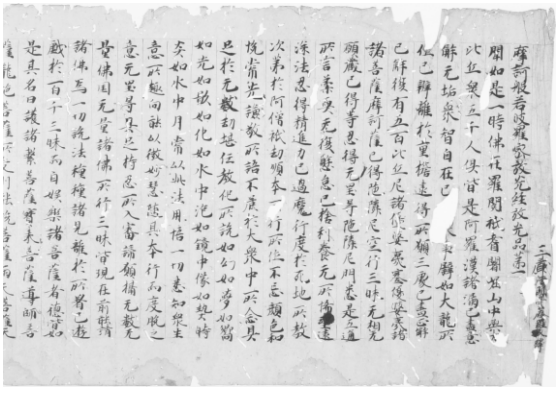

図 1: 仏典画像の例

金剛寺一切経については，大正新脩大蔵経との対応付けが行われており，大正新脩大蔵経のテキストデータはSATやCBETAにおい

て電子化,提供されている.本稿では,CBETAが配布しているテキストファイルを使用して,検索や閲覧に適したテキスト情報を抽出するための筆者らの試みを紹介する．プレーンテキストのデータではなく XML ファイルを用いることで，効果的・効率的に情報を抽出できるようになるとともに，外字の問題も考慮したテキスト処理ができるようになった.

さらに，このテキスト情報を用いたアプリケーションをいくつか試作している．その一つは，Hyper Estraier を用いた全文検索システムである．テキストデータに改行を入れ，行と列を反転させた情報を用いることで,「如炎意」といった，画像中の行頭の数文字を指定しても検索できるようにした．金剛寺一切経の画像と対応付けて閲覧するための，テキストの縦書き表示方法についても紹介する.

2. 準備

ここでは，本研究の前提となる，金剛寺一切経，大正新脩大蔵経，CBETA のテキストデータについて簡単に述べる.

大阪府河内長野市にある天野山金剛寺には,「金剛寺一切経」と呼ばれる漢訳経典約 4,500 点が今日に伝えられている．金剛寺一切経は，仏教学・国語学・歴史学などにおいて非常に重要な資料であり，その保全と活用のため,高精度ディジタルカメラにより撮影,電子化され続けている.

撮影された経典画像だけでは，文字などを探すときに大変な労力を伴うことになる．そこで，漢訳仏典で広く用いられている大正新脩大蔵経を参考にすることにした．実際，電子化作業に先立って，大正新脩大蔵経と金剛寺一切経との対応付けがなされている[2,3].

大正新脩大蔵経のテキストデータは台湾の中華佛典協会(CBETA)により整理されている．テキストデータは，Web サイト(http://www.cbeta.org/)からダウンロード可能であるほか,CD-ROM も配布されている.筆者らは，この情報をうまく使うことにより金剛寺一切経の解読を支援できるのではないかと考えた.

3 テキスト処理

3.1 対象とするファイル

まず，CBETA のテキストデータを，閲覧や検索がしやすいように加工した．用いたデータは，「CBETA Chinese Electronic Tripitaka Collection Feb. 2005」の CD-ROM のフォルダ ¥cbreader¥xml に格納されている，T01.zip から T85.zip までのファイルである．伸張すると，一つの zip ファイルあたり数十から数百の XML ファイルを得る.CD-ROM にはプレーンテキストのファイルも収録されているが，この形式を採用しなかった理由は，ファイル中の「‖」(2 重縦棒)が文字コード変換の障害となったのと，外字の表現方法が非効率であったためである.

テキスト処理を効率的に進めるため，文字コードを Big5 から UTF-8 に変換した．変換ソフトウェアには iconv を用いている．変換において，文字の欠落などは認められなかった．句点の見かけが変更されるが，句点は後に除去されるので，問題にならない.

3.2 テキスト抽出処理

放光般若經卷第一(T08n0221_001.xml)の内容の一部を，図 2 に示す．本節では，このような XML ファイルから，タグを適切に取り除いてテキストデータにする方法を述べる.

```
<note n="0001003" resp="Taisho" type="orig"
place="foot text">〔摩訶…蜜〕－【三】【宮】
</note><note n="0001003" resp="CBETA"
type="mod">〔摩訶…蜜〕－【宋】【元】【明】
【宮】</note><app n="0001003" word-
count="7"><lem><title>摩訶般若波羅蜜
</title></lem><rdg wit="【宋】【元】【明】【
宮】" resp="Taisho">&lac;</rdg></app>放光品
第一
</head>
<lb n="0001a08"/><p>聞如是。一時佛在羅閱祇耆
闍崛山中。與大
<lb n="0001a09"/>比丘眾五千人俱。皆是阿羅漢。
諸漏已盡意
<lb n="0001a10"/>解無垢。眾智自在已了眾事。
譬如大龍所作
<lb n="0001a11"/>已辦。離於重擔逮得所願。三
處已盡正解已
<lb n="0001a12"/>解。復有五百比丘尼諸優婆塞
```

図 2: XML ファイル

XML ファイルには，本文だけでなく，経巻情報や注釈などがタグとして付加されているが，これらのほとんどは，筆者らが必要とするアプリケーションでは不要であった．具体的には，ファイルの先頭から body タグが現れるまで， および note 要素と rdg 要素を削除した．ただし，lem 要素については，その中身のテキストは意味を持つので，タグのみ削除した．

さらに，XML ファイルを観察すると，後々のテキストデータベース化のために考慮するとよいタグがあることも分かった．具体的には，head 要素と jhead 要素については，その中身のテキストがタイトルとなっている．また，lg 要素は，これを独立した 1 行とみなすとよさそうである．これらのタグは除去するとともに，その前後に改行コードをつけた．ファイルにあった改行は削除した．結果として，1 行が数千字に及ぶものもある．

XML ファイルでは，外字を「&CB00165;」という形で参照している．この実体は別のファイルで定義されているが，本研究では使用していない． このような実体参照は，縦書き表示時や，全文検索システムに構成において，適切とはいえない．我々は，テキスト化に当たってはこのコードをすべて「=」(ゲタ記号)に置き換えた．そして，外字情報ファイルを別途作り，その文字コードを出現順に記載した．これにより，この文字が 1 文字であることが明確になる．一つのテキストデータに，ゲタ記号が複数回出現することになるが，それぞれの漢字が何であるかは，XML ファイルや ent ファイルに立ち返ることなく，変換されたテキストファイルでの出現位置と，外字情報ファイルから求めることができる．

この時点で，1 バイト文字(ASCII 文字)は不要となるのですべて削除する．また，ASCII 文字でない記号類(句点「。」，二重丸「◎」，および空白文字)も削除する．

放光般若經卷第一に対して処理を行い，得られたテキストについて，その一部を図 3 に示す．

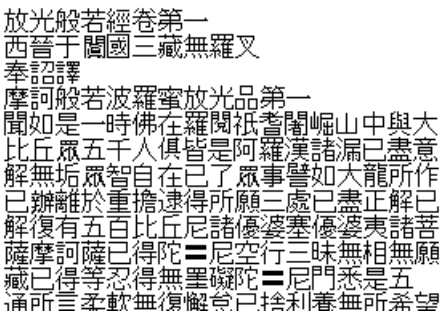
放光般若經卷第一
西晉于闐國三藏無羅叉
奉詔譯
摩訶般若波羅蜜放光品第一
聞如是一時佛在羅閱祇耆闍崛山中與大
比丘眾五千人俱皆是阿羅漢諸漏已盡意
解無垢眾智自在已了眾事譬如大龍所作
已辦離於重擔逮得所願三處已盡正解已
解復有五百比丘尼諸優婆塞優婆夷諸菩
薩摩訶薩已得陀〓尼空行三昧無相無願
藏已得等忍得無罣礙陀〓尼門悉是五
通所言柔軟無復懈怠已捨利養無所希望

図 3: テキストファイル

テキスト変換プログラムは，Ruby を用いて構築した．例えば，文字列 string の中にある句点をすべて取り除きたいとき，

```
string.gsub!(/。/u, "")
```

と書けば，文字列を UTF-8 とみなして「。」を探し，削除する．

Pentium M 2.0GHz, 1GB メモリ, Gentoo Linux の計算機を用いて，XML ファイルからプレーンファイルを精製した．時間は，iconv による文字コード変換，XML からプレーンテキストへの変換プログラムのそれぞれで，1 ファイルあたり 1 秒未満であった．

4 アプリケーション例

4.1 全文仏典検索システム

テキストにしたファイル群をもとに，ユーザが指定するキーワードを含むファイルを求めるための全文検索システムを構築した．実行環境としては，サーバには Linux と Apache および CGI (Common Gateway Interface)を用いた．ユーザは Windows 環境で Web ブラウザを起動してサーバにアクセスし，検索語をタイプする．(本稿では，ユーザが指定する文字列を「検索語」と呼び，全文検索システム内で登録されている「キーワード」と区別する．)

全文検索エンジンについては，Hyper Estraier[4]を採用した．このソフトウェアの特徴を簡単に述べる．まず，内部コードが

Unicode (UTF-8)となっている．ファイル登録時に，特に何も指定しなければファイルの内容から文字コードを推定するが，UTF-8 をコマンドオプションに指定することで，効率よく登録される．

Hyper Estraier の 2 番目の特徴は，N-gram 法を採用していることである．Namazu などで採用している「分かち書き方式」を使用する場合，形態素解析ツールを用いて処理する必要がある．しかし我々の目標の下では，語すなわち形態素に分割する必要はなく，連続する N 文字ごとにキーワードの候補として格納するほうが適切である．また，意味解析をすることのない分，N-gram 法は効率よくファイル登録ができると期待される．

9,035 個のテキストファイルに対して，インデックスを作成した．前述の Linux 計算機で，インデックス生成時間は 19 分 30 秒となった．またファイルサイズは，テキストファイルの総計が 240M バイトに対して，インデックスは 528M バイトとなった．キーワード数は 2,009,148 である．

このインデックスを用いて，検索を行った(図 4).

図 4: Hyper Estraier による検索例

ここでは既に画像化されている放光般若經卷第一と大寶積經卷第六十五を用いる．事前の調査で，放光般若經卷第一に対応するテキストは T08n0221_001 のはじめの 2/3 ほどであり，大寶積經卷第六十五に対応するテキストは T11n0310_065 の全てと T11n0310_066 の前半であることが分かっている．画像の中から，文字の並びをいくつか見つけ，検索語として入力して，期待するファイルが発見できた(ヒットした)かを確認した．さらに，期待するファイルを含めて何件のファイルがヒットしたか(ヒット数)も求めた．全文検索システムにおいて，検索語の出現回数などによる「順位」も知ることができ，Hyper Estraier でも設定ファイル次第で，スコアやそれに基づく順位を知ることができる．しかし，本研究では上位に出現することを望むものではなく，むしろ，ヒットすることを確認し，どのような検索語を与えればヒット数が下がる(理想的には 1，すなわち期待するファイルのみ)かを知ることができればよい．したがって，スコアや順位は考慮しないこととする．

検索結果を表 1 に示す．ここで，「001」の列は， T08n0221_001.xml から抽出したテキストファイルがヒットしていれば「○」としている．「065」，「066」も同様である．この表で，「066」に関して獲得できているのはいずれも，大寶積經卷第六十六すなわちその次の経巻に含まれている文字列である．

表 1: 検索結果（縦方向検索）

検索語	ヒットしたファイル			ヒット数
	001	065	066	
炎如水	○			11
炎如水 意所趣	○			1
深法	○			1455
深法忍	○			109
深法忍得	○			4
深法忍得精	○			3
深法忍得精進	○			1
深法忍 次第於	○			3
所言柔 深法忍 次第於	○			1
復有八億		○	○	4
供養如 及聞授		○	○	2
供養如来及聞授				0
供養如來及聞授		○	○	2
供 養 如 來 及 聞 授	○	○	○	3347

「深法」から 1 文字ずつ増やしていくと，ヒット数が減少していき，「深法忍得精進」としたところで，期待するファイルのみがヒットするようになった．また，「深法忍」の前後 3 行の 3 文字分(9 文字)を検索語としても，期待するファイルのみとなった．他にも検索を試行した限りでは，3 文字未満の検索語ではヒット数がかなり多いが，字数が増えていけばいくほど，ヒット数の低下，すなわち期待されるファイルの獲得確率が上昇することが分かった．また，3 文字程度の検索語であっても，複数指定することで，やはりヒット数が低下することが確認できた．

しかし，「供養如来及聞授」のように，1 文字でも間違っていると，ヒット数が 0 になってしまう．これは「来」の字が CBETA のテキストデータでは使用されていないためであり，この字を「來」に変更すると，ヒットするようになった．

検索語として「供養如來及聞授」を指定すると，この文字列が出現するファイルを求めるが，すべてに空白文字を入れて「供 養 如 來 及 聞 授」とすると，これらの文字がそれぞれ含まれているファイルを求めることになり，ヒット数は非常に大きくなる．

検索時間について触れておく．1 文字の検索語を与えない場合，それぞれの検索時間は，短いものは 0.01 秒未満，長くても 0.5 秒程度であった． 1 回の検索で指定する検索語の語数については，1 語でも，2 語以上でも時間に差は見られなかった．検索語として 1 文字の語を与えると，これは N-gram による検索には適しておらず，検索に時間がかかる．1 文字の語を複数指定すると，その語数に応じて検索時間がかかるようである．ただし，「供 養 如 來 及 聞 授」という検索語でも，1 秒以内であった．また，Web サーバや検索エンジン内のキャッシュのため，同一の検索語を用いて複数回検索すると，2 回目以降の時間が短くなった．

4.2 横方向検索が可能な全文仏典検索システム

横方向検索ができるよう，テキストデータを加工した．具体的には，3.2 節で生成したテキストデータに対して，決められた字数で改行を行った．ただし，テキストデータ中の改行は保持している．字数については，さまざまな仏典画像を見る限り，17 字が標準であるが，16 字または 18 字の行も見られた．放光般若經卷第一の各行の字数を調査したところ，17 文字が 380 行で全体(446 行)の 85%，16 字と 18 字の行を合わせると全体の 97.5%を占めることがわかった．そこで，これまでのテキストデータに，「17 文字で改行し，転置(縦横反転)してできるテキストデータ」を加えた．ここで，字数の足りない箇所には「・」の文字を挿入し，行と列の位置関係が崩れないようにした．16 文字と 18 文字でも同様に，改行・転置してできるテキストデータを追加した．出来上がったファイルの一部を図 5 に示す．

```
放光般若經卷第一西晉于闐國三藏無羅...
座結跏趺坐正受定意三昧其三昧名三昧...
法爾時三千大千國土諸盲者得視聾者得...
今現在以般若波羅蜜教化一切有菩薩名...
邊沙有世界名思樂其佛號思樂威如來無...
...
17-01:放西摩聞比解已解薩藏所深次和...
17-02:光晉訶如丘無鞞復摩已言法第悅...
17-03:般于般是眾垢離有訶得柔忍於常...
...
```

図 5：横方向検索用のテキストファイル

9,035 個のテキストファイルに対して，上記の方法で新たなファイルを生成し，それを対象として，Hyper Estraier でインデックスを(前節とは別のディレクトリに)構築した．生成時間は約 4 時間かかった．また，ファイルサイズは，テキストファイルの総計が 910M バイト，インデックスは 2,355M バイト，キーワード数は 5,508,604 となった．

このインデックスを用いて，検索を行った．まず，前節と同じ検索語を試したところ，「深

法」のヒット数が 2220,「深法忍」は 116 となったが，それ以外はすべて表 1 と同じ結果となった．次に，横方向の検索語として，放光般若經卷第一と大寶積經卷第六十五より，文字の並びをいくつか取り出して検索した．その結果を表 2 に示す.

表 2: 検索結果（横方向検索）

検索語	ヒットしたファイル			ヒット数
	001	065	066	
聞比解	○			6
諸願所	○			23
所深次	○			4
如炎意	○			1
王益常	○			1
南生真	○			3
南生真光				0
南生 生真 真光	○			5
一切行薩		○		6
我既既		○		4
誰誰誰			○	10

表に示した以外にも，3 文字で横方向検索を試したが，ヒット数が 100 を超えるものは見られなかった.

放光般若經卷第一では行頭が「…南生真光…」となっている箇所があるが，これを検索語としたところ，ヒットしなかった．経典画像を見直したところ,「南」と「生」の距離は 17(これらの間に 16 文字ある),「生」と「真」の距離も 17 であったが,「真」と「光」の距離は 18 であった. そこで,「南生 生真 真光」で試してみた．この入力は,「南と生の距離，生と真，真と光の距離がいずれも 16～18 文字以内であるファイルを求めよ」という意味である．こうすると，5 つのファイルが見つかり，放光般若經卷第一も含まれていた.

経典の中には，1 行が 14 文字(7 文字，空白，7 文字)または 20 文字(5 文字が 4 回あり，それぞれが空白で挟まれている)で数行から数十行続くという書式を持つものもある．1 行が 14 文字となっている中で, ある行頭の 3 文字を選んだのが,「我既既」という検索語である．この出力結果を見ると，それぞれのファイルに「我既既」が 3 つ含まれていることが確認できる．これは，17 文字で改行した場合，18 文字で改行した場合，16 文字で改行した場合のすべてでこの文字列が含まれているためである.

また，本検索システムでは 1 行の区切り方は最大 18 文字なので，1 行が 20 文字のものについて，そのまま横方向検索することはできない．代わりに，ある行の先頭，その行の 11 文字目の先頭, 次の行の先頭の文字を選んでできた「誰誰誰」を試したところ，大寶積經卷第六十五を含むファイルにヒットした.

4.3 縦書き表示

金剛寺一切経ディジタルアーカイブの活用方法の一つとして，仏典画像に対して，画像とテキストを対照させて閲覧できることを構想している[5]．ここで，テキスト情報は，表示の方法(フォントサイズ，行間の長さ，1 行の字数など)を変更するのが容易である. また, 1 行の字数は，17 文字固定ではなく，閲覧画面においては，仏典画像の行と対応付けられているべきである．この対応付けは，機械的処理のみでは容易ではない.

ここにわずかな人手を介することにする．具体的には，3.2 節で生成したテキストデータについて，17 文字ごとに改行し(ただし，テキストデータの改行は保持する), できたテキストデータをユーザに提示する. ユーザは, テキストエディタを用いて，画像と照合しながら，改行すべき位置に「/」の文字を置いて指定する(図 6). 改行位置の指定作業が完了すれば，さらにプログラムを実行させ，これまでの改行コードを取り除き，「/」を改行コードに置換することにする.

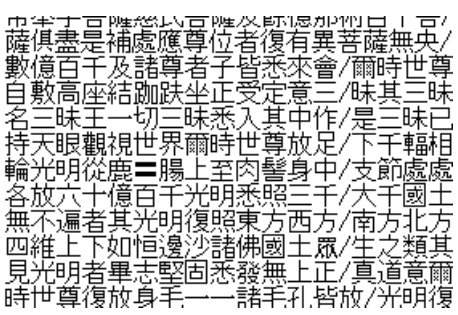

図 6: 手作業による改行位置の挿入

17 文字ごとに改行したのは，前述の通り，放光般若經卷第一の調査で 85%の行が 17 文字であり，他の経典もそれほど違わないためである 1 行が 14 文字や 20 文字で長く続く行については，テキストデータの改行を保持しておくことで，意図しない行の統合や分割が起こることはない．

446 行ある放光般若經卷第一で，筆者の一人が「/」の文字を置く作業を行ったところ，18 分を要した．1 行当たり約 2.4 秒である．1 行が 17 文字となるのが数十行続き，単純作業で配置できたところもあった．1 行が 17 文字でないところでも，すぐ下の行の前後を見ればたいてい改行位置が見つかる．

放光般若經卷第一は，T08n0221_001.xml のはじめの 2/3 であり，残りは放光般若經卷第第二の先頭となる．このように，金剛寺一切経の経巻と，CBETA 収録のテキストファイルは，一対一対応をなしていないが，この場合でも，テキストエディタの機能を活用して不要な領域を除去したり，複数のテキストファイルを連結したりすることで，画像と行単位で対応付けが可能なテキスト情報を，効率よく作成できるようになった．

自作の縦書きプログラム(JavaScript)を通して得られた結果を図 7 に示す．2 箇所出現する「隣」の字は外字であり，CD-ROM に収録されている外字画像を組み込んだ．

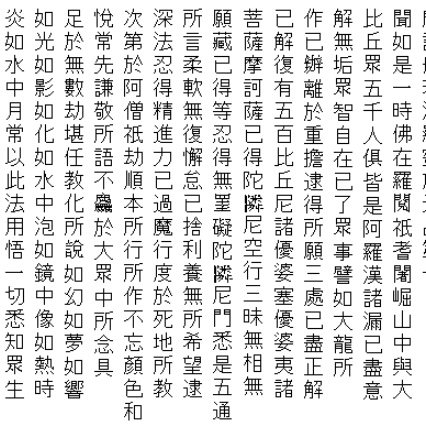

図 7: 縦書きのテキスト

5 仏典検索サイトとの比較

ここで，既存の二つの CBETA 仏典検索サイトを紹介し，本システムとの比較を試みる．

テキストデータの編集・配布元である CBETA のサイトでは，様々な方法で経典の検索が可能である．とりわけ，蔵経の種類，巻・号・頁で指定したり，経典名で検索したり，巻を限定して全文検索したりすることが可能である．

さらに，サイト上のファイルに対して，Google を用いた全文検索システムも提供されている．ただしこの検索では，文字コードが Big5 で，句点もそのままで登録されているファイルに対する検索となる．さらに，その精度は高くない．例えば「意所趣」を検索語とすると 4 件がヒットしたが，この中に放光般若經卷第一はなかった．

日本人によって管理・提供されている，大正新脩大藏經に関する全文検索システムとして，仏教典籍検索(http://www.kosaiji.org/~kyoten/)が知られている．検索エンジンに Namazu を使用し，インデックスの生成においては分かち書きツールの KAKASI に独自の仏教用語辞書を使用している．文字コードは Shift-JIS である．

ここの検索システムの最も大きな特長は，新字体で検索できる点である．これにより，例えば「如來」ではなく「如来」を検索語として利用でき，ユーザからの入力がより容易

になる．新字体への対応は，筆者らも，本稿で紹介した全文検索システムにおいて，取り入れたい機能の一つである．ただし，インデックス生成までの時点で，テキストを新字体にしておくのがよいかは検討の余地がある．すなわち，インデックス作成の時点では原テキストデータのままとし，検索時や表示時にオプションを指定することで，原文のまま，あるいは新字体に変換して表示するかが選べ，多くの利用者に好まれるのではないかと考えている．

また，検索語も分かち書きによって分割されてしまう点を指摘しておきたい．具体的には，「深法忍得精進」を検索語として結果を見たところ，9 件がヒットした．放光般若經卷第一も含まれている．結果によると，「深法忍得精進」という検索語は「深」，「法」，「忍」，「得」，および「精進」に分割されて，これが全て含むファイルを検索していた．筆者らは，これは余計な分割であり，ユーザの指定する語が「XY」の場合と「X　Y」の場合とでは，検索内容は異なるべきだと考えている．すなわち，前者は「X と Y がこの順に連接しているようなファイル」，後者は「X と Y の両方が出現するファイル」とするのが自然である．

6 あとがき

本稿では，大正新脩大蔵経テキストファイルを活用した，金剛寺一切経経典画像の検索や閲覧に適したテキスト情報を抽出するための筆者らの試みを報告した．Hyper Estraier を用いた全文検索システムを構築し，従来ある文字通りの検索(縦方向検索)だけでなく，複数行の先頭位置などを指定した検索(横方向検索)を実現し，数文字の検索語で期待する文書が獲得できることを確認した．また，縦書き表示方法の取り組みも紹介した．

今後の課題としては，検索については精度と速度のより精密な評価が，縦書き表示については省力化および閲覧システムへの組み込みが挙げられる．

謝辞

XML ファイルの活用に関して，京都大学の Christian Wittern 氏よりアドバイスをいただきました．深く感謝します．

本研究は部分的に日本学術振興会科学研究費補助金 基盤研究(A) 課題番号 15202002 の補助を受けた．

参考文献

[1] 張他: 「仏典データベースのための画像処理について」, 情報処理学会研究報告, 2005-CH-069 (2006)

[2] 落合俊典: 「金剛寺一切経の基礎的研究と新出仏典の研究」, 平成 12 年度～15 年度科学研究費補助金基盤研究(A)・(1) 研究成果報告書, 課題番号 12301001, 364p (2004)

[3] 「日本現存七種一切経対照目録(暫定版)」, 国際仏教学大学院大学 学術フロンティア実行委員会, 260p (2005)

[4] Hyper Estraier: a full-text search system for communities, http://hyperestraier.sourceforge.net/

[5] 仁野他: 「仏典画像閲覧のためのデータベースシステムの構築」, 情報知識学会誌, Vol.15, No.2, pp.15-18 (2005)