

地理情報に関連づけられたデジタル写真収集システムの提案

山本 理絵[†]吉野 孝[†][†]和歌山大学

1 はじめに

近年デジタルアーカイブは有形無形の資源をデジタル化することで、容易に多種多様な資源を収集・蓄積・保存・提供することができるサービスとして利用されている。利点として利用者・研究者・組織が持つ分散されたデータを集約できることや、データを公開することで、教育や観光、研究等への二次利用が可能となることが挙げられる。一方で国立国会図書館の報告書では、データの追加・更新の停滞や操作性の向上が見られない問題点が挙げられており、デジタルアーカイブに対するユーザの評価・支援が必要であると報告している [1]。

そこで我々はデータの収集支援を行い、更にデータを再編成することで、地域の活性化や観光支援に繋がるのではないかと考えた。デジタルアーカイブが観光事業に影響を与えた例として、函館市で実施されているデジタルアーカイブプロジェクトが挙げられる [2]。

本研究では、観光支援を目的としたデジタルアーカイブの構築及びデジタル写真の収集支援システムを提案する。本システムでは、Google 画像検索を用いてデジタル写真の収集支援を行う。また自動収集した写真を地図上にマッピングし、データの再編成を行う。本稿では、システムの概要について述べる。

2 関連研究

小田島らは、地域コミュニティ内の文化保護活動を支援するデジタルアーカイブシステムを開発した [3]。このシステムは文化資源に関する情報を配信することで、地域住民からの情報提供を促している。

また近藤らは、展示のあらすじとなるテキストに対して、メタデータを利用した画像の推薦を行うことでオンライン3次元ミュージアムの展示を半自動的に構成する手法を提案した [4]。これは大規模なデジタルアーカイブを効率的に公開することを可能としている。

本研究では、データの収集及びデータの再編成を自動的に行うシステムを提案する。

3 デジタル写真収集システム

3.1 概要

本システムはデジタル写真を Web 上から自動収集し、デジタルアーカイブに蓄積する。また収集したデータを地図上にマッピングし、データの再編成を行う。これはデジタルアーカイブにおけるデータの収集支援となる。またデータを観光資源となる形に再編成することで観光支援を行う。

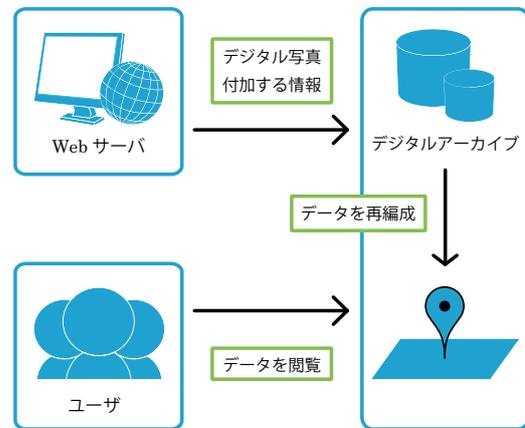


図 1: システム構成

3.2 システム構成

本システムの構成図を図1に示す。本システムはデジタル写真や情報を取得する Web サーバと取得したデータを保存するデジタルアーカイブから構成される。なお、デジタル写真の自動収集は Google 画像検索を利用する。以降の各節でデータの収集方法とデータの再編成について述べる。

3.3 情報抽出

本節ではデジタル写真に付加する情報の抽出について述べる。データ再編成時にデジタル写真の説明を付加するため、情報抽出を行う。

(1) 画像検索から得られる情報

画像検索結果から得られる情報はタイトル、画像ファイルのパス、画像が存在するページのリンク、画像が存在するページの top へ繋がるリンク、スニペット、MIMEType である。これらに加え画像ファイルのパスや MIMEType から拡張子を抽出する。

(2) 最終更新日時 of 抽出

画像がアップロードされた日時を記録する。そのために最終更新日時の情報抽出を行う。しかし最終更新日時は明記されていないため推定する必要があるが、正確な日時を推定することは困難であるため、以下に述べる3つの手法を提案する。情報の抽出はより正確に日時を取得できる手法 (A)、手法 (B)、手法 (C) の順に適用する。

(A) RSS フィードから記事の記載日時を取得

記事本文中に画像検索結果で得られた画像が含まれているかを確認し、画像が存在するならばその記事の記載日時を取得する。これはブログのみ有効な手法であるが、正確な日時を取得することが可能である。しかし、RSS フィードはある一定量の最新記事をまとめた文書であるため、画像が含まれている記事が

古い場合、その記事を取得できないという問題がある。

- (B) 画像が存在するページの最終更新日時を取得
ホームページの最終更新日時は、HTTP ヘッダから取得することができる。しかし、動的コンテンツを含むページは更新が頻繁に行われるため、正確な最終更新日時を設定しない場合に、最終更新日時を取得できないという問題がある。
- (C) 画像が存在するページのテキストに記述された日時を取得
画像が存在するページのHTML文書を取得し、画像検索結果から得られた画像付近にある、本文として記述された日付を最終更新日時として取得する。

(3) 場所の抽出

デジタル写真が撮影された場所を記録する。そのために撮影場所の情報抽出を行う。しかしデジタル写真そのものから撮影場所を特定することは困難であるため、以下に述べる2つの手法を提案する。情報の抽出は手法(A), 手法(B)の順に適用する。

(A) 画像が存在するページのテキストから場所を抽出

ページ内において画像付近にある文章の中に、撮影場所に関する情報があると考えた。この文章を「本文」と表現する。本文を特定するために、HTML文書をdivタグ, tdタグで区切り、その中から句読点を含む文章を抽出する。句読点を含む文章を抽出する理由は、日本語の文章であり、かつ本文である可能性が高いと考えられるためである。次に、画像付近にある文章の一部を取得する。画像付近にある文章の一部を取得する理由は、前述の句読点を含む文章が本文であるかを確認するためである。得られた句読点を含む文章と画像付近にある文章の一部を照合し、句読点を含む文章の中に画像付近にある文章の一部が含まれているならば、その句読点を含む文章を本文として抽出する。そして抽出した本文と観光地・地名をリスト化したテキストデータを照合し、本文中に観光地・地名が含まれているならば、その観光地・地名を場所の情報として抽出する。

(B) タイトル・スニペットから場所を抽出

画像検索結果から得られたタイトル・スニペットと観光地・地名をリスト化したテキストデータを照合し、タイトル・スニペットの文字列中に観光地・地名が含まれているならば、その観光地・地名を場所の情報として抽出する。

(4) 色の抽出

画像の色を一色で表現し、記録する。そのために画像の色を以下に述べる過程で抽出する。まず各ピクセルごとに色をRGBで取得する。RGBに比べ、HSVは人間が色を知覚する方法と類似しているため、取得した色をHSVに変換する。次に各ピクセルの色を赤・オレンジ・黄・緑・シアン・青・



図2: デジタル写真を地図上にマッピングしたデータ

紫・ピンク・白・灰・黒の11色に分類する^{*1}。最後に11色に分類した色の中で最もピクセル数が多い色を画像の色として抽出する。

3.4 データの再編成

本節ではデータの再編成について述べる。収集したデジタル写真をGoogleEarthを用いて地図上にマッピングし保存する。図2にデジタル写真を地図上にマッピングした画像を示す。デジタル写真をアイコンとして表示し、情報の抽出により得られた場所にマッピングした。またアイコンをクリックすることで、デジタル写真とそれに付加された情報を見ることを可能とした。

4 おわりに

本稿では、地理情報に関連づけられたデジタル写真収集システムを提案した。今後は、デジタル写真を色・場所・時間・感情等のカテゴリーに分類することを検討する。また情報抽出の精度評価を行う。

謝辞

本研究の一部は、和歌山大学平成25年度独創的研究支援プロジェクトの補助を受けた。

参考文献

- [1] 文化・学術機関におけるデジタルアーカイブ等の運営に関する調査研究: http://current.ndl.go.jp/FY2009_research (参照 2014-01-02).
- [2] 川嶋聡夫: 観光情報学: 10. デジタルアーカイブを活用した観光コンテンツ, 情報処理, Vol.53, No.11, pp.1192-1197 (2012).
- [3] 小田島瑞希, 竹野健夫, 植竹俊文, 菅原光政: 地域コミュニティを主体とする文化資源収集支援システムの開発, 情報処理学会研究報告, 情報システムと社会環境 (IS), 2013-IS-123(7), pp.1-6 (2013).
- [4] 近藤悠太郎, 川嶋聡夫: デジタルアーカイブのための3次元オンラインミュージアムの半自動構成, 情報処理学会研究報告, デジタルドキュメント (DD), 2010-DD-75(2), pp.1-6 (2010).

^{*1}HSVの色を表現するために、無彩色として白・灰・黒が必要であると判断した。また有彩色は色相環をもとに赤・オレンジ・黄・緑・シアン・青・紫・ピンクを利用する。