複数の歴史文書ディジタルアーカイブを対象とする年表型ユーザインタフェースの開発

越智理恵[1]、永森光晴[2]、杉本重雄[2]
1．筑波大学図書館情報専門学群
2．筑波大学図書館情報メディア研究科
〒305-8550　茨城県つくば市春日 1-2
E-mail:s0612180@u.tsukuba.ac.jp, {nagamori, sugimoto}@slis.tsukuba.ac.jp

概要

現在、いくつもの文化遺産を扱うディジタルアーカイブがあり、それらを横断に検索することもできる。その一方、資料間の意味的なつながりに基づいて、統合的に資料を扱う機能は提供されていない。本研究では、歴史文書を対象としている二つのディジタルアーカイブを用いて、年表型のインタフェースを利用した統合的アクセス支援ツールの開発を進めた。統合的アクセス支援のために文書の関連付けを行い、これには文書のメタデータを用いて、共通の分類を付与した。ユーザには文書に付与した分類と時系列に基づく年表型のインタフェースを提供することとした。

キーワード：　ディジタルアーカイブ、歴史文書、統合的利用、ユーザインタフェース

A Chronology User Interface for Multiple Digital Archives of Historical Documents

Rie Ochi[1], Mitsuharu Nagamori[2], Shigeo Sugimoto[2]
1. School of Library and Information Science, University of Tsukuba,
2. Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
E-mail:s0612180@u.tsukuba.ac.jp, {nagamori, sugimoto}@slis.tsukuba.ac.jp

Abstract

There are several digital archives of cultural heritage, which can be retrieved across each other. However, the semantic relationships between their resources are not used for resource access across the archives. This paper shows a chronological user interface to brose historical documents and recortds using two digital archives of historical documents – Newspaper Article Archive of Kobe University Library and Government Record Database of Japan Center for Asian Historical Records. In this study, we collected metadata of the articles and records from the two archives and automatically classified them to organize a chronological user interface to access the articles and records. The user interface shows the titles of the articles and records in chronological order and with classification indicators.

Keywords: Digital Archive, Historical Documents, Integrated Use, user interface

1. はじめに

現在、公文書だけでなく新聞記事や文化遺産などもアーカイブ化され、その多くが電子化によりディジタルアーカイブとしてインターネットで公開されている。歴史文書を扱っているディジタルアーカイブに、国立公文書館のアジア歴史資料センターや神戸大学附属図書館の新聞記事文庫などがある。これらのディジタルアーカイブは、それぞれ文書の画像の公開や検索機能を持っており、各機関において独自の分類やメタデータの記述が行われている。

また複数のアーカイブを扱っているポータルサイトとして、国立国会図書館のデジタルアーカイブポータルの PORTA がある。PORTA は登録されたディジタルアーカイブの横断検索機能を提供している。しかしながら、多数の異なるアーカイブを横断的に検索するため、一般的なテキスト検索しか提供しておらず、資料同士の関連や意味的つながりを見つけるといった機能は持っていない。複数アーカイブを一つのテーマをもとに統合したものとして、ディジタル・シルクロードがある。ここでは独自のアーカイブを複数作成し、それらを地図や年表などの軸による統合を行っている。単一のアーカイブではわからなかった異なる種類の資料同士の関連を視覚的に表現することが可能となっている。

本研究では、神戸大学附属図書館の新聞記事文庫と国立公文書館アジア歴史資料センターの公文書アーカイブを対象として、年表型のインタフェースを利用した異種歴史文書のための統合的なアクセス支援ツールの開発を進めた。神戸大学の新聞記事文庫、アジア歴史資料センターの公文書アーカイブはともに明治期から昭和時代前半の文書を提供している。前者は経済関係を中心とする新聞記事、後者は、外交、軍、そして内閣の公文書を提供しており、一概に歴史文書とは言っても、その内容と書き方等の面において性質はかなり異なる。その一方、ある一つの事件に関連して作られた記事と公文書もあり、それらを関連付けて閲覧することができればディジタルアーカイブとしての価値をより高めることができる。

二つのディジタルアーカイブは、国立国会図書館の PORTA を介してともにメタデータを提供している。PORTA ではメタデータの記述項目は統一されているが、記述内容は各ディジタルアーカイブが提供するメタデータの内容をそのまま提供しているため、分類や件名等、文書のナビゲーションにおいて重要な役割を持つ記述内容が統一されていない。そこで、本研究では、二つのディジタルアーカイブのメタデータを収集して、それらに対して共通の分類を付与し、時系列と分類に基づく年表型のインタフェースを提供することにした。本研究では、PORTA ならびに各アーカイブで提供されている外部提供インタフェースを用いて、必要なメタデータを収集し、収集したメタデータを用いて、タイトルと内容記述を用いて NDC に基づく分類を付与した。そして、年表型のインタフェースについては、MIT で開発されたツールを利用して、収集したメタデータを統合的に閲覧するためのシステムを作成した。

以下、本論文では、２章においてディジタルアーカイブの現状を概観する。３章では、ディジタルアーカイブの統合的利用のための機能について述べ、４章では本研究で開発したシステムについて述べる。

2. ディジタルアーカイブの現状とその利用例

2.1 歴史文書のディジタルアーカイブ

歴史文書を扱っているアーカイブはいくつかあるが、以下に、本研究で利用したアジア歴史資料センターと神戸大学新聞記事文庫について簡単に述べる。

国立公文書館のアジア歴史資料センターが提供する公文書アーカイブは世界でも有数の規模を誇るものである[1]。アジア歴史資料センターは、国立公文書館が運営しており、パソコンを通じてアジア歴史資料を提供する電子資料センターである。ここで扱われている資料は、明治維新から太平洋戦争終了までの間の公文書であり、国立公文書館、外務省外交史料館、防衛省防衛研究所図書館が保管するアジア歴史資料

である。これらは近現代における日本とアジア近隣諸国等との関係に関わる歴史資料である。

新聞記事文庫は、神戸大学経済経営研究所によって作成された新聞切り抜き資料である。ここで扱われている資料は、明治末から昭和 45 年までの新聞切り抜きであり、電子化されたものから公開されている。その多くが経済分野であるが、政治外交・法制など広範にわたって資料が集められており、複数の新聞を用いて資料が収集されているのも特徴の一つである。

2.2 ディジタルアーカイブの横断的利用支援ー国立国会図書館デジタルアーカイブポータル（PORTA）

国立国会図書館では、広く国のディジタル情報全体へナビゲーションする総合的なポータルサイトとして PORTA[3]を構築している。PORTA のサービスの目的[4]は、利用者が必要とする情報資源へのアクセス支援として、ディジタルアーカイブのコンテンツそのものへ、ワンストップで案内可能とすることにより、電子的情報資源や情報提供サービスの利活用を促進することである。国立国会図書館保有の 13 ディジタルアーカイブに加え、公共機関や民間機関等が保有する複数のアーカイブの全 52 アーカイブ（2010 年 1 月現在）に対する検索ができる。

2.3 複数のディジタルアーカイブの横断利用

ディジタル・シルクロード[5]はシルクロード地域を対象とした文化遺産ディジタルアーカイブの構築を目指すプロジェクトであり、文化資源へのアクセス性を高めるとともに、文化資源の活用を進めていくことを目標としている。このプロジェクトでは独立したアーカイブとして構築した後に、これらを統合する地理空間や時間空間などの統合空間を用いた複数アーカイブの統合を目指したものである。北本らの論文[6]によると、貴重書のコレクション、危機に瀕する文化遺産のコレクション、写真のコレクションのそれぞれに検索機能やナビゲーション機能をもたせ、それらのコレクションが個々のディジタルアーカイブとなっている。そしてこれらのアーカイブを統合するためのハブとしての統合情報空間を利用して、複数アーカイブの統合を行う。統合情報空間には、地理空間、時間空間、概念空間の 3 つの情報空間を用いて、統合を進める。これらの情報空間を使い、個々のリソースを統合空間に射影することで、同じ点や領域に射影されたリソースを取り出すことができ、複数のリソースの統合が可能となる。

2.4 歴史文書アーカイブのためのユーザインタフェース

機関リポジトリの構築について述べられた[7]によると、機関リポジトリなどのポータルサイトでは、横断的な検索インタフェースが充実しているものは少なく、多くが簡易検索や詳細検索の最小限の検索環境しか用意されていない。なぜならば検索の際に文書同士の関連を考慮する必要がない検索を想定しているためである。そのため簡易検索のような文字列一致の検索が多く、結果表示も件名の羅列にとどまっている。しかしながら文書同士の関連を把握するためには、ユーザの目的に沿ったインタフェースが必要とされる。

本研究では、複数の歴史文書へのアクセスを考えているため、時間指向のインタフェースが効率的であると思われる。これは、歴史という流れの中で、物事の変化がどのように起こったかを把握するのに、歴史学習で用いられるような年表型にすることで、文書へのアクセス性を高められると考えられるからである。

3. 神戸大学新聞記事文庫とアジア歴史資料センター公文書アーカイブの統合的利用

3.1 対象ディジタルアーカイブの概要

本研究で対象とした神戸大学附属図書館新聞記事文庫と国立公文書館アジア歴史資料センター公文書アーカイブに関して概要を示す。

(1) 神戸大学附属図書館新聞記事文庫

神戸大学附属図書館新聞記事文庫は、神戸大学経済経営研究所によって作成された新聞切り抜きを集めたものであり、電子化されたものから順次公開され、その利用が可能である。

新聞記事文庫の特色[8]として、継続性と網羅性、専門家による選択・分類、採録対象紙の多さがあげられる。資料の収集期間は、戦前から戦後にかけて約 60 年分が蓄積されている。また、収録範囲も経営・経済を主として、政治、外交、法制、教育など広範にわたって収録されている。また専門研究者の視点で独自の分類が行われており、特定分野の記事をまとめて閲覧することも可能である。さらに、採録対象紙は大阪の主要紙と経済紙を中心に、東京・大阪その他の新聞や主要地方紙、旧植民地・外地誌など幅広く採録されており、同一事件について複数の新聞から採録していることも特徴である。

また新聞記事文庫で提供されているデータ項目は、大きく分けて記事全文画像、見出しインデックス、全文テキストの 3 つである。見出しインデックスでは、記事見出し、著者情報、新聞名、記事日付、記事分類がデータ項目にあり、検索結果やブラウジングリストでの個々の記事表示情報となっている。また全文テキストは、現在の常用字体や現代仮名遣いへの置き換えなどの処理がほどこされている。新聞記事文庫のメタデータは PORTA が提供している API を用いて取得が可能である。このときに取得できるメタデータの例を図 1 に示した。

```
<item>
<title>福岡県の農業労働 ： 純労働者は僅に一万人 ： 帰農者は漸次増加の傾向</title>
<link>http://www.lib.kobe-u.ac.jp/das/jsp/ja/DetailView.jsp?LANG=JA&amp;METAID=00490091</link>
<dc:identifier
xsi:type="dcterms:URI">http://www.lib.kobe-u.ac.jp/das/jsp/ja/DetailView.jsp?LANG=JA&amp;METAID=00490091
</dc:identifier>
        <dc:identifier>00490091</dc:identifier>
        <dc:language>jpn</dc:language>
        <dc:contributor>神戸大学経済経営研究所</dc:contributor>
        <dc:date>1921.1.20 (大正 10)</dc:date>
        <dc:type xsi:type="dcndl_porta:Digitalize">1</dc:type>
        <dc:type xsi:type="dcndl_porta:Web-get">1</dc:type>
        <dc:type xsi:type="dcndl_porta:Payment">1</dc:type>
        <dc:type xsi:type="dcndl_porta:PORTAType">Article</dc:type>
        <dcndl_porta:dpid>kobe-n</dcndl_porta:dpid>
        <dcndl_porta:repository_no>R000000035</dcndl_porta:repository_no>
        <dcndl_porta:item_no>
        </dcndl_porta:item_no>
</item>
```

図 1　PORTA の API で取得したメタデータの例

(2) アジア歴史資料センター（JACAR）の公文書アーカイブ

国立公文書館アジア歴史資料センターは、国の機関が所蔵公開している歴史資料のうち、日本およびアジア諸国等の歴史に関する資料をディジタルアーカイブ化してインターネット上で公開する役割を担っている。資料の概要[9]によると、平成 20 年 12 月現在での資料公開数は約 115 万件・1,740 万画像である。

また文書には、件名標題、階層、レファレンスコード、作成者名称、資料作成年月日、規模、内容などのメタデータが付与されており、内容には、各文書の冒頭の約 300 文字が書かれている。また、レファレンスコードにより資料の所蔵元が判別可能である。

アジア歴史資料センターの公文書アーカイブも PORTA 経由で検索が可能である。本研究では、ダウンロード可能なメタデータ数の制約を回避するため、同センターに許諾を得て SRW を用いて本ディジタルアーカイブにアクセスした。アジア歴史資料センターで付与されたメタデータの例を図 2 に示す。

```
<record>
  <recordData>
    <dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns="">
    <dc:title>外国出張者旅費前渡金返納の件</dc:title>
    <dc:creator>主計課</dc:creator>
    <dc:date>大正１０年１月～大正１０年３月</dc:date>
    <dc:identifier>C03025208600</dc:identifier>
     <dc:identifier xsi:type="dcterms:URI">
       <![CDATA[http://www.jacar.go.jp/DAS/meta/MetaOutServlet?GRP_ID=G0000101&DB_ID=G0000101EXTE
       RNAL&IS_STYLE=default&IS_TYPE=&SUM_KIND=MetaFolder&IS_START=1&IS_NUMBER1=&data_ty
       pe=&IS_KIND=MetaDetail&ID=M2006090103403286861&CAT_TYPE_DISPLAYED=&XSLT_NAME=Oya
       Meta.xsl&MEDURL=&IMG_FLG=&REF_CODE=]]>
    </dc:identifier>
    <dc:description>
                陸軍省受領欧受第三八号 外国出張者旅費前渡金返納ノ件 欧経第七号 軍事欧第一九号 副官ヨリ陸
                軍東京経理部長へ通牒　　（欧発）　　陸軍砲兵少佐森田宣ヨリ別紙横浜正金銀行為替券
                ヲ以テ返納証書ノ通リ旅費前渡金残額金千四十三円七十六銭五厘（英貨換算額百十三＠十七＠九片）
                返納有之候条精算方取計相成度候也　　陸軍省送達欧発第一九号 一月二十四日 為替券ハ川瀬
                主計正伊藤書記官承知置相成度候 軍事課決行済一月二十五日 官房御中　　欧第一一二一号其
                一 仏庶第五十七号 注受第八号 森少佐外二名経費決算ノ件 大正九年十一月十五日 ＠＠
    </dc:description>
    <dc:subject>防衛庁防衛研究所</dc:subject>
    <dc:subject>件名</dc:subject>
    <dc:language>jpn</dc:language>
    </dc>
  </recordData>
</record>
```

図 2　アジア歴史資料センターで付与されたメタデータの例

3.2 統合的利用機能の検討

現在の横断検索では、文書間の意味的なつながりが無視されており、複数のアーカイブから関連する文書を見つけ出すことは困難である。これはアーカイブ毎に適切な検索語が違うため発生している問題である。この問題点を解消し、複数のディジタルアーカイブを統合的に利用する際に必要とされる機能として、対象アーカイブの内容に応じ、対象アーカイブのメタデータを横断的にブラウジングできる機能が重要であると考えた。以下は歴史文書のディジタルアーカイブの横断的なメタデータブラウジングに関して必要と考えられる機能である。

(1) 共通の分類による横断的なメタデータブラウジング機能：文字列一致を基本とする横断検索機能は PORTA でも実現されているため本研究では中心的な研究課題とはしなかった。本研究では、異なるアーカイブにおかれた文書同士の関連を見つけ、意味的なつながりを見つけることを支援するための機能が重要であると考えた。そこで、同時期の関連する文書の関連付けを行うために、共通の基準で文書を分類し、分類を利用した横断的なメタデータブラウジングの機能を実現することにした。共通の基準に基づく分類付与のために、本研究では、国立国会図書館件名標目表（以降 NDLSH）が持つ代表分類を用いた。代表分類は NDLC と NDC で与えられており、本研究では NDC を利用した。

(2) 横断的かつ時間軸に沿ったメタデータブラウジング機能：結果表示のインタフェースは重要な要素である。本研究では、文書間や文書と史実との関連を直感的に見つけられることと時系列の物事の変化を把握することの 2 点が重要なユーザインタフェース機能であると考えた。この 2 点を満たすインタフェースとして、共通の分類に基づいて文書を時系列状に表示できる年表型を採用した。また年表上に、歴史教科書等に記載されている史実を示すことで、史実との関連や時代背景の把握の手助けとした。これは、時間軸と分野の 2 つの軸を持つ表の上に文書を配置することで異なるアーカイブに所蔵される文書を一覧できるようにしたものである。また、歴史教科書で用いられる年表と同様な構造を持たせることで、文書と歴史上の出来事を結び付けやすくすることも目指した。

3.3 システムの利用

本システムは時間と分類の 2 軸からなるインタフェースを持っている。たとえば、「原敬暗殺事件が起こった際の国内の動向を知りたい」や「大正時代の政治と経済の関連を知りたい」などの、「いつ」の「どのような分野」の動向や関連文書を知るためのツールとして使われると想定している。本システムを使うことで「原敬暗殺事件」が発生したときの「経済」や「政治」分野の新聞記事や公文書を時系列で見ることができる。

4. 複数の歴史文書ディジタルアーカイブの統合利用システムとその実現

4.1 システムの概要

異なる種類の文書へのアクセス支援のため、各文書に NDLSH の代表分類を用いた分類を付与し、分類と時間の 2 つの軸を持った年表型のユーザインタフェースを開発した。システムの実現のためのメタデータの処理過程を図 3 に示す。

① アジア歴史資料センター、新聞記事文庫から外部提供インタフェースを用いて、メタデータを取得、データベース化を行う。

② 取得したメタデータからタイトル（title）または内容記述（description）を取り出し、分類のための処理を行う。

③ ②によって付与された分類をデータベースに格納する。

④ PHP を用いて、メタデータを取得しインタフェースに表示させるための XML 形式に変換する。

⑤ ④で作成した XML 文書と歴史上の出来事の一覧を記述した XML 文書を年表作成に用いるアプリケーションに受け渡し、年表型のインタフェースを表示させる。

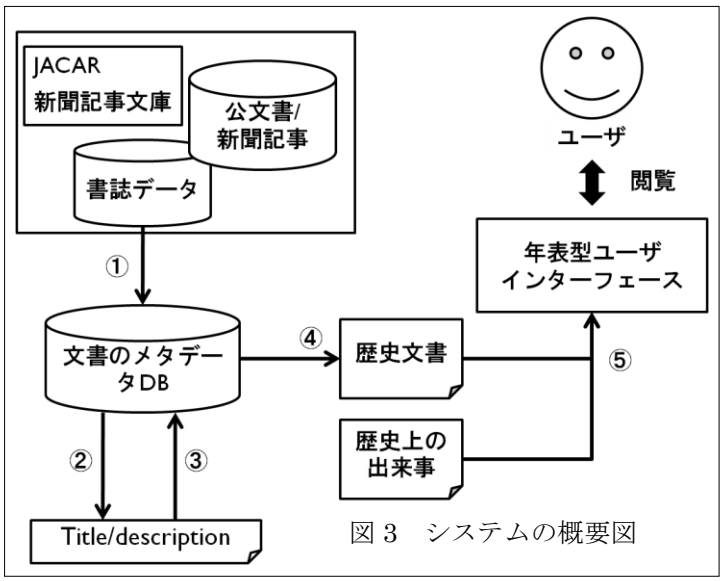

図 3　システムの概要図

4.2 メタデータの取得とデータベース化

本研究では、大正 10 年（1921 年）の 1 年分のメタデータを用いている。メタデータの取得には、それぞれ外部提供インタフェースを用いた。神戸大学附属図書館新聞記事文庫に関しては国立国会図書館 PORTA の API[10]の一つである OpenSearch を用いてメタデータの取得を行った。アジア歴史資料センターは SRW リクエストを送信することでメタデータを取得した。なお、SRW は SOAP 仕様に基づく XML の検索リクエストを送り、返戻形式は XML である。本研究で実際に収集したメタデータの量は、新聞記事、アジア歴史資料ともに大正 10 年の 1 年分で各 1 万件程度である。

4.3 メタデータを利用した文書の分類

それぞれの文書のメタデータには、分類を示すものは含まれていない。そのため、title や description エレメントの記述内容を用いて、分類を行っていく必要がある。アジア歴史資料の場合は所蔵元を判別するために ref エレメントを利用できる。

以下、文書の関連付けのためのメタデータを利用した文書の分類について述べる。分類には NDLSH の代表分類を用いた。NDLSH の代表分類の一つである NDC の 2 次区分までを用いて、分類付与を行った。NDLSH および NDC の取得には、本研究室で開発した HANAVI[11]がもつデータベースを利用した。

文書の分類付与のために、格納したメタデータから新聞記事の場合は title エレメント、アジア歴史資料は description エレメントを抽出し、形態素解析システム ChaSen[12]を用いて形態素解析を行った。ChaSen で付与された品詞のうち、固有名詞以外の名詞の単語を取り出し、単語と部分一致する NDLSH の語の代表分類 NDC を集める。そして文書毎に NDC を合計し、上位をその文書の分類として付与した。

NDC を用いた分類の妥当性の検討のために、新聞記事とアジア歴史資料から各 50 件ずつのメタデータを使って実験を行った。この実験では、title と description エレメントの内容を基に人手で付与した分類を、NDC 二次区分の分類を付与した集合を正解集合とした。そしてシステムで機械的に付与した NDC と比較した。その結果、「人」や「国」など一般的な語の場合、該当する NDLSH が多くなりすぎることなどいくつかの問題点が判明した。それらの解消のため、NDC 付与の際分類の分布や総数に一定の閾値を設けることとした。

また、アジア歴史資料センターの公文書には漢字カタカナ混じり文の文書が多く含まれる。本研究で用いたツールでは幹事カタカナ混じり文の形態素解析に対応していなかったため、形態素解析に不備が生じてカタカナが名詞とされる場合が多くみられた。その多くは、本来助詞である「ニ」や「ヲ」などのカタカナ語である。これらのカタカナ語による分類の影響を避けるために、全てのカタカナ語に対して NDC を取得させない処理を行った。これらの NDC 付与の際の設定により本システムでは、7 割から 8 割程度の文書に対して正しい分類がなされている。

NDC の二次区分は 100 項目あり、年表型のインタフェースに 100 項目の全てを表示させることは難しい。そのため実際に歴史年表[13]で使われている政治、軍事、外交、経済など 10 項目に NDC を割り振っている。また、資料の所蔵元と文書の内容とつなぐことができると考え、アジア歴史資料ではメタデータに記述されている所蔵元をこの年表の分類に割り振り、文書の分類の一つとした。以上のように、NDC や文書の所蔵元を年表の分類に割り振ることで、文書の 8 割から 9 割程度に正しい分類が付与された。

4.4 年表型インタフェース

年表型のインタフェースの開発のため、MIT の SIMILE Widgets | Timeline[14]（以降 Timeline）を用いた。Timeline は XML または JSON 形式のソースを読み込ませると、年表上にアイコンを表示させるソフトウェアであり、API が公開されている。

この API を用いて、年表型のインタフェースを開発した。読み込ませるソースは PHP を介してデータベースに接続、分類ごとにアイコンの色を変えた XML 文書を出力し、表示させている。トップページの表示画面を図 4 に示す。トップページでは、全分類が一つの年表上に表示されており、分類はアイコンの色で区別する。一番上段には、歴史上の出来事を表示させ、史実と文書の関連を見ることができる。時間軸はスクロールすることができ、指定の月にジャンプする機能を付けることで、歴史上の出来事があった前後の文書へアクセスしやすくした。

図 4　トップページ

分類の選択を行うと、年表には史実、新聞記事、アジア歴史資料の分布を一度に見ることができる。また分類ごとの画面で、文書のタイトル名をクリックすると、詳細が吹き出しで表示され、title や description を見ることができる。また文書のタイトル部分から該当ページへのリンクを張っており、クリックすると、実際の文書の画像や全文を読むことができる。

4.5 システムの利用例

「原敬暗殺事件」（1921 年 11 月 4 日）を例としてシステムの使い方を示す。

① トップページで事件が発生した 11 月の文書をみる。各ページには各月へのジャンプ機能が付いており、ここでは 1921 年 11 月をクリックする。

② 11 月 4 日前後の文書を探す。教科書に記載があるような歴史上の出来事は年表の上部に表示されている。また、史実の発生時期と文書のタイトルから関連する文書を探す。文書の表示は

「●原首相暗殺さる：政友会近畿」

のようになっており、文書のタイトルから関連文書をみつける。

③また、文書のタイトルの前についているアイコンは分野ごとに色が異なっており、ここから文書の分野がわかるようになっている。関連文書のアイコンの色から該当する分野のページをみてみると、その他の関連文書も見つけられる。分野ごとのページでは、新聞記事とアジア歴史資料の両方に関連する文書を見ることができる。

5. 考察

本研究では、種類の異なる文書に対し分類を付与することで文書の分類を行い、年表型のインタフェースにより文書へアクセスできるようにした。文書の関連付けを行うことで、一般的な文字列一致の検索では発見できなかった文書へ一度にアクセスができるようになった。また年表形式のインタフェースにすることで、時系列の文書の変化を見て取ることや国内の変遷を見て取ることができるようになったと考えている。対象に新聞記事を用いることにより、国民の生活に沿った歴史文書へのアクセスができるようになり、教科書には書かれていない昔の生活や情勢などが垣間見ることができると思う。

一方、本システム構築の際に棚上げした問題がいくつか残っており、それらの改善が今後必要である。まず、アジア歴史資料における漢字カタカナ混じり文の問題である。現在、漢字カタカナ混じり文に対応している形態素解析器はないため、現代語の形態素解析器を用いている。本研究では、漢字カタカナ混じり文に対し前処理等を行わずに形態素解析の処理を行っているが、送り仮名が全てカタカナのため、本来カタカナの国名や人名と送り仮名のカタカナの区別ができていない。また、本来助詞の部分が名詞に分類されてしまうなどの問題も発生している。本研究では、4.3 で述べたようにカタカナ語を除外した名詞で NDC 取得を行っているが、カタカナ語の名詞を含めた文書中のすべての名詞をきちんと取り出すには、形態素解析器もしくは入力文字列自体に何らかの前処理を行う必要がある。この前処理では、カタカナをひらがなに変換するだけでなく、国名や人名はカタカナのままにしなければならないなど多くの課題がある。文書で使われている言葉の意味を保ったまま、形態素解析が正しく行われるような前処理を行うことが重要である。

次に分類の精度の問題がある。4.3 で述べたように、NDC のみでの分類付与よりも年表の分類を用いたときの方が正しい分類がなされた文書は増えている。その理由はいくつか考えられるが、その一つとしては、自然科学と技術のような分類として近いものが NDC では細分化されているため、別の分類を付与されていたが、年表の分類ではそれらが一つにまとまっているためと考えられる。またアジア歴史資料の多

くは、公文書という性質からか、どのような分野の文書かということは暗黙のうちに決まっており、description エレメントに直接的な表現で「政治」や「軍事」、「外交」などの文字が出現することは少ない。そのためアジア歴史資料の所蔵元を分類に利用することで文書の分類が正しく行われるようになったと考えられる。

現在、正しいと思われる NDC の分類がなされているのは約 7 割程度で、残りの 3 割は不適切な分類となっており、NDC の分類の精度が良いとは言い切れない。NDC の分類の精度を良くすることができれば、年表の分類に割り振る際にもより正しい分類を行うことが可能となる。この精度の問題は先に述べた漢字カタカナ混じり文の問題からの影響も受けていると考えられる。また現在は分類付与の際、名詞と部分一致する NDLSH の代表分類を全て採用しているが、完全一致の分類への重みづけや共起関係を考慮した分類を行うなど分類付与手法の改善も必要だと考えられる。文書の関連付けを正しく行うためにも、形態素解析から分類にかけての処理において、今後検討が必要である。

最後に、本研究では 1921 年の約 1 年分を用いて行ったが、本研究の目的である史実との前後関係の把握を行うためには、データの量を増やことが望ましいが、これは将来への課題としたい。

6. おわりに

本研究では、種類の異なる歴史文書へのアクセス支援のためのシステムの実現を図った。そのために、文書の関連付けとインタフェースの開発を行った。これによって、文字列一致での横断検索では一度に見つけることができなかった文書を一つの画面上に発見できるようになり、新聞記事とアジア歴史資料へのアクセスが容易になった。また年表型の時系列に沿った表示により、時期によって変化する文書の分類を見て取ることができる。歴史の学習の際にも、教科書に書かれている出来事に対して、新聞報道がどのようなものだったか、どのような政治的な影響があったかなどを知ることができるようになった。

本研究の今後の課題は考察でも述べたようにいくつかあり、改善が必要である。またアジア歴史資料は公的文書であるため、言葉遣いが難解でメタデータを見ても実際の文書の内容の理解が困難なことがある。そのため、文書に対してユーザが説明を書き込めたり、インターネット上の情報にリンクしたりできるようになればよいのではないかと思われる。本研究では、文書へのアクセス支援のためのツールにとどまっているが、今後これを拡張した情報共有ツールになればよい。

謝辞

本研究を進めるにあたり国立国会図書館の PORTA を活用した。PORTA の運営に関わる方々に感謝の意を表したい。また、国立公文書館・アジア歴史資料センターの公文書アーカイブのメタデータを利用する上で、国立公文書館ならびにアジア歴史資料センターから技術的支援を頂いた。末筆ではあるが感謝の意を表したい。ディジタルアーカイブの構築には多大なコストと努力が必要であり、これまでに構築の努力をされてきた方々に感謝の意を表したい。特に、アジア歴史資料センターの大規模なアーカイブ構築に多大な貢献をされた故牟田昌平氏には心からの感謝を申し上げたい。

参考文献

[1] 国立公文書館アジア歴史資料センター. “　アジア歴史資料センター(アジ歴) | Japan Center for Asian Historical Records(JACAR) National Archives Japan.”, http://www.jacar.go.jp/, (参照 2010-01-18).

[2] 神戸大学附属図書館. “神戸大学附属図書館 デジタルアーカイブ【新聞記事文庫】”, http://www.lib.kobe-u.ac.jp/sinbun/, (参照 2010-01-18).

[3] 国立国会図書館. “PORTA(国立国会図書館デジタルアーカイブポータル)”, http://porta.ndl.go.jp/portal/dt, (参照 2010-01-18).

[4] 国立国会図書館. “PORTA のコンセプト”, http://porta.ndl.go.jp/portal/dt?action=content&provider=JSPTabContainer, (参照 2010-01-18).

[5] 国立情報学研究所ディジタル・シルクロードプロジェクト.”ディジタル・シルクロード – 文化遺産のデジタルアーカイブ -”, http://dsr.nii.ac.jp/, (参照 2010-01-18).

[6] 北本 朝展ほか. "ディジタル・シルクロード : 多彩な文化遺産を統合するディジタルアーカイブ", 人文科学とコンピュータシンポジウム じんもんこん 2005, pp. 121-128, 2005. http://agora.ex.nii.ac.jp/~kitamoto/research/publications/jinmonkon05.pdf, (参照 2010-01-18).

[7] 阿蘇品 治夫. “機関リポジトリを軌道に乗せるため為すべき仕事 ―千葉大学の初期経験を踏まえて―”. 情報管理. 2005, vol. 48, no. 8, 496-508. http://www.jstage.jst.go.jp/article/johokanri/48/8/496/_pdf/-char/ja/, (参照 2010-01-18).

[8] 神戸大学附属図書館. “新聞記事文庫とは？”, http://www.lib.kobe-u.ac.jp/sinbun/gaiyou.html, (参照 2010-01-18).

[9] 国立公文書館アジア歴史資料センター. “資料の概要”, http://www.jacar.go.jp/siryo/siryo2.html, (参照 2010-01-18).

[10] “国立国会図書館デジタルアーカイブポータル（PORTA）外部提供インタフェース仕様書 ver. 2.0(最終更新：2009.07.01)”, http://porta.ndl.go.jp/wiki/attach/%E5%A4%96%E9%83%A8%E6%8F%90%E4%BE%9B%E3%82%A4%E3%83%B3%E3%82%BF%E3%83%95%E3%82%A7%E3%83%BC%E3%82%B9%E3%81%AB%E3%81%A4%E3%81%84%E3%81%A6/externalInterface_ver2.0.pdf, (参照 2009-07-07).

[11] “HANAVI: Hybrid And Network-Assisted Vocabulary Interface.” http://raus.slis.tsukuba.ac.jp/subjects/graph, (参照 2010-01-19).

[12] 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座. ChaSen version 2.4.0, 2008-03. http://chasen.naist.jp/hiki/ChaSen/, (参照 2009-09-03).

[13] 児玉幸多編. 日本史年表・地図. 第 15 版, 吉川弘文館, 2009, 64,56,16p.

[14] Massachusetts Institute of Technology. “SIMILE Widgets | Timeline”, http://www.simile-widgets.org/timeline/, (参照 2010-01-18).