Vol.2017-CH-114 No.6
2017/5/13

# 多漢字文献の検索効率向上とマルチデバイス対応の試み

劉冠偉[†1]　李媛[†2]　池田証壽[†3]

**概要**：多漢字文献テキストデータベースを構築する際には，文献に出現する文字をコード化して正しく表示するのは重要な課題である．一方，構築したデータベースを公開する際には，収録内容の漢字を検索するために求める漢字を効率的に入力することも重要な課題である．CHISE や GlyphWiki などは漢字検索にも有益なツールとなるが，漢字検索に特化し，PC 以外の端末でも使いやすいシステムはまだないようである．この発表は漢字の構成記述文字列（IDS）情報と画数情報と組み合わせて，PC やスマートフォンに対応した新しい漢字検索システムを提案する．材料とする多漢字文献としては，平安時代漢字字書総合データベース（HDIC）に収録した漢字字書 2 種（篆隷万象名義，大広益会玉篇）を取り上げる．

**キーワード**：多漢字文献，IDS，平安時代漢字字書総合データベース，漢字検索

# Improving of Searching Efficiency and Development of Multi-Device Adaptive for a Multiple Chinese Characters Archive Database

Guanwei Liu[†1]　Yuan Li[†2]
Shoju Ikeda[†3]

***Keywords***: archive database, IDS, HDIC, Chinese characters searching

## 1. まえがき

近年，漢籍や仏典に関する多漢字文献データベースが数多く公開され，国語学の研究・学習に非常に有益となっている．また，スマートフォンやタブレット端末などの PC 以外の端末が盛んに利用されており，インターネット上の言語資源もこれらのスマートデバイスへの対応が求められている．そこで，PC にも PC 以外の端末にも使える漢字検索と表示を特化した多漢字文献データベースシステムを開発したいと考えるに至った. その課題は, 次の 5 点となる．

(1)　利用に制限のない, 多漢字文献のデジタル化データ
(2)　多漢字文献のデータの格納・検索
(3)　難字・異体字の入力・表示
(4)　PC 以外のマルチデバイス対応
(5)　開発と実装

公開された多漢字文献は徐々に増えているが，この発表では平安時代漢字字書総合データベース（HDIC）に収録・公開の『篆隷万象名義』と『大広益会玉篇』を取り上げる．前述の五つの課題は次の 2 節〜5 節にわたってその詳細を論じる．

まず第 2 節では課題（１）に関しての利用資料の詳細を述べる．第 3 節では課題（２）・課題（３）に関して漢字検索をめぐる問題とその解決方法を述べる．第 4 節では課題（4）のマルチデバイス対応について，PC との区別を中心に述べる．第 5 節では，実際に開発した際に生じた問題点と実装したフレームワークにつて述べる．

## 2. 利用資料

HDIC は日本の平安時代に編纂された『篆隷万象名義』『新撰字鏡』『類聚名義抄』の総合データベースであり，池田（2014）によると，この三つの辞書を解読するために，中国の『大広益会玉篇』などの辞書もデータベースに含めている．ほかの多漢字文献と比べると，次の二つの公開特徴がある．

(1) GitHub で公開しているため，データの点検・更新がわかりやすい
(2) ライセンスが明記されており，利用に制限が少ない

## 3. 漢字検索の効率化

### 3.1 多漢字文献における漢字の検索と入力の問題

前述のように, 多漢字文献データベースを利用する際に, かな入力で入れにくい難字･異体字が多い. それに対して, 通常の文献の内容を検索するにあたっては，次の三つのステップによって行うことが多い.

(1) 「かな入力」などの入力方式によってかなを入力する［予備入力］
(2) (1) の入力内容によって変換した漢字の候補から一つを選定する［本入力］
(3) システムは選定した漢字を検索する［内容検索］

かな入力以外の手段では，次の三つの漢字検索・入力方法がある．

†1 北海道大学文学研究科博士後期課程
Graduate School of Letters, Hokkaido University
†2 北海道大学文学研究科専門研究員
Graduate School of Letters, Hokkaido University
†3 北海道大学文学研究科
Graduate School of Letters, Hokkaido University

1　手書き入力:指や特別なペンを用いて入力する方法.漢字知識を必要としていないが，拡張漢字のサポートが不十分である．速度はキーボードほど期待できない．デバイスやソフトウェアによって認識度が大きく異なるが，すべての Unicode 漢字をサポートするシステムは管見の限りまだ見つからない.

2　字形入力 :「蒼頡」や「五筆」など，本来，中国語専用の入力方式を用いて入力する方法．字形をいくつかの小さなパーツに分け，それぞれに対応させたアルファベットを入力する方式である．特に「蒼頡」はほぼすべての Unicode 漢字をサポートするので，難字・異体字の入力には最も頼れる方式である．しかし，これらの入力方式は運用できるまでの長い学習時間を要し，台湾や香港のソフトウェアであるため，フォントが異なることなどで，国語学研究に応用するのはまだ検討が必要である.

3　部首画数検索：部首画数を用いての漢字検索方法.利用しやすいが，候補が多数を表示されるため，検索効率が低い.

表 1　Unicode 漢字入力方法の比較

| | 入力可能数 | 入力効率 |
|---|---|---|
| かな入力 | やや少ない | やや低い |
| 手書き入力 | 少ない | 低い，やや低い（機種による） |
| 字形入力 | ほぼすべて | 高いが，学習時間が必要 |
| 部首画数検索 | すべて | 低い |

上記のように，入力効率から考えると字形入力が最も優れている．一方，かな入力は学習が容易であるが，次の問題がある.

多漢字文献を内容検索する際は，難字・異体字が多いため，探したい漢字の読み方が不明であったり，変換辞書に未収であったりすると，漢字を内容検索システムに入れられない場合もある．つまり，予備入力ができず，本入力もできないこととなってくる．和訓がわからず，漢字音を用いて入力すると候補が大量となってしまい，本入力の候補を探し出すまでに相当な時間や工数がかかる.

### 3.2 IDS 漢字検索の必要性

部首画数検索は紙の辞書から継承され，利用しやすい漢字検索方法である．検索効率では次の２点が特に問題となる.

(1) 同部首同画数の字数が多い場合,求める漢字を探すのは難しい.

(2) 所属する部首が分からない場合,利用できない.

部首に依存せず，より小さい漢字構造上の要素によって検索するシステムを作ることでこの二つの問題は解決できる．そのようなシステムを実現するための漢字記述の方法として,「漢字構成記述文字列（IDS)」がある．IDS とは,漢字の構成を文字列で記述したものである. IDS は IDC と漢字の部品からなる．符号化されていない漢字を表すことのできる漢字記述言語の一種である．IDS をすでに符号化した漢字に用いて,漢字の検索方法とすることもできる．このような漢字検索システムはいくつか開発されており,もっとも代表的なものは CHISE IDS-FIND である.

表 2 入力方式の比較

| 入力方式 | 予備入力 | 本入力 |
|---|---|---|
| かな入力 | かな, または「かな」にあたるローマ字 | 候補から漢字を探す. エンターキーなどを押して検索欄に入れる |
| 字形入力 | 字形のパーツにあたるアルファベット | 候補から漢字を探す. エンターキーなどを押して検索欄に入れる |
| IDS-FIND | かな入力などを用いて入力する IDS 情報 | 候補から漢字を探す. 目標漢字をコピーして, 検索欄にペーストする |

IDS-FIND は，難字・異体字の入力問題を解決したように考えられたが，漢字入力専用のツールではないため，上記の表 2 を示したように本入力の前の予備入力での利用にとどまるのでさらに工夫の余地がある.

### 3.3 IDS 画数漢字検索の必要性とその開発

CHISE の IDS-FIND はすぐれたシステムであるが，予備入力における利便性を考慮すると,次の二つの問題がある.

(1) 漢字のパーツ自体が難字・異体字である場合，IDS 漢字検索が利用できない.

(2)漢字のパーツを間違えると目標漢字が検索できない.つまり，曖昧検索が実現していない.

例えば，IDS 漢字検索を用いて「錫」(U+935A) 字を入力したい場合では，予備入力が「金昜」となる．しかし，そもそも「昜」が入力しにくい難字である.「金日」で検索すると候補の数が大量となる.選定するには効率が落ちる.

この二つの問題点を解決するため，予備入力の段階で,漢字の総画数情報を併用する.「IDS」+「残り画数」を予備入力の情報として用いることで，漢字検索の効率の向上をはかる.

上記の例の場合では,「金日」と残りの画数「５」を予備入力すれば，候補漢字を 18 字に絞ることができる. 次の図１に示す.

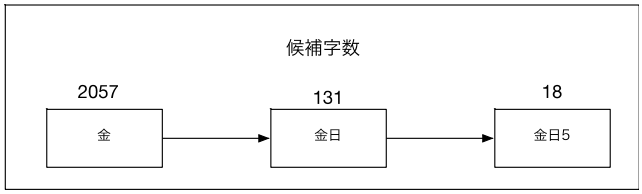


図 1　「錫」の候補字数の比較

候補字数の低減によって，目標漢字の選定速度が上がり，漢字検索の効率向上ができるようになる．

また，残り画数の幅をとるために，曖昧検索を実現できるように改善を加える．例えば，上記の残り画数「5」で検索すると，残り画数が4と6の漢字も候補欄の上下に示す．検索画面は図2に示している．

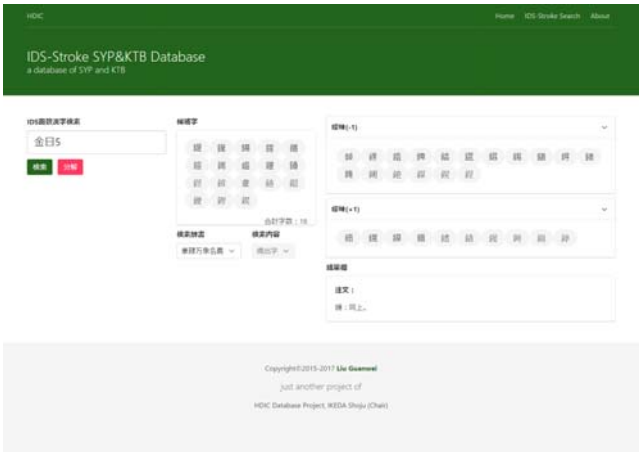


図 2　PC 端末の検索画面

利用する漢字総画数データは Unihan[a]の「kTotalStrokes」である．

## 4. マルチデバイスへの対応

### 4.1 スマートデバイスにおける問題点

多漢字文献の利用について，スマートデバイスは PC と大きく異なって，次のような違いがある．

（1） 画面のサイズが小さく，同時に表示できる情報が少ない．
（2） OS の制限で，一部の漢字が表示できない．
（3） アプリケーションとウィンドウの切り替え，テキストのコピー・ペースト操作が難しい．
（4） 画面をタッチすることによって操作する．

上記の制限を解消するため，多漢字文献をスマートデバイスで検索する際に，スマートデバイスの特徴に対応するシステムを開発することが必要となる．

しかし，PC と全く異なるシステムとするのは利用者・開発者のいずれにとっても負担となる．

a http://www.unicode.org/Public/UCD/latest/ucd/UCD.zip

### 4.2 画面サイズへの対応

レスポンシブデザインを採用して，画面の解像度によって表示の配置が自動的に変わるシステムとする．

（1）ナビメニューを隠す
（2）左右で配置された検索欄と結果欄を上下にする
（3）視認性の高い文字サイズに対応する


図 3　スマートフォンの検索画面

### 4.3 フォントシステムの補助

今回扱っている資料は Unicode を用いて公開されたテキストデータのため，Unicode をサポートする OS であればデータベースを利用することができる．格納されたデータはフォントファイルによって端末の画面に表示する．つまり，全ての Unicode 漢字が収録しているフォントがインストールされていない端末には漢字が表示できない．

この場合，代わりに文字の画像を利用して解決しようとすると，端末の解像度に応じた最善の外観で表示するのは難しい．画像となるとデータをコピーするなどの操作が不便となることも多い．本来は，クライアント側が全ての Unicode 漢字を収録したフォントをインストールすることで簡単に解決できるが，iPhone や iPad などのデバイスにおける iOS システムなどユーザが自らフォントをインストールできない OS では厄介となる．

そのため，今回のシステムはウェブフォントを利用して，クライアント側のフォントはサーバ側から提供する．利用するフォントは「花園フォント」である．「花園フォント」のサイトによると，「このフォントに含まれている文字種は ISO/IEC 10646 および Unicode 標準に収録されている 97,745

字となります」とあり，現段階の HDIC に公開されているデータベースの符号化方針を満たしている．

### 4.4 ローカルストレージの利用

ネットワークが不安定な場合があるため，漢字検索・入力に関するデータをクライアント側に保存する．

多漢字文献のインターフェイスと同一のシステムにすることで,検索画面やウィンドウの切り替えが不要である．

## 5. 開発と実装

### 5.1 フロントエンドについて

フロントエンドはユーザが直接に操作するため，主として漢字検索とマルチデバイス対応を提供する．このシステムのフロントエンドは JavaScript フレームワークの VueJS と CSS フレームワークの Bulma から構成した．全 Unicode 漢字を表示するため，花園フォントの B ファイル「HanaMinB.ttf」をウェブフォントとして利用している．少なくともローディング時間を少しでも軽減するため，TTF ファイルであったフォントファイルを WOFF へ変換した．

### 5.2 バックエンドについて

このシステムのバックエンドはユーザ管理，HDIC に所属するデータベースの検索･編集などのデータ処理を行う．PHP フレームワークの Laravel によって構成した．Laravel はドキュメントが詳細であり，非専門家でも比較的に開発しやすいフレームワークである．

4 バイト漢字を MySQL と通信するため，Laravel の /Config/database.php ファイルの MySQL に関する部分にあたって，「charset」と「collation」を修正する必要がある．

フロントエンドと通信のための API は次の表 3・表 4 のように作成した. 検索結果は JSON 形式でレスポンスする．後日の公開する予定である．

表 3　API 公開アドレス

| Scheme | Host | Path |
|---|---|---|
| https:// | hdic2.let.hokudai.ac.jp | /api/v1/search |

表 4　API のパラメータ一覧

| パラメータ | 説明 |
|---|---|
| db | 検索するデータベースである. 現在では ktb（篆隷万象名義）と syp（大広益会玉篇）のみ利用できる. |
| entry | 検索する親字である．entry か def かのいずれかが必須である. |
| def | 検索する注文である．entry か def かのいずれかが必須である. |

結果は JSON の形式でレスポンスする. 次に一例を示す.

リクエスト：/api/v1/search? db=ktb&def=北

レスポンス：[{"TBID": "5_015_A31","TB_vol_radical": "v15#223","TB_radical": " 北 ","Entry": " 北 ","Entry_type": "Regular","Entry_diff": "","TB_def": "補墨反．乖也．","SYID": "b049b041","YYID": "","TB_remarks": "¥r"}]

### 5.3 データベースについて

HDIC に所属するデータベースは RDBMS の MySQL を用いて格納されている．ただし，MySQL に拡張漢字 B 以降の 4 バイト漢字を保存する際に，文字化けになる可能性がある. これを解決するには MySQL の配置ファイル「my.ini」に相応の場所で次のような設定を追加する必要がある．

```
[mysqld]
character-set-server=utf8mb4
 [mysql]
default-character-set=utf8mb4
```

また，4 バイト漢字を検索するため，各データベースの各テーブルおよび各コラムの「Collation」を「utf8mb4_general_bin」に設定する．

## 6. あとがき

本発表ではマルチデバイス対応する多漢字文献検索システムの設計と開発に着目した．本来の目的は古辞書を翻字する際の効率的な入力方式の開発であったが，スマートデバイスの普及，オープンデータの活発化によって，さらに利便性が高い多漢字文献の公開手段が求められている．本発表はその試みを述べた．

**謝辞**　本研究は JSPS 科研費 16H03422 による成果の一部である．

## 参考文献

[1] “MySQL と寿司ビール問題”. http://blog.kamipo.net/entry/2015/03/23/093052, (参照 2017-04-10)

[2]“三生三世, Web 里字体”. http://weiwei.blog/web/web_fonts.html, （参照 2017-04-10）

[3]“Unicode® Standard Annex #38 UNICODE HAN DATABASE (UNIHAN)”. http://unicode.org/reports/tr38/, (参照 2017-04-10)

[4]池田証壽.平安時代漢字字書総合データベースの構築. 北海道大学文学研究科紀要. 2014, vol. 142, p. 79-90.

[5]池田証壽. “『大広益会玉篇』データベースの構築と利用”. 情報科学と言語研究. 加藤重広, 佐藤知己編. 現代図書, 2016, p. 65-83.

[6]池田証壽, 李媛, 申雄哲, 賈智, 斎木正直. 平安時代漢字字書のリレーションシップ. 日本語の研究. 2016, vol. 12, no. 2, p. 68-75.

[7]劉冠偉, 李媛, 池田証壽. 平安時代漢字字書総合データベースの拡張と和訓対応. 研究会報告人文科学とコンピュータ. 2015-CH-106, no. 4, p. 1-8.

[8]劉冠偉.『大字典』和訓データベース構築の現状と課題. 研究報告人文科学とコンピュータ. 2016-CH-110, no. 9, p. 1-4

[9]劉冠偉, 李媛, 池田証壽. スマホで古字書−『篆隷万象名義』の IDS 検索を例に−. 言語資源活用ワークショップ 2016 発表論文集. 2017, p. 140-147.