N-002

歴史 DB における検索語の分析と検索行動特性の推定
Estimation of Search Action by Analyses of Search Term in Historical Database

小野田 賢人† 安達 文夫‡ 徳永 幸生† 杉山 精*

Kento Onoda Fumio Adachi Yukio Tokunaga Kiyoshi Sugiyama

1. はじめに

現在，様々な機関において歴史・文化の研究に資する資料をデジタル化し，データベース（以下.DB）として公開する取り組みが活発に行われている．これらの DB は，Web 上に公開されている専用の検索インタフェースを用い資料を検索する．しかし，検索インタフェースの設計は対象とする資料により異なり，歴史資料を対象とした場合の最適なインタフェースの設計方法は明らかでは無い．

そこで，最適なインタフェースの設計方法を検討するため，利用者の検索ログデータを用い，現在の利用者の検索行動を分析する．これにより，利用者の検索行動に沿ったインタフェースの設計方法を検討する．

筆者らはこれまでに，DB が保持しているフィールド毎に入力される検索語を分析し，利用者がどのような目的を持ち検索をしているかを分析してきた[1]．そこで，本稿では分析をするフィールドを「資料名称」と「フリーワード」に絞り，「物・事・所・時・人」に検索語を分類し，どのような観点から資料を検索しているかを分析したため，報告する．

2. 分析に用いるログデータ

2.1 分析対象の DB

本稿では，国立歴史民俗博物館が Web 上に公開しているデータベース「れきはく」という DB 群の検索ログデータを用い，利用者の検索行動特性を分析する．

本稿で対象とする DB は，所蔵資料の目録を収録した「館蔵資料 DB」と民俗学についての文献資料の目録を収録した「日本民俗学文献目録 DB」である．

2.2 検索インタフェースとログデータ

データベース「れきはく」の館蔵資料 DB で検索を行うためのインタフェースを図 1 に示す．

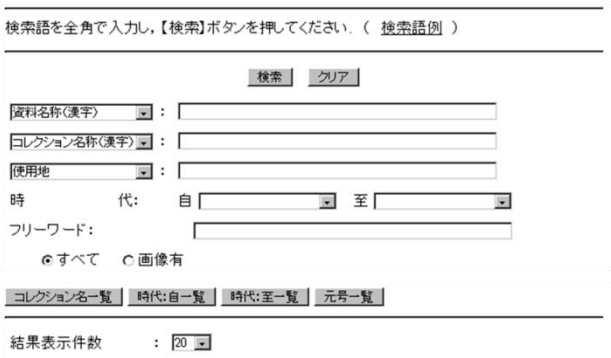

図 1 データベース「れきはく」検索インタフェース

この検索インタフェースは，プルダウンリストを用いて検索を行うフィールドを選択し，その隣にある入力フォームに検索語を入力することで検索を行う．データベース「れきはく」に含まれる DB 群の検索インタフェースは同様の設計となっている．

本稿で分析に用いるログデータは，「プルダウンリストの項目」ごとに検索された「検索語」の検索回数がリストとなったものである．

3. 検索語の分類

3.1 対象とする項目

本稿では，「資料名称」「フリーワード」という 2 つの項目についての分析を行った．対象とした 2 つの項目についてだが，「資料名称」は資料の名称を扱うフィールドを検索する項目となっており，「フリーワード」は DB に存在する全てのフィールドを対象とし検索を行う項目となっている．歴史資料において資料名称は，「物・事・所・時・人」などの要素により与えられており，入力される検索語は他の項目と比較すると雑多になっている．そこで，「資料名称」「フリーワード」という 2 つの項目にて入力される検索語を本稿では「物・事・所・時・人」という要素に分類する．

これにより，利用者がどのような目的を持ち検索語を入力しているか分析する．

3.2 検索語の分類方法

利用者の検索の目的を探るため，入力される検索語を 3.1 で述べた 5 つの要素に分類する．この作業は検索語の中から一部をサンプリングし，手作業にて行う．

しかし，これらの要素に分類が出来ない語も存在する．そこで，複数の語を繋ぎ文となっている検索語を「複合語」，記号や数字などの分類にも属さない語を「分類不能」とする．

分類の詳細を以下に記す．

物
有形，無形のものを表す語（文書，建築物，動植物）

事
行為，操作，属性を表す語

所
場所を表す語（国名，地名，荘園名）

時
時間について表す語（時代，和暦，元号）

人
人名を表す語（人名，役名）

複合語
上記の語が 2 語以上組み合わされた文

分類不能
上記の語以外の語または文字

† 芝浦工業大学，Shibaura Institute of Technology
‡ 国立歴史民俗博物館，National Museum of Japanese History
＊東京工芸大学，Tokyo Polytechnic University

また，これらの分類に 2 つ該当する検索語も存在する．そのため，それらは 2 つの分類に属する検索語として扱う．

複数の分類に属する検索語の例を以下に記す．

物・事に分類される語

はやし，盆など

所・人に分類される語

石川，伊達など

所・時に分類される語

江戸，鎌倉など

3.3 分析結果

3.2 の手法により分類を行った．館蔵資料 DB の「資料名称」「フリーワード」について分類ごとの検索数を正規化したものを図２に示す．

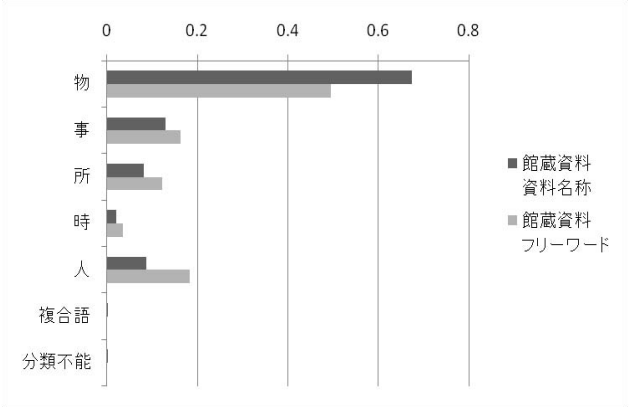

図２　館蔵資料 DB の分類結果

図２より館蔵資料 DB では，「物」についての検索が多いという傾向が見える．これは館蔵資料 DB が所蔵資料の目録を収録した DB であるため，DB が対象としている「物」自体に着目し検索が行われたためであると考えられる．2 つの項目について比較すると，「資料名称」に比べ「フリーワード」では「事」「所」「時」「人」の分類の検索数が多くなっている．これはフリーワードでは全ての項目に対して検索を行うことが出来るため，「物」以外の検索の比率が伸びたと考えられる．

また，もう 1 つの分析対象である，日本民俗学文献目録 DB の「資料名称」「フリーワード」について分類ごとの検索数を正規化したものを図３に示す．

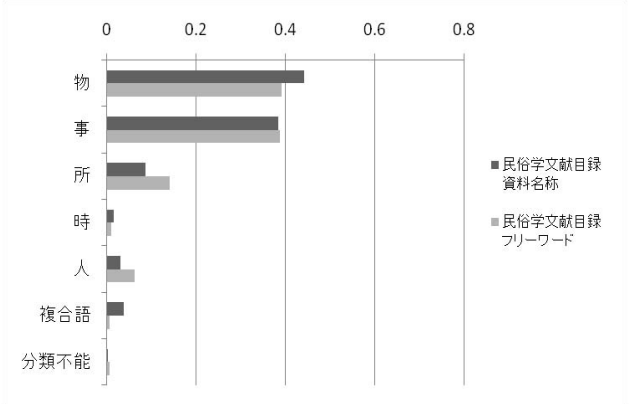

図３　日本民俗学文献目録 DB の分類結果

図３より日本民俗学文献目録 DB では，「物」「事」についての検索が多いという傾向が見える．これは日本民俗学文献目録 DB が，民俗学の文献資料の目録を収録 DB であるため，当時の風俗を表す「物」や，各地方で行われていた風習や行事を表す「事」に着目し検索が行われたためであると考えられる．2 つの項目について比較すると，館蔵資料 DB の場合と同様に「フリーワード」では入力される検索語の比率が分散する傾向にある．

また，分類ごとに 1 語彙あたりの平均検索回数を分析した。分析にあたり、「分類の検索数」を「分類の語彙数」で除算した．結果を図４に示す．

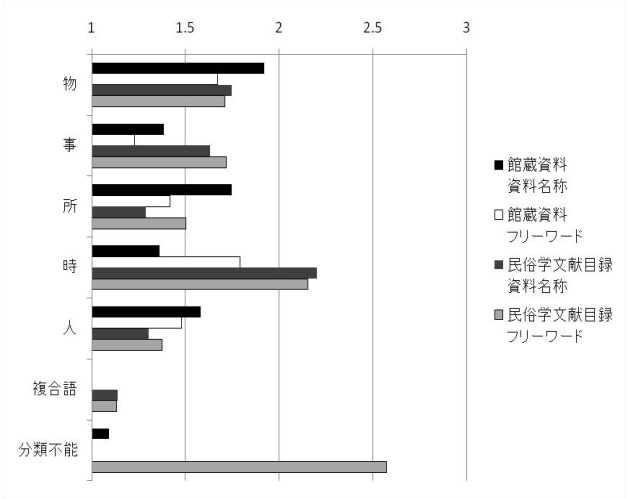

図４　分類ごとの平均検索回数

図４より，特に平均検索回数が多い分類は，民俗学文献目録 DB のフリーワードの分類不能である．入力されている検索語は「2000」など西暦で検索をしていると考えられる検索語であり，それぞれが複数回検索されている．また，語彙数が 7 件と極めて少ないため平均検索回数が他より高くなったと考えられる．他の分類に関しては，平均検索数が 2 回前後と少ない値になっている．これは，殆どの語彙の検索回数が 1 回となっており，複数回検索される語彙が多くは存在しないためである．

4. おわりに

本稿では，利用者が検索を行う際にどのような視点から検索語を入力しているかを分析した．分析にあたり，検索語を「物・事・所・時・人・複合語・分類不能」に分類しどのような目的を持ち検索を行っているかを分析した．

その結果，DB が対象とする資料の違いにより，検索に用いられる語彙の分類に差がある事が明らかとなった．そして，資料名称に比べフリーワードを用いた検索では，利用される語彙の分類が分散する傾向にある．

また，分類ごとに 1 語彙あたりの平均検索回数を算出し，分類ごとに平均検索回数の差を分析した．その結果，全ての分類において平均検索回数が，2 回前後となり大きな差は見られなかった．これは，殆どの語彙の検索回数が 1 回となっているためである．今後は，特に検索回数の多い語彙について分析し検索フィールドの違いによる検索行動の違いを明らかにする．

参考文献

[1] 小野田賢人，安達文夫，徳永幸生，杉山精 “歴史 DB の検索インタフェース設計に向けた検索語の分析”， 第 9 回画像ミュージアム研究会， (2011).