京都学ディジタル図書館の構築と多言語情報アクセス

前田 亮 *　佐古 愛己 **　杉橋 隆夫 ***

*立命館大学 理工学部情報学科

**立命館大学 COE 推進機構

***立命館大学 文学部史学科

本稿では，我々が現在進めている京都学ディジタル図書館の構築について紹介し，それに対する多言語情報アクセスを実現するための手法について述べる．ここでは，平安時代の貴族の日記である「兵範記」を採り上げ，この刊本からのテキスト化，メタデータの付与，および概念検索の実現手法についての構想を述べる．さらに，これらのテキストに対して他言語からの情報アクセスを実現する，言語横断情報検索技術の適用の可能性について考察する．本研究では，古文書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．

Building a Digital Library of Kyoto Studies and its Multilingual Information Access

Akira Maeda *　Aimi Sako **　Takao Sugihashi ***

*Department of Computer Science, College of Science and Engineering, Ritsumeikan University
**The Organization for the Promotion of the COE Program, Ritsumeikan University
***Department of History, College of Letters, Ritsumeikan University

This paper introduces a digital library of Kyoto studies and discusses issues in providing multilingual information access to this digital library. As an example, we pick up "Heihanki (or Hyouhanki)" which is a diary written by an aristocrat during the late Heian era (1132-1184), and describe our plans to digitize its typefaced paper edition, to attach metadata, and to realize conceptual information retrieval method. Moreover, we consider the possibility of applying cross-language information retrieval method to the digitized text, which will realize information access to the text using other language such as English. This research project aims at providing information access to the contents of Kyoto studies across both ages and languages, by uniting the retrieval of paleography using modern language and cross-language information retrieval, based on a common framework of conceptual information retrieval.

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．

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

図 1：「兵範記」の一部

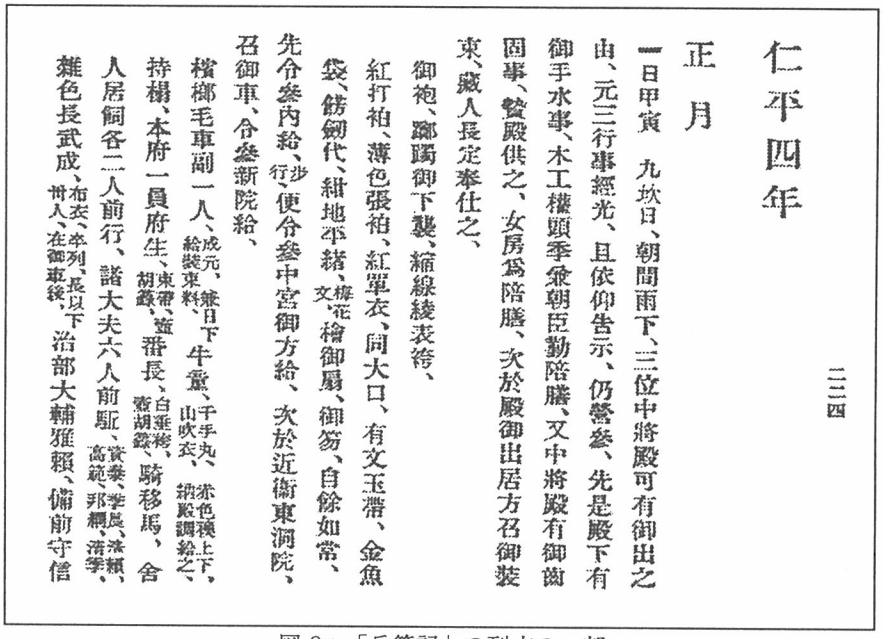
り，現代語によって検索する機能や，現在の文字コードに含まれない文字を含む文書を検索する機能を実現する．また，文中に様々な表現で現れる人名・地名・建造物名などの自動抽出，および人名索引や古地図との対応付けを行う．

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．

本稿では，現在基礎的な検討の段階にある本研究プロジェクトの構想について述べ，古文書の検索と言語横断情報検索の実現に関わる問題点およびその解決の見通しについて考察する．

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（1171）までの間，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[1]．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．



仁平四年　二三四

正月

一日甲寅　九坎日、朝間雨下、三位中將殿可有御出之由、元三行事經光、且依仰告示、仍營參、先是殿下有御手水事、木工權頭季兼朝臣勤陪膳、又中將殿有御齒固事、贊殿供之、女房爲陪膳、次於殿御出居方召御裝束、藏人長定奉仕之、

御袍、躑躅御下襲、縮線綾表袴、

紅打袷、薄色張袷、紅單衣、同大口、有文玉帶、金魚袋、飾劍代、紺地平緒、梅花文、檜御扇、御笏、自餘如常、

先令參內給、歩行、便令參中宮御方給、次於近衛東洞院、召御車、令參新院給、

檳榔毛車副一人、成元、兼日下給裝束料、牛童、千手丸、赤色襖上下、山吹衣、納殿調給之、

持楊、本府一員府生、束帶、壺胡簶、番長、白垂袴、壺胡簶、騎移馬、舍人居飼各二人前行、諸大夫六人前駈、資泰、季長、清頼、高範、邦綱、清季、

雜色長武成、布衣、卒列、長以下卅人、在御車後、治部大輔雅頼、備前守信

図２：「兵範記」の刊本の一部

仁平四年

正月

一日甲寅九坎日、朝間雨下、三位中將殿可有御出之由、元三行事経光、且依仰告示、仍営参、先是殿下有御手水事、木工権頭季兼朝臣勤陪膳、又中將殿有御歯固事、贊殿供之、女房為陪膳、次於殿御出居方召御装束、蔵人長定奉仕之、

御袍、躑躅御下襲、縮線綾表袴、

紅打袷、薄色張袷、紅単衣、同大口、有文玉帯、金魚袋、飾剣代、紺地平緒、〔梅花文〕、檜御扇、御笏、自余如常、

先令参内給、〔歩行〕、便令参中宮御方給、次於近衛東洞院、召御車、令参新院給、

檳榔毛車副一人、〔成元、兼日下給装束料〕、牛童、〔千手丸、赤色襖上下、山吹衣、納殿調給之〕、持楊、本府一員府生、〔束帯、壺胡簶〕、番長、〔白垂袴、壺胡簶〕、騎移馬、舍人居飼各二人前行、諸大夫六人前駈、〔資泰、季長、清頼、高範、邦綱、清季〕、雑色長武成、〔布衣、卒列、長以下卅人、在御車後〕、治部大輔雅頼、備前守信

図３：テキスト化した刊本の一部

本研究では，この「兵範記」の刊本（活字本）[3]をもとに，テキスト化の作業を進めている．

3. 刊本のテキスト化

図 1 に兵範記の影印本[2]の一部を，図 2 に該当部分の刊本の抜粋を示す．すでに一部テキスト化を進めているが，すべて人手で行っている．図 3 に刊本のテキストを示す．残された刊本についてすべて人手でテキスト化を行うのには限界があるため，今後は OCR を用いてテキスト化を行うことを検討している．しかし，活字には旧字体が含まれ，また文の途中で 2 行に分かれる形の「割書」も多く存在する．このため，既存の活字 OCR では認識が困難である．

そこで我々は，市販の OCR ライブラリを用い，割書の文字領域の自動認識，および旧字を外字として登録を行うシステムの構築を検討している．

4. 古文書の検索

本研究の主要な目的の一つは，古文書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. 人名・地名・建造物名の抽出

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[4]，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．

4.2. 古文書の概念検索

古文書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

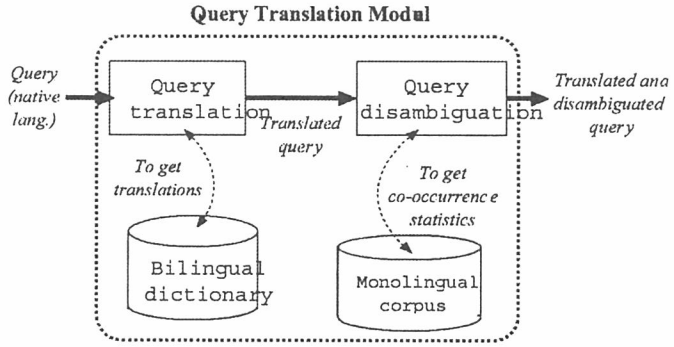


図 1：問合せ翻訳の流れ

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なない．

本研究では，古文書の概念検索への第一歩として，漢和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，次節で述べる単語共起傾向を用いることで，古文書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．

5. 言語の壁を越える検索

前節までは兵範記を例に古文書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[5]．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[6].

5.1. 問合せ翻訳の流れ

我々が提案している訳語曖昧性解消手法における問合せ翻訳の流れを図 1 に示す．利用者の母国語による問合せは，まず対訳辞書を用いて検索対象言語に翻訳される．得られた訳語候補は，単言語コーパスから得られた単語共起傾向のデータを用いて曖昧性が解消され，検索モジュールに渡される．最終的に検索結果が利用者に返される．

5.2. 対訳辞書による訳語候補リストの生成

利用者から入力されたある言語による問合せは，まず形態素解析システムなどを用いて

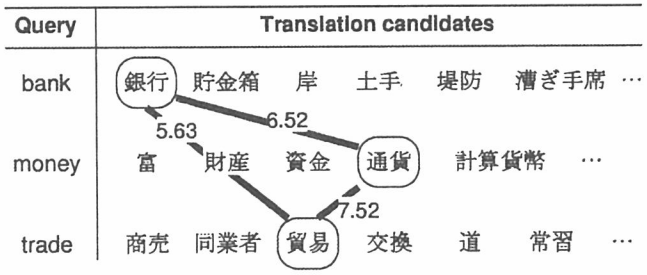


図 5：単言語コーパスによる曖昧性解消の例

単語に分割する．その後，機械可読辞書を用いて検索対象言語に翻訳する．

この際，あらかじめ辞書との照合を行い，最長一致する語を用いて翻訳する．たとえば，「電子図書館」という問合せは，「電子」および「図書館」に分割されるが，「電子図書館」というフレーズが辞書に含まれる場合は，その訳語を用いる．最長一致部分が重複する場合は，その両方の訳語を用いる．

5.3. 単言語コーパスによる訳語曖昧性解消

対訳辞書により得られた訳語の候補は，検索対象言語のコーパスにおける 2〜n 単語間の共起頻度の情報を用いた次のような手法で曖昧性の解消を行う．

5.3.1. 共起傾向の尺度

語の共起傾向の強さを測る尺度として，相互情報量（mutual information）を用いる手法がある[7].

相互情報量 MI は以下の式で得られる．

$$MI(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)}$$

ここで $p(x, y)$は事象 x と y が同時に生起する確率であり，$p(x)$および $p(y)$は x または y がそれぞれ単独で生起する確率である．この相互情報量を文書中の単語に対して用いることによって，単語自体の出現頻度に依存せずに，単語間の結び付きの強さを測ることができる．本研究ではこれを，2 単語が共起する傾向を測る尺度（以下これを共起傾向と呼ぶ）として用いる．文書中に現れるある二つの単語 w_1 と w_2の 2 単語共起傾向 COT_2を次のように定義する．

$$COT_2=\log_2\frac{\frac{f(w_1,w_2)}{N}}{\frac{f(w_1)}{N}\frac{f(w_2)}{N}}$$

ここで N はコーパス中の文の総数，$f(w)$は単語 w の出現回数，$f(w_1, w_2)$は単語 w_1 と w_2が同時に出現する回数である．

表 1 に，無作為に収集した約 3 万件の Web 文書コーパスから得られた日本語の単語間における 2 単語共起傾向の例を示す．

表 1：日本語の単語間における 2 共起傾向の例

w_1	w_2	$COT_2(w_1, w_2)$
情報	ニュース	1.41
情報	検索	1.46
情報	図書館	1.23
情報	団子	0.25
情報	陶磁器	0.16
情報	負傷	0.77

5.3.2. 訳語の選択

最終的に訳語として選択される単語は，次の手順で得られる．

1. 問合せ中の各単語のすべての訳語候補の組合せについて 2 単語共起傾向を求める．
2. 出現頻度が閾値 T_{freq} より大きい訳語候補のすべての組合せについて，共起傾向の平均を求める．
3. 共起傾向の平均が閾値 T_{COT} より大きいすべての組合せを問合せの訳語として選択する．

図 5 に，この手法による曖昧性解消の具体例として，3 単語からなる英語の問合せ“bank”, “money”, “trade”を日本語の問合せに翻訳する場合の例を示す．

ここでは，「銀行」，「通貨」，「貿易」の組がもっとも共起傾向の平均が高くなっている．実際には，これを含む，共起傾向の平均が閾値 T_{COT} より大きい訳語候補の全組合せを OR 演算子で結合したものが日本語の問合せとして用いられる．

なお，手順 2 において出現頻度が極端に低い語を除去しているが，これは相互情報量の特性である極端に出現頻度が低い場合の影響を避けるためである．たとえば図 5 の例では，閾値を設定しない場合には，低頻度語の影響によって，まったく無関係な「漕ぎ手席」，「富」，「常習」の組が共起傾向の平均が最大になってしまう．

6. 断簡の復元

兵範記は，40 年の期間のうち 17 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）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．

文字の大きさや癖のパラメータ化のためには，古文書文字認識の技術を応用することを検討している．また，紙質の特徴量の抽出のために，武田ら[8]によるフラクタル次元を用いた特徴量抽出技術などを用いることを検討している．

これらの推定は，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，膨大な数の断簡をある程度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．

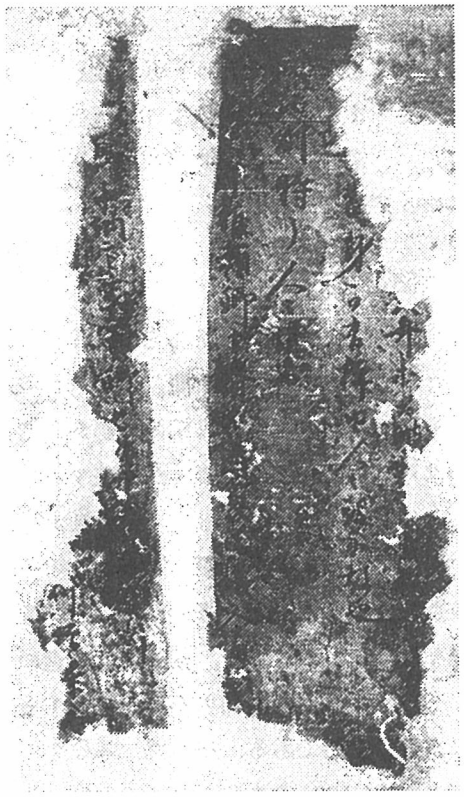


図 6：兵範記の断簡の例

7. おわりに

本研究は，古文書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．

謝辞

本研究は，文部科学省の 21 世紀 COE プログラム「京都アート・エンタテインメント創成研究」のプロジェクトの一つとして行っている．

参考文献

[1] 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）

[2] 京都大学文学部国史研究室編：「兵範記 二」，京都大学史料叢書 2，思文閣出版（1988）

[3] 『増補史料大成　兵範記一』，臨川書店（1965）

[4] 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）

[5] 桶谷猪久夫，才藤千津子，Delmer Brown：簡易型タグを利用した歴史史料の英日全文連携検索システムの設計と開発 －日本書紀、古事記における事例－，人文科学とコンピュータシンポジウム論文集，pp.65-72（2001）

[6] 前田 亮, 吉川 正俊, 植村 俊亮：言語横断情報検索における Web 文書群による訳語曖昧性解消，情報処理学会論文誌：データベース，Vol. 41，No. SIG 6 (TOD 7)，pp. 12-21（2000）

[7] Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol.16, No.1, pp.22-29（1990）

[8] 武田 哲也，村山 健二，岡田 至弘，坂井 利之：大規模古文書画像データベース構築のためのパターン解析手法の検討，京都大学大型計算機センター 第 57 回研究セミナー報告，pp.11-15（1997）