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

小牟礼 雅之
立命館大学
理工学研究科

前田 亮
立命館大学
情報理工学部

佐古 愛己
立命館大学
文学部

杉橋 隆夫
立命館大学
文学部

近年，古記録や古文書などの古典史料は，経年劣化などへの対策としてディジタル化による保存が試みられている．現在は，画像による保存やテキストの単純な保存が主となっており，その内容にまで踏み込んで保存を行っているものは少ない．本研究では，内容を重視した保存方法として，古記録のテキストデータ内の単語や人物に，説明となるメタデータを自動で付加した上でデータベースを構築し，それを取り出すための閲覧インタフェースおよび検索システムを構築した．また，検索手法の一つとして，現代語と古語の二つの辞書を利用し，対応するキーワードを見つけ出して検索を行う「時代横断型検索」を提案する．予備実験の結果から，現代語による古記録の検索において一定の効果を見込めることがわかった．

An approach to Browsing Interface and Retrieval Methods for Historical Documents

Masayuki Komure
Graduate School of Science and Engineering, Ritsumeikan University

Akira Maeda
College of Information Science and Engineering, Ritsumeikan University

Aimi Sako
College of Letters
Ritsumeikan University

Takao Sugihashi
College of Letters
Ritsumeikan University

Old documents in paper are deteriorating by time. Most of the existing digital archive system for such old documents stores only image and/or text. This paper proposes a digital archive system that exploits the contents. Specifically, we add metadata of words and person names that appears in the text content. For utilizing the metadata, we devised a browsing interface and retrieval techniques. One of the new approaches to retrieval is named “Cross-Age Information Retrieval” which uses two dictionaries, i.e. for modern language and for ancient language, to find an archaic word equivalent of a modern word. Preliminary experiments show promising results for retrieving old documents using modern language query.

1. はじめに

図書館等に保管されている古文書や古記録などの古典史料は近年，経年劣化などへの対策としてディジタル化による保存が試みられている．しかし，画像による保存やテキストの単純な保存が主であり，その内容にまで踏み込んで保存を行っているところは少ない．本研究では，内容を重視した保存方法として，古記録のテキストデータ内の各人物や単語に，説明となるメタデータを自動で付加してデータベースを構築することで，必要な情報を簡単に取り出し，理解することができる閲覧インタフェースおよび検索手法を考案した．

なお，データベースを構築する対象として，古記録の『兵範記』[1]の本文をテキストデータ化したものを用いる[2]．データベースの構築には全文検索システム OpenText[1]を用いている．

2．古記録『兵範記』

『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（1171）までの記録が伝わり，自筆浄書本 54 巻が現存している．平信範は朝廷の要職を長期間勤め，鳥羽・後白河院の院司，また摂関家累代の家司（家政機関

[1] http://www.infocom.co.jp/das/open/

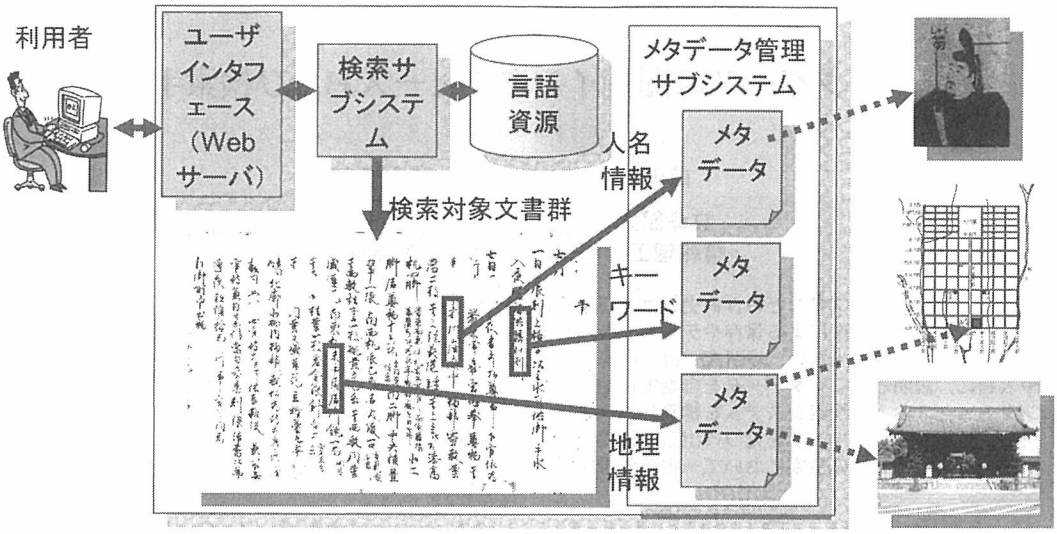

図１：システムの概要

職員）としても活躍した人物である．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．そのため，歴史資料としての価値が高い．

『兵範記』は日記であるため，内容は日付ごとに分けて書かれている．そこで，日付ごとに一文書として，データベースの構築を行った．

3. システム概要

本システムでは，専門家だけではなく専門的な知識を持たない一般の利用者の便にも考慮する．専門知識を持たない人たちが古典史料を閲覧する際に一番問題となることは，対象に関する知識が足りないため，書かれていることを理解するのに辞書などを用いる必要があることだと考える．それに伴い，理解に時間が掛かってしまう．そこで，本文中に現れる人物や単語の意味をメタデータとして付与することとした．また，これらを容易に利用できる閲覧インタフェースを実装することで一般利用者でも古文書の理解力を専門家に近づけられ，専門家は自身の知識に足りない部分を埋めて研究などにおけるロスを無くすことができるであろう．

図１にシステムの概要を示す．ユーザがキーワードを用いて検索をかけると，それに応じて検索結果を出力する．その結果，表示された文書中の人物や単語，地名などのテキスト部分からその人物や単語の情報を提供する．

4. 閲覧インタフェース

本文のテキストデータに付加する情報は，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といったものを付加することにする．

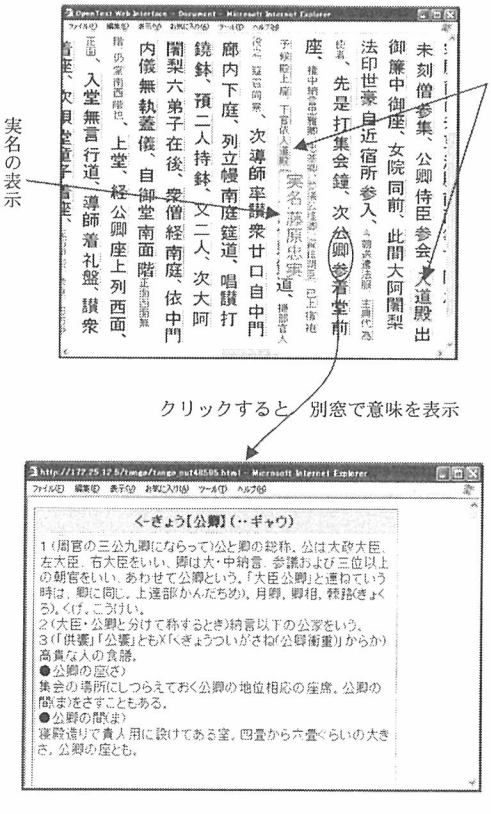

図２：閲覧インタフェースの例

ただし，各人物の情報については，検索エンジンの Google で当該人物の実名で検索した結果へのリンクとした．また，単語については「国語

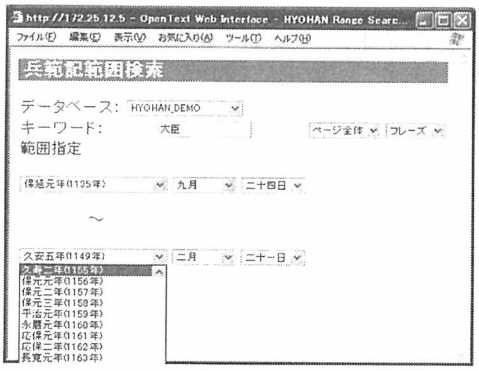

図 3：範囲検索

大辞典」[3]から単語ごとにファイルを作り出し，そのファイルへのリンクを張ることにした．
文章から人名を探し出すため，「兵範記人名索引」[4]を，単語に関しては上述の「国語大辞典」を用いてテキスト中の該当部分をそれぞれ見つけ出し，リンクを自動で付加する．また，人物については，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，さら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される人物の実名が容易に確認できる．図 2 に閲覧インタフェースの一例を挙げる．

5. 検索システム

文書データベースから必要な情報を正しく得るためには，情報を探すための検索部分が重要である．単純なキーワードによる検索だけでは上手く情報を見つけることができない場合も多い．そこで，本研究ではいくつかの検索手法の実装を行った．

5.1 範囲検索

対象文書が日記であり，一日ごとのデータとして分割してデータベースの構築を行っているため，狙った日付のみを調べた方が対象文書を見つけやすいのではないか，と考え，検索する日付の範囲を絞り込む機能を実装した．（図 3）

5.2 KWIC 検索

キーワードだけでなく，周囲の文脈を同時に表示させることで探している部分を見つけやすくする機能を実装した．（図 4）

5.3 人物検索

古記録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，単純に実名のみを入力して検索をかけようとしても上手く見つけ

図 4：KWIC 検索

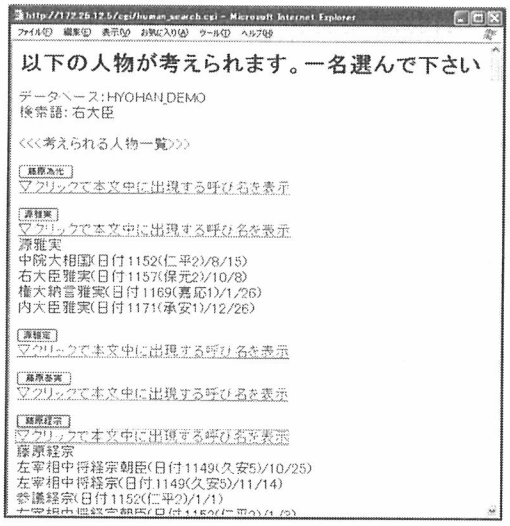

図 5：人物検索

ることができない．そこで，本文中に通称で記載された人物の考証作業を行い，人物を示す表記全てに実名データを付加した検索専用の文書を別に作成した．

さらに，このままでは，人物の実名を知っている上でそれを正確にキーワードとしなければ見つけることができない．そこで，人物の本文中の表記と実名をまとめたリストを作成しておき，入力されたキーワードをリストに含む人物を全員利用者に提示する．その上で，探したい人物のみを検索にかけることで，探している人物が書かれている文書を見つけ出す機能を実装した．また，各人物の本文中の全表記と出現す

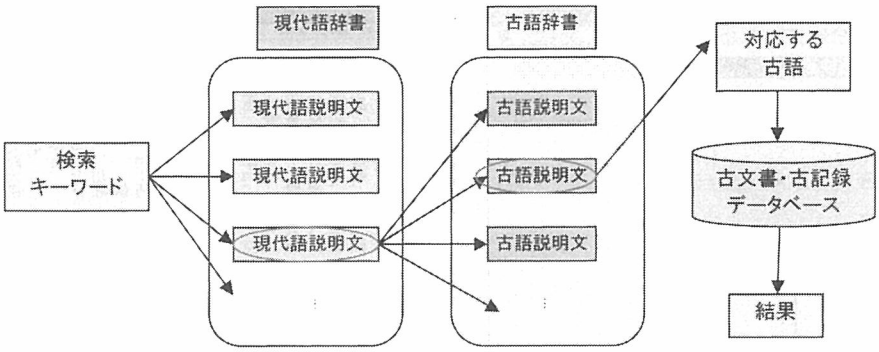

図 6：時代横断型検索のモデル

る日付のデータも同時に提供することで対象人物を見つける補助とした．

なお，本文中に現れる人物の特定には文献[4]を用いた．（図 5）

5.4 時代横断型検索

「時代横断型検索」とは，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によって古文書を検索する機能である．詳細は 6 章にて述べる．

5.5 検索システムのまとめ

上記のような様々な検索機能を用意することにより，目的の文書を発見する補助となる．また，これらの検索機能を統合することで，より検索精度を向上できると考えられる．

6. 時代横断型検索

古文書・古記録などの古典史料から必要な情報を探す際に重要となるのが，「古語」である．現在では使われていない言い回しや表現が用いられているために，検索（特に全文検索）においてこれらの古語に対する知識が必要となる．

そこで，現代の言葉を用いても古い文書から適切な情報を得られるような検索として，「時代横断型検索」の手法を提案する．本手法における検索の流れを図 6 に示す．

図のように，まず利用者が入力した検索キーワードを現代語の辞書と照らし合わせて一致するものを探し出す．今回は，キーワードと完全一致する見出し語を探しだすこととする．

次に，一致した見出し語の説明文を元にして古語辞書のそれぞれの説明文との類似度を調べる．

説明文間の類似度が一番高かった古語の見出し語を検索キーワードに対応する古語として取り出し，それを用いてデータベースに検索をかけ，結果を取得する．

この流れを，例を挙げて説明する．例えば，入力された検索キーワードが「かわいい」という単語であった場合を考える．現代語において，「かわいい」という単語は以下のような意味を持つ．

かわいい(可愛い)
愛すべきである。深い愛情を感じる。 「―・い我が子」「―・い声で歌う」 小さくて美しい。「―・いスズランの花」

この説明文の意味と一致する古語を探すと次のような語になる．

うつくしい(美しい・愛しい)
主として肉親に対する愛情を表わし、現代の「いとしい」「かわいい」などの意に使われ、「めぐし」「かな(愛)し」などに近い。 1 かわいい。いとしい。愛らしい。 2 様子が、いかにもかわいらしい。愛らしく美しい。可憐である。 3（美一般を表わし、自然物などにもいう。美麗である。きれいだ。みごとである。立派だ。

つまり，現代語の「かわいい（可愛い）」は古語において「うつくしい（美しい・愛しい）」と意味が近い語となる．

また，古語における「かわいい」は以下の意味を持っている．

かわいい(可愛い)
あわれで、人の同情をさそうようなさまである。かわいそうだ。ふびんだ。

このように意味合いとしては現代語よりもネガティブな用途で扱われる語となっている．

このような現代語と古語の差異を考慮し，現代語を元にして古語で検索を行うシステムを時代横断型検索と名付け，実装を試みた．

7. 実験

時代横断型検索を実装するために，要となる現代語－古語間の類似度による一致について実験を行った．今回は，現代語辞書として「広辞苑」[5]を用い，古語辞書として古語情報も収録されている「国語大辞典」を用いた．

まず，それぞれの辞書から単語の見出しとその説明文のデータを抜き出す．今回は漢文の形式で書かれている『兵範記』に対応するため，見出しが漢字のみで構成されている単語のデータを抽出した．抽出数は以下の通りである．

表 1：辞書からの抽出単語数

	抽出単語数
広辞苑	147,384 語
国語大辞典	174,429 語

これらの抽出した各単語とその説明文を基にして，広辞苑の単語と国語大辞典の単語間の類似度を求める．

各単語の説明文を形態素解析し，その結果から TF-IDF 値を求める．TF-IDF 値を元にして説明文間の類似度をコサイン距離で求める．

説明文の形態素解析には ChaSen を用い，コサイン距離の計算には以下の式を用いる．

$$\cos(x, y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \bullet \sum_{i=1}^{N} y_i^2}}$$

なお、x は広辞苑の説明文中の形態素解析された単語ごとの TF-IDF 値が格納されており，y には国語大辞典の説明文中の形態素解析された単語ごとの TF-IDF 値が同様に格納されている．また，N は x に格納されている値の総数である．

こうして求めたコサイン距離の値が一番高い単語を現代語と対応する古語とみなして検索キーワードとして扱うこととする．

7.1 実験 1

この実験では抽出したデータをそのまま用いて類似度を計算する．説明文をそのまま形態素解析し，値を調べる．

7.2 実験 2

説明文中から単語の用例の部分を削除し，一つの単語に複数の意味がある場合，それらを項目ごとに分けてから処理を行い，値を調べる．

なお、意味ごとに分けたため，全体数が以下のように変化した．

表 2：意味ごとに分けた単語数

	単語数
広辞苑	190,016 語
国語大辞典	242,093 語

7.3 実験 3

実験 2 に追加して，得られた TF-IDF 値を正規化する．正規化された値は TF-IDF 値に（その単語の TF 値/ その説明文の全単語数）を掛けたものとする．

7.4 実験結果

実験の結果は以下のようになった．

表 3：実験結果

	正しい単語と結び付けられた率（%）
実験 1	0
実験 2	65
実験 3	35

それぞれ，広辞苑の単語 114 個分に対して正しいと思われる単語と一致している数を数えた結果である．ここでいう正しいと思われる単語とは，現代語と古語それぞれの説明文を読み，意味が同じものであるかどうかを自分自身で判断したものである．

8. 考察と問題点

実験の結果から用例などの無駄な部分が含まれていると上手く一致させることができないことがわかった．また，単純な正規化では精度を向上させることができないこともわかった．

実際に検索に用いるとして，現状では使えるだけの精度を実現できていない．この原因の一つとして TF-IDF 値の特徴がある．TF-IDF は全体での出現回数が少なく，ある文書における出現数が多い場合に高くなる．これは，文書内に多く出現した単語が重要であるとの考え方に

基づくものだが，出現数が少なくとも重要な単語である場合もある．指標に用いる値を変えることで違った結果が得られるのではないかと考えられる．

9. 閲覧インタフェースの利用者評価

日本語が読める人を対象に 20 人にシステム閲覧インタフェースの利用者評価を行った．その結果を表 4 に示す．

結果より，どの項目においても評価が上がっていた．また，この項目だけでなくシステム適用後の本文を読んだ感想は，「古文の知識や読めない漢字があっても単語帳を見ることで少しは意味が分かった」，「自分で辞書を引かなくていいので楽だった」，「同一人物がすぐ分かるのが良い」などのプラス評価が多かった．

10. おわりに

時代横断型検索はまだ現実に使えるものではないが，精度を向上させる余地は多いと思われる．今後，システムへの実装と評価実験を行い，精度の向上を試みたいと考えている．また，現代語－古語間以外にも適用させることができる可能性も考えられる．

参考文献

[1] 京都大学文学部国史研究室，京都大学資料叢書兵範記一，二，三，思文閣出版，1988.

[2] 前田 亮, 佐古 愛己, 杉橋 隆夫. 京都学ディジタル図書館の構築と多言語情報アクセス. 人文科学とコンピュータシンポジウム論文集, pp. 195-202, 2003.

[3] 小学館　スーパー・ニッポニカ　日本大百科＋国語大辞典，小学館出版，2002.（CD-ROM）

[4] 立命館大学文学部人文学会 兵範記輪読会編，兵範記人名索引，思文閣出版，2007

[5] 広辞苑　スーパー統合辞書 99, 富士通，（CD-ROM）

表 4：閲覧インタフェースの利用者評価の結果

評価項目	システム適用前					システム適用後				
	1	2	3	4	5	1	2	3	4	5
本文の古文らしさ	6	11	3	0	0	0	0	4	14	2
本文に登場する人物がわかる	7	12	1	0	0	0	0	0	0	20
本文に登場する呼び名の違う同一人物が発見できる	18	2	0	0	0	0	0	0	0	20
大まかな本文の意味がわかる	15	5	0	0	0	0	1	15	3	1
一行（一部分）でも意味がわかる	13	5	1	0	0	0	0	8	10	2
得られる情報量	16	3	1	0	0	0	0	6	12	2