京都学ディジタル図書館の構築

―京都学デジタル図書館プロジェクト―

前田 亮

Abstract: Abstract: In this project, we research for the techniques to realize the high information access aims mainly at the text information.
Concretely, for the ancient writings and resources, we provide the functions such as for searching them in contemporary languages by analyzing the meanings not just in character string matching, but in whole document, word and character units.

In this regard, the techniques of XML, metadata, and the Semantic Web are essential.
Since there contains many characters unused in contemporary character code among these ancient writings and resources, we research for the techniques for searching the writings which include these "external characters" effectively.

On the other hand, even if it was a research on Kyoto, we need to consider the connections not only inside the domestic, but also among the East Asian neighboring countries.
Therefore, it is considered that the provided contents are not necessarily only in Japanese.
Also, research results of COE program should be accessible easily not only to Japanese but also to many people from all over the world.
For this realization, we research for the cross-language information retrieval technique which enables us to search without the preparation of the translation of these contents.

1. はじめに

近年, ディジタル図書館やディジタルアーカイブが注目され, さまざまな文化的資料のディジタル化や保存に関する研究が盛んに行われている. しかしながら, それらのコンテンツに対して容易で効率的なアクセス手段を提供するという観点からの研究はまだ多くはない. コンテンツの量が膨大になればなるほど高度なアクセス手段が要求されることは, 現在のWebの状況を見ても明らかである. 本研究プロジェクトでは, 主に京都に関する古文書／古記録の文字情報を対象として, 高度な情報アクセスを実現する手法について研究を行っている[1][2].

具体的には, 平安時代の貴族の日記である『兵範記』を例として, 単なる文字列マッチングではなく, 文書全体あるいは単語単位, さらには文字単位で意味を解析することにより, 現代語によって検索する機能や, 現在の文字コードに含まれない文字を含む文書を検索する機能を実現する. また, 文中に様々な表現で現れる人名･地名･建造物名などの自動抽出, および人名索引や古地図との対応付けを行う.

さらに, 本研究プロジェクトの成果を広く世界に向けて発信するために, コンテンツの翻訳版を用意することなく検索を可能とする言語横断情報検索技術について研究を行っている[3].

2. 『兵範記』について

兵範記（「へいはんき」もしくは「ひょうはんき」と読む）は, 平安時代後期の長承二年（1132）から元暦元年（1184）までの間, 公卿の平信範が記した日記であり, 54巻が現存する. 『人車記』『平洞記』『北隣記』などとも呼ばれる. 平信範は, 朝廷実務の要職である蔵人･弁官を長期間勤め, 鳥

羽·後白河院の院司，また摂関家累代の家司（家政機関職員）としても活動した人物である．彼の中級貴族·実務官僚という立場に基づき，院政期の行政，たとえば政策決定にいたる推移や行政文書の写し，要人の見解などの情報と，公家有職，たとえば朝廷·院·摂関家に関する儀式次第などに関する精確·詳細な記述が見られる[4].

本研究では，この『兵範記』の刊本（活字本）[5]をもとに，テキスト化の作業を進めている．

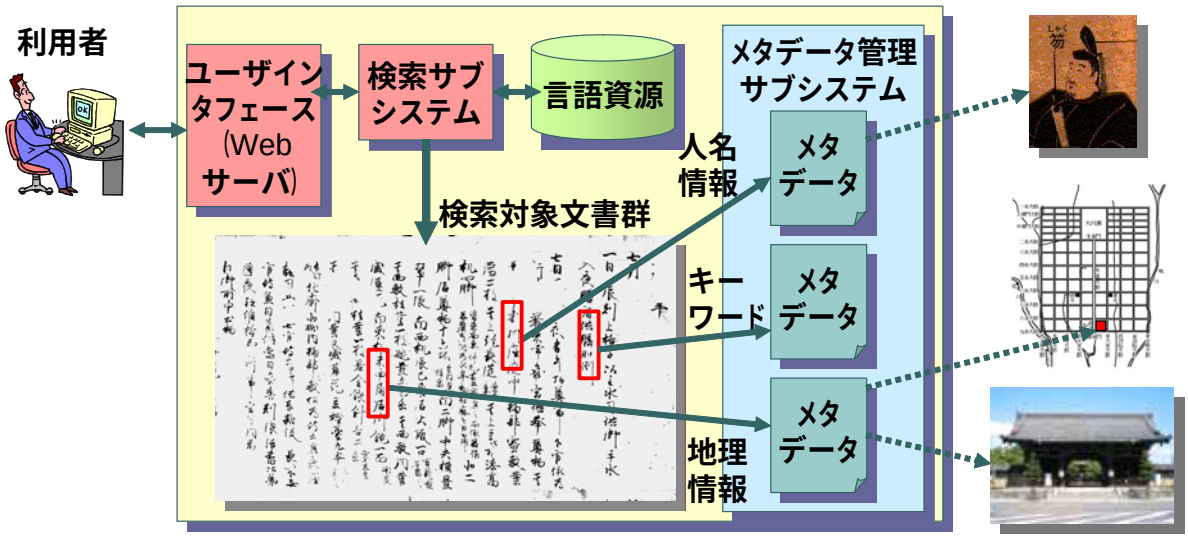

図 1：京都学ディジタル図書館システムの概要

3. 京都学ディジタル図書館システム

京都学ディジタル図書館システムの概要を図1に示す．利用者は，Webブラウザからシステムにアクセスする．検索サブシステムでは，利用者が入力した問合せ（検索語）を，辞書などの言語資源を用いて解析し，検索対象文書群に対して検索を行う．

検索対象文書群からは，あらかじめ人名，地理情報，キーワードなどのメタデータを抽出しておく．メタデータの抽出の詳細は次節で説明する．これらのメタデータからは，人物情報，地図，建造物の写真などへのリンクが張られており，利用者は検索結果の本文から直接関連情報を参照することができる．

4. 古文書／古記録の検索

本研究の主要な目的の一つは，古文書／古記録の文字情報を対象として，高度な情報アクセスを実現する手法の確立である．具体的には，古文書／古記録に対して，単なる文字列マッチングではなく，文書全体あるいは単語単位，さらには文字単位で意味を解析することにより，たとえば現代語による検索や，人名や地名，建造物名などを自動抽出し関連情報へリンクする機能などを提供することを目指している．この実現のために，XML，メタデータ，セマンティックWebなどの技術を用いる．また，古文書／古記録には現在の文字コードに含まれない文字が多く含まれるが，これら「外字」を含む文書に対して効率的な検索を行う手法についても検討している．現在のところ，古文書／古記録および外字を含む文書に対する概念検索の技術について基礎的な検討を行っている段階である．

4.1. メタデータ

メタデータとは，「データに関する（構造化された）データ」のことであり，たとえば図書における書誌情報（タイトル，著者，主題，出版社，出版年など）に相当する．

電子化された情報資源のためのメタデータ規格として，Dublin Core（http://dublincore.org/）があり，基本的な15のメタデータ要素（タイトル，著者，キーワード，内容説明，発行者，寄与者，日付，種別，フォーマット，識別子，出典，言語，関係，時空間範囲，権利）が定義されている．Dublin Coreはメタデータの事実上の業界標準となっており，これに従ってメタデータ記述しておくことで，システム間での相互運用性が確保される．

『兵範記』のメタデータとしては，日付ごとに一つの情報資源として扱い，キーワードとして，人名，地名，建造物名，重要語などを付与することを考えている．

4.2. 固有表現の抽出

固有表現とは，人名，地名，建造物名，日付などを指す．文書中に現れる固有表現は，情報検索において非常に重要なキーワードとなる．

今回対象とする『兵範記』に関しては，人名については「立命館文学」において兵範記の人名索引が出版されており[6]，すでにテキスト化が行われている．これは表形式のデータベースとして格納されており，本文中に現れた人名に対して，それが表す人物の本名，出現した日付，その他付加情報などが記載されている．兵範記に現れる人物の数は膨大であり，また人名は実名で書かれることはほとんどなく，様々な表記で記述される．現段階では，本文中に現れた約3,600名について，約2万件の出現のデータを抽出している．

また，地名・建造物名については，現段階で約270の建造物について，読み・分類・現在地名などの付加情報が記載されている．

本研究では，これらを基に，本文中に様々な表記で現れる人名・地名・建造物名に対して，そのメタデータや地図上の位置へのリンクを自動的に付与する手法を検討している．

4.3. 古文書／古記録の概念検索

古文書／古記録を現代語で検索するためには，文書中に現れる単語の意味を知る必要があるが，これを現在の自然言語処理技術で自動的に行うことは難しい．しかしながら，通常の情報検索においても，文書あるいは単語の意味をシステムが理解した上で検索しているわけではなく，語義の曖昧性を残したままで検索を行っているのが現状である．情報検索では質問に対する完全な答えを求める必要はなく，関連すると思われる文書あるいは文書中の部分を返すものであるため，曖昧性を必ずしも完全に解消する必要はない．

本研究では，古文書／古記録の概念検索への第一歩として，大規模な国語辞典などの既存の辞書を用いて，すべての文字あるいは単語について可能性のある語義をすべて索引に登録し，これと質問との文字列マッチングを行うことで，関連する可能性のある文書中の部分を検索結果とすることを考えている．また，単語の共起傾向を調べることで，古文書／古記録における語義の曖昧性を解消することが可能かどうか検討を行っている．

4.4. 言語の壁を越える検索

前節までは兵範記を例に古文書／古記録のディジタル化について述べたが，一般に京都に関するコンテンツの研究では，国内だけではなく東アジアなどの近隣諸国との関係も重要である．そのため，研究対象となる資料が書かれている言語が必ずしも日本語だけとは限らない．また，研究成果を公開する際には，日本だけではなく世界のより多くの人々が容易にアクセスできる手段を提供するべきである．この実現のために，コンテンツの翻訳版を用意することなく検索を可能とする言語横断情報検索技術に関する研究を行っている．

5. 断簡の復元

兵範記は，50余年の期間のうち29年分しか現存せず，その中でも「断簡」（何らかの事情で本来つながっていた日記の一部が切断され，ばらばらになったもの：図2参照）が存在する．テキスト

化を進めるにあたって, この断簡を復元することは重要であるが, 手がかりの少ない多数の文書の断片から復元するのは非常に困難な作業である. また, 史学研究を進める上で, 日記間の前後関係や年代を特定することは非常に重要である. そこで, 本研究では, 個々の断簡における文字の大きさや癖, 紙の質感などをデータ化し, これを断簡復元の際のヒントとして用いることを検討している. つまり, これらのパラメータを特徴量とし, その類似度の高いものが, おそらく同じ文書中の断簡であろうという推定を行う.

このような推定は, 最終的には人手で確認する必要があり, コンピュータによる処理はあくまでもそれに対する補助的なものであるが, 膨大な数の断簡を対応付ける現実的な手段として有効であると考えている.

6. おわりに

本研究は, 古文書／古記録の現代語による検索と, 言語間を跨って検索する言語横断検索とを, 概念検索という共通の枠組みを用いて融合することで, 京都学コンテンツに対して時代や言語の壁を越えた検索を可能にすることを目指している.

現在ディジタル化が進められている, 京都に関わる過去から現在の膨大かつ多様な文化･アートの集積にアクセスする手段を提供するには, 単なる文字列マッチングではなく, より高度な検索技術が必要不可欠である. 本研究の成果によって, 世界中のより多くの人々が, 京都に関わる過去から現在に至る膨大な知識の集積に容易にアクセスすることが可能になる.

参考文献

[1] 前田亮. 京都学デジタル図書館プロジェクト — 京都学コンテンツに対する情報アクセスの研究. 21世紀COEプログラム･シンポジウム「PCクラスタとアート･エンタテインメント研究」予稿集, p. 97, 2003.

[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] Akira Maeda. Multilingual Information Processing for Digital Libraries. In *Proceedings of the PNC Annual Conference and Joint Meetings 2002*, pp. 80-81 (abstract), Osaka, Japan, 2002.

図 2:兵範記の断簡の例

[4] 杉橋隆夫.「『人車記』とその周辺」, 陽明叢書記録文書篇第5輯『月報』13, 1986

[5] 『増補史料大成　兵範記一』, 臨川書店 (1965)

[6] 兵範記輪読会編：「兵範記人名索引」Ⅰ～Ⅲ,
『立命館文学』別巻, 1987, 1991, 1999