2ZB-01

震災関連資料デジタルアーカイブのための半自動メタデータ付与・検索機能の開発

今川泰基† 富澤浩樹† 市川 尚† 阿部昭博†
岩手県立大学ソフトウェア情報学部†

1. はじめに

2011 年 3 月 11 日に発生した東日本大震災津波は各地に甚大な被害をもたらした．被災した県の図書館では，震災関連資料（以下，資料）を収集，保存，公開する取り組みが行われている．岩手県立図書館では，OPAC（Online Public Access Catalog）を用いて管理することで，収集した資料を迅速に公開することが可能となった．しかし，利用者が関心のある資料を検索し，目的の資料を見出すことはとても難しいが，各図書館の震災アーカイブではメタデータにより検索支援が行われている．

そこで本研究では，岩手県立図書館震災図書館担当者（以下，図書館担当者）と共同開発した震災関連資料デジタルアーカイビングシステムを対象に，検索にメタデータを利用することで検索の容易性を担保することとし，半自動メタデータ付与と検索を行うための機能を実装する．その上で，機能の有用性について評価する．

2. 調査

2.1 先行システム

先行システムとして，新聞見出し分析から導出したキーワードを用いて利用者がメタデータの登録と検索を行う研究がある [1]．メタデータのサンプルとして，岩手日報社「＜特集＞3．11 東日本大震災～立ち上がろう岩手～」等の記事見出しを内容分析した結果を用いている．拡張性の乏しい OPAC に対しても，資料にメタデータを付与することで検索性の向上に繋がることが示されたが，膨大な資料に対しては登録者の負担が大きいことが課題である．

一方，中越災害アーカイブ（長岡震災アーカイブセンター）では，カテゴリ選択とカテゴリ別のメタデータ等による絞込みが可能である．しかし資料に関連するキーワードやメタデータはシステム管理者が登録しており，設定のために相応のコストがかけられている．

Semi-automatic Metadata Creation and Search Functions of Digital Archives System for Earthquake Disaster-related Documents
†Taiki Imakawa, Hiroki Tomizawa, Hisashi Ichikawa and Akihiro Abe
†Faculty of Software and Information Science, Iwate Prefectural University

2.2 図書館担当者へのヒアリング

2017 年 8 月 3 日，岩手県立図書館会議室において，図書館担当者 4 名に対して，本研究課題に関するヒアリング調査を実施した．その結果，現在も震災関連資料は増加し続けていること，人員に余裕がないことから，引き続き重要なテーマであることを確認できた．また，資料の特性から地名をメタデータに含めること，全文検索を可能とすること，PDF で公開されている資料を対象とするメリットについて合意された．

2.3 関連研究

メタデータによって検索効率を向上させる研究は数多く行われている．たとえば，横山ら [2]は，本研究と同様の問題意識の下，写真資料に含まれる個別のメタデータを用いて関連資料を集約する手法を提案している．しかし，資料へのメタデータ付与に関しての研究はほとんどなく，PDF を用いた研究も見当たらない．

3. システム設計

3.1 設計方針

以上の調査を踏まえて，本システムの設計方針を定めた．すなわち，対象を PDF 公開資料とすること（方針 1），メタデータを「一般」と「地名」に分けて登録すること（方針 2），全文検索を可能とすること（方針 3），図書館担当者がメタデータを付与すること（方針 4）の 4 点である．

3.2 システム構成

本システムの構成を図 1 に示す．本システムは，PDF として公開されている資料を対象に，図書館担当者がメタデータを半自動で付与する．具体的には，システムが PDF からメタデータ候補を抽出し，図書館担当者がメタデータを修正・登録するとともに OPAC と紐づけして登録する．

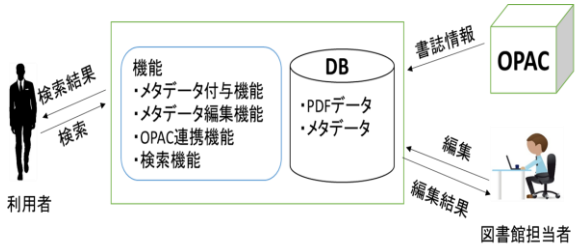

図 1 システム構成

4. 主要機能の開発と実装

本システムは PC での利用を想定して開発した．開発言語は PHP，Python，HTML5 を用いた．また，データベースは MySQL を使用する．主要機能は以下の 4 つである．

1) メタデータ付与機能：アップロードされた PDF に対して，PDFMiner を用いて PDF からテキストを抽出し，更にそのテキストから MeCab を用いて名詞のみを頻出順で抽出する．それらを「一般」10 件，「地名」5 件のメタデータとして付与する．なお，「地名」はシステム内の地名リストと照合して振り分ける．

2) メタデータ編集機能：抽出されたメタデータについて編集可能とする．

3) OPAC 連携機能：抽出・編集されたメタデータを OPAC の書誌情報と紐づける．

4) 検索機能：一般検索，メタデータ検索の他に，全文検索機能を可能とする．また，メタデータ既付与リストを表示する．

5. 評価

5.1 一次評価

震災学習に関心のある学生 6 名に本システムの使い勝手を確認してもらった．本システムの操作への理解度をそれぞれの機能について確認したところ，全員が理解できたと回答した．また，データの重複の解消や，UI に関する改善要望等が挙がった．

5.2 二次評価

一次評価での要望について改善を図った上で，メタデータ付与機能の有用性評価を目的として，図書館担当者 6 名を対象に二次評価を実施した．具体的には，被災地域の 2 つの復興計画（資料 A：21 ページ，資料 B：100 ページ）を渡した上で，「一般」と「地名」それぞれのメタデータを実際に付与してもらい，その所要時間を記してもらった．なお，本システムを用いない場合についても同様に実施した．

(1) メタデータ付与の作業効率について

6 名のメタデータ付与にかかった所要時間に基づいて作業効率を計算した．その結果，本システムを用いた場合，資料 A については約 1.8 倍，資料 B については約 1.7 倍，作業効率が高まったことが分かった．

(2) 適合率，再現率について

メタデータ付与機能によって推薦された「一般」と「地名」のメタデータに対して，図書館担当者による編集結果を正解群として，検索結果の妥当性評価に用いる適合率，再現率を算出して評価した(表 1)．平均を比較すると，「一般」に関しては，適合率と再現率ともに 4 割以下となり，総じて低い結果となった．一方「地名」に関しては，資料 A の適合率が他と比べて低い結果となったものの，その他については 5 割を超えた．図書館担当者からは，「ワードの置き換え機能が欲しい」，「単体で意味のないワードがある」，「どの要素にウエイトが置かれるかの参考になる」といった意見がみられた．

表 1　適合率と再現率

資料	区分	担1		担2		担3		担4		担5		担6		平均	
		P	R	P	R	P	R	P	R	P	R	P	R	P	R
資料A	一般	60	67	0	0	60	86	20	29	50	71	0	0	33	40
	地名	100	100	60	60	100	100	20	100	100	100	20	100	37	58
資料B	一般	40	44	10	20	60	100	0	0	60	75	10	25	33	36
	地名	80	100	100	100	80	100	60	100	100	100	20	100	67	53

（P：適合率，R：再現率，担：図書館担当者，単位：%）

5.3 考察

二次評価より，本システムは図書館担当者のメタデータ付与の作業効率向上に貢献可能であることが明らかとなった．しかし，付与するメタデータの傾向が図書館担当者ごとに異なるため，「一般」については全く貢献しない場合もある一方，「地名」はそのまま採用される場合も多い．以上のことから，本システムによるメタデータ付与について，「一般」に関してはまだまだ改善の必要があるが，「地名」に関しては一定程度有用と考えられる．

6. おわりに

本研究では，震災関連資料デジタルアーカイビングシステムに，半自動メタデータ付与・検索機能を開発・実装した．メタデータ付与機能については多くの課題が残るものの一定程度の有用性を確認できた．今後はメタデータ付与の精度向上，登録者による差異の緩和，資料の種類による特徴の抽出等についても検討する必要がある．

参考文献

1) 高谷晃生：震災関連資料デジタルアーカイブシステムのための利用者参加型メタデータ登録・検索機能の開発，岩手県立大学ソフトウェア情報学部卒業論文要旨集，pp. 168-169（2016）．

2) 横山雄哉，積祐典，三原鉄也，永森光春，杉本重雄：シンプルなメタデータが付与された東日本大震災アーカイブの写真資料のための時空間情報を利用したコンテンツ集約手法，情報処理学会第 79 回全国大会，1ZF-02（2017）．