2008－CH－78 (7)
2008／5／23

特集セッション:『日本文化デジタル・ヒューマニティーズ』とその展開

# 古典史料データベース検索システムの提案

木村 文則 †, 小牟礼 雅之 ††, 前田 亮 †, 佐古 愛己 ‡, 杉橋 隆夫 ‡

† 立命館大学 情報理工学部 †† ユニバーサルコンピューター株式会社 ‡ 立命館大学 文学部

本論文では，古典史料のディジタルデータを基にした古典資料データベースとその検索システムを構築することにより，古典資料のディジタル図書館システムの実現を試みた．用いる古典史料として平安時代の古記録『兵範記』を使用している．専門家には研究の補助のための知識提供として「日付指定検索」など各種の検索方法を提供した．一方，一般の利用者に対しては，メタデータのデータベースを構築することにより，古典に関する理解の補助を行った．また，新たな検索手法の一つとして，現代語を用いて古典史料テキストから求める情報を探し出すための「時代横断型検索」を提案した．本論文において，上記の検索システムの検索実験を行い，本システムの有用性の評価を行った．

# Proposal of Retrieval System for Historical Documents Database

Fuminori Kimura † Masayuki Komure †† Akira Maeda † Aimi Sako ‡ Takao Sugihashi ‡

† College of Information Science and Engineering, Ritsumeikan University

†† Universal Computer Inc. ‡ College of Letters, Ritsumeikan University

This paper proposes a retrieval system for historical documents database. We have constructed a database system for the historical document "hyouhan-ki". In our system provides several retrieval method in order to support for specialists. Our system also provides metadata database system in order to support general users. Besides, we propose a new retrieval approach named "Cross-Age Information Retrieval". We conducted retrieval experiment of Cross-Age Information Retrieval in order to verify effectiveness of Cross-Age Information Retrieval System. Moreover, we conducted user assessment of our system. These results show that our retrieval system for historical documents database is very effective.

## 1 はじめに

古来より日本という地域が歴史を織り成していく中で，様々な文献や遺物などの史料が記され，残されており，歴史を紐解く鍵となっている．これらの史料は古典史料として主に，図書館や博物館などで保管されたり，研究機関で歴史研究に用いられるなどしている．

しかし，近年，それら図書館などに所蔵されている古典史料は経年劣化等による破損が進み，その保存方法を考える必要が出てきている．一般には専門の技術を持った職人による史料の修理・補修が行われることが多いが，時間と手間がかかり，なおかつ，そのような技術を持つ人が多くないことから，全ての史料をカバーすることは難しい．最近では，ディジタル技術の進歩から，史料のディジタル化による保存も方法の一つとして，大いに進められている．例として，国立国会図書館[1)]が所蔵する，明治・大正期に刊行された図書の書誌情報および目次情報，資料本文のディジタル画像などをまとめてデータベース化し，公開している近代デジタルライブラリー[1]や，歌舞伎・文楽や錦絵などを中心として，日本の伝統芸能，芸術に身近に触れる機会を提供する，教育目的のプロジェクトである日本芸術文化振興会[2)]の文化デジタルライブラリー[2]などが挙げられる．

このように，近年，史料のディジタル化による保存は大きく広まっている．しかし，現在では主に古典史料の画像データやテキストデータの保存が中心に行われているが，多くは保存のみが目的であり，そのデータの利用についてはあまり詳しく考えられていない．先に挙げた２例のシステムでは史料の画像を表示するだけであったり，説明が付いていても，著者や年代などの簡素なもののみである場合がほとんどである．

そこで本研究では，古典史料テキストのディジタル保存において，利用の点を考慮して，テキストデータ中の人名や単語の部分に説明となるメタデータを付加し，メタデータ・データベースの構築を試みた．また，求める情報をより適切に取得するための閲覧インタフェースと検索システムの作成を行った．検索については，古典史料に対する新しい検索手法として「時代横断型検索」を提案し，構築した[11) 12)]．

[1] 近代デジタルライブラリー http://kindai.ndl.go.jp/

[2] 文化デジタルライブラリー http://www2.ntj.jac.go.jp/dglib/

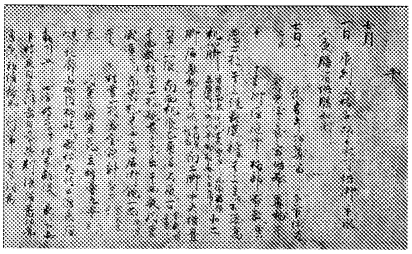

図 1 ：兵範記

## 2 古記録『兵範記』について

『兵範記』は「人車記」や「平洞記」などとも呼ばれ，平安時代後期の貴族，平信範（たいらのぶのり，1112〜87）が記した日記である．天承二年（1132）から元暦元年（1184）までの記録が伝わっており，自筆浄書本としては，現在では天承二年（1132）から承安元年（1171）までの約 40 年間分 54 巻のみが現存している．しかし，現存している巻も破損，汚損，紙の劣化等により完全な形では残っていないものがある．

平信範は朝廷の要職を長期間勤め，鳥羽・後白河院の院司 (事務職員) や摂関家の藤原忠通，藤原基実らに家司 (家政機関職員) として仕えるなど当時の政治の中枢にいた人物である．そのため『兵範記』には，政策決定に至る推移や行政文書の写し，要人の見解，朝廷・院・摂関家に関する儀式次第など，当時の朝廷や政治の様子が詳細に描かれている．特に，保元の乱 (1156) や乱後の後白河院・平家などについての記述は他の諸記録よりもはるかに詳細に描かれている．このため，歴史資料としても価値が高い史料となっている．図 1 は京都大学電子図書館 [3] で公開されている『兵範記』の自筆・古写本の第一巻「長承元年秋冬」の画像データである．また，国立歴史民俗博物館 [4] の Web サイトにおいて，『兵範記』のテキストデータを登録したデータベースが公開されている．

本研究では，『兵範記』刊本 [5] [6] を基にテキスト化し，立命館大学文学部杉橋研究室で校訂されたものを用いる．

## 3 関連研究

桶谷らは XML を利用した英日全文連携検索システムの設計と実現を行っている [9]．この研究では，文書の論理構造や属性を定義可能なマークアップ言語 XML を用い，日本の古典史料を日本語でも英語でも検索可能にすることで，日本文化の世界発信と英語圏の研究者や学生の日本史・国文学の研究に貢献することを目的としている．

『日本書紀』と『続日本紀』について日本語の本文と英訳した本文のデータを用意し，XML を用いてデータを関連付けることで，英・日どちらの言語を用いても対応する文書を探し出せる．また，注釈となる英文メモも同時に関連付けてあり，英語圏の利用者がより理解しやすくなっている．

つまり，ある言語で書かれたテキストに対して，日本語と英語のように複数の言語で検索を可能とする言語横断検索を実現するための手段として，それぞれの言語で書かれたテキストを全て用意することで実現を目指している．この研究では，言語横断検索を試みているが，本研究では，一つの言語内での時代をまたいだ時代横断検索の作成を試みている．

## 4 古典史料データベースシステム

### 4.1 システム概要

本システムでは，専門家のための研究補助となるよう不足分の知識を提供する側面と，反対に専門的な知識を持たない一般の利用者であっても書かれていることがある程度までは理解できるようになる理解補助の側面の両方を目的に構築を目指した．

専門知識を持たない人が古典史料を理解する上でまず問題になるのは，対象についての知識が足りないため，書かれていることを理解するには辞書などを用いて一つずつ語の意味を確認しなければならないことである．そこで，本文中に現れる人名や単語部分にその語の意味をメタデータとして付与した．また，これらを容易に利用できるような閲覧インタフェースを実装することで，知識を持たない一般の利用者であっても古典史料をある程度理解することができ，専門家は足りない情報をすぐに取得できるため，手間を省くことができる．本研究において構築したシステムの全体図を図 2 に示す．インタフェースを介して利用者が入力した検索キーワードを古典史料データベースに対して検索し，結果を表示させる．表示された本文中の人名部分を選択することでその人物の情報を表示させ，単語を選択することでその単語の意味などの情報を表示させるなどすることで，利用者がテキストを理解する補助とする．

### 4.2 メタデータ・データベース

日付ごとに分割された本文のテキストデータに対して，人名であればその人物の説明，単語であればその単語の意味を付加する．人名は『兵範記人名索引』[7] を用いて探し出す．『兵範記人名索引』には，各人物の「実名」，「本文中に出現する日付とその文中での表記」，「改名した場合の前後の名」，「僧であるか否か」，「校訂」，「史料の表に書かれているか裏に書かれているか」，といった情報が記されている．このうち，「実名」と「本文中に出現する日付とその文中での表記」の情報を利用する．出現する日付と表記の情報より，本文中の人物部分を発見し，発見した人物に実名と説明へのリンクを付ける．今回は検索エンジン Google で当該人物の実名で検索した結果へのリンクとした．単語は，『国語大辞典』[8] を用いて探し出す．辞書から「見出し語」の情報を抽出してリスト化し，

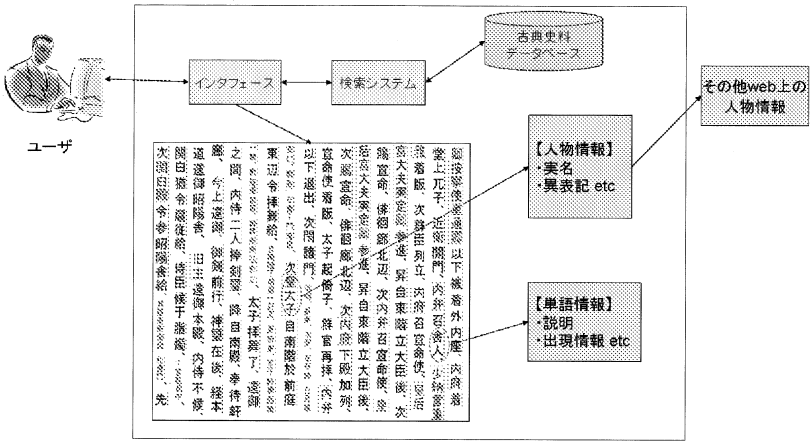


図 2 ：システム全体図

『兵範記』本文のテキストに対して，見出し語を最長一致方式で一致させて単語を探し出す．見つかった単語には単語の説明ファイルへのリンクを付加する．

## 5 閲覧インタフェース

作成した閲覧インタフェースの例を図 3 に示す．前節までに述べたように，単語の説明ファイルへのリンクを，人名には検索エンジン Google で当該人物の実名検索を行った結果へのリンクを付加している．さらに，人物についての情報の補助として，本文中の該当部分にマウスポインタを重ねることで実名が表示され，同時に同一文書中に現れる同一人物の表記をハイライト表示する機能を組み込んだ．これにより，本文中で様々な表記がなされる人物の実名や出現箇所を容易に確認することができる．

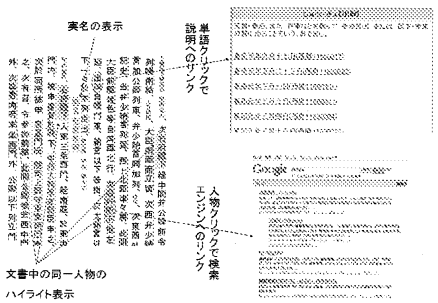


図 3 ：インタフェース例

### 5.1 検索システム

テキストデータベースから必要な情報を得るためには，検索が重要となる．しかし，単純なキーワードによる検索だけでは上手く情報を得ることができない場合がある．そこで，単純な検索以外の検索手法をいくつか提案し，実装を行った．

#### 5.1.1 範囲検索

本研究で用いるテキストは日記であり，一日ごとの日付で一文書として，データベースを構築した．そこで，検索する日付の範囲を限定する機能を付加し，検索対象を絞り込むことができる検索機能を実装した．

検索キーワードと共に，検索する文書の日付の範囲を指定して検索を行う．例えば，保延三年七月八日 (1137/07/08)〜仁平二年九月一日 (1152/09/01) までの範囲で，「寝殿」という語が含まれる文書の検索を行うと，日付指定が無い場合は 166 件，277 語が検索されるが，範囲検索を行うと 32 件，86 語に絞り込むことができる．

#### 5.1.2 KWIC 検索

KWIC(KeyWord In Context) とは，キーワードだけでなく，キーワードの前後の文脈も取り出して索引を作ることである．前後の文脈を知ることができるため，求める部分を見つけやすいのが特徴である．本システムでは，本文の表示を刊本にしたがって縦書き表示に変更し，キーワード部分は強調表示を，さらに，その部分が書かれている日付へのリンクを付加している．

#### 5.1.3 人物検索

古典史料に記載される人名は，実名の他に，時期によって変遷する通称（官職名）などで記される場合が多い．そのため，単純に実名で検索を行ったとしても，その人物を上手く見つけ出すことができない場合がある．逆に，通称

を用いて検索を行った場合，対象人物以外も検索されてしまう．そこで，『兵範記人名索引』を基に各人物の考証作業を行い，本文中の人物表記全てに対して実名のデータを付加した検索専用文書を別に作成した．これにより，実名を用いればその人物が書かれている箇所を全て見つけ出すことができる．また，人物の本文中の表記と実名をまとめたリストを作成し，入力されたキーワードを含む人物をすべて，利用者に提示する機能を付加した．また，入力されたキーワードを含む人物の本文中に出現する全表記と，それが記されている日付のデータも同時に提供することで，対象の人物を見つける補助とした．

#### 5.1.4 時代横断型検索

「時代横断型検索」は，本研究で提案する新しい情報検索の手法であり，現代語のキーワードを用いて古典史料を検索する機能である．詳細は 6 章にて述べる．

## 6 時代横断型検索

古典史料から情報を探す際に重要となるのが，「古語」である．古典史料に対する検索において古語の知識を持っていないと目的の情報を上手く取得することができない．そこで，古語の知識がなくとも，現在の言葉である「現代語」を用いて古典史料などの古い文書から適切な情報を得ることができる検索手法として，「時代横断型検索」手法を提案する．

### 6.1 現代語と古語

「現代語」とは，現在，日常的に用いられている言葉のことであり，「古語」とは，昔は使われていたが，今では一般には使われなくなった古い時代の言葉のことである．古典史料はこの古語で書かれており，現代語とは表現や言い回しなどに大きな差がある．また，昔と今とでは同じ言葉であってもその意味するところが全く違う言葉になってしまっているものも多く存在する．

例えば，「かわいい（可愛い）」という言葉は現代では専ら「愛しい」や「美しい」などといったプラスの感情を表すのに用いられている．しかし古い時代においては，「かわいい（可愛い）」は「ふびんだ」，「あわれだ」といった同情や憐憫の感情を表す語，またそれらを誘うような状態を表す語として用いられており，現代の用法とは異なっている．また，現代でいう「かわいい」に相当する語は古語では，「うつくしい」である．

このように，昔と今とでは言葉の意味や使われ方が違うため，単純に古典史料から現代の言葉で情報を探すのは難しい．

### 6.2 検索手法の概要

6.1 節で述べたように，現代語と古語には差異があるため，古典史料に対して検索を行う場合にはその差異を考慮しなければならない．しかし，そのような差異を考慮して検索することは，古語やその時代の知識が無い一般の人では難しい．そのような困難を乗り越え，現代の言葉を入力しても古い文書から適切な情報を取得できる検索手法として提案するのが「時代横断型検索」である．図 4 に本手法の概要図を示す．

本手法では，現代語辞書と古語辞書の二つを用いる．まず，利用者が入力した検索キーワードを基にして，現代語辞書の見出し語と完全一致するものを探し出す．見出し語が見つかれば，次に，その語の説明文の文書ベクトルを取り出す．続いて，古語辞書のすべての見出し語の説明文に対して同様に文書ベクトルを取り出し，現代語説明文のベクトルと古語説明文のベクトルを基に文章間類似度を計算していく．そして，最も類似度が高かった古語説明文の見出し語を，入力された検索キーワードに対応する古語と決定する．最後に，決定した古語を用いて，古典史料データベースに対して検索を行い，結果を出力する．

## 7 現代語と古語の対応実験

### 7.1 実験概要

時代横断型検索において核となる現代語説明文と古語説明文の類似度による一致について，どれだけ正しい語と結び付けられているか正解率を調査した．今回は，現代語辞書として『広辞苑』[10] を用い，古語辞書として，現代語だけでなく古語の情報も収録されている『国語大辞典』[8] を用いた．

まず，それぞれの辞書から各単語の見出し語とその説明文のデータを抽出する．今回は，漢文の形式で書かれている『兵範記』に合わせて，見出しが漢字のみで構成されている単語のデータを用いた．抽出した説明文には，その語の用例や対義語，特殊記号文字，品詞情報なども含まれているが，提案手法においてはノイズとなるため，それらを削除した．上記の情報を削除することにより，正解率が約 6%向上する（表 1）．

また，複数の意味を持つ語は説明がいくつかの項目に分かれているため，それらを項目ごとに分割した．抽出数は広辞苑からは見出し語 147,384 語，抽出単語 190,016 語，国語大辞典からは見出し語 174,429 語，抽出単語 242,093 語となった．この抽出データを基にして，現代語辞書の単語と古語辞書の単語の類似度を調べた．

まず，二つの辞書両方に対して，各単語の見出し語と説明文を合わせたテキストを形態素解析器 ChaSen [3] を用いて解析し，テキスト中に出現している単語を抽出する．次に，抽出した単語の重み付けを行う．重み付けは各見出し語ごとに行うため，別の見出し語中に出現する場合は同じ単語であっても重みは異なる．こうして得られた各単語の重みを用いて，各見出し語の単語ベクトルを構成する．最後に，現代語辞書の各説明文と古語辞書の各説明文の類似度をコサイン距離で計算する．コサイン距離は，二つのベクトルの内積を両ベクトルの大きさで割

[3] ChaSen's Wiki http://chasen.naist.jp/hiki/ChaSen/

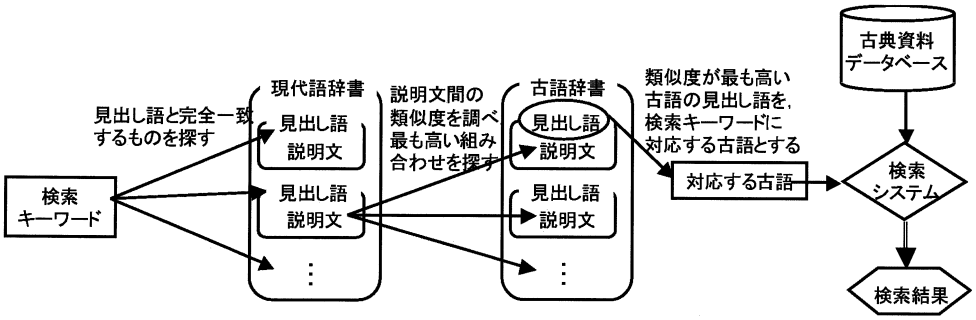


図 4 ：時代横断型検索の手法

ることにより求まる．

例えば広辞苑のある単語 A に対して，国語大辞典の全単語を対象にコサイン距離を計算し，一番高い組み合わせとなる単語が，単語 A に対応する古語として決定されることとなる．

### 7.2 抽出データの重み付け

7.1 節で述べた手法により，現代語と古語の対応付けがどの程度正確に行えるのか実験を行った．本実験では，説明文のベクトルの要素の重み付けに「出現頻度（TF）」，「IDF」，「TF-IDF」，「正規化した TF-IDF」の 4 種類を用いた場合について実験を行った．

出現頻度 TF（term frequency）は，ある一つの単語の説明文において注目している単語 $t$ が出現する回数である．IDF（inverse document frequency）は，出現する説明文が少ないほどその単語の重みを大きく評価する指標である．IDF が高い単語はその単語が出現する説明文を絞り込むことができため，特定性が高いことを示している．IDF は以下の式により求められる．

$$IDF(d,t) = \log \frac{N}{DF(t)} + 1$$

$N$ は抽出単語数，$DF$ は，ある一つの単語の説明文 $d$ において，説明文中にでてくる単語 $t$ が抽出単語数 $N$ 個のうち何個に出現したかを示した数である．

TF-IDF は，出現頻度（TF）と IDF の積である．TF と IDF を掛けることにより，その単語が頻出し，かつ説明文の特定性が高い単語の重みを高く評価することができる．ただし，説明文が短い語よりも長い語の方が重みが高く評価される傾向がある．そのため，単語数で割ることにより正規化を行うこともある．本実験では正規化を行っていない TF-IDF と正規化を行った場合の両方について実験を行った．

### 7.3 実験結果

それぞれの実験結果をまとめたものが表 1 である．調査単語はランダムに選び，対応させた現代語と古語が正解かどうかは，対応させた見出し語および説明文を見比べて，同じ意味の語と結び付けられているかを，著者の一人が確認し，判断した．

表 1 ：実験結果

| 実験名 | 調査単語数 | 正解数 | 正解率（%） |
|---|---|---|---|
| 除去前 | 114 | 75 | 65.7895 |
| TF | 114 | 81 | 71.0526 |
| IDF | 114 | 82 | 71.9298 |
| TF-IDF | 114 | 82 | 71.9298 |
| 正規化 | 114 | 81 | 71.0526 |

4 種類の重み付けのいずれにおいても約 71%の正解率となった．説明文中のノイズを排除する前に対して，6 %程度精度が向上した．ただし，4 種類の重み付けの間では，正解率の差はあまりみられなかった．ある重み付けでは正しく一致していた単語が，別の重み付けでは間違っていることもあり，どの重み付けが良いとは断定できない結果となった．

### 7.4 検索実験とその考察

前節では，ある現代語に対する古語の対応付けを行う際に，最も類似度の高い古語のみをその現代語の訳語としていたが，類似度が 2 番目以降の古語が正解となる可能性もある．そういう可能性の有無について調査するために，「本」「石」という 2 例の現代語の具体例の場合について追加の検索実験を行った．まず，書や刊行物を意味する「本」という単語で検索を行った．

表 2 より，「本」に対応する最も類似度の高い語は「巻」であると分かる．意味的にも同じ意味であると判断できる．類似度が 2 番目 3 番目の項目は「本」の意味の一つである，「もとで」と「もとからあるもの」に対応する語を選んでいれば良いのだが，固有名詞が選ばれており，正しいとは言えない．

ここで，それぞれの古語で本システムにおいて『兵範記』に対して検索を実行した結果，得られた検索結果の件数を表 2 に示す．件数は検索された日付の数，ワードは検索された語の総数を示す．またそれとは別に，元々の入力語である「本」，すなわち現代語のままで検索を実行した結果も表 2 に示す．

表 2 ：現代語「本」に対する古語への翻訳実験

| 検索語 | 類似度 | 検索件数 | 検索語数 |
|---|---|---|---|
| 巻 | 0.9983 | 270 | 581 |
| 円満院 | 0.9968 | 0 | 0 |
| 華厳五十五所絵巻 | 0.9728 | 0 | 0 |
| 本 | 0.9020 | 514 | 1316 |

「円満院」と「華厳五十五所絵巻」は固有名詞であることから，1 件も検索されなかった．そもそも，目的の意味とは何も関係が無い．「巻」と「本」の単純な検索数で比べたところ，「本」の検索数の方が上となっている．しかし，全てが適合しているとは限らないため，実際に検索された文書が適切であるか確認する必要がある．

「巻」について検索された日付の本文を確認したところ，「五十巻」，「上巻」など，ほぼすべてが探す目標の「書物」の意味を指していることが確認できた．ただし，いくつかは「巻く」などの意味で使われている部分が検索されていた．「本」について同様に調べたところ，「本数」や「本家」など書物とは関係のない部分ばかりが検索され，目標である「書物」として使われている部分は，調べた限りでは発見できなかった．これは，本検索手法が有効に動作した好例と言える．

次に「石」について「本」の場合と同様の検索実験を行った．その結果を表 3 に示す．上位 3 件は「青石」，「赤玉・赤珠・明珠」，「石取」となっているが，その意味を確認しても，現代語の「石」の訳語としてはあまり良い結果であると考えられない．その下，4，5 番目の方が類似度は劣るものの求める意味としては正しいものを選べていると思われる．

実際に本システムで検索した結果を見ると，上位 3 語では 1 つも検索されず，「石」で検索を行った方が検索数が多かった．また，「石」の本文中での意味を確認したところ，「いし」としての意味では全く見つからず，全て「米などを数える“こく”」として使われている部分であった．なお，この「数える単位」の意味と一致した項目は類似度は 0.69 から 0.86 と低い値であった．

表 3 ：現代語「石」に対する古語への翻訳実験

| 検索語 | 類似度 | 検索件数 | 検索語数 |
|---|---|---|---|
| 青石 | 0.9942 | 0 | 0 |
| 赤玉 or 赤珠 or 明珠 | 0.9929 | 0 | 0 |
| 石取 | 0.9911 | 0 | 0 |
| 石 | 0.9897 | 206 | 458 |
| 箴 | 0.9293 | 0 | 0 |

この結果から，単純に類似度が高い語を検索キーワードとして用いても上手くいかないことが見て取れる．他の語についてもあまり意味的に一致しない語が上位にくることが多くあった．そういった語を比べたところ，それらの語が上位に来てしまう原因として，類似度計算に用いている現代語説明文が短い場合が多いことがわかった．説明文が短く，ほとんど一単語程度しか類似度計算に用いることができないため，その一単語のみでは一致が上手くいかなくなっていると考えられる．

また，求める意味が上位に来ない場合もあり，それを自動で選ぶことはまだ難しい．しかし，対応する古語の候補をある程度の件数選択することを許せば，適切な古語を得ることは可能である．以上のことを勘案すると，システム側は対応する古語の候補を提示し，その中から利用者自身に古語を選んでもらうことが，現状では現実的な方法であると考えられる．

## 8 利用者評価

日本人の大学生 16 名 (古語知識を持たない情報理工学部生 (以後「一般人」と表記) : 8 人 , 古語知識を持つ文学部生 (以後「専門家」と表記) : 8 人) を対象にシステム全体の利用者評価を行った．結果を表 4 に示す．評価は 1 から 5 の 5 段階で，1 が最も低く，5 が最も高い評価である．なお，平均の項目は小数点以下 2 桁目を四捨五入している．また，無回答の項目を含む回答者がいたため，合計が 16 名になっていない項目も存在する．

結果より，一般人の評価では，時代横断型検索の項目が高い評価を得ており，時代横断型検索が古典資料に対する検索手法として有用であることを示唆している．また，人物の情報量やインタフェースの使いやすさなどの項目も高くなっており，知識の補助という面ではある程度の結果がでている．逆に，低い評価となったものとして，人物の探しやすさや本文の意味の理解が挙げられる．特に本文の意味の理解の項目が低いということは，目的の一つである「知識を持たない一般人の理解補助」が達成できていないこととなる．これは，単語や人物など一つ一つの情報を提供しても，それらを連携するための知識が乏しく，一般の人は理解しづらいのではないかと考えられる．解決法としては，本文の現代語訳を提供することなどが挙げられる．

次に，専門家の評価を見ると，一般人と比較して時代横断型検索の評価は低くなっている．これは，すでに古語の知識を持っているので，意味が正しい語と結び付けられていないなどのシステムの不十分な点がよく分かる

表 4 ：利用者評価

| 評価項目 | | 一般人評点 (人数) 低 1 | 2 | 3 | 4 | 5 高 | 専門家評点 (人数) 低 1 | 2 | 3 | 4 | 5 高 | 平均 一般 | 専門家 | 総合 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (全体的な) インタフェースの使いやすさ | | | | 3 | 5 | | | 2 | 3 | 3 | | 3.6 | 3.1 | 3.4 |
| 範囲検索の使いやすさ | | | | 3 | 4 | 1 | | 4 | | 4 | | 3.8 | 3.0 | 3.4 |
| KWIC 検索の使いやすさ | | | 2 | 6 | | | | 1 | 3 | 2 | 2 | 2.8 | 3.6 | 3.2 |
| 人物検索 | 人物検索の使いやすさ | | 1 | 4 | 2 | 1 | | | 1 | 5 | 2 | 3.4 | 4.1 | 3.8 |
| | 対象の人物の探しやすさ | 1 | 3 | 1 | 2 | 1 | | 1 | 6 | 1 | | 2.9 | 3.0 | 2.9 |
| | 人物の情報量 | | | 3 | 2 | 3 | | 3 | 2 | 3 | | 4.0 | 3.0 | 3.5 |
| 時代横断型検索 | 時代横断型検索の使いやすさ | | 1 | 2 | 3 | 2 | | 2 | 3 | 3 | | 3.8 | 3.1 | 3.4 |
| | 現代語に対する古語の一致度 | | | 2 | 4 | 2 | | 2 | 2 | 2 | 1 | 4.0 | 3.3 | 3.7 |
| | この検索の実用性 | | 1 | 1 | 3 | 3 | | 1 | 4 | 1 | 2 | 4.8 | 3.5 | 3.8 |
| 必要な情報をどれだけ得ることができたか | | | 1 | 3 | 2 | 1 | | 1 | 1 | 4 | 1 | 3.4 | 3.7 | 3.6 |
| 本文の意味の理解にかかる手間 | | | 4 | 2 | | | | 1 | 1 | 5 | | 2.3 | 3.6 | 3.0 |
| 検索結果の満足度 | | | 1 | 4 | 2 | | | | 4 | 4 | | 3.1 | 3.5 | 3.3 |
| 平均 | | | | | | | | | | | | 3.5 | 3.4 | 3.4 |

ためだと考えられる．時代横断型検索システムが今のところは専門家の知識に及んでいないことも示唆しており，検索精度の改善に取り組む必要がある．

また，一般人の評価よりも高い評価を得られた項目に，本文の意味理解にかかる手間がある．これもまた，自身の持つ古語知識が下地にあるため，辞書を引く手間を軽減し，単語等の情報を提供するだけで理解ができるようになったためだと思われる．これ以外に特徴的な点として，KWIC 検索の使いやすさの項目が高くなっている．専門的な立場からすると，求める文書を探すには，出現箇所の一覧があったほうが見つけやすいとの考え方によるものと思われる．

総合で見てみると，各検索手法については現時点でも一定の評価が得られていると思われる．しかし，双方から低い評価を得た項目として，求める人物の探しやすさなど，インタフェースの扱いやすさに関連するものが挙げられる．感想には，インタフェースに対する肯定的な意見もあるが，全体では不満な点が目立つようである．

表 5 は，本システムに対する利用者の意見である．処理が早くて良いといった評価点もあるが，人物に関しての情報不足や人物・単語の発見のしづらさなどといった要望，不満点が多く出た．また，一般人からは本文理解がしづらく，現代訳がほしいという意見が見られ，専門家からは，インタフェース面での不満が多かったように感じられる．評点はある程度の値がでているが，意見・感想を見る限り，まだまだ改善すべき点は多い．

表 5 ：提案システムに対する利用者の意見

| | 一般人 | 専門家 |
|---|---|---|
| 長所 | 検索所要時間が短くて良い | 人名比較が早くて良い |
| | 語句の意味がわかりやすい | 全文が表示されるのが良い |
| | 結果が見やすい | 辞書を引く手間が省けるので楽 |
| 短所 | 人物検索で提供されたリストから目的の人物を見つけづらい | 表示が右寄りになったりするので，中央揃えにするなど見やすくしてほしい |
| | 検索した文字が文書のどこにあるかを知りたい | システム自体の使い方が分かりづらかった |
| | 検索語が人名中の語，単語中の語関係なしに検索されるので分けてほしい | 二人以上を指す文字 (「両人」など) で一人しか実名・ハイライト表示されない |
| 要望 | 本文理解には時間がかかった | 検索した文字がどこにあるか分かりづらい |
| | 人物の読み方があるといい | |
| | 現代語訳がほしい | |

## 9 考察

利用者評価の結果を見るに，やはり，一般人と専門家とでは求める情報，機能に違いがあり，一般人は現代語訳などの直接的な情報を，専門家は個々の人物や単語，地名などの詳細な情報を求めていて，両方を満たすシステムの構築の難しさが再確認された．しかし，インタフェースについて不満はでているものの，構築した検索手法についてはおおむね良好で，特に提案した「時代横断型検索」の実用性については，双方から期待がもたれていることが伺える．

続いて，評価実験の結果から，時代横断型検索は，現状では実際の検索に用いることができるほどの精度を実現できていないこと，また，単純な正規化では精度向上

は見込めないことがわかった．あまり高い精度が得られなかった原因として以下のようなことが考えられる．

1. 類似度計算の指標に用いている値が，単語の出現回数を基にした単純な値であるTF，IDF，TF-IDF値のため，出現回数の増えやすい長い説明文を持つ単語が対応する単語として選ばれやすくなる．

2. ノイズ等を除去した結果，説明文が簡潔になりすぎて情報量がほとんどない語も存在する．

3. 日本語の書き方の違い，表記ゆれによる差のため，類似度計算の際に同じ語として認識されないものがある．(「あはれ」と「哀れ」や「土ぼこり」と「土ほこり」など)

4. この時代横断型検索は，「ある単語の説明文はどの辞書においても大抵似た表記で解説されている」ことを前提としており，類似度は辞書の表記に依存してしまう．辞書の説明文の書き方に大きな差があれば，対応させるべき語との類似度は低くなり，全く違う語が選ばれてしまう．

また，本研究では古語辞書として『国語大辞典』を用いているが，「古語としての情報も載っている」だけであり，古語辞書として適当とは言えない．このことも要因になっていると考えられる．

1. に関しては，単語の重みとしてTF-IDF値が用いるのに適当であるかどうかという問題であり，対処としてエントロピーやベイズの定理の利用などが考えられる．2. は何語以下の説明文は用いないなど，しきい値を設定して制限する方法が考えられる．また，複数の辞書を用いることにより，単語数を増やし，十分な統計情報が得られるようにすることも考えられる．3. は各語の対応辞書を作成することで対処でき，4. に関しては，正式な古語辞書を用いることで対処できるのではないかと思われる．

## 10 おわりに

本論文では，利用を考慮した古典史料のディジタル図書館システムとして，説明情報を付加した古典史料のテキストデータベースと閲覧インタフェースおよび検索システムの構築を行い，新たな検索手法として「時代横断型検索」を提案した．また，「時代横断型検索」における現代語と古語の類似度による一致について正解率調査の実験を，実際の使用に関して検索実験を，それぞれ行った．最後に，構築したシステム全体に対して利用者評価を行い，古典史料のディジタルデータがどれだけ扱いやすくなっているかを調査した．

利用者評価より，システム全体では，ある程度の評価が得られたと言える．しかし，インタフェース面での不親切さについての不満も出ており，システム自体の使いやすさを向上させる必要がある．目的の一つである古語知識のない一般人の本文理解については，現状では十分ではないと言える．現代語訳データを載せるなどの方法を考える必要がある．

新しい検索手法として提案した「時代横断型検索」は実用に耐えうるだけの精度をまだ実現できていない．しかし，改良の余地は多く残されている．また，一般人による利用者評価では好意的に受け止められており，時代横断型検索システムに対する期待も伺える．

時代横断型検索の検索精度向上のための今後の課題として，類似度計算に用いる指標を変更することや，本研究で使用している辞書が正式な古語辞書ではなかったため，正式な辞書のデータを用いることなどが挙げられる．また，今回は『兵範記』に合わせるため，見出し語が漢字のみで構成された単語だけを用いていた．今後，『兵範記』以外の古典史料，ひらがな等が含まれる単語も対象にしていく必要がある．その際には各語の活用形などの詳細な部分を考慮することなどを検討する必要がある．

## 参考文献


1) 国立国会図書館, http://www.ndl.go.jp
2) 日本芸術文化振興会, http://www.ntj.jac.go.jp
3) 京都大学付属図書館．京都大学電子図書館，http://edb.kulib.kyoto-u.ac.jp/minds.html
4) 国立歴史民俗博物館. http://www.rekihaku.ac.jp
5) 株式会社 臨川書店内 増補「史料大成」刊行会, “増補 史料大成 兵範記１・２・３・４.” 第７刷, 臨川書店, 1998.1998.1998.1998, 350p.294p.423p.373p.
6) 株式会社 臨川書店内 増補「史料大成」刊行会, “増補 史料大成 兵範記５、江記、平知信朝臣記.”, 第７刷, 臨川書店, 1997, p.1-237
7) 立命館大学人文学会兵範記輪読会編,『兵範記人名索引』, 思文閣出版, 2007.
8) 杉原正利, DVD-ROM 版 スーパー・ニッポニカ 2002 日本大百科全書＋国語大辞典, 小学館出版, 2002 (DVD-ROM)
9) 桶谷猪久夫, Delmer Brown, 藤本雅彦, 大久保祐子, “XML を利用した英日全文連携検索システムの設計と実現.” 情報処理学会研究報告 2004-CH-64(6), p.39-46, 2004
10) 新村出 編, スーパー統合辞書 99 広辞苑第五版, 富士通, 1999 (CD-ROM)
11) 小牟礼雅之, 前田亮, 佐古愛己, 杉橋隆夫, “古記録データベースの閲覧インタフェースおよび検索手法の提案.” 人文科学とコンピュータシンポジウム論文集（じんもんこん 2007）, p.283-288, 2007
12) 小牟礼雅之, 前田亮, “古典史料テキストの時代横断型検索手法の提案.” 第 70 回情報処理学会全国大会講演論文集, 第 4 分冊（1ZJ-3）, p.835-836, 2008