古典史料テキストの時代横断型検索手法の提案

小牟礼雅之†　　前田亮‡

立命館大学　理工学研究科†　　立命館大学　情報理工学部‡

1．はじめに

近年，図書館等に保管されている古典史料の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，それらは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，古典史料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と，その情報を得るための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*を用いている．

2．『兵範記』

『兵範記』は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．「人車記」や「平洞記」などとも呼ばれる．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．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．古典史料のディジタル図書館システム

本研究では，古典史料などに対する知識を持った研究者だけでなく，知識を持たない一般の人々にも扱うことができ，文書の内容を理解することができるデータベースの構築を目指した．

知識を持たない一般の人が古典史料を理解するために必要となるのは，書かれている単語の意味や出現する人物，地名，建物などについての情報であると考えられる．そこで，それらの情報のメタデータを該当する部分に付加し，参照することで理解の補助となるようインタフェースを構築した．図1にその例を示す．

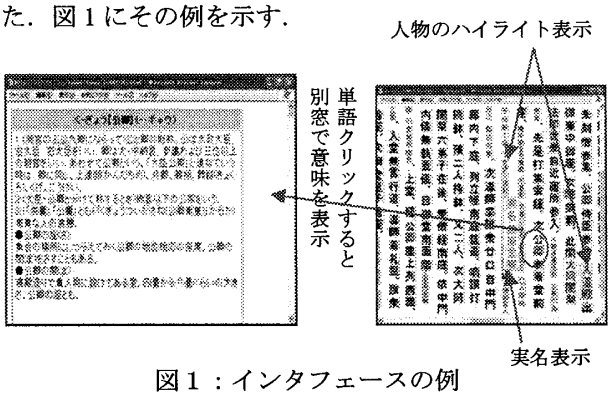

図1：インタフェースの例

また，必要な情報を得る手段として，いくつかの検索手法を実装した．

3．1．範囲検索

登録してある日記の日付を絞り込んで検索をする．

3．2．KWIC検索

検索でヒットした本文の周囲の文字列を同時に表示させる．

3．3．人物検索

テキスト中に出現する人物は実名だけでなく，官職名（右大臣，大納言 etc）など通称で書かれていることが多く，単純な検索では見つけづらい．そこで，本文中に出現する人物の表記のリストを作り，利用者に提示することで，対象の人物を見つけやすくする．

例えば，「大臣」と入力すると，「右大臣」や「太政大臣」などの表記がされている人物をその実名とともに一覧表示し，その中から目当ての人物で検索を行うことにより，その人物が記されている本文を全て見つけ出すことができる．

4．時代横断型検索

古典史料から必要な情報を探す際には「古語」

An Approach to Cross-Age Information Retrieval of Historical Documents

†Komure Masayuki, Graduate School of Science and Engineering, Ritsumeikan University

‡Akira Maeda, College of Information Science and Engineering, Ritsumeikan University

*http://www.infocom.co.jp/das/open/

が重要となる．古典史料のテキストは現在では使われていない言葉で書かれていたり，現代語とは違う意味で扱われていたりする語が存在する．そのため，検索を行い，正しく情報を得るためには利用者が古語の知識を持っていることが必要となってくる．

本研究では，古語の知識を持たない人でも求める情報を見つけることができる方法として，「時代横断型検索」を提案する．検索の流れを図 2 に示す．この手法では，現代語の辞書と古語の辞書から各見出し語とその説明文を用いる．

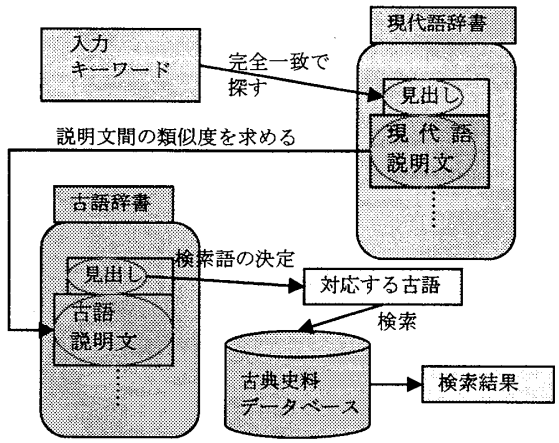

図 2：時代横断型検索のモデル図

まず，利用者が入力したキーワードと現代語辞書の見出し語とを完全一致で調べる．該当する語が見つかるとその説明文を取り出す．次に，その説明文と古語辞書のすべての語の説明文との類似度を調べ，最も類似度が高い説明文を見つけ出す．その説明文の見出し語を入力されたキーワードに対応する古語として決定し，その語を検索キーワードとして検索を行う．

なお，本研究では，類似度を求める手段として，説明文を形態素解析器 ChaSen を用いて形態素解析して TF-IDF 値を求め，その値を基に説明文のベクトルのコサイン距離を求めることで説明文間の類似度を計算している．

5．予備実験

時代横断型検索の中心部分である現代語説明文と古語説明文間の類似度を用いた現代語の古語への変換について実験を行った．

今回は現代語辞書として「広辞苑」[3]を，古語辞書として「国語大辞典」[4]を用いる．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．さらに，説明文を意味の項目ごとに分割した．見出し語数と抽出単語数は表 1 の通りである．

表 1：抽出結果

	見出し語数	抽出単語数
広辞苑	147,384 語	190,016 語
国語大辞典	174,429 語	242,093 語

こうして抽出したデータを基に類似度を調べ，決定された一番類似度の高い組み合わせが，現代語と古語で同じ意味を持っているもの同士を選べているかどうかを調べた．今回は広辞苑の単語 114 語分を調べた．結果は表 2 の通りとなった．なお、正しいかどうかは自身で確認を行った．

表 2：実験結果

単語数	正解数	正解率(%)
114 語	74 語	65

6．考察

実験の結果，現代語の古語への変換の精度は実際に用いるには低い値となった．原因としていくつかが考えられる．例えば，類似度を調べる際に用いた指標が単語の出現回数を基にした単純な値である TF-IDF 値のため，長い文や出現率の高い単語を含む説明文が対応するものとして選ばれやすいことと，説明文中に語の説明としてはあまり重要でない部分（対義語や品詞情報，読みなど）が含まれており，類似度を調べるのにノイズとなることである．この 2 つを改善することで精度の向上が見込めると思われる．

7．おわりに

本研究では，古典史料のテキストに対する新たな検索手法として，単語の説明文間の類似度を用いて，検索キーワードを現代語から古語へ変換する検索を提案した．

今後の課題として，類似度による単語の一致の精度向上が挙げられる．

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] 広辞苑　スーパー統合辞書 99，富士通，1999，（CD-ROM）

[4] 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）