4S－4

歴史系文献資料の索引作成作業支援のための知識型分析ツール†

高橋謙†† 藤田茂†† 菅原研次†† 八重樫純樹†††

††千葉工業大学情報工学科 †††国立歴史民俗博物館

1. はじめに

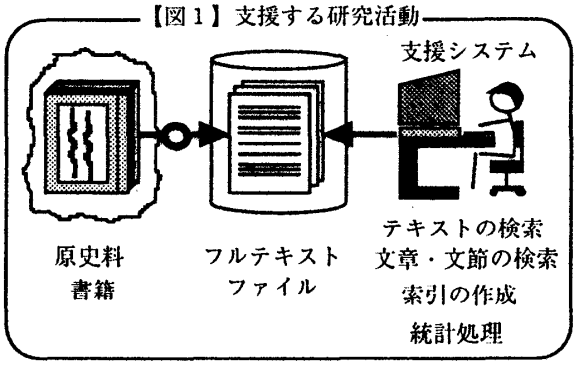
古文書などの研究を支援する作業の一つに、コンピュータに取り込まれたテキスト形式資料から索引情報を抽出することが挙げられている。しかし、これまでの索引作成では、個々の利用者の利用要求を必ずしも満足させるものではなく、それぞれの利用者の利用方法に対応可能な索引作成とその利用環境を与えることが必要である。

本稿では、テキストに含まれる必要な知識へのアクセスの実現のために、研究者の対象のドメインに対する概念や、各研究者ごとに異なる固有の知識に対する概念に基づいた索引付けが有用であるという前提の基に、それを実現する高度な索引の作成と利用の支援について述べる（図1参照）。

2. 索引作成支援ツールのモデリング

テキストファイルに変換済みの古文書についての索引作成作業の支援を目的とした知識型分析ツールの開発を行う。その際に意味解析の手段として、自然言語処理系の技術は使わずに、研究者個人の領域知識を予め分析したものを用いるというアプローチをとる。

【図1】支援する研究活動



本発表ではその実験対象領域として、日本書紀第５巻の原文のテキスト（プレーンテキスト）を対象とした実験を試みた（図2参照）。

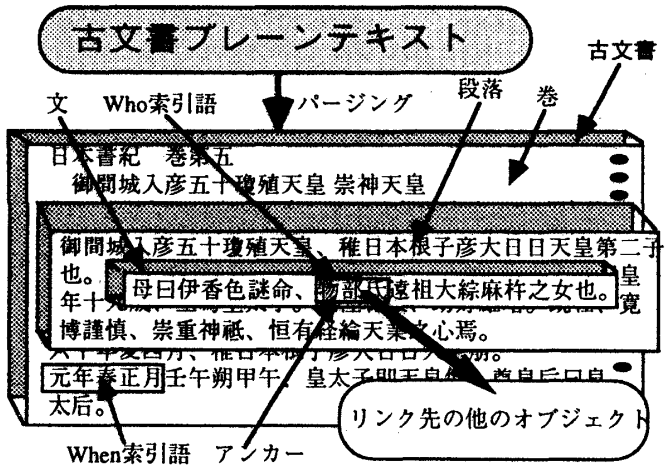
3. 高度な索引の定義とその効用

歴史系文献資料は自然言語としての固有の文書構造を持つ。その性質に着目して検索に利用する索引が文書構造型索引（巻・章、段落、節、文など）である。また、歴史系の研究に頻繁に利用される知識は、5W1H（いつ・何処で・誰が・何を・どうして・どの様に）型の質問で得られるものが多い。その性質に着目した索引が5W1H型索引である。

文書構造型索引と5W1H型索引をフレーム型の知識として定義した。これらの性質に基づいて構造化されたテキストは多岐にわたる観点からの検索手段を持ち、各索引フレームどうしがリンクされたハイパーテキストを構成する（図2・3参照）。

そして、利用者が個々の知識に基づいてこのハイパーテキストのリンクを張ることにより、従来の方式に比べて、個人利用者がより高度な利用上の便宜をはかれるような索引を形成することができる。

【図2】古文書資料の高度索引によるハイパーテキスト化の概念図



4. 索引作成支援の手法（技術面について）

計算機支援において、自動化できる部分と対話形式で分析しなければならない部分がある。以下に、それぞれについてまとめる。

4.1. 自動化過程

(1)各種索引構造の切り出し

プレーンテキストから、文書構造型索引と5W1H型索引を定義されたルールに従ってパージングする。図4にパーザで使うパージングルールの定義を示す。

†Knowledge-based Analizing Tools for Indexing of Historical Document
††Ken TAKAHASHI ††Shigeru FUJITA
††Kenji SUGAWARA †††Junki YAEGASHI
†† Chiba Institute of Technology
††† National Museum of Japanese History

【図3】古文書資料の索引フレーム（ハイパーテキストオブジェクト）の定義

<古文書>=<<古文書ID>、<古文書題目>、
<古文書属性>、<古文書関係>、<古文書実体>、
<アンカー>>
<古文書実体>= {<巻>}
<(巻～段落)索引語>=<<索引語ID>、
<索引語番号>、<索引語題目>、<索引語属性>、
<索引語関係>、<索引語実体>、<アンカー>>
<索引語実体>=
{<（巻～段落）索引語>のサブクラス}
<文>=<<文ID>、<文番号>、<文属性>、
<文関係>、<文テキスト>、<アンカー>>
<(5W1H)索引語>=<<索引語ID>、<索引語属性>、
<索引語関係>、<索引語意味情報>、
<索引語文字列>、<アンカー>>
<アンカー>=<<アンカーID>、
<アンカー文字列>、<アンカー属性>、
<リンクオブジェクトID>、<リンク属性>>
<リンクオブジェクトID>=
<<(文書構造型)索引語ID>or<(5W1H)索引語ID>>

(2)用語系索引の同定

対象文献資料のドメイン分析から索引の候補用語のリスト（索引用語リスト）を用意する。この索引用語リストをプレーンテキストと照合し、索引語（フレーム）やその属性値（スロット）の抽出を行う（キーワード辞書照合方式）。またその際、指定した不要語を索引用語リストからの除去（不要語除去）、計数処理の結果利用、シソーラスの利用なども行う。

4.2. 対話過程

(1)必要な索引フレームの再定義の支援

パージングされた索引フレームの修正を行う（修正用エディタ）。

(2)リンク支援

いくつかの索引フレームの多次元イメージの合成とその射像による分析（関係演算型）、出現頻度や出現分布などの統計処理結果の分析（計数処理型）、情報フィルタ（利用過程をプロフィールとして保存して活用する）の利用などを行う。

4.3 計数処理結果の分析過程

分析系の前処理として索引用語に対する計数処理を行う。その結果を索引に採用する属性値を抽出する際の目安とする。

(1)登録用語の出現頻度分析

文書構造型索引の各フレーム単位に索引用語リストに載っている用語の出現数を計測した出現頻度表を出力する処理。文書構造型索引の各フレームごとのキーワードの抽出などに利用する。

(2)指定語の出現分布分析

文書構造型索引の各フレームごとに指定語（索引用語リストから一語を指定）の出現箇所と出現数を計測した出現分布表を出力する処理。指定語をキーワードとする文書構造型索引のフレームの検索などに利用する。

【図4】パージングルールの定義

(EntityModule-subclass)
<KanEntity> テキストファイル単位
(AbstractModule-subclass)
(TextStractureModule-subclass)
<KanModule>：<KanEntity>フレーム単位
<ShouModule>：連続した「改行」の検出点区間
<DanrakuModule>：「改行」の検出点区間
<BunModule>：「句点」の検出点区間
<SetuModule>：「読点」の検出点区間
(5W1HIndexModule-subclass)
<WhenModule> ,<WhereModule> ,<WhoModule>
,<WhatModule> ,<WhyModule> ,<HowModule>
：5W1H用語リストに登録してある語の各検出点単位

(3)複数指定語の相関

索引用語リストから二語を指定し、出現箇所を抽出し、二語間の距離を計測した指定語相関表を出力する処理。リンクプロセスでの知識に利用する。

5. インプリメント

開発言語はObjectworks/Smalltalkを使用した。計算機環境は、SPARC /IPX/ELC、Apple Macintoshなど、Objectworks/Smalltalkの動作する機種である。

6. おわりに

本研究では、特定のドメインのテキストの索引付けに有効なヒューリスティクスの表現形式の提案とその妥当性の検証を行うための実験を行っている。研究者ごとにその知識が異なるため索引付け過程は各個人の領域知識に基づくべきであると考え、この知識の獲得を利用者の検索事例の観測・分析により試みた。この様な、利用者の領域知識に基づくハイパーテキストの構造定義と索引付けを行うことにより、従来の方式に比べて個人利用者にとってより高度な利用を促進する文献資料検索システムの実現が期待できる。

【参考文献】

[1]菅原研次、伊與田光宏、八重樫純樹、「歴史系テキストデータとリバースエンジニアリング」、国立歴史民俗博物館研究報告、第53集、1993
[2]高橋謙、菅原研次、八重樫純樹、「歴史系文献資料管理のための知的データベース」、電子情報通信学会秋期大会講演論文集（情報・システム）、D-56(p59)、1994
[3]安永尚志、「日本古典文学の本文データベース」、情報処理、Vol.35、No.7、p642~650、1994
[4]小島憲之、他、校注・訳「新編日本古典文学全集2、日本書記1」、小学館、1994